NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

The Situational Features and Textual Dimensions of Electronic Language

Milena Collot

A Thesis
in
The Department
of
Applied Linguistics

Presented in Partial Fulfilment of the Requirements
for the Degree of Master of Arts at
Concordia University
Montreal, Quebec, Canada

June 1991

ISBN 0-315-68708-8

ABSTRACT

The Situational Features and Textual Dimensions of Electronic Language

Milena Collot

This study examines the language which people use to conduct dialogues across computerized telecommunications networks. This variety, which I have called electronic language, is characterized by a set of situational constraints which sets it apart from other varieties of English. Messages delivered electronically are neither "spoken" nor "written", in the conventional sense of those words. They do not fall easily into the category of spoken English, since the participants neither see nor hear each other during the language event. Nor can they be strictly interpreted as written English, since many messages are composed directly on-line, thereby ruling out the use of planning and editing strategies which are at the disposal of even the most informal writer.

Following the approach developed in corpus linguistics, the sample which was analysed was rather large, measuring approximately 200,000 words. With the aid of an automatic tagging program, CLAWS 1, each word in the corpus was assigned a part-of-speech label. An automatic text retrieval program was then used to establish the frequency of a number of linguistic features, particular configurations of which have been shown to characterize specific text types, e.g., narrative versus non-narrative disourse. Finally, multivariate statistical techniques were used identify the

position of electronic language with respect to such variables, and to compare this genre with other genres previously analysed.

The results indicate that electronic language displays both some of the linguistic features which have been associated with certain forms of written language and others which are more usually associated with spoken language. For example, it uses contractions and first and second person pronouns which typically occur in informal, face-to-face conversations; and, at the same time, features which occur in formal written language with a high information content. Referents are usually situation-dependent, i.e., they are not explicitly or elaborately identified but must be inferred from the general context of the message, just as in ordinary conversation among people who share knowledge. It is also highly fragmented. The genres which it most resembles are public interviews and letters, personal as well as professional.

ACKNOWLEDGEMENTS

First of all, I would like to thank my thesis supervisor, Nancy Belmore, for providing the inspiration for this project and countless words of wisdom throughout. I would also like to thank Jack Usphur, who prevented this thesis from turning into a statistical nightmare. Thank-you also to Margery Fee for her valuable, and prompt, comments. Many thanks are also due to Professor G.N. Leech, Professor of English and Director of the Unit for Computer Research on the English Language at the University of Lancaster for permission to use the CLAWS 1 tagging programs; to Prof. Ossi Ihalainen, professor of English at the University of Helsinki and L. Sadeniemi of the University of Helsinki's Computing Centre, for making available their copies of CLAWS 1. Finally, I would like to thank Ashley Saldanha, without whose technical expertise and encouragement this project surely could not have been done; and my son, Adrian Saldanha, without whom this project could have been done in half the time but would have been far less of a challenge.

Table of Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Historical Perspective

The study of language variation has been one of the major forces behind the development of linguistics proper. Indeed, it was the study of variation in consonant sounds that led the 18th century philologist William Jones to suspect that Sanskrit, Greek, Latin and the Germanic languages were all interrelated. Further tracing this thread of variation, scholars such as Grimm, Bopp and Rask established the connection between Indian and European languages, and as such launched comparative linguistics as a systematic field of study (Pedersen, 1931). Its principal preoccupation was the reconstruction of proto-languages and the determination of language families. Its driving force was the observation that languages experienced systematic changes over time, or diachronic variation.

What had been largely ignored, however, was the kind of variation that took place in the same period in time, or synchronic variation. As a reaction to the historicism of his contemporaries, de Saussure (1916) was one of the first to underline the importance of describing contemporary language, thus marking the beginning of synchronic, or what later came to be known as descriptive linguistics. Through the work of Boas (1911,1940), Sapir (1921) and Bloomfield (1931), descriptive linguistics developed as a separate discipline, primarily concerned with the study of languages that had no written form. When the object of study was a language with a written form, the focus was on a standard dialect.

With the geopolitical upheavals in this century, ranging from wars of independence in Europe's colonies to two world wars, the question of what form of a

language should be considered "standard" arose. In the case of English, for example, could RP be considered the standard when Great Britain was no longer the seat of an English-speaking empire? Could some variety of American English be described as the standard, when significant populations in areas like India, the Caribbean, or Australia also spoke educated varieties of English which could certainly be called "standard"? The development and establishment of such varieties of English, both native and non-native, led to the realization that the study of a particular language required a clear specification of the geographic dialect, or dialects, to be considered.

If language varies according to its historical and geographic context, so too does it vary with the socio economic background of its speakers. The study of socio-economic variation in language, pioneered by Fries (1940) and developed by Labov (1971) and Bernstein (1973), led to a new domain of linguistics, sociolinguistics, which was redefined as "the study of language within the context of a speech community" (Labov, 1971).

Language varieties which are determined by history, geography or social class are relatively permanent in individual speakers. A speaker cannot switch among such varieties, unless he intentionally sets out to mimic someone from another time, place or socio-economic background (Crystal and Davy, pp. 67-68). However, there is a type of variation through which a native speaker can and does navigate easily: the variation dictated by situational constraints. Speakers will adopt different varieties of a language according to the social and communicative setting, or the speech situation. The purpose of communication, the medium, the relationship between speaker and listener, are all features which shape a given situation. The recognition of a vast number of communicative situations, and a corresponding range of language varieties used on different occasions by the same individual, was both the cause and effect of

the development of a new school of thought within linguistics. From the early sixties on, this kind of variation gained increasing importance in language analysis (Crystal and Davy, 1969; Hymes, 1972).

This is the type of variation to which Crystal and Davy (1969) allude when they write:

> Each of us, as an educated speaker of English, is, in a sense, multilingual; for in the course of developing our command of language we have encountered a large number of varieties, and, to a certain extent, have learned how to use them. A particular social situation makes us respond with an appropriate variety of language, and, as we move through the day, so the type of language we are using changes fairly instinctively with the situation. (1969:4).

Though the native speaker's choice of one particular variety over another is not usually a conscious decision, the same cannot be expected of a second language speaker, who has not had the kind of exposure required to allow him to master contextually appropriate forms.

To compensate for this lack of intuitive awareness, recent trends in language pedagogy have advocated the implementation of syllabuses that are notionally based (Wilkins,1976), and objectives which are more directly concerned with sociolinguistic appropriateness than linguistic correctness (Littlewood, 1981; Harmer, 1983). However, for a notional syllabus to be truly representative, one would expect it to be based on empirical studies which identify the features of the specific varieties of English which ESL speakers need to learn. This type of study is best illustrated by Biber (1988), who has made two important contributions to the field of descriptive linguistics. First, he proposes a sophisticated framework of analysis which allows researchers to determine if samples of language produced under particular circumstances can be identified by a statistically significant co-occurrence of linguistic features. Second, he offers a model for the description of varieties which is

so refined that it can accommodate the study of even the least charted varieties of English.

1.2 Purpose and Content of Thesis

One variety of English which, to my knowledge, is still largely unexplored, is the language which people use to conduct dialogues across computerized telecommunications networks, via a piece of hardware called a modem. Telecommunications are being used by more and more people, and from all parts of the world. In 1970, there were 15,000 modems in operation in the world. In 1980, the number had risen to 250,000, representing a thirteen-fold increase. Just 7 years later, in 1987, the number of modems used was estimated to be 10 million, a forty-fold increase over the 1970 figure (Rosch, 1987). These figures give us an indication of the importance that this variety is gaining throughout the world.

This variety, which I have called electronic language, is characterized by a set of situational constraints which sets it apart from other varieties of English. For example, messages delivered electronically are neither "spoken" nor "written", in the conventional sense of the words, and yet they can be classified as either. In some respects they may appear to be spoken, since there is an easy interaction of participants and alternation of topics, similar to that exhibited during cocktail parties. However, they cannot be strictly labelled as spoken messages, since the participants neither see nor hear each other. By the same token, they cannot be considered strictly as written messages since many of them are composed directly on-line, thereby ruling out the use of planning and editing strategies which are at the disposal of even the most informal writer.

Because electronic language has unique situational features, it seems reasonable to assume that it embodies a distinctive set of linguistic features as well. This assumption has been partially corroborated by a pilot study which I recently conducted (Collot, in press). The study compared a privately collected corpus of electronic language to the Survey of English Usage corpus (hereafter SEU), a public corpus of written and spoken English (Quirk 1960, Svartvik and Quirk 1980). The study revealed clear differences in the use of comparative adjectives in the two corpora.

Although the pilot study indicated how a description of this new variety might be approached, it did not of course provide a complete description of electronic language because it focused on a single linguistic feature. Nor could the study show to what extent electronic language differs from other varieties of English. This is the purpose of my thesis. The question that has guided my work, therefore, is: given the unique set of situational features which characterize electronic language, is there a corresponding unique set of linguistic features? If so, how do they compare with other varieties already analyzed? The research is based on a computerized corpus of electronic language, and the analysis follows Biber's approach (1988).

Chapter 2

PREVIOUS RESEARCH ON LANGUAGE VARIATION

2.1 Descriptive studies of language variation

Studies concerned with diachronic variation have been briefly mentioned in the preceding section. Though seminal in their own right, these works provided little direct benefit to the study of language use, as they looked at language in isolation, divorced from the context in which the utterances were actually produced. In this section, I would like to focus on studies of synchronic variation, particularly those which will guide us here in finding an appropriate model for the description of the situational features of electronic language.

2.1.1 Group variation: Geography, class and ethnicity

The view of linguistics as "the study of language within the context of a speech community" has inspired and given direction to a large number of variety-specific studies. One area of investigation which has found enormous scope since the sixties has been the study of geographical varieties. Thus, much academic attention has been devoted to Jamaican and Haitian Creole (Valdman, 1977), to Indian English (Kachru, 1983), African English (Spencer, 1971; Sey, 1973) and South East Asian English (Llamzon, 1969; Noss, 1983; Tongue, 1974). In general the aim of these studies is to describe and classify the specific linguistic features which distinguish these varieties. The authors approach their objects of study from a descriptive perspective, withholding value judgement on the variety at hand. Beyond its academic interest, much of this work has immediate practical value, as exemplified by the compilation of numerous variety-specific dictionaries (Winer, 1989).

In addition to geographical variation, a number of scholars have looked at variation determined by the speaker's social class. One of the first attempts at this type of sociolinguistic description is Fries (1940), who examined the language used by people of different socio-economic backgrounds in letters written to their draft boards during World War II. For spoken English, this type of investigation has been conducted by Labov (1972), who identified certain phonological features which systematically distinguished upper class from lower class New Yorkers. Though all the subjects in his study manifested a tendency to modify their speech according to the formality required by the situation, he found that lower class speakers had a greater propensity to use stygmatized forms, such as *dem* and *dat* for *them* and *that*. In the same period, in Britain, Bernstein (1973) published a study suggesting that members of the working class use a restricted variety of English, which relies heavily on exophoric references, and as such is context dependent, while middle class children use an elaborated code, which is less dependent upon the immediate context and therefore allows for greater abstraction of thought.

The ethnic background of speakers has also been a major concern of works on language variation. In a study reminiscent of Bernstein's, Scollon and Scollon (1981) showed how Athapascan children speak within an oral tradition of concreteness, while the discourse of white children is more heavily steeped in a literate, abstract tradition. Black English has also attracted great scholarly attention, resulting in numerous descriptive studies. The major contribution of these studies has been to isolate the linguistic features which characterize this variety and distinguish it from more standard varieties of English. Thus, double negation, the invariant form of *be*, and the systematic omission of the plural *-s* in some nouns have all been identified as

distinctive features of Black English (Sutcliffe, 1982: Dillard, 1972; Smitherman, 1977).

To be sure, Electronic Language contains features which are specific to different geographic, social and ethnic dialects. Telecommunication networks, as we have seen, are becoming accessible to more and more users, from increasingly varied locations. However, and precisely because of this heterogeneity, the distinguishing situational features of this type of language cannot be traced to the specific background of the user. Rather, it would seem that any difference that will be uncovered between electronic language and the "common core" of English, will be more closely associated with the context in which this language is produced.

2.1.2 Individual variation: context of situation

The first acknowledgement of the effect of situation upon language can be traced to the work of Malinowski, according to whom: "An utterance becomes only intelligible when it is placed within its context of situation... " (1923: 306). For Malinowski, the context of a speech event could be derived from the purpose that a specific utterance was meant to fulfill. He identified four major functions: phatic, narrative, speech of action, and the ritual use of words. Thus the first situational feature, purpose, was identified.

In the 1930s, Firth elaborated on the concept of 'context of situation' by combining Malinowski's sociological perspective with the more formal aspects of linguistic analysis. Utterances had to be analyzed for their phonetic, lexical and syntactical characteristics, as well as for their contextual manifestations. The study of language was to be conducted within a framework which would identify the context of situation of a particular speech event. The context was seen as a

configuration of four specific features: a) the relationships among the participants in the speech event; b) the action, both verbal and non-verbal, of the participants; c) the relevant objects and events involved in the situation; d) the effects, or changes, brought about by what is said (Firth, 1935).

The study of context was further developed in the 1960s by Halliday, McIntosh and Strevens (1964). Benefitting from the advances in descriptive linguistics that had occurred during the years separating them from Firth, they refined the concept of context of situation into a fuller account of language variation, known as 'register'. The notion of register is derived from two observations: that language has specific functions, and that language events take place in given contexts of situation. Register describes the systematic relationship between these two aspects of language use along three dimensions. The first is 'field of discourse', which refers to what the speaker is doing during the language event. For example, a language event can be seen as happening in the general context of 'discussion', or the more restricted context of 'doctors in surgery'. The second dimension is 'mode of discourse', which refers to the medium. Generally, the medium is either speech or writing. The mode can be subclassified into specific genres, like the language of newspapers, advertising, or conversation. The third dimension is 'style of discourse'. This refers to the relations among the participants, which can be seen in terms of formality or status difference, as in the case of a parent speaking to his child.

The notion of register suffers from two limitations. The first is the premise that "the crucial criteria of any given register are to be found in its grammar and its lexis" (Halliday, McIntosh & Strevens, 1964: 88). Later scholars have shown that prosody is at least as important as grammar and lexis in describing language. In fact, an entire field of study has emerged around the issue of prosodic features and their relationship

to meaning (Crystal, 1969, 1975; Ladd, 1978). Secondly, though the concept of register was important in demonstrating the interrelation of situational and linguistic features, it did not provide a complete model of description. "There are many aspects of the way in which English is used which... cannot be handled adequately by such categories as register, tenor , field, mode , and so on in any of their current senses" (Crystal & Davy, 1969:61). One important situational variable, largely ignored by the Neo-Firthians, was the particular form that a given message could adopt within one mode. Thus, the notion of register did not subsume the differences between, say, letters and memos.

Reacting to the vagueness of register, Crystal and Davy (1969) published their pioneering study of varieties of English, *Investigating English Style*. The aim of the study was to identify, "from the general mass of linguistic features common to English, those features which are restricted to certain kinds of social context" (1969:10). The study proposed a framework for the description of social context which was very comprehensive. Generally, it described social context in terms of three groups of components. The first group consists of features which are relatively permanent and cannot be easily manipulated by language users, unless they are especially talented at imitating other speakers. These include idiosyncratic features such as the handwriting or voice quality of an individual, or the use of pet words; dialect, which is expressed by phonological, lexical and grammatical features that are unique to a specific geographical community; and temporal provenance,which refers to the historical period in which the speech event takes place. It is interesting to note that, although the authors acknowledge the importance of geographic and historical variation, they do not consider the socio-economic determinants of language variation which have preoccupied scholars like Labov.

The second group of components identified in Crystal and Davy can be manipulated by interlocutors, and is determined by the situation in which language is being produced. Under the broad category of discourse, Crystal and Davy include the medium and the nature of participation. 'Medium' describes the way in which language is presented, and is generally classified as either speech or writing. 'Participation' refers to the interaction between a speaker and his audience; the basic distinction is between monologue, in which there is "no expectation of a response" and dialogue, which is "an utterance with alternating participants, usually, though not necessarily, two in number" (1969:69). The model also makes provisions for language events which cross these two categories, by distinguishing between simple and complex relations. Thus, language which remains within one category, for example language which is spoken to be heard or written to be read, is simple. Language which crosses the boundaries, for example a written text which is intended to be spoken, as in plays or newscasts, is complex. Similarly, language which crosses the participation boundaries, as when monologue is introduced in a conversation when someone is telling a joke, has complex participation.

The third group of situational features refers to localized or temporary variations in language. One of the features in this group is 'province', which refers to the kind of professional occupational activity being engaged in. Province includes, but is not restricted to, subject matter. A second feature, 'status', describes "the systematic linguistic variations which correspond with variations in the relative social standing of the participants in any act of communication" (1969:73-74). This includes notions such as formality, politeness, and intimacy. Thirdly, 'modality' refers to the form in which language is presented. In writing, for example, the choice of linguistic features will vary according to whether one is writing a letter, a

postcard, or a memo. The term 'genre' would fall under this category. The last component in this group, 'singularity', refers to the linguistic idiosyncrasies of particular individuals.

The concepts of context, register and style, and the models of situational descriptions which they implied, paved the way for systematic descriptions of a wide range of predictors of linguistic variation. For example, Joos (1961) examined the various levels of formality which a speaker can use to indicate his relationship to the listener, and classified them on a five point scale: frozen, formal, consultative, casual and intimate. Similarly, Brown and Ford (1961) looked at the same type of variation in role relationships, as reflected in the use of different terms of address.

Many studies have looked at the patterns of interaction among language users, a subject which later came to be known as discourse analysis (Coulthard, 1973; Sinclair and Coulthard, 1975). Within this tradition, Berko-Gleason (1973) has described the speech of mothers to children; Candlin et al. (1976) have described that of doctors to patients; and Torode (1976), that of teachers to students. In an interesting study, Fielding and Fraser (1978) show that speakers who are bound to their listeners by sentiments of affection use more verbs and pronouns, while speakers who dislike their listeners tend to use passives and nominalizations. Thus they associate a verb-heavy text with a vivid, personal style, and a highly nominalized text with impersonal, unemotional characteristics.

Situational variation of this kind has also been studied by specialists in natural language information processing who have wanted to determine the lexical and grammatical features of what they call sublanguages (Kittredge and Lehrberger, 1982), subsets of English restricted to narrowly and precisely-defined subject areas and document types. The identification and analysis of sublanguages has resulted in

numerous applications, including the automatic translation of specialized texts such as weather reports and information retrieval from medical documents.

As the study of situationally-determined language variation has evolved, the theoretical models which provide a basis for an adequate description have been successively refined. One of the more recent analyses of language variation, Biber 1988, marks an advance over previous analyses in at least two respects. The first difference is procedural. Rather than examining a body of language and then describing it in terms of the features that are intuitively felt to be stylistically significant, as Crystal and Davy had proposed, he starts with a set of linguistic features, and determines under what circumstances particular subsets of those features co-occur. Needless to say, the choice of which linguistic features to examine is contingent on a pre-existing body of research which has already identified those features which are likely to have predictive value. Crystal and Davy, preceding Biber by two decades, were clearly at a disadvantage in this respect.

The second difference lies in the descriptive power of Biber's model of speech situations. Biber identifies eight components of the speech situation. Though many of these components overlap with Crystal and Davy's model, there are a number of interesting innovations, some of which I have found particularly suited to a description of electronic language. For example, Biber's model not only considers the effect of time and space upon a given variety, but "the extent to which time and space are shared by the participants" (1988: 30). As we will see, temporal and spatial disembodiment is one of the characteristic features of an electronic language corpus (hereafter ELC), and may play a crucial role in explaining the linguistic patterns which will be observed. Another important innovation is the component labelled 'relations of the participants to the text'. This component refers to the

question of whether participants are required to produce language extemporaneously or whether they can interact with their text at a more leisurely pace. This is an important variable in electronic language because a participant may be operating within real time constraints, similar to those experienced by a speaker in face-to-face conversation. Because of these innovations, Biber's model is especially appropriate for analyzing electronic language.

2.1.3 Variations across speech and writing

Within the wider concept of "context of situation" one area which has attracted much scholarly interest has been the relationship between speech and writing. Historically, linguists have considered one medium primary; the other, secondary. From antiquity to the very recent past, writing was considered primary. With the advent of synchronic linguistics and the development of phonetics as an established discipline, speech was considered primary. Finally, since the advent of sociolinguistics in general, and discourse analysis in particular, speech and writing have come to be viewed as equally important varieties. More recently, however, the question has been raised as to whether the relationship between speech and writing is indeed as dichotomous as earlier research would have suggested (for example, see Biber, 1988, pp. 160-164).

From the multitude of studies that have been conducted on differences between speech and writing, we can extract some of the principal features which have been thought to distinguish the two. Firstly, there seems to be a consensus on the different functions which speech and writing are likely to fulfil. Brown and Yule (1983) see speech as primarily interactional, fulfilling "that function involved in expressing social relations and personal attitudes" (p.1), and writing as transactional, where the

term is defined as "that function which language serves in the expression of content" (p. 1). This functional distinction is roughly parallel to a number of similar dichotomies reported in the literature: referential/emotive (Jakobson, 1960), ideational/interpersonal (Halliday, 1970), descriptive/social - expressive (Lyons, 1977). These functional differentiations have, to a greater or lesser degree, found support in more detailed descriptive studies. Thus, a number of studies have suggested that, in general, writing is characterized by a higher concentration of new information than speech (Stubbs, 1980; Brown & Yule, 1983).

A second generalization implied by studies comparing the two varieties is that the syntax of spoken language is much less structured than that of written language. Thus, spoken language is seen to contain many incomplete sentences, and often simply sequences of phrases (Crystal & Davy 1969). There is also a suggestion that spoken language typically contains little subordination (O'Donnell, 1974; Kroll, 1977), shorter sentences (Drieman, 1962) and few examples of it-clefts or wh-clefts (*It was John who..., What should be mentioned is...*) (Brown, Currie & Kenworthy, 1980).

Thirdly, speech is generally seen to be less systematic in its topic development, tending to meander and stray from the subject at hand. There are fewer logical connectors like *moreover, besides, however* and *in spite of*, and coordination is typically very loose (Leech & Svartvik, 1975; Beaman, 1984). In writing, discourse would seem to be more deliberately planned and topics more consistently pursued, as reflected in the presence of rhetorical organizers such as *firstly, more important* and *in conclusion.* These are very rare in spoken language (Ochs, 1979; Rubin, 1980; Akinnaso, 1982).

Fourthly, a number of studies have reported that speech is more personally involved than writing. Thus, much spoken language makes abundant use of first person pronouns, and references to thought processes in the forms of *I think* and *I mean* (De Vito, 1967). Conversely, writing is more detached and abstract (Blankenship, 1962; Chafe, 1985).

Finally, many studies have suggested that speech is less explicit than writing. Speech is seen to rely heavily on the intersubjectivity of speaker and listener (Markova, 1978), making references which would be obscure to a speaker who did not share the same knowledge or situation (Crystal & Davy, 1969; Kay, 1977; Olson, 1977)). In addition, speakers tend to use vocabulary which is very generalized, like *a lot of, do, things like that, stuff* (Brown & Yule, 1983).

The proliferation of information and ideas on the differences in speech and writing have generated a need for a more global analysis, which could account for all of the differences mentioned above. One such typology of discourse types has been developed by Chafe (1982). His model rests on two underlying assumptions: "speaking is faster than writing (and slower than reading)" and "speakers interact with their audience directly, whereas writers do not" (p. 36). These differences account for manifestations of a distinctive set of linguistic features in the two varieties, and the absence or presence of these features can help to classify a particular discourse type along two functional dimensions: fragmentation/ integration and involvement/ detachment. Some characteristic features of integration are nominalizations, attributive adjectives, and sequences of prepositional phrases and subordinate clauses. Fragmentation is characterized by the lack of these features. A characteristic feature of detachment is the use of passives, while involvement is marked by first person pronouns, references to mental processes, emphatic particles and direct quotes.

Measuring samples of speech and writing for these features, he finds that speech is generally more fragmented and involved, and writing is more integrated and detached. These findings are explained in terms of the process by which speech and writing are produced. Thus, speech is fragmented because there is little planning time, and it is more involved because speakers have the opportunity to interact directly with their audience.

The studies discussed in this section have two major weaknesses. First, they are based on a very small number of texts. For example, Chafe's observations were derived from 9911 words of spoken language and 12,368 words of written language (1985 36). It is highly unlikely that such a small sample can adequately capture all the features that co-occur in a given variety. In fact, it has recently been estimated that a prosodic analysis requires about half a million words of text, while an adequate syntactic analysis requires a sample of at least one million words (Leech & Beale, 1984). A larger corpus is needed to ensure that a sample is representative, and especially that no idiosyncratic features of a particular speaker/writer would bias the findings.

Secondly, they assign undue weight to the genres chosen for analysis. According to Biber, "most studies have compared only two genres, one spoken and one written, and many of these have not controlled for the communicative task represented by those genres" (Biber,1988:52-53). Thus, the observation that speech is unplanned may apply to face-to-face conversation, but certainly not to public speeches and academic lectures. Similarly, the observation that writing is detached may apply to prose, but not to personal correspondence. Biber's own analysis, which does control for the various genres within speech and writing, leads to the conclusion

that "there is no single, absolute difference between speech and writing in English" (1988: 199).

2.2 The corpus linguistics approach

To study bodies of language of the magnitude suggested by Leech and Beale (1984) and to control for communicative variables is simply not feasible if the texts are not computer-readable and the analytic procedures at least partially automated. In the past thirty years, a branch of linguistics called corpus linguistics (Leech and Beale, 1984) has been preoccupied with collecting statistically adequate samples of language, and developing the computational tools that are required for such large-scale formal empirical research to be possible.

In the context of linguistics, a corpus can be defined as "a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis" (Francis, 1982:7). A recent survey lists 39 such corpora which are available for analysis (Taylor & Leech, 1989). These corpora are useful for research in at least three respects. Most are large enough to be considered statistically adequate representations of the varieties sampled (Sinclair, 1982). They are computer-readable, thus enabling automatic identification of linguistic features. Finally, they represent a wide range of genres which are clearly classified according to communicative purpose.

The two pioneering corpora of written English were compiled separately according to the geographical variety they represent: American English is represented by the one million-word Brown corpus compiled at Brown University (see Francis & Kucera, 1964; Kucera & Francis, 1967), and British English is represented by a corpus of identical size, compiled at the universities of Lancaster, Oslo and Bergen

(Johansson, Leech & Goodluck, 1978; Johansson, 1986), known as the Lancaster-Oslo-Bergen corpus (hereafter LOB). Within this geographical division, text samples are further classified by genre. Thus both corpora, which were compiled in parallel fashion to ensure maximum comparability, contain 500 text samples of about 2000 words each, representing the following 15 genres: A. Press (Reportage); B. Press (Editorial); C. Press (Reviews); D. Religion; E. Skills and Hobbies; F. Popular Lore; G. Belles Lettres, Biography, Memoirs; H. Miscellaneous; J. Learned; K. General Fiction; L. Mystery and Detective Fiction; M. Science Fiction; N. Adventure and Western Fiction; P. Romance and Love Stories; R. Humor.

In addition to written corpora, there are several corpora of spoken language. Perhaps the most important is the London-Lund Corpus of Spoken English (Svartvik & Quirk, 1980; Svartvik 1990), consisting of 500,000 words of British spoken texts derived from the Survey of English Usage (Svartvik, Eeg-Olofsson, Forsheden, Orestrom & Thavenius, 1982). The London Lund corpus (hereafter LLC) is also classified into genres, with 87 texts of 5,000 words each from the following varieties: private conversations, public conversations (for example interviews and panel discussions), telephone conversations, radio broadcasts, spontaneous speeches and prepared speeches.

These and similar corpora have been used for a number of purposes. The most widely known perhaps is the compilation of word frequency lists. For example, Kucera and Francis (1982) have compiled a frequency list on the basis of the Brown corpus. This has clear educational significance, particularly for the grading of language teaching materials. Frequency lists are not only restricted to individual words, but to grammatical structures. Based on the *Corpus of English Conversation* (Svartvik & Quirk, 1980) and various other texts, a recent study has shown that the

future is more often expressed, in British English conversation, by the form *will* rather than the expected form *going to* (Mindt, 1986). When one compares these findings with the sequence in which future expressions are usually taught in British ESL materials, the potential value of this kind of research becomes clear. Other studies which have clear educational implications are the compilation of the COBUILD Dictionary, based on a 20 million-word corpus developed at Birmingham University (Sinclair, 1982; Renouf 1987). Recent studies have also looked at collocational patterns (Kjellmer, 1984) and word class combinations (Johansson & Hofland, 1989), which show not only how often words are used, but the linguistic environment in which they most often occur.

If corpus linguistics has made valuable contributions to the study of English in general, the same holds true for the study of specific varieties of English. Developing in a manner similar to non-corpus research, variety-specific studies have looked at variation determined by the language user, such as diachronic and geographical varieties, and variation determined by language use, such as written language, spoken language, and the range of written and spoken genres discussed above.

One major corpus which is concerned with diachronic variation is the Helsinki collection, which contains 1.5 million words of texts written from the 8th to the 18th century (Ihalaien, Kyoto & Rissanen, 1987). This corpus has inspired a number of studies, in particular some concerned with syntactical features such as referential pronouns, determiners and the auxiliary *do* (Kyoto & Rissanen, 1984; Rissanen, 1988). More recently, diachronic varieties have been examined by Biber and Finnegan (1988), who look at the changes in three genres, narratives, essays and personal letters, over the 18th, 19th and 20th centuries. The study shows that

personal letters have become much more involved and less informational over time; narrative texts have changed from extremely elaborated and explicit to much more context-dependent, and both narratives and essays have become less abstract.

The study of geographical variation has benefitted greatly from the parallel compilation of the Brown and LOB corpora, discussed above. For example Hoffland and Johansson (1982) have compiled parallel frequency lists for the Brown and the LOB corpus, revealing interesting differences in vocabulary use between British and American English. Numerous studies have looked at the syntactical differences between British and American English, such as the use of ellipsis (Meijs, 1984) and the distribution of personal pronouns (Kjellmer, 1986). In addition, there are several corpora which have been designed with the explicit intention of isolating a particular geographical variety. Among these, we count the Strathy corpus of Canadian English (Lougheed 1986; Fee, 1989), the Kolhapur corpus of Indian English (Shastri, 1988), and the Queen's University corpus of Spoken Northern Ireland English (see Taylor and Leech, 1989).

The third type of language variation, determined by language use rather than language users, has also attracted considerable attention in corpus linguistics. One aspect which has been extensively scrutinized has been the influence that medium exercises upon language use. Unlike many non-corpus studies, the samples analyzed are strictly controlled for communicative situation. Thus, the project *English in speech and writing* (Tottie & Backlund 1986) is based on two extremes within each variety: spontaneous conversation, derived from the London-Lund corpus, and expository prose, from the LOB corpus .

One interesting study has been conducted by Bengt Altenberg (1986), who examines the frequency and distribution of contrastive linking in the two media. He

shows that there are more contrasts in speech than in writing noting, in particular, a greater use of the word *but*. He explains this in terms of the interactive nature of conversation, which is often carried forward not in the form of questions and answers, but in contrasts in which speakers express differences of opinion, objections, and concessions. In other words, the coordinator *but* is a tool for defensive maneuvering in conversation. He also shows that speakers favour coordination over subordination; and when speakers do use subordination, they put it at the end of an utterance, using the subordinator *though*. This he explains in terms of real-time planning: developing Chafe's argument, he postulates that the speed at which speech is produced cannot easily sustain anticipatory constructions.

Ingegerd Backlund (1986) examines conjunction headed abbreviated clauses (*after seeing him, when in Rome*). She finds they are less common in speech than in writing. Again the notion of planned discourse is invoked as an explanation: "the same social act verbalized in planned discourse will be more compact than the unplanned version" (1986:53). Extending her analysis to other varieties within writing, she finds that abbreviated clauses are especially frequent in the Skills and Hobbies category of the LOB corpus. and may be seen as a distinctive stylistic feature for written instructions.

The notion that speech is more involved than writing finds support in a number of studies. For instance, Backlund (1986) finds that conditional *if*-clauses, which express speculation about past or future events rather than pure informational exchanges, are more frequent in conversation. And Tottie (1986) finds that contingency adverbials expressing cause and reason dominate in speech. This is explained in terms of the interactive nature of speech: "when an interlocutor is present, we are simply more solicitous of providing explanation, reasons, and causes.

We are anxious to justify not only our speech acts but any kind of statement we make concerning actions, thought, etc." (1986:112). Hermeren (1986) finds that there are many more expressions of volition in conversation. This is assumed to reflect the speakers' propensity to give free rein to their feelings in a face-to-face situation, where the audience can respond to these emotions. It confirms an earlier corpus based study (Tottie, 1982), which found that there were almost four times as many "mental" verbs in speech than in writing.

Britt Erman (1986) has studied the functions of conversation-specific expressions such as *you know, you see* and *I mean*. She sees them as revealing the relationship among interlocutors, or the speaker's assumption of the knowledge that he and the listener share. The notion of shared information also finds support in a study by Aijmer (1986), which focuses exclusively on the use of the word *actually*. She finds that it occurs 10 times as often in speech, especially in British speech, and explains it as follows: "the information which is available to both the speaker and the hearer represents the shared world. In this model *actually* conveys that the speaker's world and the shared world are the same. It establishes contact or intimacy and signals group solidarity" (1986:125).

A number of generalizations can be drawn from these studies, as well as non corpus-based descriptive studies. We have seen that speech is generally considered interactive, unplanned, involved, and situation-dependent; conversely, writing has been characterized as informational, planned, detached, and context independent. Drawing on previous studies, Biber (1988) has illustrated how each of these dimensions is associated with the co-occurrence of particular clusters of linguistic features. Because his study is the most comprehensive to date, I have decided to

adopt his framework, which he labels "multidimensional-multifeature", for the purposes of this study.

2.3 The multidimensional-multifeature approach

In his framework, Biber identifies six textual dimensions, each characterized by the presence or absence of a set, or cluster, of linguistic features. For example, there seems to be a consensus in the literature that writing is generally the more informational of the two media, and speech the more involved. Biber's first dimension is thus labelled "Involved versus Informational Production". A text which rates highly on the 'involved' end of the scale would presumably display the following linguistic features: many private verbs, contractions, first and second person pronouns, hedges like *sort of* and *kind of*, if-clauses, and emphatics like *for sure* and *a lot*. Conversely, a text which rates highly on the 'informational' end of the scale would be characterized by many nouns, attributive adjectives, prepositions and place adverbials.

The second dimension identified by Biber is labelled "Narrative versus Non-Narrative Concerns". Here, texts rating highly on the "Narrative" end display features such as past tense verbs, third person pronouns and perfect aspect verbs, to name but a few, while "Non-Narrative" texts display the absence of these features. Dimension 3 is labelled "Explicit versus Situation-Dependent Reference" and distinguishes between texts displaying phrasal coordination and nominalizations on the one hand, and time and place adverbials on the other. Dimension 4 is labelled "Overt Expression of Persuasion", featuring infinitives, modals and suasive verbs, among others. Dimension 5 is "Abstract versus Non-Abstract Information", and is based on the presence (for Abstract) or absence (for Non-Abstract) of features such as

conjuncts, passives and adverbial subordinators. Finally, Dimension 6 is entitled "On-Line Informational Elaboration", and is classified according to the presence or absence of features such as relative clauses in object or complement positions.

In his study, Biber applies this multidimensional model to a considerable body of English texts, consisting of the LOB corpus, the LLC corpus, and a collection of personal and professional letters. The results can be seen in Figures 1-6. The existence of such a comprehensive model makes it possible to re-phrase the question initially posed: "How do the linguistic features of electronic language compare with other varieties of English?" can now be understood to mean "Where does electronic language fit on each of the dimensions outlined by Biber?". In order for any predictions to be possible, it is now necessary to look into the situational features of electronic language in greater detail.

Figure 1 (Biber 1988)
Dimension 1:
Involved vs Informational Production

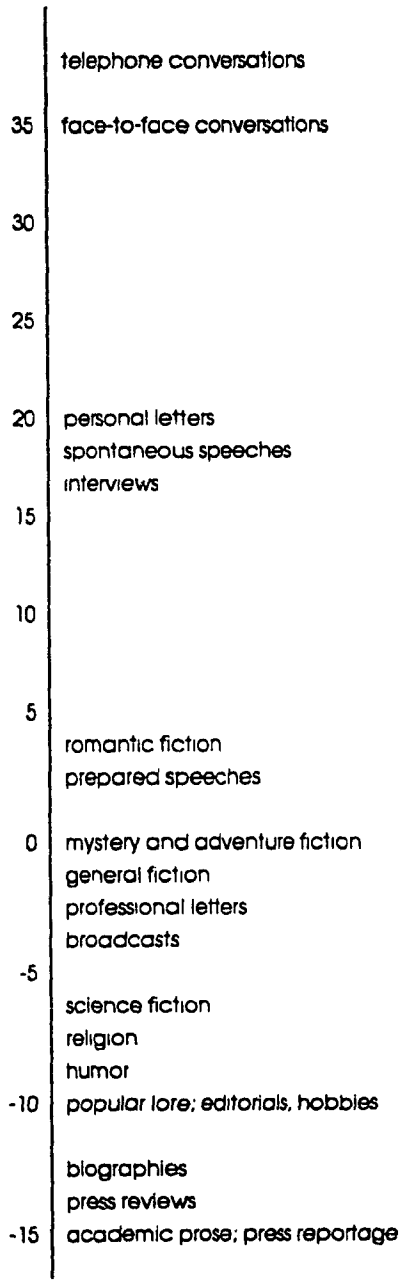


Figure 2 (Biber 1988).
Dimension 2:
Narrative vs Non-Narrative concerns

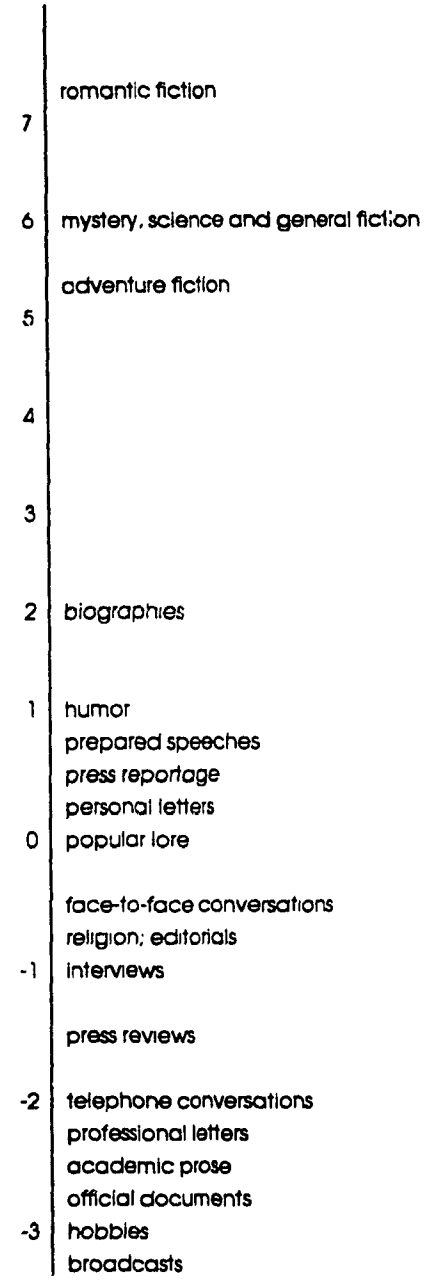


Figure 3 (Biber 1988)
Dimension 3:
Explicit vs Situation-Dependent.

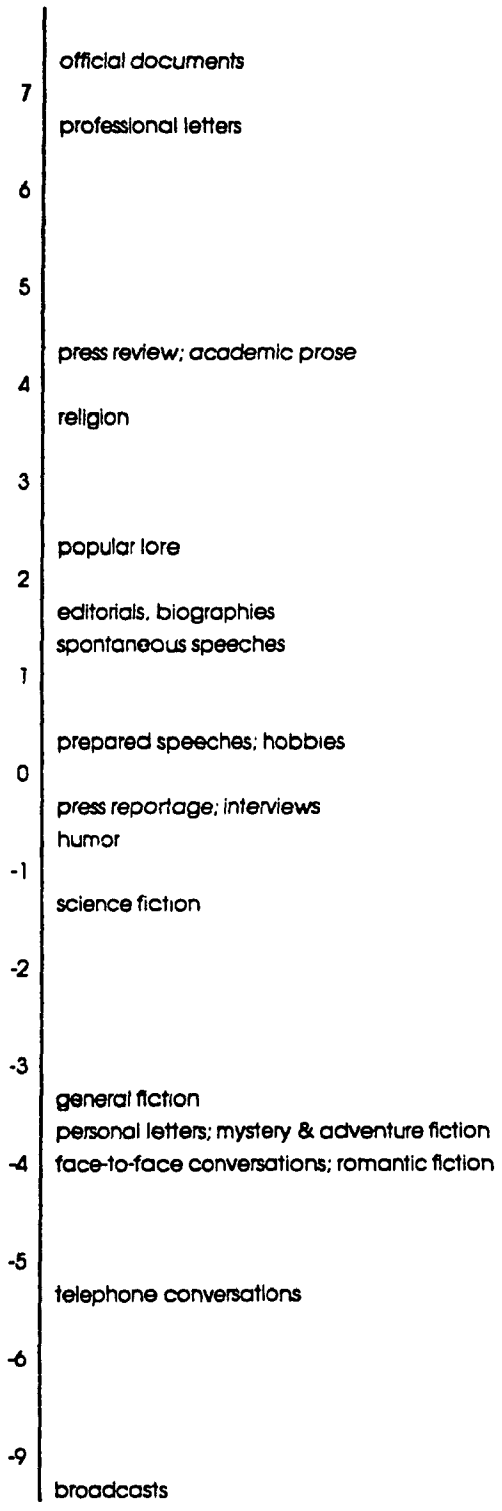


Figure 4 (Biber 1988)
Dimension 4:
Overt Expression of Persuasion

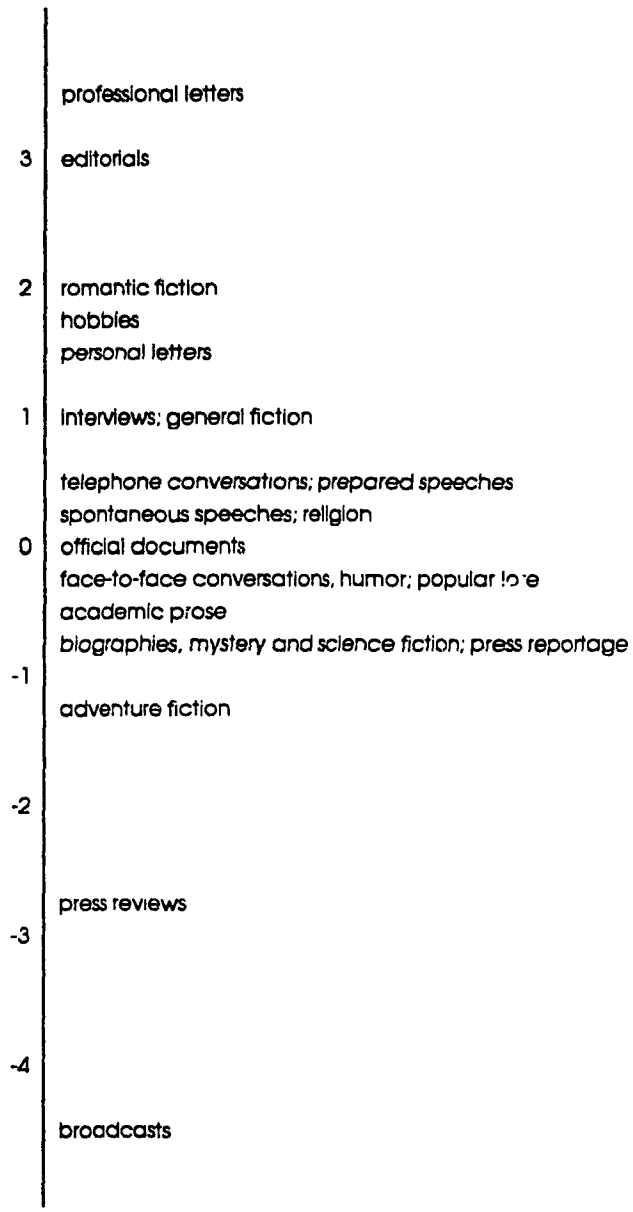


Figure 5 (Biber 1988)
Dimension 5:
Abstract vs Non-Abstract Information.

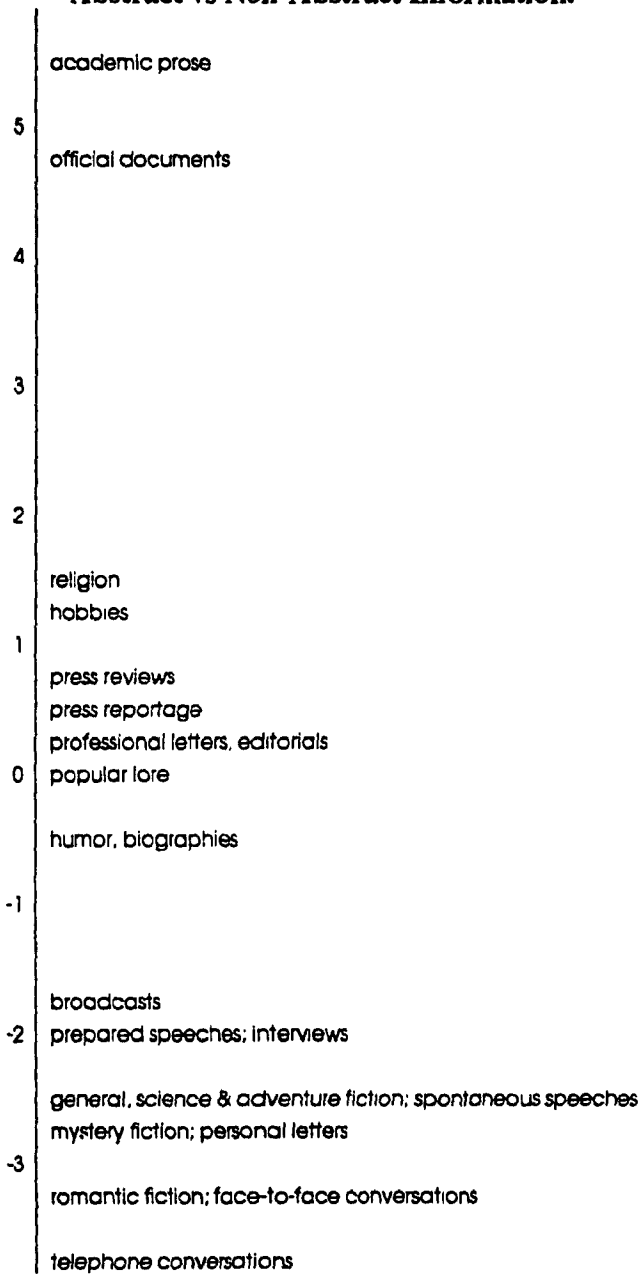
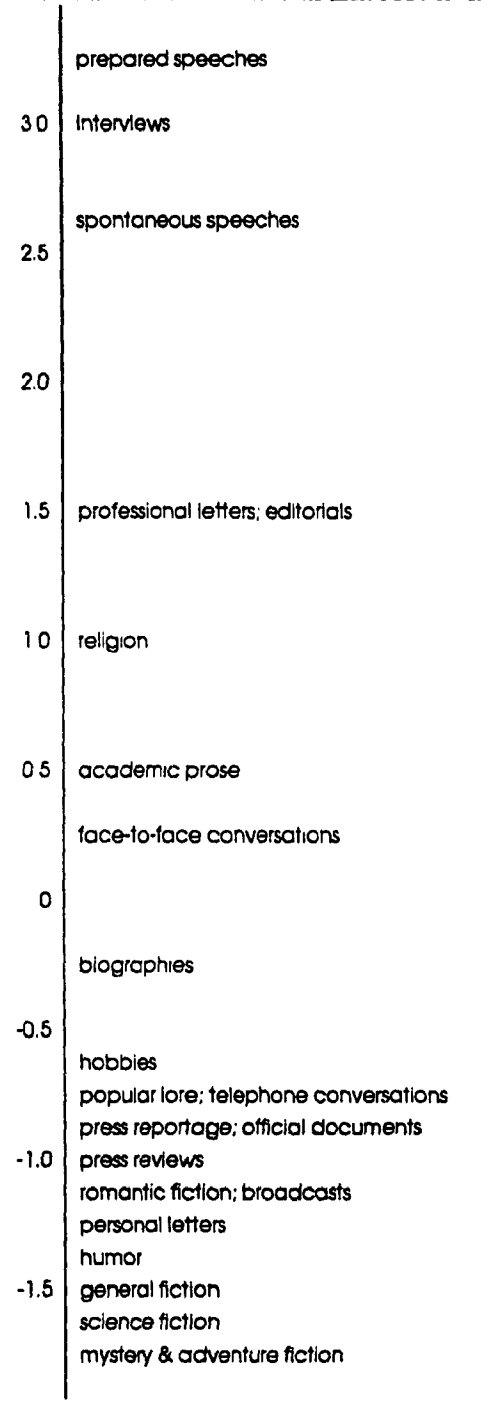


Figure 6 (Biber 1988)
Dimension 6:
ON-Line Informational Elaboration.



Chapter 3

FEATURES OF ELECTRONIC LANGUAGE

3.1 General Description

Writing uses some form of paper, speech uses sound waves and air. This study will examine language as transmitted by a relatively new medium: the computer. Besides a computer, the instruments required for electronic communication to take place are a modem, the appropriate communications software, and an electronic network to connect to. A modem is a piece of hardware which links a computer to a telephone line. When properly programmed, it will "dial" the phone, "listen in" to ensure that contact has been made, and let one transmit across the telephone line. Modems transmit by converting digital data into modulated, audio-like signals that can be carried over telephone lines as if they were voice communications.

A modem gives access to a vast array of so-called 'on-line' services. Generally, there are two major types: database services and communications services. Databases are updated banks of information that can vary from airline schedules to zinc prices. Communications services offer users an opportunity to exchange advice, opinions and information on a variety of subjects. This is the type of on-line service with which we are concerned.

Communications services allow users to post messages on the electronic equivalent of bulletin boards, known as bulletin board systems (BBS), to be read by people who dial in. All computers authorized to access the same service constitute a 'network', where individual computers function as 'terminals' to the 'host' computer. To connect, or 'log on', users simply program their modem to dial the number of the

host computer. While connected, network members type a message from their terminals, designate an audience by indicating the name of the addressee, and 'post' their message. After some time has elapsed, senders can log back on to see how many people, if any, have responded to the original message.

Contrary to what one might expect, on-line services are not intended only for computer professionals. While this association may have been partially accurate a few years ago, today the benefits of communications services can be, and are, enjoyed by business, professional, and home users alike. The subject matter is no longer restricted to computer technology, but covers a vast array, ranging from discussions on art, to political debates, to personal advice columns. The proliferation of subjects has led the people responsible for BBSs, called systems operators (or 'sysops' for the aficionados), to create the concept of a 'conference'. Each BBS conference, identified by an appropriate title, gathers all the messages with a common theme. For example, in one BBS, all messages concerning advice on personal matters have been grouped under the conference title "Dear Sue".

Messages are entered into the BBS chronologically, but participants always identify the message to which they are referring. This makes it possible to keep track of the "thread" of a conference. Most telecommunications software has a command that makes it possible to trace a thread: one types the reference number, and gets only the messages related to the original query, in sequence and without a miss. The topics often tend to wander and eventually wane, as a spoken conversation would. An entertaining and accurate account of electronic language recently appeared in Harper's Magazine, where the author describes telecommunications networks as a modern version of the old fashioned telephone party line (Kenner, 1989). The

notable difference, of course, is that interlocutors are not only speaking from different places, but at different points in time.

Following is a sample of an electronic conversation, excerpted from a conference entitled Chit Chat. Note how the speakers identify themselves and the audience they wish to address. Also worthy of notice is the way in which the conversation tends to meander, and the manner in which spatial disembodiment affects the participants' perceptions of each other.

```
Date: 02-04-90 (10:11)      CHITCHAT Number: 3 (Echo)
  To: SKIP BERTSCH                  Refer#: 679
From: CLIFF WATKINS                   Read: NO
Subj: HELLO

    Hello Skip! Are you from the Riverside area or another one of those
beautiful SoCal cities/counties?  I'm looking forward to visiting Cal
and hope to make it to the southern portion as well. Mostly I'll be from
S.F. and north. Anyway, your testing echo made it to NY.   bye...


Date: 02-05-90 (22:33)      CHITCHAT Number: 4 (Echo)
  To: CLIFF WATKINS                 Refer#: 3
From: JONATHAN NEAL                   Read: 02-06-90 (12:33) (Has Replies)
Subj: HELLO

I'm on the same board as Skip's and I can say that the cities here are
NOT beautiful..... with the exceptions of Riverside (my hometown), Palm
springs, and some mountain towns. The weather is not spectacular,
either. It has been raining the last three days, a high of 63. Oh
well....    Also, I am 12.
                          see ya....
                                    Jonathan

Date: 02-06-90 (07:48)      CHITCHAT Number: 6 (Echo)
  To: CLIFF WATKINS                 Refer#: 3
From: SKIP BERTSCH                    Read: 02-07-90 (09:58) (Has Replies)
Subj: HELLO

Thanks for the chit-chat! I do appreciate it. I am really excited about
this echo mail thing. Pretty new adventure for me really. I have been
using the modem for about 2 years, and never really got involved with
echo mail before. Well hope you enjoy yourself out here when you come
out. Is it gonna be biz, or pleasure when you come out? I was living in
```

the Palm Springs area, but recently moved back to Riverside. (Home town and all, and Grandma couldn't stand the baby being so far away)

Date: 02-06-90 (07:51) CHITCHAT Number: 8 (Echo)
To: JOHNATHAN NEAL
From: SKIP BERTSCH Read: NO (Has Replies)
Subj: HEY

Hey fella, how ya doing? I think it is great that your 12 and using the computer and modem and all. Great for you!!! I really love seeing kids use their minds and learn about computers. Your gonna need it in the future.

Date: 02-06-90 (21:47) CHITCHAT Number: 10 (Echo)
To: SKIP BERTSCH Refer#: 685
From: JONATHAN NEAL Read: NO (Has Replies)
Subj: HEY

Thanks. I nave been involved with computers for about 6 years, but I've only been involved with modems about 8 months. I usually try to hide the fact that I am 12 over the computer. It fools a lot of people who have been on the BBSs when I tell them my age. I get a kick out of it.....

Date: 02-07-90 (10:01) CHITCHAT Number: 12 (Echo)
To: JONATHAN NEAL Refer#: 687
From: CLIFF WATKINS Read: NO
Subj: HEY

>fact that I am 12 over the computer. It fools a lot of people who have
>been on the BBSs when I tell them my age. I get a kick out of it.....
Try the TEENS conference. I'm sure they'll enjoy having you in there.

Date: 02-08-90 (05:18) CHITCHAT Number: 14 (Echo)
To: ANDY PEA Refer#: 690
From: CLIFF WATKINS Read: NO
Subj: HELLO

Welcome Andy! Glad to see you made it. Talk with you soon...
Now you just need to get the NET.ID (not to be confused with the ID.NET) file to put in this space the network tag. You can request it from Dave.
\/

Date: 02-08-90 (03:01) CHITCHAT Number: 20 (Echo)
To: ALL
From: WILLIAM PADILLA Read: YES
Subj: INTERLINK

On my Bank of America monthly checking account statement, under "other debits", I noticed the following entry:

"Interlink Network trans 660555 on 02-01; customer payment to Lucky store no. 465; Glendale, CA; $82.27"

After staring at it thoughtfully for a moment or two, it finally hit me that
the reason I had been so "fascinated" by it was because of its reference to "Interlink Network".

Hmmm.... wonder if the Interlink folks are getting into electronic banking services, perhaps to help raise some much-needed funds....???

<Grin>

The example above illustrates how electronic language combines some characteristics traditionally associated with speech and others associated with writing: it can be as interactive as spoken conversation, but the speech event takes place at different points in time and space, as is the case with writing. Given this situation, a study of an ELC would seem to call for a comprehensive model of language description, which goes beyond the traditional distinction between reading and writing. As indicated by the review of the literature, the best candidate would seem to be the model proposed by Biber (1988). To my knowledge, it is the most recent model, and also the most comprehensive.

3.2 Situational features of the ELC

Biber distinguishes eight components of the speech situation. Table 1 summarizes my description, adapted from Biber, of the speech situation in electronic language. The first component concerns the roles of the participants, as well as their personal

and group characteristics. Electronic language involves three types of participants: an addressor, an addressee and an audience. Messages can be sent by individual addressors to a specific addressee, or to the BBS public at large. In the first case, the speaker will simply type in the name of the person he wishes to address; in the second he will press the ALL option on his menu. In either case, the messages will appear for all to be seen. There are therefore no private messages as such.

The roles of the participants are clearly defined. As illustrated by the example above, each message is introduced by a summary of who the message is for, who it is from, and which previous message, if any, is referred to. However, very little is known about the personal characteristics of the participants. Since the corpus contains messages from a vast number of people, no participant's personal style is expected to have a salient effect on the overall language used.

The group characteristics of the participants, for example geographic, demographic and socio-economic distinctions, are also quite varied. The corpus includes messages from across Canada and the United States, so that no one dialect is favoured over any other within the wider variety known as North American English. There are no obvious restrictions of age: anyone old enough to use a computer can have access to the boards. One would assume that the subject matter would restrict it to more adult participants, but as shown by message #4 in the excerpt above, this is not always the case. As for occupational status and social class, computers have become such a pervasive feature of many peoples' lives, that the ownership of one would include practically all strata of mainstream society. The minimal educational requirements are for people to be literate and to know how to type.

Table 1.
Components of the speech situation in electronic language, adapted from Biber (1988).

PARTICIPANTS	
roles	addressor, addressee, audience
personal characteristics	diverse
group characteristics	Canadian and American
RELATIONS AMONG PARTICIPANTS	
social relations	egalitarian
personal relations	friendly
degree of shared knowledge	high
SETTING	
physical context	participants' own desktop
temporal context	any time of day or night
extent to which space and time are shared by participants	none
TOPIC	topics are classified into "conferences", and vary accordingly
PURPOSE	To request and give information; to make announcements, to engage in discussion
SOCIAL EVALUATION	
attitude towards communicative event	acknowledgement of the unique nature of the event
attitude towards content	varies according to topic
RELATIONS OF PARTICIPANTS TO THE TEXT	
type 1	planned text, prepared beforehand
type 2	unplanned text, composed on-line
CHANNEL	writing

A second component involves the social and personal relations among the participants. In this respect, the ELC is fundamentally different from spoken language. The kinds of relations which in ordinary speech are established when the person's age group or social status is known are non-existent here. For one thing, the participants do not see or hear each other. Thus, it is impossible to guess one's age, or infer one's status, as we so often do, from physical clues. In the excerpt above, the

surprise caused by the revelation that one user is only 12 years old illustrates this point quite clearly. Furthermore, participants who decide to join a "conversation" are not requested to introduce themselves, or be introduced by a mutual friend who will inform the others about a person's station in life. Rather, participants are joined primarily by a common interest in the subject matter, and social and demographic features rarely seem to play any role. Even gender, though it can usually be correctly inferred, may have an attenuated role.

Personal relations are kept at an intermediate level of friendliness, rarely touching the extremes of like and dislike. Because of the invisible but ever present audience, it would seem improbable for two participants to engage in intimate, passionate discourse. Similarly, there are rarely occasions where participants display an obvious dislike for each other. Unlike face to face conversation. participants are rarely forced into contact with people they do not like. If someone bothers them, they simply do not respond to their message. The signals we tend to send when we are with someone we don't like and would like to shut down communication are not necessary here. There are occasional flare ups on controversial issues, but sysops generally intervene to moderate them and keep the tone civil.

Inevitably, some form of friendliness develops. People who have been using a bulletin board for a while know each other's nicknames, mannerisms and ideas. They have followed each other's arguments on many different subjects, and have accumulated a wealth of shared knowledge. Even people who are new to the board know that their audience will be generally sympathetic because they are bound to them by common interests. The BBS makes for a special kind of intimacy, not often found in other varieties. The messages are similar to personal correspondence because of the shared knowledge and friendly tone. Yet they differ from it because

there is always an audience, similar perhaps to those who read the editorial page of a newspaper, especially the audience for letters to the editor.

Another important component is the scene, which can be described in terms of setting, topic and purpose. The conversations are highly discontinuous in terms of time and place. By definition, no two participants can be at the same place, since they are connected by a telephone line. For each participant, the "place" in which the conferences occur is one's own desktop. Similarly, there is no telling how much time will elapse between one message to the next. A case was recently cited in a computer magazine where someone typed his query, got up for a glass of water, and came back to find one response already on his screen. Although the messages are never separated by more than a year, neither are they simultaneous.

The topics vary with each conference; those included in this thesis will be described in greater detail in chapter 4. The purposes are as varied as the topics. The conventional purposes are generally to request and give information, to make announcements, and to engage in discussion about specific issues. The personal goals are often to establish human contact. Many advertisements for BBS systems emphasize their friendly and relaxed atmosphere, as well as the expertise of their members.

The next component is social evaluation, which refers to the attitude of the participants to the communicative event and the content of the message. We can infer that participants will confront the event with a positive attitude, as there is no social obligation (as in many phone calls and conversations) or professional obligation (as in written reports) to participate. Moreover, one detects an overall feeling of enthusiasm for this new form of communication, a sense that it sets the participants apart from most other language users. For example, the author of

message #3 above regards using the BBS as an "adventure". This sense of exclusiveness occasionally leads to sectarian, sometimes snobbish attitudes towards participants who do not conform to the norms established by BBS user. An example of this kind of attitude is given in Appendix A, an excerpt from the CHAT conference in which regular BBS users react to the intimations of a newcomer that their grammar is substandard.

The participants' attitude towards the content will vary from conference to conference, from keen interest in the more technical subjects, to passionate involvement in the more controversial issues, to jovial light-heartedness in the more conversational issues. In general, these messages display the degree of commitment often found in spoken conversation, which varies according to the seriousness of the topic. For example, the message reproduced above have much of the flavour of spoken chit-chat, where speakers easily stray from the subject.

A component which had been largely ignored prior to Biber refers to the relations of the participants to the text. Developing the observation that speech is faster than writing (Chafe, 1982), Biber states "The writer can write as slowly and carefully as (s)he wishes; the reader can read as quickly or as slowly as (s)he wishes; but speakers and listeners must produce and comprehend language 'on-line', with little opportunity for interaction with the text" (1988: 33). In the ELC, this relation is complex. Writers have two options. They can read and write their messages directly on-line, in which case they are operating under real time constraints. Or they can copy those messages onto their computer (a process know as "downloading"), read them and write a reply at their own pace, and then send the finished messages to the BBS when they are ready. For this thesis, I have separated the on-line messages from

the pre-written ones, in order to determine the impact of this particular component on the speech situation.

Both of these options have a feature which distinguishes electronic language from ordinary spoken discussion. The messages are invariably completed and concluded before any answer can be expected, as each writer has to 'sign off' at the end of an intervention. Thus many devices that are common in conversation for turn-taking, interrupting, and other kinds of maneuvering are rendered impossible here.

The last component in Biber's model is the channel of communication. The primary medium is the electronic network, and the form which language takes is writing. Because messages are sent in written form, there are very few sub-channels available. For example, prosodic features are necessarily absent. Even the traditional written equivalents to prosody, such as bold facing, italicizing and underlining are not available in this variety, as these devices do not appear on network boards. However, the paralinguistic sub-channel, usually associated with spoken language and expressed by gestures or facial expressions, is occasionally used in the ELC. For example, in message #20 above, the sender supplies a paralinguistic feature by actually typing the word "grin" and offsetting it in quotation marks.

Chapter 4

METHOD

In very general terms, the global design of this study can be classified into five major steps. The first step was to gather the data. This task was greatly simplified by the fact that an ELC is, by definition, already in computer-readable form, and no time needs to be wasted in manually entering the data into the computer. The question was mostly one of data selection, rather than collection proper. Secondly, the corpus was automatically formatted in such a way that it could be read by CLAWS 1 (Garside, 1987), an automatic parsing program which assigns each word in the corpus to a word class. Step three consisted of coding the corpus, manually or semi-automatically, to facilitate automatic parsing. Step four consisted of actually running the corpus through the series of programs which make up CLAWS 1. The final step involved extracting the frequencies of all of the features which characterize the textual dimensions of language described in Biber.

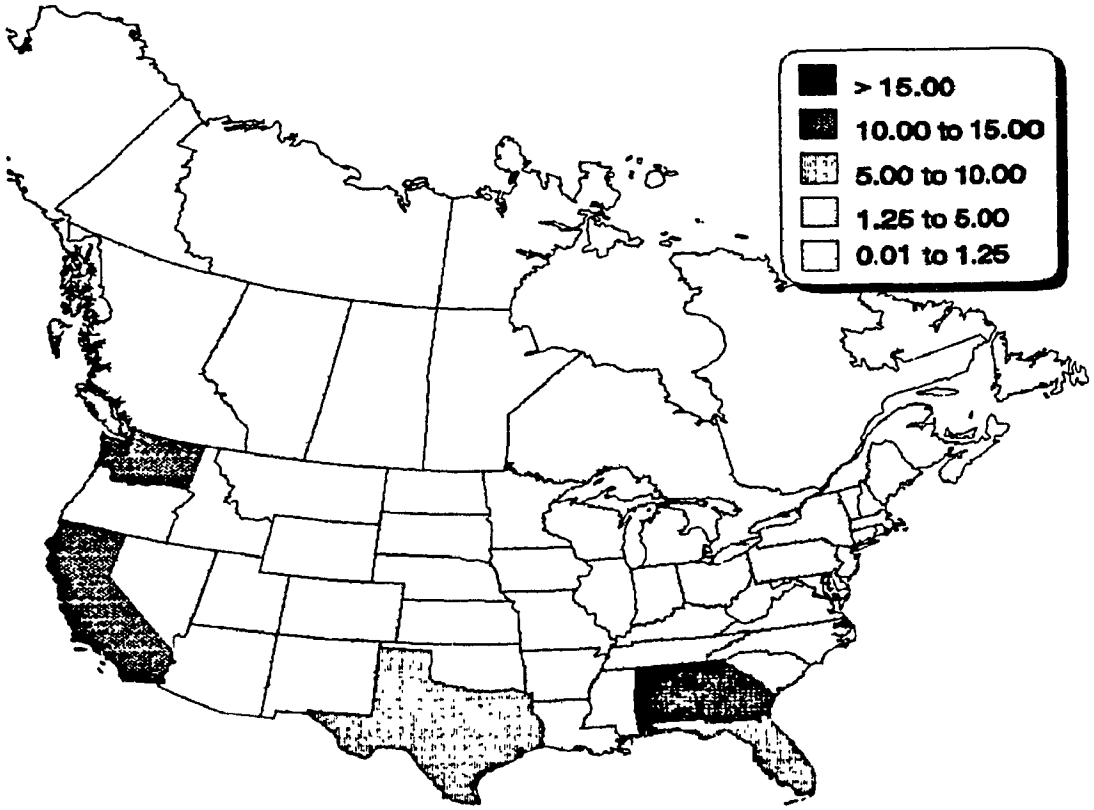
4.1 Text selection

Because the aim of this study was to compare the ELC with other varieties of English, and more specifically with the varieties analyzed by Biber, it was important to ensure that the ELC would indeed be as comparable as possible to the LOB corpus, the LLC corpus, and Biber's collection of personal and professional letters. To ensure maximum comparability, a number of features had to be controlled.

The first feature to control for was regional distinctiveness. Though unlikely, it is conceivable that any differences between Biber's corpora and the ELC would be a result of a highly localized, regional dialect. In order to minimize this possibility, I

decided to use only those texts which were downloaded from an international BBS called Input Montreal. Although the BBS is located in Montreal, it is international in the sense that it gathers and posts messages from across Canada and the United States, channeled through the larger networks, Fidonet and Metrolink. As a result, Montreal speakers happen not to be represented in this corpus. Figure 7 illustrates the distribution of the messages across North America.

Figure 7: Geographical provenance of BBS messages (percentages)



The second feature to control for was the topic. This was done in order to avoid the possibility of attributing any difference which may be found between the ELC and other varieties to the fact that they differed in subject matter. From the 86 conferences which appeared on the menu of Input Montreal, I selected those which

would be as comparable as possible to the LOB and LLC corpus in terms of subject matter. To this end, I deliberately rejected any conferences that were exclusively concerned with computer hardware or software. It is important to note, however, that this measure minimized, but did not eliminate, the amount of computer jargon used in the corpus. In fact, computers remain an important concern for many of the participants in BBS conferences and the general conferences, such as 'Chit-Chat', are replete with queries and discussions related to computer equipment.

With these constraints in mind, I selected the following 9 conferences. The first conference is entitled *Chit-Chat*, and is really an agglomeration of four minor conferences: Netchat, Gossip, Late Nite and Students. As the conference title indicates, its major purpose is to allow participants to chat, and there is no major restriction on the subject matter. As such, the messages here are probably most comparable to the face-to-face or telephone conversations included in the LLC.

The second conference, *Current Events*, includes discussion about issues in the news. The topic is popular, and the aim is primarily suasive. Many of the messages contain well developed arguments for or against particular issues, and as such is comparable with the LLC category entitled "Public Debates". The third conference *Science*, is a general discussion on the physical sciences by science students, professionals and laymen. The topic is academic and the purpose is expository; as such, it is loosely comparable to the academic prose genre included in the LOB corpus. Another topic which is also represented in the LOB corpus is the conference *Science Fiction* which, by specifications of the sysop, features light conversation about science fiction books, movies, authors and television shows excluding Star Trek (which forms the subject of a special conference). The fifth conference,

Finance, is a highly informational forum on the stock market, investments and interest rates, a category also represented in LOB.

The sixth conference is entitled *Film and Music*, and consists of discussions about the latest developments in these two areas. Its concern is primarily cultural, and it is represented in the LOB corpus as a sub-category, entitled "cultural", of the press genre. The seventh conference consists of two minor conferences entitled *Photo* and *Cooking*. "Photo" includes information and discussion on camera equipment, film development and printing; "Cooking" concerns questions and information regarding food preparation. The primary purpose of this conference is for non-professionals in the two fields to seek information and give instructions. It is represented in the LOB corpus by the category "Skills and Hobbies".

The eighth conference is entitled *Medical*. Topics here include all health related issues as well as a discussion of new methods of treatment for various ailments. The conference is open to everyone, and the issues addressed are consequently not highly specialized. Although there are some instances in which participants give each other advice, the purpose seems to be primarily expository. The topic of medicine is also included in the LOB corpus, under the general heading of "academic prose". Finally, the ninth conference is entitled *Sports*. This conference includes discussion of sports in general, as well as specific information about particular teams. As such, it is comparable to both the sports section of LOB's press category and the sports section of the LLC's broadcast category.

Table 2 indicates the size of each conference and of the corpus as a whole; it also gives an indication of the relative popularity of the conference, indicated by the number of participants, as well as an impression of the average size of the messages, indicated by the ratio of number of words to total messages. In determining the total

size of the corpus, it was important to ensure that the ELC was large enough to be representative of electronic language in general. In the LOB corpus, the average size for each of the 15 genres is 66,000 words, for a total of 1 million words in the entire corpus. In the LLC, each of the 6 spoken genres is represented by 83,000 words, for a total of 500,000 words in the corpus. An appropriate size for the ELC, which incorporates two genres, would therefore seem to be at least 150, 000 words. The finished size of my corpus, once it was formatted and stripped of redundant information as described below, was 195,693 words.

Table 2.
Conference profiles of the raw corpus.

CONFERENCE TITLE	MESSAGES	AUTHORS	WORDS
Chit-Chat	656	76	17709
Current Events	301	28	21836
Science	272	77	28311
Science Fiction	388	82	26472
Finance	265	75	25103
Film and Music	453	94	25055
Photo and Cooking	568	109	24764
Medical	273	101	10443
Sports	404	96	26000
TOTAL	3580	738	205693

A final feature to control for was the question of medium or, more specifically, the relation of the participants to the text. This is especially important in light of the observation, conceptualized by Malinowski (1935) and developed by Firth (1959), that a variety is determined not by single, isolated features, but by the entire context of situation. In our case, the context of situation involves, *inter alia*, features such as spatial and temporal disembodiment, the anonymity of the participants and the friendly relations among them, and cannot be reduced to the single question of medium. In order to avoid confounding medium with the overall context of the

speech event, it was considered essential to separate those messages which were written on-line from those which were written beforehand, or off-line.

Table 3.
Conference profiles broken down into Other and Off-line categories.

OTHER				OFF-LINE		
MESSAGES	AUTHORS	WORDS	CONFERENCE	MESSAGES	AUTHORS	WORDS
464	55	12594	Chit-Chat	192	21	5115
257	22	9245	Current Events	44	6	12591
137	56	12587	Science	135	21	15724
198	48	15535	Science Fiction	190	34	10937
166	45	14460	Finance	99	30	10643
220	67	14333	Film and Music	233	27	10722
396	81	12510	Photo and Cooking	172	28	12254
210	69	5639	Medical	63	32	4804
288	79	18715	Sports	116	17	7285
2336	522	115618	TOTALS	1244	216	90075

Off-line messages can be positively identified by the presence of a tag, inserted after the last line of a message, which indicates that an electronic "mail reader" was used. The function of these mail readers is to compress a message in order to speed up transfer to the host computer, and they are used by people who have a pre-written message to deliver. The most common types of off-line readers, which act as signatures to off-line messages, are readers like EZ Reader, Megamail, and Quickmail. The second category, consisting of on-line messages, is not as robust as the first. Here, it was assumed that when messages did not have positive evidence of being written off-line, they were on-line. This is a reasonable assumption to make since off-line mail reader programs always add their "signature" to messages written with them, as a form of advertisement. However, there is always the possibility that certain messages were pre-written using an ordinary word processor or editor in a multi-tasking environment, in which case a mail reader "signature" would not appear. In an attempt to solve the problem, I decided to label the two categories as "Off-Line"

and "Other", where the latter category includes, but is not limited to, on-line texts. The resulting configuration of the corpus is depicted in Table 3.

4.2 Automatic Formatting of the Corpus

Once the corpus was selected and transferred from the network to my personal computer, it was necessary to follow a long series of steps to adapt the corpus so that it could be read by CLAWS 1. Figure 8 is an example of what the messages looked like when first downloaded from the BBS.

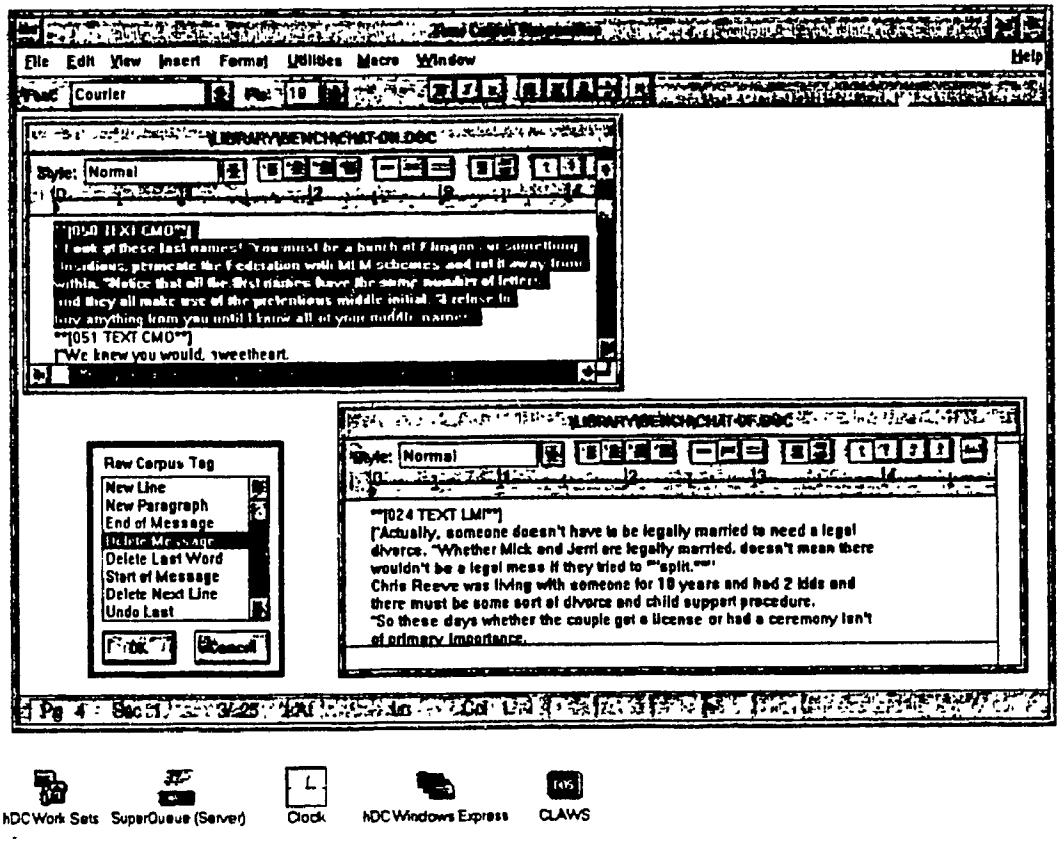
Figure 8: Extract from the Science conference as first downloaded from Input Montreal.

```
==========================================================================
=
Date: 01-12Ã91 (23:15)          Number: 9240      ⊕ Input Montreal BBS
  To: MAURY MARKOWITZ           Refer#: NONE
From: DAVE KNAPP                  Read: YES
Subj: RE: EVERETT                 Conf: (42)
--------------------------------------------------------------------------
>  A while ago (a month I'd say), we were involved in a short
>discussion in which you showed me an experiment which could very well
>show that Everett's "multiverse" was not possible.

  I did?  Are you sure?  This is surprising to me since I have no
idea what Everett's multiverse is.  Maybe we talked about it without
using the name.  Anyway, if you explain and I remember I'd be glad to
post anything I can remember.  -- Dave
--- TBBS v2.1/NM

 * Origin: Diablo Valley PCUG-BBS, Walnut Creek, CA 415/943-6238
```

At this stage, the raw data were run through a series of programs, specifically written for this study and illustrated in Figure 9, which automated much of the editing necessary to prepare the texts for CLAWS 1. Written in WordBasic, the programming language built into Microsoft Word for Windows, the functions of these programs can be divided intro three major steps. First, each message was

checked individually for the presence of an off-line mail reader "signature", which always indicates that a message was composed prior to connecting with the network. Signatures were cross-checked against a list of popular off-line mail readers and, when found, such messages were moved to a separate file.

Figure 9: WordBasic raw corpus preparation (semi-automated procedures)



The second step was to remove all excess information, specifically quotations of previous messages, excess spaces and lines, and the message "headers" which appear before each message. The deletion of quotations from prior messages was a relatively simple matter to automate. Since only a few characters are used to indicate an extract or "quote" from a prior message, it was sufficient to simply check the first two or three characters at the beginning of a line of text for the presence of a

"quote" character (such as ">" or "»"), and, when found, to delete the entire line of text. It is also common to use the high-order "extended" ASCII character set as "quote" characters, and since these characters were automatically converted into their equivalent low-order ANSI characters, typically as foreign language symbols, the location and deletion of quotations from prior messages was easily automated. The removal of excess characters was, likewise, a simple programming task, requiring little more than ordinary "Search and Replace" procedures (e.g., search for multiple exclamation points and replace with a single exclamation point).

The next stage in the automated procedure involved extracting and indexing telephone area codes for the originating bulletin board where a message was first posted. Although not mandatory, it is common for a bulletin board to attach its own "signature" to every message originating from it, prior to "echoing" the message across the international network, as a form of advertisement. For my purposes, these area codes serve as a rough index of the geographical diversity of authors. These area codes were indexed as hidden text, so that they form no part of the corpus itself.

The next step was to delete the message "header" which appears at the beginning of every message, indicating its date, the sender and intended recipient, the conference and the specific subject with which the message is concerned. Again, for the purposes of this study, the only significant information in the "header" is the author, whose name was automatically indexed by the macro-program by extracting the first letter of the author's first name and the first two letters of the last name. Once concatenated and indexed in this format, the resulting three characters were incorporated as part of the "Begin Corpus Text" tag as prescribed by CLAWS 1, along with a sequential corpus text number, and the message "header" was deleted in its entirety.

Once all extraneous information was either classified or deleted, the program began inserting three of the many symbols and meta-symbols which CLAWS 1 expects: paragraph markers, begin quotes and quotes, and markers for the beginning of sentences. For paragraph markers, the program inserted the symbol | at the beginning of every line that had been deliberately preceded by a blank line. Similarly, the macro-program located all quotation marks, determined which should be interpreted as begin quotes and which as end quotes, and then replaced them by the CLAWS 1 symbols for begin quotes (*") and end quotes (**"). Quotation marks would be manually confirmed at a later point to determine whether they enclosed entire sentences or simple words or phrases. For sentence initial markers, the program inserted the symbol ^ after every question mark, exclamation mark and period which was followed by a deliberate space and additional characters. Thus a series of ellipses would not be automatically tagged as signalling the end of a sentence, since a previous routine of the macro-program had already standardized all ellipses to consist of only three periods, with no spaces before or after their occurrence, even though ellipses sometimes do indicate the trailing end of sentence. On the other hand, the signature of an author (eg. "Dave") or sign-off (eg. "TTYL" -- an abbreviation for "talk to you later") would be marked as the beginning of a new sentence, and would require manual deletion at a latter point as grammatically irrelevant text. I will discuss these issues in more detail below, under the heading "Semi-Automatic Coding".

A final series of routines which cleaned up the corpus, by searching for and deleting such things as contiguous paragraph marks (|^|^) and the strange symbols authors sometimes use to "personalize" their messages (eg. boxed names), completed

the fully automatic section of the preparation. Figure 10 shows the message in figure 8 after it had been edited by the formatting program.

Figure 10: Extract from the Science conference after automatic formatting. It should be compared with Figure 8, which shows the same message before formatting.

```
==========================================================================
**[001 TEXT DKN**]
|^I did?  ^Are you sure?  ^This is surprising to me since I have no
idea what Everett's multiverse is.  ^Maybe we talked about it without
using the name.  ^Anyway, if you explain and I remember I'd be glad to
post anything I can remember. ^Dave
```

4.3 Semi-automatic coding of the corpus

Once the corpus had run through this initial program, the task of coding could begin. My principal source of information was the *Lob Manual PRE-EDIT Handbook* (Atwell, 1981), which indicates which symbols and meta-symbols the CLAWS tagging programs expect. At this point, a terminological distinction, not often made in the literature, seems to be in order. CLAWS 1 is a set of programs which assigns each individual word in a corpus to a specific grammatical word class; its function will therefore be referred to as "tagging". However, before CLAWS 1 can read and tag a given corpus, the corpus itself must include a set of symbols which will, for example, tell it where one sentence stops and the other begins, which words it should ignore, and so on. This function, i.e. adding the symbols and meta-symbols required by CLAWS 1, will be referred to as "coding". The decisions I made during this stage will therefore be referred to as "coding decisions", and are quite distinct from the "tagging decisions" that I made after the corpus had been run through the tagging programs.

Although much of the coding can be done after CLAWS has begun its work, I decided to follow Belmore (1991) and code as much of the corpus as possible before running CLAWS 1. Table 4 lists the symbols and meta-symbols which I had to include.

Table 4: CLAWS 1 symbols

FEATURE	CODE	EXAMPLE
1. new sentence	^	^High speed modems are new for me...
2 new paragraph	|	| ^High speed modems are new for me. .
3 dash	*-	^I am a programmer *- can I offer my services?
4. quotation marks	** ***	^It's strange to hear Disch referred to as a **big name***.
5. included quotations	-** ***	^Cuffy says: -**Because of the critical role which transitional fossils played in convincing scientists of .***
6. abbreviations	\0	^He could give \0Mr Spock a hard time
7. sequence of initials	|0 |	^A |0CFC| can be volatile but not explosive
8. non-standard word	\2	| ^You \2gotta be kidding
9.non-standard sequence	|2 |	|2Usen's here ain't got no reel 'spect fer no dem queen imposin' her english on usen's|
10 technical jargon	\4	| ^Personally, I prefer \4ZModem.
11 nonce forms	\5	| ^<snort>. ^You always did have a low \5offensensitivity index
12 paralinguistic	**(PARALINGUISTIC**)	| ^That is a misspelling, you twit | ^<grin>) <grin> **(PARALINGUISTIC**)
13. foreign word	\6	^Cook the \6roux for 5 minutes
14 foreign sequence	|6|	^Remove the |6 bouquet garni|

As mentioned earlier, steps 1, 2 and 4 were implemented during the formatting phase and required no manual intervention. Step 1, marking the beginning of sentences was done largely by machine, except for cases where sentences were followed by three periods (...). In these cases, it was impossible for the program to know whether the following utterance constituted a new sentence or not. Consider for example the following sentences, excerpted from the Chit-Chat conference:

someone here is putting words in my mouth ... or my modem.

So I bought her a new pair... By the way, I don't suppose you've heard from Chevron lately...

Ellipses points often represent a trailing off of thought, and are used quite frequently in the ELC. The problem I was confronted with was to decide whether this pause was momentary or permanent. In the first case, what follows the ellipses points (read by the computer as periods) is merely a continuation of the previous utterance, and cannot be considered a new sentence. In the second case, the thought trails off, but this time it is not picked up. Given this situation, I decided to mark as new sentences all those utterances which represented a thematic shift from the utterance which preceded the ellipsis points. Conversely, all those which continued the preceding thought were left uncoded.

Steps 3 and 5 were also done semi-automatically by a search and replace routine. For step 3, all dashes which appeared in the corpus were replaced by the symbol *-. These were then manually inspected to determine whether they were indeed used as a punctuation mark, or simply as a hyphen with extraneous space. In the latter case, the asterisk and spaces surrounding the hyphen were removed.

Decisions regarding included quotations, or step 5, were somewhat more complicated. Initially, a search and replace routine simply replaced the existing quotation marks by the symbols for begin and end quotes, as described above. The quotation marks were then called up again to determine whether they enclosed simple words or expressions, or entire sentences. In the latter case, the sentence was marked as an included sentence. This distinction should make it possible to distinguish between utterances produced by participants in the ELC from those which are quoted from other sources. This was especially a concern for the science conference, in which participants tend to substantiate their claims by quoting from previously published material, thereby creating a possible source of confusion between "true" electronic language and language which was created in other contexts.

An additional complication was that, whereas the Manual calls for a distinction between single and double quotes, participants in the ELC tend to alternate between the two conventions quite arbitrarily. In order to avoid unnecessary complications, I decided to code both single and double quotes with the same symbol.

The remaining steps, 6 to 14, are the ones which required the most manual intervention, and were consequently the most time consuming. To correct all obvious spelling and typographical errors, as required by section 16 of the manual, also took a lot of time. Much of the coding in this section was done with the aid of a software program called *Grammatik Windows*, which is designed to check a given text for grammatical appropriateness. This program was invaluable not only in flagging obvious spelling and typographical errors, which occurred very frequently in the corpus, but also in isolating words which it did not recognize, such as abbreviations, nonce forms, and so on. Furthermore, it corrected simple grammatical errors which would have been impossible for CLAWS 1 to interpret, such as the use of *your* where *you're* is intended, as in the sentence, previously cited, *I think it's great that your 12 and using the computer.*

Although *Grammatik* helped speed up the coding process, it did not exhaust the work to be done. In the case of abbreviations, for example, *Grammatik* will not flag words such as Dr., Mr., or U.S.A., and yet all of these are seen by the authors of CLAWS 1 as abbreviations (see manual, section 13). Therefore, after all the spelling was corrected, I had to return to the corpus to code it for steps 8 to 14, individually. For steps 6 and 7, the major problem was to decide whether a given sequence constituted an abbreviation or a sequence of initials. In conformity with the manual, I decided to code as abbreviations clipped words which are not pronounced as such in

speech. Thus, words like "Capt." and "para." were coded as abbreviations, while words like "info" and "rep" were not. According to the manual, however, it is important to distinguish between these words and sequences of initials. Thus, words like U.S.A., whether they contained abbreviation points or not (USA), were coded as sequences of initials.

Steps 8 and 9 were rather time consuming at first, since I had to rely on *Grammatik* to discover all of the words to be classified as "non-standard". Fortunately, as my work with the style checker progressed, I was able to accumulate substantial lists of recurring words, which could be inserted in automatic Search and Replace routines for subsequent conferences. In this manner, after several hours in the style checker, I was able to add the following words to the non-standard category: *thar, yer, fer, gonna, kinda, Gawd, ya, youse*, etc. Following the conventions stated in section 12.6 of the manual, I did not code as non-standard words which were shortened with an apostrophe. Thus, words like *d'you, 'em* and so on were left uncoded.

Step 10 consisted of coding specialized, computer-related words as "technical jargon". Coding these words as \4 represents somewhat of a departure from the instructions in the manual, where the code \4 is reserved for science fiction. Upon manually inspecting my own science fiction category, I found that the few non-English words that were used were enclosed in quotation marks. This is presumably different from the science fiction segment of LOB, which represents actual science fiction texts, rather than discussion about science fiction texts. Since it seemed redundant to code words which were already set off from the text, I decided to reserve the code \4 for computer jargon. This category includes all the technical words that the Spell Checker did not recognize, including abbreviations.

Step 11 consisted of coding nonce words, or words specifically coined for a given occasion. Nonce words are distinguished by the fact that they are either completely invented, as in the case of "offensitivity" above, or are spelled in an unconventional fashion, such as "I'm not n-n-n-n-nervous at all.". The use of intentionally unconventional spelling is quite frequent in the ELC, and can be seen as a device intended to compensate for the fact that participants do not hear each other. Step 12 consisted of coding paralinguistic comments. The code for paralinguistic comments was inserted for all those comments, usually enclosed in brackets, which the participants use to compensate for the fact that they do not see each other. Finally, steps 13 and 14 consisted of coding words borrowed from other languages and not recognized by the style checker as foreign words or sequences.

After all of the coding was done, the final step was to automatically insert line numbers in the entire corpus. Figure 11 is an excerpt from the corpus after this step was completed.

Figure 11: Extract from the Science Fiction conference immediately before applying CLAWS 1.

```
**[001 TEXT DDU**]
M01 0001 |^Thank-you John,
M01 0002 |^I feel pretty secure in this bit of **strange*** casting. ^I
have a deep
M01 0003 belief that Hunt could pull it off admirably. ^She was marvelous
in
M01 0004 the above film & although she wouldn't be able to pull off a 13
\0yr.
M01 0005 old Brazilian Boy cartwheel, don't we use long-shots & stunt-
people
M01 0006 (Body-Doubles) for lots of films these days?
```

4.4 Running the CLAWS 1 program suite

CLAWS 1, as noted earlier, is a suite of programs which assigns each word in a corpus to a specific grammatical word class. The system is available on a VAX at Concordia University's Computer Centre, and was accessed from my personal computer via modem. The entire system is subdivided into five component programmes, three of which (WORDTAG, IDIOMTAG and CHAINPROBS) have the specific function of increasing the certainty with which grammatical tags are assigned. The two other, non-core, components of CLAWS (PREEDIT and LOBFORMAT) pre- and post-format the corpus and were ultimately ported from the VAX and run on my personal computer.

The first program within the CLAWS system is PREEDIT. Its major function is to "verticalize" the corpus, or to arrange it so that each word or punctuation mark is placed on a separate line. In this phase, the coding symbols described above are placed on a subsidiary position on the line, so that they can be referred to in the later stages of the tagging process. Additionally, this program splits contracted forms into two words. A final function of this program is to remove sentence-initial upper case letters which are later restored, if justified by grammatical constraints (as in the case of proper nouns or the pronoun *I*). Figure 12 is an example of one of the sentences in the extract above, as it appears after it has been run through PREEDIT. Note that *wouldn't* has been split into two words.

Normally, the stage immediately following PREEDIT would have required substantial manual intervention. However, since I had decided to code the corpus beforehand, the only manual intervention required of me was to restore those upper case letters which should have been retained, as in the case of sentences beginning with proper nouns that could be mistaken as common nouns (*Mark, Rose*, etc.).

Figure 12: Output from PREEDIT.

```
*001 TEX020 ddu**]
M01 0002112 ------------------------------------------------
M01 0002120 I
M01 0002130 have
M01 0002140 a
M01 0002150 deep
M01 0003010 3
M01 0003020 belief
M01 0003030 that
M01 0003040 Hunt
M01 0003050 could
M01 0003060 pull
M01 0003070 it
M01 0003080 off
M01 0003090 admirably
M01 0003091 .                                              .
M01 0003092 ------------------------------------------------
M01 0003100 she
M01 0003110 was
M01 0003120 marvelous
M01 0003130 in
M01 0004010 4
M01 0004020 the
M01 0004030 above
M01 0004040 film
M01 0004041 &                                          CC
M01 0004050 although
M01 0004060 she
M01 0004070 would                       >
M01 0004071 n't                         <
M01 0004080 be
M01 0004090 able
M01 0004100 to
M01 0004110 pull
M01 0004120 off
M01 0004130 a
M01 0004140 13
M01 0004150 \0yr                        \0
M01 0005010 5
M01 0005020 old
M01 0005030 Brazilian
M01 0005040 Boy
M01 0005050 cartwheel
M01 0005051 ,                                              ,
```

The next program, WORDTAG, begins the task of tag assignment proper. Appendix B contains the list of symbols used by the system to designate specific word classes. WORDTAG checks each word in the corpus to see if it matches one of the words in a special dictionary file or contains a suffix in a special suffix list. If it does, the tag or tags associated with the word or suffix are listed as potential tags of that word. For example, the word *deep* can occur as an adjective (represented by the symbol JJ), an adverb (RB) or, in rare circumstances (represented by @), a noun (NN). Accordingly, the word *deep* in the first sentence in Figure 12 was tagged as follows:

```
M01 0002150 deep                                    02 JJ RB NN@
```

The third program, IDIOMTAG, reduces the list of tagging possibilities by checking groups of words against a list of pre-programmed sequences. For example. if the program found the sequence *one another*, WORDTAG would tag the words individually, as cardinal number (CD) and singular determiner (DT), respectively. IDIOMTAG, however, would then label the entire sequence as a reciprocal pronoun (PPLS).

The fourth program, CHAINPROBS, disambiguates words which still have multiple tags by calculating the probability that a given tag is the correct tag on the basis of the context in which the word appears (Marshall, 1987). Each tag is followed by a figure which indicates the probability that it is the correct tag. Thus, Figure 13 shows that there is a 40% chance that the word *deep* is an adjective, a 40% chance that it is an adverb, and only a 20% chance that it is a noun.

Figure 13: Output from CHAINPROBS

```
M01 0002112 ----------------------------------------------------
M01 0002120 I                                    02 PP1A
M01 0002130 have                                 02 HV
M01 0002140 a                                    02 AT
M01 0002150 deep                                 02 [JJ]/ 40 RB/ 40 NN@/ 20
M01 0003010 3                                    10 CD
M01 0003020 belief                               02 NN
M01 0003030 that                                 02 [CS]/ 32 DT/ 32 WP/ 32
                                                    QL%/  4
M01 0003040 Hunt                                 52 NP
M01 0003050 could                                02 MD
M01 0003060 pull                                 02 [VB]/ 50 NN/ 50
M01 0003070 it                                   02 PP3
M01 0003080 off                                  02 [RP]/ 50 IN/ 50
M01 0003090 admirably                            54 RB
M01 0003091 .                                    01 .
M01 0003092 ----------------------------------------------------
```

At this stage, it was necessary to revert to extensive manual intervention in order to post-edit the entire corpus. This was done by examining all the words and expressions which had been previously coded as *foreign, non-standard, technical*, and so on. For example, many non-standard words were tagged incorrectly by CHAINPROBS, on the basis of context alone. Thus, in the post-editing phase, it was common to find expressions like *c'mon* and *gonna* tagged as common nouns; like *Ahh*, as proper nouns and *E-mail* in *E-mail me as soon as you can* as an attributive adjective. It was therefore necessary to correct all such erroneous tags. When in doubt, I relied on Quirk *et al.* (1985).

The function of the fifth program, LOBFORMAT, is to remove all tags which were not selected by CHAINPROBS as the best alternative whenever a word was assigned more than one tag. Before running this program, I spot-checked those words in the output of CHAINPROBS which still had multiple tags, and modified the decisions made by CHAINPROBS whenever there appeared to be an error in either rank order or, rarely, in listing the correct tag as one of the alternatives. The corpus

was then reformatted by means of customized LOBFORMAT programmes which produced four versions of the tagged corpus: a tagged horizontal version, a tagged vertical version, only the tags and only the words. All four versions were used to extract the specific linguistic features which were necessary to apply Biber's multidimensional-multifeature model.

4.5 Extracting the linguistic features

The multidimensional-multifeature approach rests on the assumption that language genres can be plotted along several textual dimensions, as explained in Chapter 2. In turn, each of these dimensions is characterized by the presence or absence of a set of linguistic features. Overall, Biber identifies 67 linguistic features which, together, represent the totality of the various dimensions. Of these features, only 59 were retained in Biber's actual computation of factor scores. For the ELC, the 59 features were extracted with the aid of an information retrieval program called *GOfer,* which is illustrated in Figure 14.

In order to allow *GOfer* to automatically identify the linguistic features, it was necessary to write algorithms which the program could understand. For example, although the program could not understand the command "Search for present tense", it could understand "Search for all words tagged VBS (third person singular present) and VB (base form of the verb) except where VB is preceded by TO (in order to eliminate the infinitive form). The algorithms listed below were adapted from Biber and "translated" in such a way that the tag symbols assigned by Biber would coincide with the symbols used by CLAWS 1. In addition, where features were defined as lists of lexical items, certain words had to be added which could be expected in an ELC in a manner that Biber could not predict. For example, Biber defines the

category "second person pronouns" as containing the following words: *you, your, yourself, yourselves.* However, in order to thoroughly examine this category in the ELC, it was necessary to add the following words to the list: *ya, youse, y'all, yer, u.* Similarly, the algorithm for the present tense stating that VB must not be preceded by TO needed to be modified in order to capture non-standard expressions ending with a suffix which serves the same function as TO, for example *wanna, gotta, gonna, hafta* and *otta.*

Figure 14: *GOfer* information retrieval program

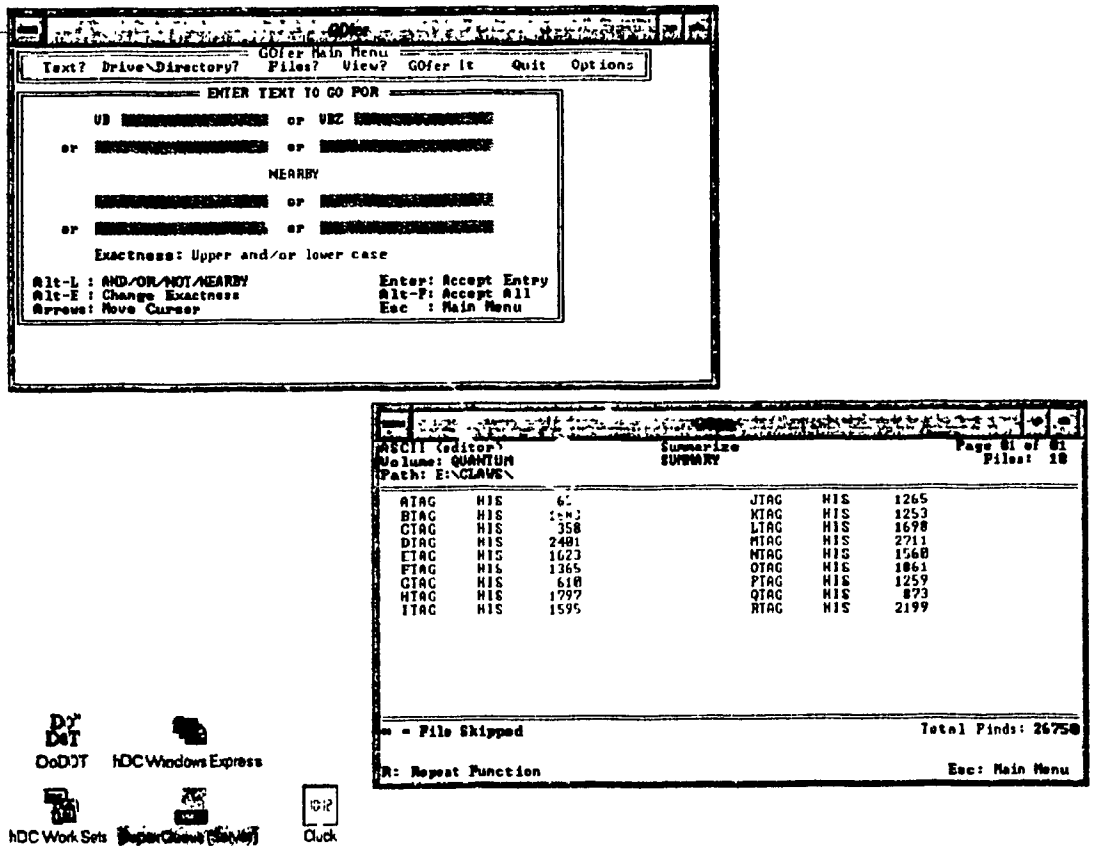


Table 5 is a list of the 59 linguistic features and their corresponding algorithms. Words in upper case refer to the tag symbols used in the CLAWS tagset. Tags followed by an asterisk indicate that all of the forms of a given tag are included. For

example, the category VB* includes VBD (past tense of verb), VBG (present participle, gerund), VBN (past participle) and VBS (third person singular).

Once all of the linguistic features were extracted, the research was complete. What remained to be done was a frequency count of the 59 linguistic features, and a statistical analysis which would allow me to compare my findings with those in Biber (1988). These are the issues which will be addressed in the following chapter.

Table 5. Linguistic features and algorithms used in the study.
(per-case letters designate tags used in the CLAWS 1 tagset)

1 past tense	VBD/ BED/ BEDZ/ DOD/ HVD
2 perfect aspect	HV/ HVD/ HVN/ HVZ + xxx + xxx + VBN
3 present tense	VBZ/ VB Subtract TO/ *gonna, gotta, wanna, hafta, otta* + VB
4 place adverbial	*aboard, above, abroad, across, ahead alongside, around, ashore, astern, away, behind, below, beneath, beside, downhill, downstairs, downstream, east, far,hereabout, indoors, inland, inshore, inside, locally, near, nearby, north, nowhere, outdoors, outside, overboard, overland, overseas, south, underfoot, underground, underneath, uphill, upstairs, upstream, west*
5 time adverbials	*afterwards, again, earlier, early, eventually, formerly, immediately, initially, instantly, late, lately, later, momentarily, now, nowadays, once, initially, originally, presently, previously, recently, shortly, simultaneously, soon, subsequently, today, tomorrow, tonight, yesterday*
6 1st person pronouns	*I, me, we, us, my, our, myself, ourselves, mah, usen s*
7. 2nd person pronouns	*you, your, yourself, yourselves, ya, youse, y'all, yer, u*
8 3rd person pronouns	*she, he, they, her, him, them, his, their, himself, herself, themselves, 'er, 'em*
9 pronoun IT	*it*
10 demonstrative pronouns	DT/ DTS + _/ ./ ./ /?/ !/ HV*/ VB*/ BE*/ DO*
11 indefinite pronouns	PN Subtract *so* PN
12 pro-verb DO	DO/ DOD/ DOZ. Subtract DO/DOD/ DOZ + xxx + VB DO/DOD/ DOZ + xxx + XNOT W* + DO/DOD/ DOZ
13 WH-questions	^/ ./ */ W*
14 nominalizations	NN* ending in *-tion(s), -ment(s), ness(s), ity(ies)*
15 total other nouns	NN* Subtract nominalizations. Subtract nouns ending in *-Ing*
16 agentless passive	BE* + xxx + xxx + VBN
17 BY-passive	BE* + xxx + xxx +VBN + xxx + xxx + *by*

18 BE as main verb	BE* Subtract BE* + xxx + xxx + VBG BE* + xxx + xxx + VBN
19 THAT verb complement	*that* CS (I hope that you can come)
20 THAT adjective complement	JJ* + *that* CS (I'm glad that you came)
21. WH-clauses	VB* + W*
22 infinitives	TO
23 present participial clauses	./ :/ ,/ /?/ !/ + VBG + IN/DT/ W*/ PP/ RB (Stuffing his mouth with cookies, he ran out the door)
24 past participial clauses	/ ./ ,/ /?/ !/ + VBN + IN/ RB (Built in a week, this house would last)
25 past prt. WHIZ deletion	NN*/PN* + VBN + IN/ RB/ BE*/ DO*/ HV* (the solution produced by this process)
26 THAT relative clause on object position	NN* + that_CS + PN*/ DT*/ AT*/ NN*/ JJ* (the dog that I saw)
27 WH relative on subject position	W*+ xxx + xxx + VB*/ HV*/ BE*/ DO* (the man who likes popcorn)
28 WH relative on object position	W* + PN*/ DT*/ AT*/ NN*/ JJ* (the man who Sally likes)
29 pied-piping	IN + W* (the manner in which he was told)
30 sentence relative	./ -* + W* (Bob likes mangoes, which is the most disgusting thing I've ever heard of
31 causative adverbial subordinator	*because*
32 concessive adverbial subordinator	*if, unless*
33 other adverbial subordinators	*since, while, whilst, whereupon, whereas, whereby, such that, so that, insasmuch as, forasmuch as, insofar as, insomuch as, as long as, as soon as*
34 total prepositional phrases	*against, amid, amidst, among, amongst, at besides, between, by, despite, during, except, for, from, in, into, minus, notwithstanding, of, off, on, onto, opposite, out, per, plus, pro, re, than, through, throughout, thru, to toward, towards, upon, versus, via, with, within, without*
35 attribute adjectives	J* + JJ*/ NN*
36. total adverbs	RB/ RBR/ RBT Subtract place adverbials (list #4) time adverbials (list #5) downtowners amplifiers (list 41) emphatics (list 42)
37 type-toke ration	number of different words over total number of words
38 word length	mean length, in orthographic letters

39. conjuncts	*alternatively, altogether, consequently, conversely, eg, e g ,else, furthermore, hence, however, i.e , instead, likewise, moreover, namely, nevertheless, nonetheless, notwithstanding, otherwise, rather, similarly, therefor, thus, viz* *in/by/as a/ + comparison, contrast, particular, addition, conclusion, consequence, sum, summary, any event, any case, other words* *for + example, ex., instance,* *on the + contrary, other hand*
40 hedges	*at about, something like, more or less, almost, maybe, xxx +sort of, sorta, xxx + kind of, kinda* (where xxx is not DT* / AT*/ PP - excludes *sort* and *kind* as nouns)
41 amplifiers	*absolutely, altogether, completely, enormously, entirely, extremely, fully, greatly, highly, intensely, perfectly, strongly, thoroughly, totally, utterly, very*
42 emphatics	*for sure, a lot, such a, real/so* + JJ*, *just, really, most, more*
43 discourse particles	/ ./ ./ ./ -*/ ?/ !/ + well, now, anyway, anyhow, anyways
44 demonstratives	DT/ DTS + NN*
45 possibility modals	*can, may, might, could* (+ contractions)
46 necessity modals	*ought, should, must* (+ contractions)
47 predictive modals	*will, would, shall* (+ contractions)
48 public verbs	*acknowledge, admit, agree, assert, claim complain, declare, deny, explain, hint, insist, mention, proclaim promise, protest, remark, reply, report, say, suggest, swear, write*
49 private verbs	*anticipate, assume, believe, conclude, decide, demonstrate, determine, discover, doubt, estimate, fear, feel, find, forget, guess, hear, hope, imagine, imply, indicate, infer, know, learn, mean, notice, prove, realize, recognize, remember, reveal, see, show, suppose, thing, understand*
50 suasive verbs	*agree, arrange, ask, beg, command, decide, demand, grant, insist, instruct, ordain, pledge, pronounce, propose, recommend, request, stipulate, suggest, urge*
51 contractions	all contractions
52 subordinator THAT deletion	public/ private/ suasive + AT*/ NN*/ P* (I think he wen to _
53 stranded prepositions	IN + / / ./ ./ -*/ ?/ !/
54 split infinitives	TO + xxx + xxx + VB*/ HV* / DO*/ BE* (he wants to convincingly prove)
55 split auxiliaries	BE*/ HV*/ DO* + xxx + xxx + VB* (they are objectively shown to)
56 phrasal coordinations	RB* + *and*_CC + RB* JJ* + *and*_CC + JJ* VB* + *and* _ CC + VB* N* + *and* _ CC + N*
57 independent clause coordination	*and* _ CC Subtract results from #56
58 synthetic negation	*no, neither, nor*
59 analytic negation	*not*

Chapter 5

DATA ANALYSIS

5.1 Frequency counts of the linguistic features

The main task fulfilled by *GOfer* was to compute the absolute frequencies in the ELC texts of the linguistic features listed in Table 5. In most cases, this calculation was done automatically, with *GOfer* indicating the frequency for each conference, as well as the total frequency for the corpus as a whole. In cases where the algorithms for the features to be extracted required multiple operations, it was necessary to run separate passes, and to perform the arithmetic operations (additions or subtractions) with the aid of the spreadsheet programme, *Microsoft Excel*. For example, the algorithm for feature #3 (present tense verbs) reads "Subtract, from all VB, the sequence TO + VB" . It was therefore necessary to run two passes, one counting all instances of VB and VBZ, the other counting the sequence TO + VB, and then to subtract the second count from the first. Similarly, where the algorithms contained a list of words which was too long for *GOfer* to handle in a single pass, as in the case of feature #4 (place adverbials), it was necessary to run several passes, each containing a partial list of the targeted words, and then to add the figures together.

In order to allow for a comparison with Biber's results, it was then necessary to normalize the frequency counts to a text length of 1000 words. The normalization was arrived at by applying a standard normalization formula, also used by Biber: **(total frequency / length of text) x 1000**. From the list of normalized frequencies, I extracted the following statistics: 1) the mean frequency, 2) the maximum and

minimum frequencies, i.e., the lowest and highest number of occurrences in any text, 3) the range, or the difference between the maximum and minimum values, 4) the standard deviation, and 5) the feature deviation score. These statistics, calculated on the basis of the entire ELC, were compared to Biber's statistics for his corpus, and are reproduced in Table 6.

As we have seen, Biber's study is very comprehensive, as it examines texts which would seem to represent the entire spectrum of major genres in the English language, including all the published texts from LOB, the spoken texts in the LLC, and his private collection of personal and professional correspondence. If Biber's study is indeed representative of the state of the language, we would expect the ELC to be largely comparable. In terms of the overall distribution of particular features, this would seem to be the case. Features which are very common in Biber's corpus are also very common in the ELC, while features which are rare in his corpus are also rare in the ELC. Thus, prepositions occur quite frequently in both corpora, with a mean of 110 per 1000 words in Biber and 118 per 1000 words in the ELC. Similarly, an infrequent feature like present participial clauses is roughly comparable in both corpora, 1 per 1000 words in the former and 1.2 per 1000 words in the latter.

However, as would be expected in a subset of English with unique situational constraints, there are some striking differences. Among the more frequent features, there are considerable differences between the frequencies of past tense verbs, nouns, attributive adjectives and adverbs. Similarly, among the infrequent features, there are important differences between the number of perfect aspect verbs, second person pronouns, indefinite pronouns and analytic negations (the use of *not* rather than *no* or *neither*).

Table 6. Frequency of linguistic feature for Biber's corpus and the ELC

Linguistic feature	Mean Biber	Mean ELC	Min. Biber	Min. ELC	Max. Biber	Max. ELC	Range Biber	Range ELC	Std.Dev. Biber	Std.Dev. ELC	FDS
past tense	40.1	28.0	0	17.8	119	64.9	119	47.1	30.4	11.6	-0.4
perfect aspect verbs	8.6	4.0	0	2.4	40	11.9	40	9.4	5.2	2.1	-0.9
present tense	77.7	68.0	12	57.5	182	170.2	170	112.7	34.3	26.3	-0.3
place adverbials	3.1	4.2	0	2.4	24	6.6	24	4.2	3.4	1.2	0.3
time adverbials	5.2	5.9	0	2.7	24	17.4	24	14.7	3.5	3.1	0.2
first person pronouns	27.2	57.5	0	32.0	122	130.2	122	98.1	26.1	21.2	1.2
second person pronouns	9.9	18.3	0	12.0	72	50.2	72	38.2	13.8	10.5	0.6
third person pronouns	29.9	25.5	0	11.8	124	40.6	124	28.8	22.5	9.4	-0.2
pronoun IT	10.3	19.3	0	14.0	47	56.4	47	42.4	7.1	9.6	1.3
demonstrative pronouns	4.6	6.9	0	4.7	30	24.5	30	19.8	4.8	4.6	0.5
indefinite pronouns	1.4	4.5	0	3.1	13	9.4	13	6.3	2	1.5	1.6
DO as pro-verb	3	4.2	0	0.6	22	9.2	22	8.7	3.5	1.8	0.4
WH questions	0.2	1.9	0	1.4	4	6.2	4	4.8	0.6	1.2	2.9
nominalizations	19.9	12.6	0	6.9	71	68.6	71	61.8	14.4	14.0	-0.5
nouns	180.5	225.1	84	189.9	298	669.1	214	479.2	35.6	107.2	1.3
agentless passives	9.6	7.0	0	4.3	38	27.3	38	23.1	6.6	5.2	-0.4
BY passives	0.8	1.2	0	0.6	8	6.9	8	6.3	1.3	1.4	0.3
BE as main verb	28 3	36.9	0	22.8	72	94.9	65	72.0	9.5	14.6	0.9
THAT verb complements	3.3	7.0	0	4.5	20	16.8	20	12.3	2.9	2.9	1.3
THAT adj. complements	0.3	0.4	0	0.1	3	1.0	3	0.9	0.6	0.3	0.2
WH clauses	0.6	1.2	0	0.4	7	4.3	7	3.9	1	0.8	0.6
infinitives	14.9	17.6	1	10.4	36	46.5	35	36.0	5.6	7.3	0.5
present participial clauses	1	1.2	0	0.5	11	4.8	11	4.3	1.7	1.0	0.1
past participial clauses	0.1	0.6	0	0.2	3	1.3	3	1.1	0.4	0.3	1.3
past prt. WHIZ deletions	2.5	2.5	0	1.4	21	11.3	21	10.0	3.1	2.2	0.0
THAT relatives: obj. position	0.8	1.7	0	0.8	7	4.1	7	3.3	1.1	0.9	0.8
WH relatives: subj. position	2.1	1.4	0	0.5	15	4.3	15	3.7	2	0.8	-0.4
WH relatives: obj. position	1.4	0.1	0	0.0	9	0.4	9	0.4	1.7	0.1	-0.8
WH relatives: pied pipes	0.7	1.2	0	0.4	7	3.0	7	2.6	1.1	0.7	0.4
sentence relatives	0.1	1.9	0	1.0	3	6.6	3	5.5	0.4	1.4	4.5
adv. subordinator - cause	1.1	1.4	0	0.2	11	2.8	11	2.6	1.7	0.6	0.2
adv. sub. - condition	2.5	7.9	0	5.7	13	23.8	13	18.1	2.2	4.0	2.4
adv. sub. - other	1	3.2	0	1.5	6	11.3	6	9.9	1.1	2.2	2.0
prepositions	110.5	118.2	50	105.1	209	380.2	159	275.1	25.4	62.3	0.3
attributive adjectives	60.7	46.6	16	28.2	115	193.7	99	165.5	18.8	35.8	-0.8
adverbs	65.6	38.9	22	26.2	125	102.7	103	76.5	17.6	16.3	-1.5
type/token ratio	51.1	56.8	35	42.2	64	66.4	29	24.3	5.2	5.8	1.1
word length	4.5	4.4	3.7	4.0	5.3	4.8	1.6	0.8	0.4	0.2	-0.2
conjuncts	1.2	3.1	0	1.9	12	5.4	12	3.5	1.6	1.1	1.2
hedges	0.6	2.2	0	1.4	10	3.0	10	1.6	1.3	0.5	1.3
amplifiers	2.7	4.0	0	2.3	14	8.5	14	6.2	2.6	1.4	0.5
emphatics	6.3	11.2	0	4.2	22	16.2	22	12.1	4.2	2.6	1.2
discourse particles	1.2	2.1	0	0.4	15	6.2	15	5.7	2.3	1.8	0.4
demonstratives	9.9	6.4	0	5.3	22	21.1	22	15.8	4.2	3.7	-0.8
possibility modals	5.8	8.9	0	6.7	21	30.0	21	23.3	3.5	5.1	0.9
necessity modals	2.1	2.0	0	1.1	13	4.3	13	3.1	2.1	0.8	-0.1
predictive modals	5.6	9.0	0	5.0	30	22.3	30	17.4	4.2	3.9	0.8
public verbs	7.7	5.5	0	3.5	40	9.6	40	6.1	5.4	1.6	-0.4
private verbs	18	20.0	1	13.1	54	43.6	53	30.5	10.4	6.4	0.2

									Biber	ELC	
Linguistic feature	Mean	Mean	Min.	Min.	Max.	Max.	Range	Range	Std.Dev.	Std.Dev.	FDS
suasive verbs	2.9	1.2	0	0.6	36	3.5	36	3.0	3.1	0.7	-0.5
contractions	13.5	16.0	0	7.3	89	28.2	89	20.9	18.6	6.0	0.1
THAT deletion	3.1	8.8	0	4.3	24	17.9	24	13.6	4.1	3.1	1.4
stranded prepositions	2	0.9	0	0.5	23	3.4	23	2.8	2.7	0.7	-0.4
split infinitives	0	0.0	0	0.0	1	0.3	1	0.3	0	0.1	0.0
split auxiliaries	5.5	2.3	0	1.6	15	6.0	15	4.4	2.5	1.0	-1.3
phrasal coordination	3.4	5.2	0	3.0	12	21.5	12	18.5	2.7	4.2	0.7
non-phrasal coordination	4.5	16.5	0	12.0	44	38.5	44	26.4	4.8	5.8	2.5
synthetic negation	1.7	0.4	0	0.0	8	1.1	8	1.1	1.6	0.3	-0.8
analytic negation	8.5	15.5	0	9.0	32	37.9	32	28.9	6.1	6.2	1.2

Several factors may be called upon to explain these differences. As Table 1 showed, the speech situation in electronic language is radically different from the situation surrounding other types of language. For example, participants in an ELC share a high degree of common knowledge, the language event occurs in different places and at different times and the messages, although written, are often produced "on-line". One area of investigation which could be readily investigated was the question of whether any of the characteristics of the ELC could be attributed to the fact that the relationship of the participants to the text involves much on-line production. In an attempt to answer this question, I have kept separate, wherever possible, texts which were definitely produced off-line, from texts which were assumed to be produced directly on-line. Table 7 separates the statistics for the corpus as a whole into an off-line category and an "other" category, which includes, but is not limited to texts produced on-line.

Table 7. Frequency of linguistic features for OFF-line and OTHER corpora

Linguistic features	Mean OTHER	Mean OFF	Min. OTHER	Min. OFF	Max. OTHER	Max. OFF	Range OTHER	Range OFF	Std.Dev. OTHER	Std.Dev. OFF	FDS OTHER	FDS OFF
past tense	29.3	27.5	17.8	19.4	64.9	49.4	47.1	30.0	14.3	8.9	-0.4	-0.4
perfect aspect verbs	4.4	3.9	3.2	2.4	11.9	6.7	8.7	4.2	2.6	1.3	-0.8	-0.9
present tense	67.6	66.3	57 5	57.6	170.2	105 4	112.7	47.9	35.0	14.6	-0.3	-0.3
place adverbials	4.2	4.2	2.4	3.1	6.0	6.6	3.7	3.5	1.3	1.1	0.3	0.3
time adverbials	7.1	5.1	5.2	2.7	17.4	7.7	12.1	5.0	3.7	1.4	0.5	0.0
first person pronouns	57.8	49.8	32 0	39.8	130.2	69.0	98.1	29.2	27.7	12.6	1.2	0.9
second person pronouns	17.6	19.0	14.5	12.0	50.2	25.4	35.7	13.3	12.8	7.6	0.6	0.7
third person pronouns	26.9	24.2	13.6	11.8	36.4	40.6	22.8	28.8	8.3	10.9	-0.1	-0.3
pronoun IT	21.6	17.1	15.3	14.0	56.4	27.0	41.1	13.0	12.6	4.9	1.6	1.0
demonstrative pronouns	6.5	6.9	4.7	4.9	24.5	11.2	19.8	6.3	6.2	2.3	0.4	0.5
indefinite pronouns	4.6	4.5	3.1	3.3	9.4	6.9	6.3	3.6	1.9	1.3	1.6	1.5
DO as pro-verb	4.8	3.7	0.6	2.9	9.2	4.6	8.7	1.7	2.4	0.7	0.5	0.2
WH questions	2.9	1.5	1.9	1.4	6.2	1.9	4.3	0.6	1.3	0.5	4.6	2.2
nominalizations	12.9	12.4	7.5	6.9	68.6	23.8	61.1	16.9	18.9	6.2	-0.5	-0.5
nouns	225.3	225.0	196.4	189.9	669.1	249.8	472.7	59.9	150.3	19.9	1.3	1.2
agentless passives	6.1	7.0	5.5	4.3	27.3	10.0	21.8	5.8	7.1	1.8	-0.5	-0.4
BY passives	1.2	1.1	0.6	0.9	6.9	1.9	6.3	1.0	2.0	0.4	0.3	0.2
BE as main verb	36.6	39.7	31.4	22.8	94.9	42.5	63.5	19.6	19.9	6.6	0.9	1.2
THAT verb complements	7.6	6.8	5.2	4.5	16.8	12.3	11.6	7.8	3.5	2.3	1.5	1.2
THAT adj. complements	0.4	0.5	0.3	0.1	0 9	1.0	0.6	0.9	0.2	0.3	0.2	0.3
WH clauses	1.2	1.2	0 4	0.7	4.3	1.5	3.9	0.8	1.1	0.2	0.6	0.6
infinitives	17.7	17.8	15.0	10.4	46.5	20.5	31.5	10.0	9.8	2.9	0.5	0.5
present participial clauses	1.2	1.2	0.5	0.8	4.8	2.8	4.3	2.0	1.3	0.6	0.1	0.1
past participial clauses	0.6	0 6	0.3	0.2	1.3	1.1	1 0	0.9	0.4	0.3	1.4	1.2
past prt WHIZ deletions	2.5	2.5	1.5	1.4	11.3	4.0	9.9	2.6	3.0	0.7	0.0	0.0
THAT relatives: obj. position	1.7	1.7	1.3	0.8	4.1	3.5	2.7	2.7	0.8	0.9	0.8	0.8
WH relatives: subj. position	1.4	1.4	0.5	0.8	4.3	2.9	3.7	2.1	1.1	0.6	-0.4	-0.4
WH relatives: obj. position	0.1	0.1	0.0	0.0	0.3	0.4	0.3	0.4	0.1	0.2	-0.8	-0.7
WH relatives: pied pipes	1.2	1.0	0.7	0.4	3.0	2.4	2.3	2.0	0.7	0.7	0.5	0.3
sentence relatives	2.4	1.6	1.0	1.2	6.6	3.1	5.5	1.9	1.7	1.0	5.7	3.9
adv. subordinator - cause	1.5	1.0	1.0	0.2	2.8	1.9	1.8	1.7	0.6	0.5	0.2	-0.1
adv. sub. - condition	7.9	7.8	5.9	5.7	23.8	10.4	17.9	4.7	5.5	1.7	2.5	2.4
adv. sub. - other	3.2	3.2	1.5	1.5	11.3	4.8	9.9	3.3	2.9	1.2	2.0	2.0
prepositions	116.9	119.1	105.8	105.1	380.2	127.4	274.4	22.3	68.3	7.1	0.3	0.3
attributive adjectives	45.3	47.9	40.3	28.2	193.7	59.3	153.4	31.2	49.3	10.1	-0.8	-0.7
adverbs	37.1	39.9	29.6	26.2	102.7	48.5	73.0	22.3	22.4	7.6	-1.6	-1.5
type/token ratio	56.8	56.8	46.8	42.2	66.4	65.0	19.7	22.8	5.9	8.1	1.1	1.1
word length	4.5	4.3	4.2	4.0	4.8	4.6	0.6	0.6	0.2	0.2	0 0	-0.5
conjuncts	3.7	2.3	2 4	1.9	5 4	5.2	3.0	3.3	1.0	1.2	1.5	0.7
hedges	2.1	2.5	1.4	1.6	3.0	2.8	1.6	1.2	0.5	0.4	1.2	1.5
amplifiers	4.2	3.9	2.9	2.3	5.3	8.5	2.5	6.2	0.8	1.8	0 6	0.5
emphatics	11.2	11.2	4.2	7.9	13.3	16.2	9.1	8.3	2.7	2.4	1.2	1.2
discourse particles	2.3	1.9	0.6	0.4	6.0	6.2	5.4	5.7	1.7	2.1	0.5	0.8
demonstratives	7.1	6.3	5.3	5.4	21.1	12.0	15.8	6.5	4.9	2.1	-0.7	-0.9
possibility modals	8.9	8.8	6.7	7.7	30.0	12.5	23.3	4.8	7.2	1.8	0.9	0.9
necessity modals	2.0	1.9	1.4	1.1	4.3	3.0	2.9	1.9	0.9	0.7	0.0	-0.1
predictive modals	9.6	8.8	6.1	5.0	22.3	12.9	16.3	7.9	4.8	2.6	0.9	0.8
public verbs	5.8	4.9	3.5	3.6	9.6	7.0	6.1	3.4	1.9	1.2	-0.3	-0.5
private verbs	20.7	19.6	14 9	13.1	43.6	26 2	28.7	13.1	8.4	3.5	0.3	0.2

											OTHER	OFF
Linguistic features	Mean	Mean	Min.	Min.	Max.	Max.	Range	Range	Std.Dev.	Std.Dev.	FDS	FDS
suasive verbs	1.4	1.0	0.6	0.6	3.5	2.2	3.0	1.6	0.8	0.5	-0.5	-0.6
contractions	16.6	14.9	7.3	7.4	28.2	23.3	20.9	15.9	5.6	6.6	0.2	0.1
THAT deletion	8.8	6.9	4.5	4.3	17.9	11.7	13.4	7.3	3.9	2.2	1.4	1.4
stranded prepositions	1.0	0.9	0.6	0.5	3.4	1.9	2.8	1.4	0.9	0.5	-0.4	-0.4
split infinitives	0.1	0.0	0.0	0.0	0.3	0.2	0.3	0.2	0.1	0.1	0.0	0.0
split auxiliaries	2.4	2.2	1.6	1.6	6.0	3.9	4.4	2.2	1.3	0.7	-1.3	-1.3
phrasal coordination	5.5	5.1	3.5	3.0	21.5	9.8	18.0	6.8	5.6	2.0	0.8	0.6
non-phrasal coordination	16.8	16.0	12.0	12.3	38.5	21.1	26.4	8.7	7.8	2.7	2.6	2.4
synthetic negation	0.4	0.4	0.2	0.0	1.1	1.0	0.9	1.0	0.3	0.3	-0.8	-0.8
analytic negation	15.1	16.2	11.1	9.0	37.9	16.6	26.9	7.6	8.2	3.2	1.1	1.3

Table 7 shows that it is not possible to explain any differences between the ELC and other varieties of English solely in terms of the peculiar nature of the relations of the participants to the text. In fact, although there are some differences in the frequencies of linguistic features in the "off-line" and "other" corpora, those features which clearly set the ELC apart from Biber's corpus are remarkably similar. For example, the figures for past tense verbs, nouns, attributive adjectives and adverbs reveal no major differences between the two categories.

The explanation would therefore have to be based on some other situational constraint. However, to attempt an explanation for each of the differences in linguistic features between Biber's corpus and the ELC would only provide a highly fragmented view of this language variety, and tell us nothing about the textual dimensions which distinguish this subset of English from the language as a whole. A more appropriate system of analysis is the multivariate statistical approach which Biber applied to his corpus. For this reason, Table 6 will be further considered in section 5.4 below, where the normalized frequencies will be interpreted in terms of factor scores.

5.2 Factor Analysis

The principal technique used in Biber's multivariate approach is factor analysis. The purpose of factor analysis is to reduce a large number of variables, for example linguistic features, to a small set of derived variables, or factors. In turn, the purpose of a factor is to determine which linguistic features tend to co-occur, and which tend to be mutually exclusive. The details of Biber's operations will not be reported here, as what concerns us most are the results of these operations. Figure 15 describes the six factors that Biber's analysis produced. The figure next to each linguistic feature represents the "loading" or "weight" which that feature carries in a given factor. Features with high positive loadings indicate strong co-occurrence relationships while features with negative loadings indicate complementary relationships. For example, a text which is representative of Factor 3 will exhibit a high frequency of WH-relative clauses (*the man who eats popcorn*) and pied-piping relative clauses (*the manner in which he was told*), while time and place adverbials will be notably absent. The purpose of loadings is to allow the researcher to determine the importance of particular features within a factor. Their value will become apparent in the interpretation of the findings, as they allow us to discriminate between less and more important features.

Figure 15: Summary of Biber's factorial structure

Factor 1

private verbs	.96
THAT deletion	.91
contractions	.90
present tense verbs	.86
2nd person pronouns	.86
DO as pro-verb	.82
analytic negation	.78
demonstrative pronouns	.76
general emphatics	.74
1st person pronouns	.74
pronoun IT	.71
BE as main verb	.71
causative subordination	.66
discourse particles	.66
indefinite pronouns	.62
general hedges	.58
amplifiers	.56
sentence relatives	.55
WH questions	.52
possibility modals	.50
non-phrasal coordination	.48
WH clauses	.47
final prepositions	.43
nouns	-.80
word length	-.58
prepositions	-.54
type/token ratio	-.54
attributive adjs.	-.47

Factor 2

past tense verbs	.90
third person pronouns	.73
perfect aspect verbs	.48
public verbs	.43
synthetic negation	.40
present participial clauses	.39

– no negative features –

Factor 3

WH relative clauses on object positions	.63
pied piping constructions	.61
WH relative clauses on subject positions	.45
phrasal coordination	.36
nominalizations	.36
time adverbials	-.60
place adverbials	-.49
adverbs	-.46

Factor 4

infinitives	.76
prediction modals	.54
suasive verbs	.49
conditional subordination	.47
necessity modals	.46
split auxiliaries	.44

– no negative features –

Factor 5

conjuncts	.48
agentless passives	.43
past participial clauses	.42
BY-passives	.41
past participial WHIZ deletions	.40
other adverbial subordinators	.39

– no negative features –

Factor 6

THAT clauses as verb complements	.56
demonstratives	.55
THAT relative clauses on object positions	.46
THAT clauses as adj complements	.36

-- no negative features --

5.3 Computing factor scores

With the factorial structure clearly defined, it was a relatively simple matter to compute the factor score for each of the linguistic features. This calculation required three operations: computing the feature deviation score for each linguistic feature and for each conference in the corpus, adding the feature deviation scores for each factor, and calculating the average of these results for each conference. The feature deviation scores were calculated on the basis of the figures for mean frequencies and standard deviations in Biber's corpus, reported in Table 6, as well as the frequency per 1000 words in each of the conferences in the ELC. The formula used was **feature deviation score = (Frequency - Mean) / Std. Dev**. For example, the frequency of prepositions in the Off-line Chat conference was 116. In Biber's corpus, the frequency of prepositions is 110.5 per 1000 words, and the standard deviation is 25.4. If we apply the formula above, the feature deviation score will be **(116 - 110.5) / 25.4.** The feature deviations score is therefore **0.2**. Table 8 below reports all of the feature deviation scores (FDS) for each of the 18 conferences examined.

Table 8. Feature Deviation Scores (FDS) across conferences by linguistic feature

Conference/	Chat	Chat	Current	Current	Science	Science	Medical	Medical	Film	Film	Finance	Finance	Cooking	Cooking	SciFi	SciFi	Sports	Sports
Linguistic feature	OFF	OTHER	OFF	OTHER	OFF	OTHER	OFF	OTHER	OFF	OTHER	OFF	OTHER	OFF	OTHER	OFF	OTHER	OFF	OTHER
past tense	-0.4	-0.4	-0.1	-0 4	-0 7	-0.7	-0 5	0 8	0.3	0.0	-0 4	-0.7	-0.5	-0.6	-0.1	-0.1	-0.4	-0.2
perfect aspect verbs	-1 0	-1.0	-1 0	-0 7	-1 0	-1 0	-0 4	0 6	-0.7	-0 7	-1 2	-0 8	-0.9	-0.9	-0.6	-0 5	-0.9	-0 9
present tense	0 2	0.4	0.0	-0 3	-0 6	-0 6	-0.1	2.7	-0.3	-0 5	-0 3	-0 1	0 8	0 3	-0 5	-0.4	-0.3	-0.3
place adverbials	0.4	0 5	0.0	0.4	0.4	0.2	0.3	0 9	0.2	0 8	0 0	-0.2	0.0	0 1	0 4	-0 2	1.0	0.3
time adverbials	0.0	0.6	-0.7	0 7	-0.2	0.0	0 0	3 5	0 4	1.0	-0.1	0.0	-0 2	0 1	0.3	0 5	0.7	0.5
first person pronouns	1.8	1 7	0 9	0 8	0 6	0.2	1 5	3 9	1 6	1 5	0 5	0.7	0.8	1 2	1 5	1.5	0.8	1.2
second person pronouns	1.9	2 2	0.3	0 4	0 2	0.3	0 9	2 9	0.7	0 4	1 1	1.3	0.7	1 0	0 3	0 6	0.2	0.3
third person pronouns	-0.1	-0.3	0 3	0 1	-0 7	-0 7	-0 5	0.0	0 1	-0.1	-0 8	-0.7	-0.8	-0 7	-0.3	0.0	0.5	0.3
pronoun IT	2.2	1.1	2.4	1 6	0 8	1.0	0.6	6.5	2.0	2 1	1 0	0.9	0.9	1.7	1.4	1.6	0.5	0.7
demonstrative pronouns	1.1	0.5	1.4	0 8	1.1	1 2	0.2	4.1	0.5	0 4	1 2	0.1	0.3	0.3	0.1	0.2	0 5	0.0
indefinite pronouns	2.7	2.0	2.8	2.1	1.0	1 2	1 5	4.0	1.8	1.3	1.3	1.6	1 5	0.9	2 2	2.1	1.5	1.2
DO as pro-verb	0.4	0.9	0 5	0 7	0 4	-0.7	0 0	1 8	0 1	0 6	0 2	0.3	0.3	-0.1	0.2	0.4	0.0	0.5
WH questions	4.9	4.8	2.9	3 9	2.1	4.6	2.1	10 0	2.3	5 5	2 2	2.9	2.1	2.9	2.0	4 9	2.6	3.1
nominalizations	-0.7	-0.6	-0.5	-0 1	0 1	0 2	0 0	3.4	-0 8	-0.7	0.3	0.0	-0 9	-0.7	-0.4	-0.5	-0.8	-0.9
nouns	0.4	0 6	0.3	0 4	0 7	1.2	1 6	13 7	1.2	1 4	1.3	1 4	1.9	1.3	0.9	1.0	1.4	1.5
agentless passives	-0.4	-0.5	0 1	-0 4	0 0	0.5	-0 2	. 7	-0 5	-0.6	-0.4	-0.6	-0.6	-0 6	-0.4	-0.3	-0 8	-0.6
BY passives	0 1	0 1	0.9	0 5	0 3	0 8	0.7	4 7	0 2	0 5	0 2	0.3	0.1	-0.1	0.1	0.3	0.4	-0.1
BE as main verb	1.6	0.6	1.5	0.9	1 5	1 3	0 7	7.0	1 2	0.8	0 9	0.4	-0.6	0.3	0.5	1 0	1.3	0.9
THAT verb complements	1.2	1.1	2.3	1 7	3 1	2.3	1 2	4 7	1.0	0 9	0 4	1.5	1.0	0.7	1.6	1.9	1.1	1.3
THAT adj. complements	0 5	0.3	0.8	0 1	0 3	0 2	0 2	1 0	-0 2	0.2	0.0	0 1	-0 4	0.2	1.2	1.0	0.6	0.8
WH clauses	0 4	0.6	0 6	1 0	0 6	-0.2	0 6	3 7	0.5	1.4	0.7	0 5	0.1	0 5	0.5	0 4	0.9	0.8
infinitives	0 5	0.9	1.0	0 9	0 6	0 6	0 5	5 6	-0 8	0 0	0 3	0.2	0 5	0.3	0.1	0.5	0.8	0.2
present participial clauses	0 0	0 2	-0.1	0 1	0 5	0 0	0 1	2 2	0 4	-0.1	0 2	0.1	1.0	0 2	0 0	-0 3	0.0	0.1
past participial clauses	1.2	3 1	0 7	1 2	1.2	1.9	0 3	2 9	1.4	0 6	0 9	1.0	2 6	2.1	2.5	1.4	2.2	0.6
past prt. WHIZ deletions	-0.4	0.0	0.1	0 0	0.0	0 4	0 0	2 9	0.0	-0 1	0 3	0.0	0.5	0.0	-0.1	-0.2	0.1	-0.3
THAT relatives: obj. position	0.0	0.9	2.4	0 9	2.1	1 3	0 8	3 0	0.8	0 5	0.1	0.7	0.5	0.5	1.0	0.7	0.4	0.8
WH relatives: subj. position	-0.7	-0.6	-0.5	-0 2	-0 2	-0.3	0 4	1.1	-0 5	-0.4	-0 1	-0 4	-0.6	-0 3	-0 2	-0 3	-0.4	-0.4
WH relatives: obj. position	-0.8	-0.3	-0.8	-0 8	-0 7	-0.6	-0 6	-0 8	-0 8	-0 7	-0.7	-0 8	-0.7	-0 8	-0.7	-0.7	-0.8	-0.8

Conference/ Linguistic feature	Chat OFF	Chat OTHER	Current OFF	Current OTHER	Science OFF	Science OTHER	Medical OFF	Medical OTHER	Film OFF	Film OTHER	Finance OFF	Finance OTHER	Cooking OFF	Cooking OTHER	SciFi OFF	SciFi OTHER	Sports OFF	Sports OTHER
WH relatives: pied pipes	1.1	0.2	-0.3	0 5	0 8	0 7	-0 1	1.3	0.3	1 2	0 8	0 4	0 0	0.0	1.5	2 1	-0 1	0.2
sentence relatives	10.0	10.1	7.5	4 2	2 9	6 3	3 9	16 2	5.3	7 3	3 0	3 6	3 4	4.9	2 7	5 7	3.9	2.3
adv. subordinator - cause	-0 5	0.1	-0.2	0 2	0.0	0.4	0 5	0 7	0.2	0 6	-0.1	1 0	-0 3	0.1	0 3	0 0	-0.1	0.2
adv. sub. - condition	2 3	2.4	3.6	3 1	2 9	2 4	2 5	9 7	2 4	1 6	3 5	3 1	1.5	2.5	1.6	1 5	1.7	2 6
adv. sub. - other	1.0	1.6	2.0	0 4	2 0	1.9	3 3	9 4	2 7	2.3	3 4	2 0	1 2	2 8	1.6	3.8	0.5	1.7
prepositions	0.2	-0.1	0.4	0 4	0 5	0 8	0.7	106	-0 2	0.2	0 6	0 3	0.3	0.2	0 3	0.3	0 1	-0 2
attributive adjectives	-1.4	-1.0	-1.7	-0 9	-0 3	0.0	-0 1	7.1	-1 2	-0.9	-0 6	-0 6	-0 6	-0 6	-0 7	-0 8	-1.2	-1 1
adverbs	-1.5	-1 5	-1 3	-1 8	-1 0	-2 0	-1 7	2 1	-1 0	-1 5	-1 8	-1.7	-1 5	-1 6	-1 0	-1.5	-2.2	-1.8
type/token ratio	-1.7	2 3	1.1	1.1	1.2	0 0	1.1	0 6	2 7	0 9	1.1	2 9	0.6	1.2	1.8	-0.8	0 9	1.1
word length	-1.3	0.0	-0 1	0 5	0 1	0 2	-0 2	0 1	-0.2	-0 8	-0 5	-0.2	-0.7	-0 9	-0 5	-0 7	-1.0	0 8
conjuncts	0.6	0.8	0.5	1.4	2 5	1 8	0.7	1 8	1 9	1.6	1 0	2 6	0 4	0 9	1.7	1.5	0 5	0.7
hedges	1.8	1.5	1.3	1.2	1.2	0 6	1 5	1 2	1 5	0.9	1.2	0 9	1.5	1.0	1.7	1 6	0.8	1.9
amplifiers	0.6	0 3	2.2	0 7	0 9	0 7	-0 2	0 3	0 5	0.9	0 4	0.5	0 0	0.1	0 5	1.0	0.9	0 6
emphatics	1 0	1.2	2.4	-0 5	1 2	0.4	0 4	0 9	1 7	1 7	0 8	1 3	1 2	1.2	1 6	1 2	1.1	1.2
discourse particles	1.7	2.1	0 3	0 2	-0 1	0 3	-0 3	0 8	2 2	1 2	-0 2	-0 3	0 5	0 5	-0.1	0.1	1 2	1.3
demonstratives	-0.9	-0 3	0 5	-0.3	-0 3	-0 5	-0.9	2.7	-0 9	-0.9	-1 1	-0.9	-1.0	-1.0	-0 8	-1 1	-0 8	-0 7
possibility modals	1.8	1 2	0.5	0 8	1 1	0 2	1 8	69	0 6	0.5	1.3	1.8	0.9	1 3	0.8	0 7	0.7	0.9
necessity modals	-0.3	-0.2	-0 1	0 2	0 1	0 3	0 0	1 0	-0 2	-0.3	0 3	-0.1	-0 5	-0 2	-0.4	0.0	0.4	0.3
predictive modals	0 9	0 8	1.3	0 9	0 8	1 0	0 2	4 0	0 0	0.1	0.8	1 0	-0 1	0 4	0 7	0 3	1.7	1 3
public verbs	-0 5	-0 4	-0 3	-0 3	-0 7	-0.6	-0.6	0 3	-0 2	-0 1	-0 4	-0 3	-0.8	-0.8	-0 1	0.1	-0 6	-0.6
private verbs	0 2	0 3	0.3	0 0	0 0	-0 1	0 1	2 5	0 8	0 4	0 0	0 1	-0 5	-0 3	0 3	0 4	0 2	0.3
suasive verbs	-0.7	-0 4	-0.6	-0.5	-0.5	-0 5	-0 5	0 2	-0 6	-0 6	-0 2	-0 4	-0.8	-0.5	-0 7	-0 5	-0.7	-0.7
contractions	0.7	0.3	-0.2	0 1	-0 3	-0 3	-0 1	0 8	0 5	0.2	-0.2	0.0	0.1	0 0	0.3	0.2	0.2	0.2
THAT deletion	1.6	1.6	1 4	0 7	0 6	0 3	1 4	3 6	2 1	1 4	1.5	1.5	0.3	0.6	1.5	1.7	1.2	1.3
stranded prepositions	-0.2	-0.2	0.0	-0 2	-0.5	-0 5	-0 5	0 5	-0.2	-0.4	-0 3	-0 5	-0 5	-0 4	-0 4	-0 4	-0.5	-0.1
split infinitives	0.0	0 0	0 0	0 0	0.0	0 0	0 0	0 0	0 0	0.0	0.0	0.0	0 0	0 0	0 0	0 0	0 0	0 0
split auxiliaries	-0.9	-1 3	-0.7	-1 1	-1 3	-1 2	-1 1	0 2	-1 5	-1 3	-1 4	-1 3	-1 5	-1.5	-1 1	-1.0	-1 4	-1.5
phrasal coordination	-0.1	0.0	0 6	0 1	0 7	0 8	0 7	6 7	0 6	0.8	0 0	0 5	2 4	1.4	0.7	0.9	0.1	0.1
non-phrasal coordination	2.8	2 6	2.0	2 8	1 7	1 7	2 4	7.1	3 5	2 6	2 4	1 6	2 4	2 6	1 6	3.0	2.3	2 6
synthetic negation	-0.5	-0.4	-0.8	-0 8	-0 8	-0.8	-1.1	-0 8	-0 4	-0 8	-0.9	-0 9	-1.0	-1.0	-0 8	-0.7	-0 8	-0.9
analytic negation	1 5	0 8	1 3	1.7	0 9	0 7	0 4	4 8	1 3	1.1	0 1	0.5	0.3	0 4	1 3	1.4	1.3	1.2

The second step in the computation of factor scores was simply to add up the FDS's for each of the linguistic features in a factor, and for each of the conferences in the corpus. To illustrate, let us consider factor 2 again. The operation was a simple addition, and read as follows: **past tense FDS + third person FDS + perfect FDS + public FDS + synthetic negation FDS + present participial FDS.**

The third, and final step consisted of calculating the average factor score for each conference.

5.4 Interpretation of factor scores

As we have seen in chapter 2, Biber interprets each of the factor scores as representing specific textual dimensions. Figures 1-6 described these textual dimensions, and situated each of the genres analyzed in a specific position along them. It is now possible to plot the factor scores extracted from the ELC onto the same graphic representation of these dimensions, thereby answering the original thesis question : "Where does electronic language fit on each of the dimensions outlined by Biber?". Figures 16-21 represent the new textual dimensions, and include both off-line and "other" ELCs for a comparison with Biber's results.

Figure 16.
Dimension 1:
Involved vs Informational Production

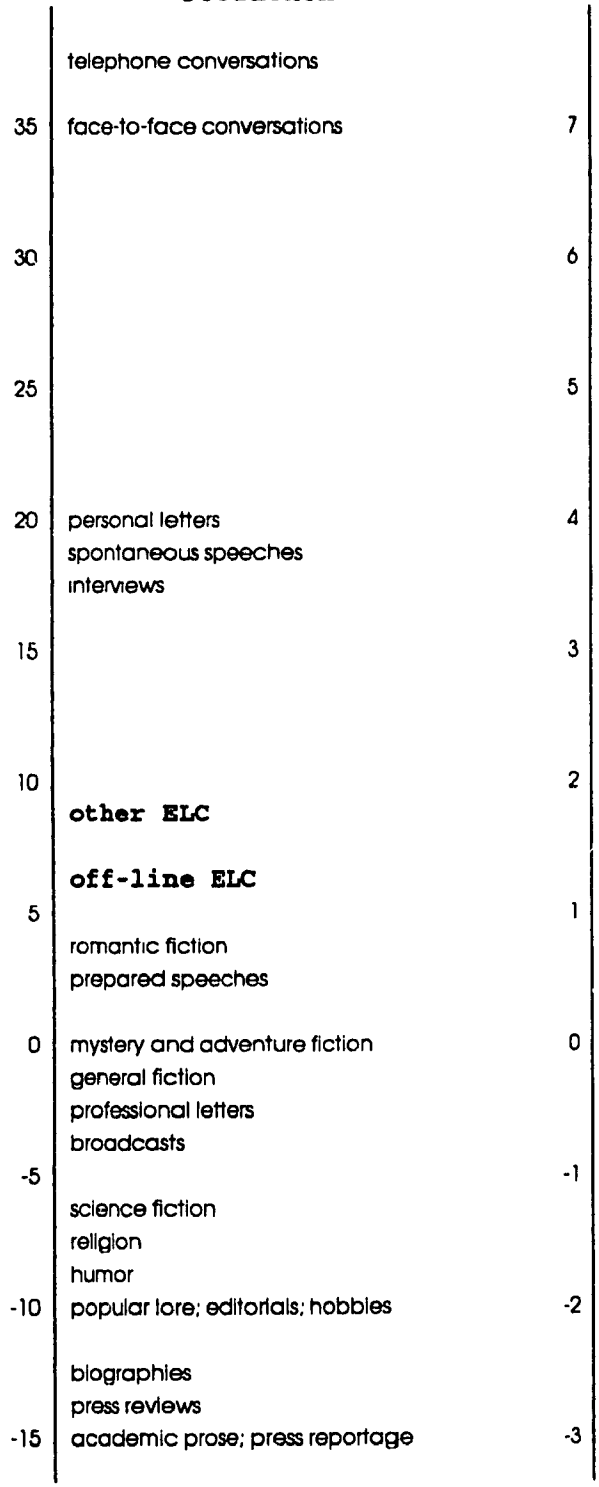


Figure 17.
Dimension 2:
Narrative vs Non-Narrative concerns

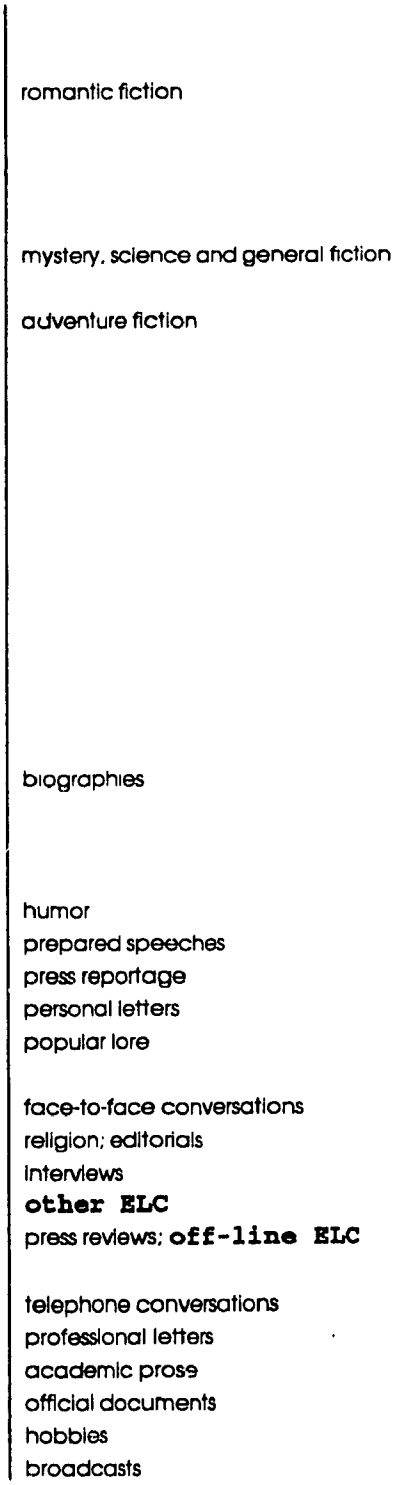


Figure 18.
Dimension 3:
Explicit vs Situation-Dependent.

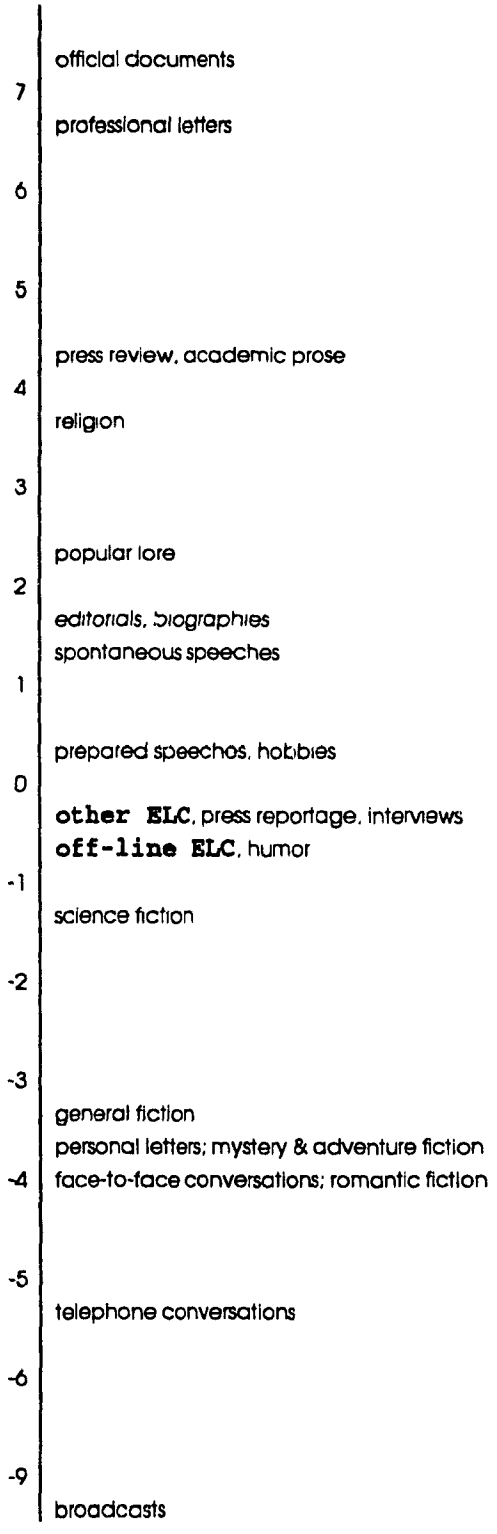


Figure 19.
Dimension 4:
Overt Expression of Persuasion

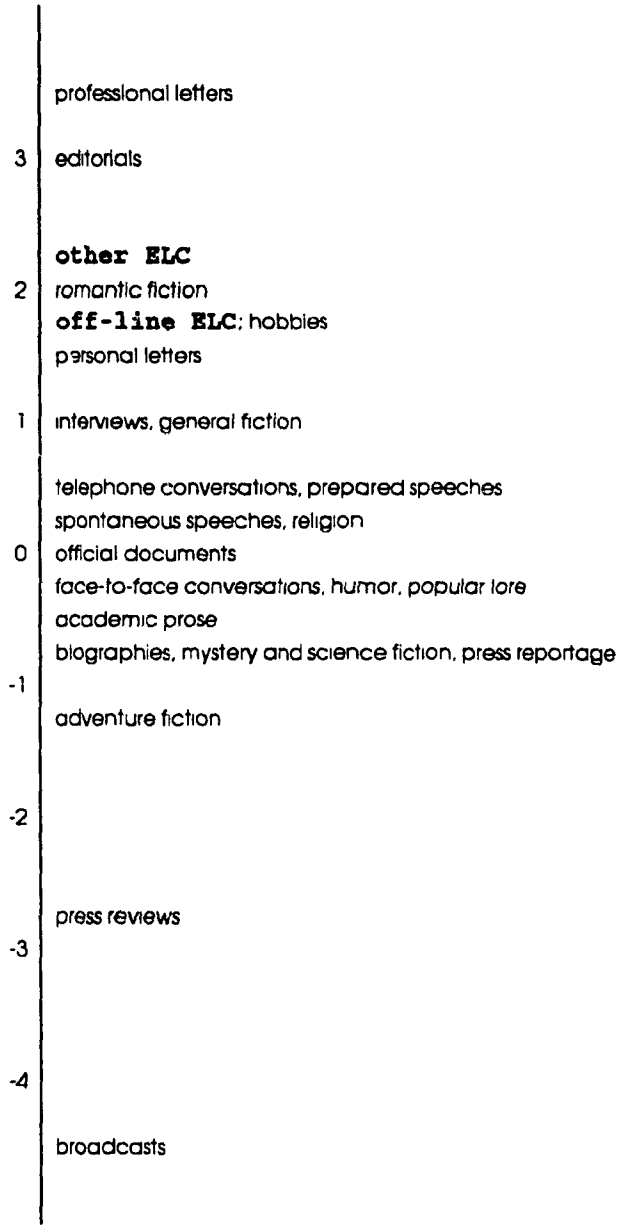


Figure 20.
Dimension 5:
Abstract vs Non-Abstract Information.

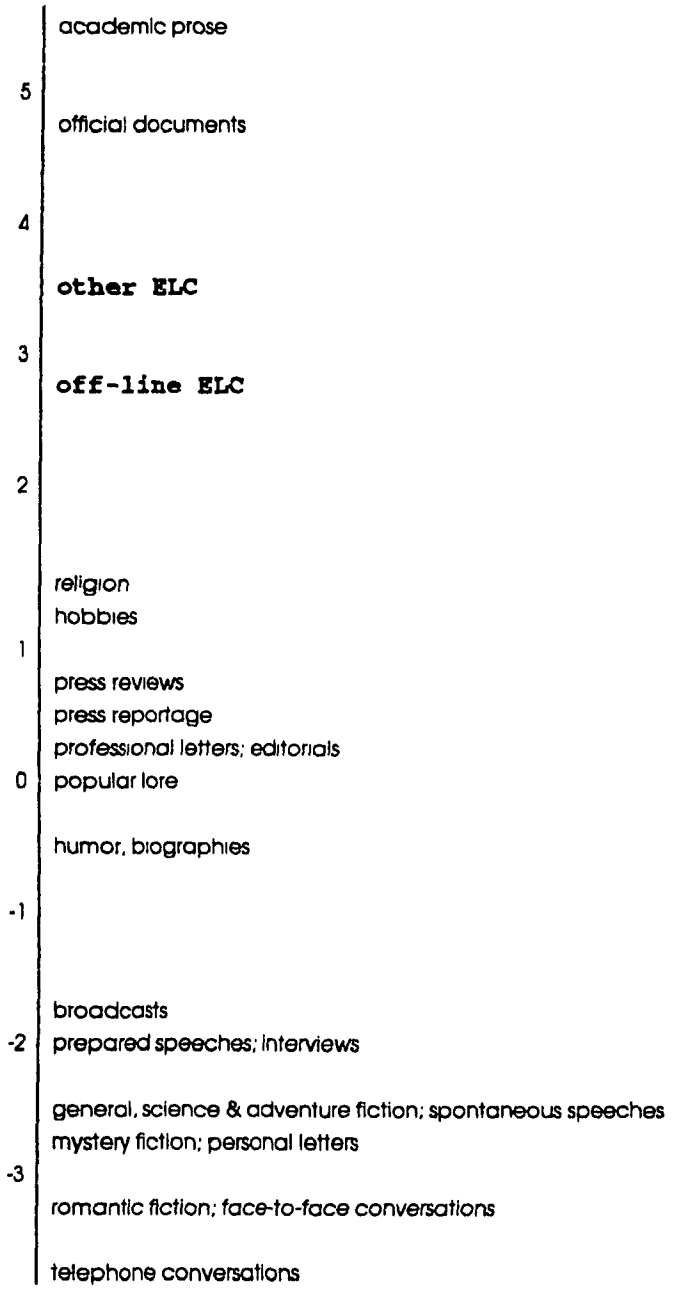
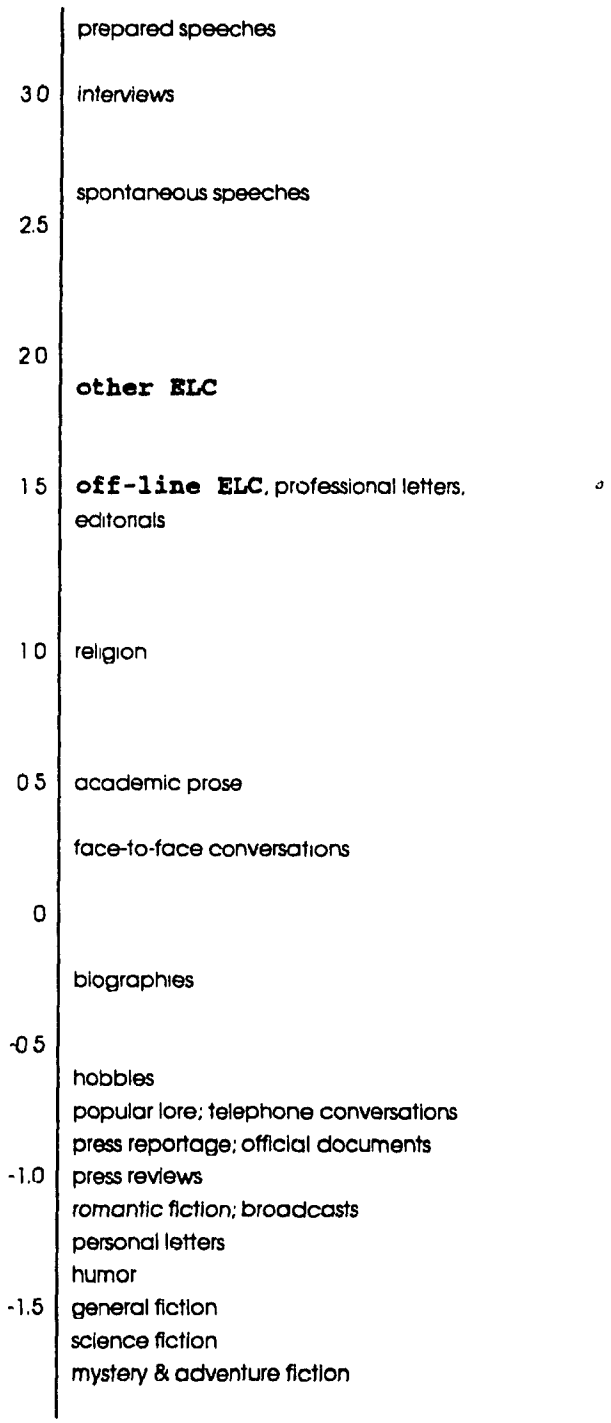


Figure 21.
Dimension 6:
ON-Line Informational Elaboration.



The first dimension, derived from Factor 1 in Biber's factorial structure, is labelled "Involved vs informational production". In Biber's factorial structure, nouns, word length, prepositions, type token ratios and attributive adjectives were all assigned substantial negative weights. A high frequency of these features would therefore represent a great density of information. For the ELC, Table 6 indicated that the frequencies for these features were largely comparable to Biber's corpus, with the exception of nouns, which were far more frequent in the ELC (225 per 1000 words) than in Biber's corpus (180 per 1000 words). This might have pushed the ELC towards the bottom end of Figure 16. However, if we consider the features in Factor 1 with positive weights, the balance is redressed. Some of the important features with positive weights are private verbs, THAT deletions, present tense verbs, and 1st and 2nd person pronouns, all of which indicate a highly verbal, and personally involved style. Again, Table 6 reveals figures which are largely comparable for the ELC and Biber's corpus as a whole. However, the frequency of many of the features in the ELC exceeds those in Biber's corpus. Thus, private verbs have a normalized frequency of 20 in the ELC and 18 in Biber, THAT deletions rate at 8.8 in the ELC and 3.1 in Biber, first person pronouns have a frequency of 57.5 in the ELC and 27.2 in Biber, and second person pronouns are measured at 18.3 in the ELC and 9.9 in Biber. The effect of these statistics was to place the ELC much closer to the informational end of dimension 1 than it would have been had we relied on the frequency of nouns and other negative features alone. Thus, in this dimension, the ELC is very close to personal letters, spontaneous speeches and interviews.

On Dimension 2, "Narrative vs non-narrative concern", there were a number of features with strong positive loadings, and no features with negative loadings. Biber sees all of these features as markers of narrative action. For example, past tense and

perfect aspect verbs describe past events, while third person pronouns (other than *it*) refer to animate, typically human, referents. Public verbs (*admit, say*, and so on) are also important in this dimension, because they function as markers of reported speech, and synthetic negation is often seen as the preferred style in literary narrative (*he said nothing* versus *he did not say anything*). In the case of the ELC, all of these features are considerably less frequent than they are in Biber's corpus. Table 6 shows a frequency of 28 past tense verbs per 1000 words in the ELC versus 40.1 in Biber, 4 perfect aspect verbs versus 8.6 in Biber, 25.5 third person pronouns versus 29.9, 5.5 public verbs versus 7.7, and 0.4 synthetic negation versus 1.7. This configuration places the ELC very near the bottom of Dimension 2, on the same level as telephone conversations and professional letters.

Dimension 3, "Explicit versus Situation-Dependent", is primarily concerned with the presence or absence of relative constructions. WH-relative clauses in object position (*the man who Sally likes*), pied piping relative clauses (*the manner in which he was told*), and WH-relative clauses in subject position (*the man who likes popcorn*) are all seen as devices for the explicit, elaborated identification of referents in a text. Thus, they carry large positive weights in this dimension. In contrast, the three features with negative weights, time adverbials, place adverbials and other adverbs, are seen as markers of reference to places and times outside the text itself, and rely on the ability of the listener or reader to infer, on the basis of context, the message which is implied. The scores for Factor 3 reveal that the ELC lies somewhere between these two extremes. Again, this is consistent with Table 6, which shows that there are relatively fewer WH-relatives than in Biber's corpus, and roughly comparable frequencies for place and time adverbials. Thus, for this

dimension, the ELC is situated on the same lines as the categories "interviews" and "humor".

In Dimension 4, all of the features are seen to function together to mark either the speaker's own point of view, or argumentative discourse intended to persuade the addressee. This function is clearly fulfilled by prediction modals and necessity modals. Similarly, suasive verbs (*command, demand*, and so on) are seen to imply an intention to bring about certain events in the future, and conditional subordination is interpreted as specifying the conditions required in order for certain events to occur. Finally infinitives, most commonly used as adjective and verb complements, are seen as marking the speaker's attitude towards the proposition encoded in the infinitive clause (*happy to do it, hope to see you*). Thus, Dimension 4 is labelled "Overt expression of persuasion". In the ELC, many of these features have a higher frequency than in Biber's corpus, as shown in Table 6. For instance, prediction modals have a frequency of 9 in the ELC and 5.6 in Biber, conditional subordination is rated at 7.9 in the ELC against 2.5 in Biber, and infinitives have a frequency of 17.6 against 14.9. However, there are three features which are less frequent in the ELC: suasive verbs, at 1.2 against 2.9, necessity modals, at 2.0 against 2.1, and split auxiliaries, at 2.3 against 5.5. As a result of these differences, the ELC is situated in the dimension somewhere between editorials and personal letters.

The underlying features of Dimension 5 are agentless passives, BY-passives, past participial clauses (*built in a single week*) and past participial clauses with relative deletion (*the solution produced by this process*). According to Biber, these forms are frequently used in procedural discourse, where emphasis on the agent is reduced because the same agent is presupposed across several clauses. Consequently, the dimension is labelled "Abstract vs Non-Abstract Information". Table 5 shows

that these features are generally as rare in the ELC as they are in Biber's corpus. For example, agentless passives have a frequency of 7.0 and 9.6, respectively, BY-passives rate at 1.2 and 0.8, past participial clauses are rated at 0.6 and 0.1, and past participial clause with deletions are rated at 2.5 for both corpora. However, Figure 20 would lead us to believe that the ELC contains many of these constructions. A possible explanation lies in the presence, in Factor 5, of two features which factor analysis reveals as frequently co-occurring with passive constructions: conjuncts (*however, in contrast*) and adverbial subordinators (*since, while*). Neither of these features are seen to play a major functional role, but they are included in the factor because they co-occur with the features mentioned above, which have the function of marking reduced emphasis on the agent. Both these features occur more frequently in the ELC than in Biber's corpus, with conjuncts rated at 3.1 against 1.2, and adverbial subordinators at 3.2 against 1. Moreover, one of the features, conjuncts, has the highest loading in the structure of that factor. It seems, then, that this dimension would have to be interpreted with caution, since the only features which do occur often are those which do not play a crucial functional role, and therefore have little interpretive value.

There are four features underlying Dimension 6: THAT clauses as verb complements, demonstratives, THAT relative clauses in object position, and THAT clauses as adjective complements. Biber sees the co-occurrence of these features as marking informational elaboration in relatively unplanned types of discourse. Hence, the label "On-line informational elaboration". Table 6 shows that the ELC rates more highly than Biber's corpus all four features: THAT verb complements have a frequency of 7.0 in the ELC and 3.3 in Biber, demonstratives are rated at 6.9 versus 4.6, THAT relative clauses on object position are rated at 1.7 against 0.8, and THAT

clauses as adjective complements are 0.4 against 0.3. Accordingly, the ELC appears towards the upper end of this dimension, slightly below spontaneous speeches and at the same level as editorials.

Chapter 6

DISCUSSION AND CONCLUSION

6.1 Main findings

The purpose of this study was to determine how the ELC compared to other varieties of English which have already been analyzed. The research question was operationalized as "Where does electronic language fit on each of the dimensions outlined by Biber?". The operations described in chapter 5 have provided an answer to this question. Let us now consider this answer as it pertains to the specific textual dimensions of the ELC, and to the more general relations between the ELC and the other varieties analyzed by Biber. In both cases, I will attempt to relate the findings to the components of the speech situation in Electronic Language, first presented in section 3.2.

6.1.1 Textual dimensions in the ELC

In terms of the specific textual dimensions of the ELC, the study has yielded the following results. On the continuum between involved and informational texts represented by Dimension 1, the ELC falls somewhere in the middle. This indicates that there is a tendency for texts to be highly informative, as signalled by the frequent use of nouns, the variety of lexical items, and the use of longer words. Considering the range of topics included in the ELC, this is not surprising. It would be difficult to imagine conferences with titles like "Science", "Medical" and "Finance" as containing little information. At the same time, however, participants in the ELC are reluctant to give up that personal involvement which is often associated with less informative prose (Chafe, 1982). In fact, the findings reveal that the ELC is replete with

indicators of involvement like first and second person pronouns, contractions, hedges (*sort of, kind of*) and amplifiers (*utterly, very*). From this, we may conclude that although one of the primary purposes for participating in a BBS is to seek and impart information, the language in which this information is couched is more similar to that of spontaneous genres, such as spontaneous speeches and private letters, than it is to that of informative genres like academic prose or press reportage.

A second finding is that the ELC contains very few instances of narrative discourse. This would suggest that ELC participants are not content with simple descriptions of past events, but tend to project their thoughts towards the present or the future. Concomitant with the finding that the ELC is non-narrative is the observation, interpreted from Dimension 4, that discourse in the ELC is highly persuasive, in much the same way as editorials are. A cursory glance at a string of BBS messages will quickly confirm this observation. In fact, many of the threads contain active, and often vociferous discussions on a wide range of subjects, which makes the ELC as rich in argumentation as it is in information.

Besides yielding information on the purpose of the ELC and the nature of the discourse types it embodies, the analysis also reveals that the ELC is situation-dependent, relying quite heavily on the reader-listener to infer meaning from context. An explanation for this may be found in the high degree of shared knowledge among participants in the ELC. Because participants in BBSs often know each other and their subject quite well, it is not necessary for them to make explicit reference to all of the concepts, issues and objects they are discussing. However, messages remain more explicit than discourse types in which participants are expected to know each other more intimately, such as face-to-face or telephone conversations among friends. After all, it must be recalled that although BBS users may know each other's opinions

on certain topics, they are likely to know little about each other's personal histories, and in most cases have never met their fellow BBS users in person.

The last finding relates to the manner in which ELC texts are structured. Dimension 6 reveals that ELC texts are largely produced under real time constraints. Participants tend to overcome this constraint by backtracking, making impromptu clarifications, and adding information in a sequence which is not strictly thematic. In turn, this constraint results in a fragmented presentation, which uses many short clauses, rather than an integrative presentation in which information is planned and presented in fewer constructions. This generalization is confirmed by the finding that messages written on-line contain more on-line elaboration than messages written beforehand. However, the difference is not great enough to explain this characteristic solely on the basis of the participants' relation to the text. Perhaps a more appropriate explanation can be found in Biber, who suggests that the same techniques used for informational elaboration (THAT clauses, demonstratives, and so on) can also be used to mark attitude o stance. In the case of the ELC, as we have seen, participants are frequently involved in discussion and persuasion. The observation that the ELC contains much on-line elaboration can therefore be explained in terms of the propensity of its participants for clearly expressing their attitude or stance.

6.1.2 Textual dimensions across genres

The findings have revealed that Electronic Language is both informative and involved, that its discourse is non-narrative and persuasive, that its referents are situation-dependent and that its organizational structure is highly fragmented. In terms of much of the research that has been done on the subject of speech and writing, this would situate the ELC closer to speech than to writing. For example, the

finding that the ELC is involved corresponds to a trait which is widely accepted in the literature as characterizing speech (Blankenship, 1962; Chafe, 1985). Similarly, its reliance on contextual reference can be traced to the commonly accepted description of spoken language as situation-dependent (Olson, 1977; Markova, 1978). Finally, its high degree of fragmentation conforms closely to the notion that speech is less structured and less systematic than writing (Crystal & Davy, 1969, Leech & Svartvik, 1975; Beaman, 1974).

There are, however, some characteristics of Electronic Language which would lead us to conclude that it is closer to writing than to speech. For example, the observation that the ELC is highly informational corresponds closely to various studies which have interpreted writing as "transactional" (Brown & Yule, 1983), "referential" (Jakobson, 1960) or "informative" (Stubbs, 1980). Similarly, the finding that it is non-narrative would seem to draw it closer to literate than to oral discourse (Scollon & Scollon, 1981).

It seems obvious that any attempt to interpret the findings of this study on the basis of previous research on speech and writing would have to heed Biber's admonishment that there is no absolute difference between speech and writing. In fact, Biber's own analysis shows that there is considerable overlap among written and spoken genres with respect to every dimension described. In order to interpret the ELC more accurately, then, it would seem more appropriate to isolate the specific genre, or genres, which resemble the ELC more closely, regardless of the broader distinction between speech and writing.

One interesting observation that emerges from this approach is that, in four of the five dimensions discussed here, the ELC is situated very close to the category "interviews". Specifically, the ELC can be compared to public interviews as being

slightly more involved, more persuasive and less narrative (Dimensions 1, 4 and 2), and equally situation-dependent (Dimension 3). An explanation for the similarity between the ELC and interviews may be sought in the situational constraints which define both types of speech events. As we saw in chapter 3, one of the components which characterizes the ELC and distinguishes it from many other varieties of English is the configuration of participant roles. Language interaction in a BBS typically involves a triad of roles: messages are produced by an addressor, directed to an addressee, and invariably posted for a wider audience to see. This situation is very similar to public interviews, in which the interviewee responds specifically to the interviewer, but also does so for the benefit of a wider audience.

The tripartite nature of participant roles would seem to strongly affect the ELC in a way that previous models of language description, which typically limited the description to a distinction between monologue and dialogue, could not predict. Even Crystal and Davy's model (1969), which allowed for language events to cross participation boundaries, as when monologue is introduced in a conversation for humorous effects, was not comprehensive enough to capture this subtlety.

The second genre to which the ELC is easily compared is letters, both personal and professional. In fact, the ELC is equivalent to letters in its non-narrative and persuasive concerns and in its on-line informational elaboration (Dimensions 2, 4 and 6); furthermore, it is only slightly less involved (Dimension 1). In order to interpret this similarity, we draw on the specific nature of the setting surrounding the ELC speech event, more particularly the extent to which space and time are shared by the participants. In much the same way as personal and professional correspondents, participants in the ELC share neither the same physical nor the same temporal context. Once again, the importance of this situational feature would seem to have

been obscured in previous models of language description. Yet, it seems to play an important role in explaining the peculiarities of the ELC, as predicted in chapter 2.

To sum up, there seem to be at least four situational features which play an important role in explaining the linguistic manifestations of the ELC. The first is the degree of shared knowledge among the participants, which allows ample opportunity for readers of BBS messages to infer meaning from context. The second is the purpose of communication, which is to request and impart information, and to engage in discussion of specific issues. Both kinds of purpose seem to play an important role in shaping Electronic Language as a highly persuasive and non-narrative discourse type. The third important component of the speech situation is the tripartite nature of the roles played by the participants, which include an addressor, an addressee and an audience. Finally, the fact that time and space are not shared by the participants may be at the root of the resemblance between the ELC on the one hand, and personal and professional letters on the other.

6.2 Secondary findings

One component which was expected to play a large role in the make-up of Electronic Language was the relationship of the speakers to the text. Thus, from the very beginning of the study, steps were taken to ensure that messages produced directly on-line be kept separate from those produced beforehand, or off-line.

In general, the relations between on-line and off-line texts would seem to confirm the more general findings which pertain to the corpus as a whole. However, there were slight differences between the two text types. For example, on-line texts were found to be slightly more involved than off-line texts, slightly less narrative,

and equally situation-dependent. Furthermore, they were slightly more persuasive and contained more on-line elaboration.

Two general conclusions can be drawn from these findings. The first relates to the tendency of on-line messages to be typically placed along each textual dimension in a position which brought them slightly closer to other spoken genres. In terms of involved or informational production, for example, the on-line ELC is slightly closer to interviews, spontaneous speech and letters, while the off-line ELC is more directly comparable to romantic fiction. Similarly, the tendency to offer impromptu elaborations would seem to draw the on-line ELC closer to spontaneous speeches than to professional letters. From this, we can infer that when BBS participants type their messages directly onto their computers, they are somewhat influenced by the pressure of real time constraints.

However, and this is the second conclusion, the differences between on-line and off-line were very slight. This may mean that it is not sufficient to attribute differences in the ELC merely to the fact that some of the messages are produced under real time constraints. If the time factor did play a large role, we would expect much greater differences between the two sub-corpora. The findings, however, reveal that on-line and off-line messages are separated by only a few points. Thus, other components of the ELC speech situation, such as the presence of an audience, the high degree of shared knowledge among participants, and the disembodiment of messages in space and time would seem, together, to exercise an influence which outweighs that of time constraints alone.

6.3 Limitations of the study

One important limitation of the study is directly related to the secondary findings discussed above. The categories used in this study, "off-line" and "other", do not in fact coincide perfectly with the distinction between off-line and on-line messages. As noted in chapter 4, there is a distinct possibility that the "other" category may not be restricted to on-line messages alone, but may include a number of off-line messages as well. This may have biased the results somewhat, as there may have been long, pre-planned messages in a category of texts which are assumed to be produced under real time constraints. As a result, the real impact of the situational component entitled "relationship of the participants to the text" may have eluded us here.

A second limitation pertains to the sizes of the various subsamples used. Although an effort was made to keep the sample sizes constant across conferences, certain conferences remained largely under-represented. I had allotted a certain amount of time for the smaller conferences, such as the medical and sports BBSs, to "grow", but after three weeks there were still very few new messages on these boards. One solution would have been to eliminate these conferences altogether. However, I feared that this would result in an overall picture which was not representative of electronic language in general. Another solution would have been to reduce the other conferences so they would be comparable with the smaller one. However, to reduce the size of the overall corpus would have impinged on the reliability of the study. Faced with this dilemma, I opted for reliability in favour of representativeness.

A third limitation lies in the fact that the ELC and Biber's corpus represent two geographical dialects which are not fully comparable. The vast majority of messages in the ELC came from North America. In Biber, on the other hand, over half the

texts are derived from the LOB corpus, which represents British English. Therefore, some of the differences found between the ELC and the genres analysed by Biber may have been due to differences in geographical variety, in addition to differences in purpose, shared knowledge, participant roles and spatial-temporal disembodiment.

The limitations discussed above all pertain to the design of the study. There were also some problems related to the tool used for parsing the corpus, CLAWS 1. Firstly, this tagging program was not the most appropriate, since it was designed to analyze the printed and published texts contained in LOB. As we have seen, electronic language abounds in abbreviations, substandard forms and nonce words. All of these words had to be coded, in order to allow the program to run at all, and then manually tagged, since CLAWS often tagged them incorrectly. Also, while the CLAWS tags are remarkable accurate, some errors do occur and it is difficult to find them all. In addition, all spelling errors had to be located and corrected before running CLAWS 1. Needless to say, all these operations involved quite a large degree of manual intervention which, in turn, may have increased the likelihood of error.

A final limitation arose from the fact that the algorithms in Biber were not always entirely accurate. For example, the formula for stranded prepositions, which reads as "Prep + all punctuation" would certainly capture instances like *the candidate that I was thinking of* or *who did you go with?*, but also includes some erroneous structures, like *I told him to come in*. Similarly, the formula for present tense verbs, which reads as "VBZ (verb marked for third person) and VB (base form of the verb) except when followed by TO" will capture, erroneously, verbs preceded by modals, as in *will eat*. Despite oversights of this sort, I decided not to correct the algorithms, as I wanted to avoid measuring something which was different from what Biber had

measured. The consequence, however, may be that the frequency counts for the linguistic features, in this study and in Biber's, may have a slight margin of error.

6.4 Extending the description

This study has succeeded in identifying the situational features which have an impact on the linguistic make-up of Electronic Language. The purpose, the degree of shared knowledge, the role of the participants, and the extent to which time and space are shared by the participants, all seem to play a major role in shaping the linguistic structure of this genre. However, what this study has not determined is the degree to which each situational feature affects the overall linguistic configuration of electronic language.

Future studies should examine some of these situational features more closely. A particularly interesting feature, and one which has eluded us here, is the relationship of the participants to the text. This feature was particularly difficult to control for in this study, because it was not always possible to tell whether messages had been written on-line or off-line. If carefully planned, a future study will be able to make a more accurate distinction. For example, a number of sysops could be contacted across the country, and told to create a password which would clearly indicate whether people are writing on line or off. This study could then be replicated, but this time an important variable would be strictly controlled.

Another interesting extension of this study would be to examine the variation within the genre I have labelled ELC. Rather than grouping all of the conferences together, as was done here, it may be interesting to examine the relations among the different conferences along the dimensions outlined by Biber. This kind of study

would produce a description of the internal coherence of the ELC, from which it would be possible to infer the importance of other situational variables, such as topic.

Regardless of the direction that future studies may take, it should be borne in mind that telecommunications are steadily and dramatically gaining in importance the world over. Electronic language, which gives voice to such communication, would therefore seem to be worthy of further exploration.

REFERENCES

Aarts, J. & W. Meijs (eds). (1984). **Corpus linguistics: recent developments in the use of computer corpora in English language research**. Amsterdam: Rodopi.

Aijmer, K. (1986). Why is actually so popular in spoken English? In G. Tottie & I. Backlund (Eds.), **English in speech and writing: A symposium**. Uppsala: Acta Universitatis Uppsaliensis.

Akinnaso , F. N. (1982). On the differences between spoken and written language. **Language and speech, 25**, 97-125.

Altenberg, B. (1986). Contrastive linking in spoken and written English. In Tottie,G., & Backlund, I (Eds.), **English in speech and writing: A symposium**. Uppsala: Acta Universitatis Uppsaliensis.

Atwell, E. (1981). LOB manual PRE-EDIT handbook. Unpublished manuscript, Departments of Linguistics and Computer Studies, Lancaster University.

Backlund, I. (1986). Beat until stiff: Conjunction-headed abbreviated clauses in spoken and written English. In Tottie, G & Backlund,I. (Eds.), **English in speech and writing: A symposium**. Uppsala: Acta Universitatis Uppsaliensis.

Beaman, K. (1984). Coordination and subordination revisited: syntactical complexity in spoken and written narrative discourse. In Tannen, D. (ed.), **Coherence in spoken and written discourse**, 45-80. Norwood, N.J.:Ablex.

Belmore, N. (1991). Tagging Brown with the LOB tagging suite. **ICAME Journal, 15**, 63-86.

Berko-Gleason, J. (1973). Code switching in children's language. In Moore, T.E. (ed.), **Cognitive development and the acquisition of language**. New York: Academic Press.

Bernstein, B.B. (1973). **Class, codes and control. Vol 2: Applied studies towards a sociology of language**. London: Routledge and Kegan Paul.

Biber, D. (1988). **Variation across speech and writing**. Cambridge: Cambridge University Press.

Biber, D. & Finegan, E. (1986). An initial typology of English text types. In J. Aarts and W. Meijs (eds.), **Corpus Linguistics II**, 19-46. Amsterdam:Rodopi.

Biber, D. & Finegan, E. (1988). Drift in three English genres from the 18th to the 20th centuries: A multidimensional approach. In M. Kyoto, O. Ihalainen & M. Rissanen (eds.), **Corpus linguistics, hard and soft,** 83-101. Amsterdam:Rodopi.

Blankenship, J. (1962). A linguistic analysis of oral and written style. **Quarterly journal of speech, 48,** 419-422.

Bloomfield, L. (1933). **Language.** New York: Holt.

Boas, F. (Ed.). (1911). **Handbook of American Indian languages.** Bureau of American Ethnology, Bull. 40. Washington, D.C.: Government Printing Office.

Boas, F. (1940). **Race, language and culture.** New York: MacMillan.

Brown, R., & Ford, M. (1961). Address in American English. **Journal of Social Psychology, 62,** 375-385

Brown, G., Currie, K.L., & Kenworthy, J. (1980). **Questions of intonation.** London: Croom Helm.

Brown, G. & Yule, G. (1983). **Discourse analysis.** Cambridge: Cambridge University Press.

Candlin, C.N., Burton, C.J., and Leather, J.L. (1976). Doctors in casualty: applying communicative competence to components of specialist course design. **IRAL, 14,** 245-272.

Chafe, W. (1982). Integration and involvement in speaking, writing and oral literature. In, Tannen, D. (ed.) **Spoken and written language: Explaining orality and literacy.** 35-55. Norwood, N.J.: Ablex.

Chafe, W. (1985). Linguistic differences produced by differences between speaking and writing. In Olson, D.R., Torrance, N., & Hildyard, A (eds.), **Literature, language, and learning: the nature and consequences of reading and writing,** 105-123. New York: Academic Press.

Crystal, D. & Davy, D. (1969). **Investigating English style.** Bloomington & London: Indiana University Press.

Crystal, D. (1975). **The English tone of voice: Essays in intonation, prosody and paralanguage.** London: Edward Arnold.

De Vito, J.A. (1967). Levels of abstraction in spoken and written language. **Journal of communication, 17,** 354-361.

Dillard, J. L. (1972). **Black English: Its history and usage in the United States.** New York: Random House.

Drieman, G.H.J. (1962). Differences between written and spoken language. **Acta psychologica, 20,** 78-100.

Fee, M. (1989). The Strathy language unit, **Canadian humanities computing, 3**(3):8.

Fielding, G., & Fraser, C. (1978). Language and interpersonal relations. In Markova, I. (ed.), **The social context of language**. London: Wiley.

Firth, J. R. (1959). **Papers in linguistics: 1934-1951**. London: Oxford University Press.

Francis,W.N. (1980). A tagged corpus: problems and prospects. In S. Greenbaum, G. Leech, and J. Svartvik (eds.), **Studies in English linguistics for Randolph Quirk**. London and New York: Longman 192-209.

Francis, W.N. (1982). Problems of assembling and computerizing large corpora. In Johansson, S. (ed.), **Computer corpora in English language research**. Bergen: Norwegian Computing Centre for the Humanities.

Francis, W.N. & Kucera H. (1964). **Manual of information to accompany a standard sample of present-day edited American English, for use with digital computers**. Revised ed. 1971; revised and augmented (with Henry Kucera) 1979. Providence: Department of linguistics, Brown University.

Francis, W.N. & Kucera, H. (1982). **Frequency analysis of English usage: Lexicon and grammar**. Boston: Houghton Mifflin.

Fries, C.C. (1940). **American English grammar: The grammatical structure of American English with especial reference to social differences of class dialects.** New York: Appleton.

Garside, R. G. (1988). The CLAWS word-tagging system. In Garside, R.G., Leech, G., & Sampson, G. (Eds.), **The computational analysis of English.** London: Longman

Garside, R.G., Leech G. N., & Sampson G. R., eds. (1988). **The computational analysis of English**. London: Longman

Halliday, M.A.K., McIntosh, A. & Strevens, P. (1964). **The linguistic sciences and language teaching**. London: Longman.

Halliday, M.A.K. (1970). **Language structure and language function.** In Lyons, J. (ed.), New horizons in linguistics. Harmondsworth: Penguin Books.

Harmer, J. (1983). **The practice of English language teaching**. London: Longman.

Hermeren, L. (1986). Modalities in spoken and written English: An inventory of forms. In Tottie,G., & Backlund, I. (Eds.), **English in speech and writing: A symposium**. Uppsala: Acta Universitatis Uppsaliensis.

Hofland, K. & Johansson, S. (1982). **Word frequencies in British and American English**. Bergen: Norwegian Computing Centre for the Humanities/ London: Longman.

Hymes, D. ('972). On communicative competence. In Pride, J.B., & Holmes, J. (eds.) (1982). **Sociolinguistics: Selected readings.** Harmondsworth: Penguin Books.

Ihalaien, O. , M. Kyoto and M. Rissanen. (1987). The Helsinki corpus of English texts: diachronic and dialectical report on work in progress. In W. Meijs (ed.), **Corpus linguistics and beyond: Proceedings of the seventh international conference on English language research on computerized corpora.** Rodopi: Amsterdam

Jakobson, R. (1960). Closing statement: linguistics and poetics. In Sebeok, T.A. (ed.), **Style in language.** Cambridge: M.I.T. Press.

Johansson, S., Leech, G.N., & Goodluck, H. (1978). **Manual of information to accompany the Lancaster-Oslo/Bergen corpus of British English, for use with digital computers**. Oslo: Department of English, University of Oslo.

Johansson, S. & Hofland, K. (1987). The tagged LOB corpus: description and analyses. In Meijs, W. (Ed.), **Corpus linguistics and beyond**. Amsterdam: Rodopi, 1-20.

Johansson, S. & Hofland, K. (1989). **Frequency analysis of English vocabulary and grammar**. Oxford: Clarendon Press.

Joos, M. (1961). **The five clocks: A linguistic excursion in the five styles of English usage.** New York: Harcourt, Brace and World.

Kachru, B.B. (1983). **The Indianization of English: the English language in India.** New Delhi and New York: Oxford University Press.

Kay, P. (1977). Language evolution and speech style. In Blount, B.G. and Sanches, M. (eds.), **Sociocultural dimensions of language change,** 21-33. New York: Academic Press.

Kittredge, R., & Lehrberger, J. (eds.) (1982). **Sublanguage: Studies of language in restricted semantic domains.** Berlin; New York: de Gruyter.

Kjellmer, G. (1984). Some thoughts on collocational distinctiveness. In Aarts, J. & Meijs, W. (eds.), **Corpus linguistics: Recent developments in the use of computer corpora in English language research.** Amsterdam: Rodopi.

Kroll, B. (1977). Ways communicators encode propositions in spoken and written English: a look at subordination and coordination. In Keenan, E.O. & Bennett, T. (eds,), **Discourse across time and space.** Los Angeles: University of Southern California.

Kucera, H & Francis, W.N. (1967). **Computational analysis of present-day American English.** Providence, Rhode Island: Brown University Press.

Labov, W. (1971). The study of language in its social context. In J.A. Fishman (ed.) **Advances in the sociology of language,** 152-216. The Hague: Mouton

Labov, W. (1972). **Sociolinguistic patterns.** Philadelphia: University of Pennsylvania Press.

Ladd, D.R. (1978). **The structure of intonational meaning.** Bloomington & London: Indiana University Press.

Leech, G. & Svartvik, J. (1975). **A communicative grammar of English.** London: Longman.

Leech, G. & Beale, A. (1984). Computers in English language research. **Language teaching and linguistics, 17,** 216-229.

Littlewood, W. (1981). **Communicative language teaching: An introduction.** Cambridge: Cambridge University Press.

Llamzon, T.A. (1969). **Standard Filipino English.** Manila: Ateneo University Press.

Lougheed, W.C. 1986. **In search of the standard in Canadian English.** Kingston: Strathy Language Unit.

Lyons, J. (1977). **Semantics.** Cambridge: Cambridge University Press.

Malinowski, B. (1935). Coral gardens and their magic. London: Allen & Unwin. Reprinted as **The language of magic and gardening: Indiana University studies in the history and theory of linguistics.** (1967). Bloomington: Indian University Press.

Marshall, I. (1987). Tag selection using probabilistic methods. In Garside, R., Leech, G. & Sampson, G. (eds.), **The computational analysis of English.** London: Longman.

Meijs, W. (1984). You can do so if you want to - Some elliptic structures in Brown and LOB and their syntactic description. In Aarts & Meijs (eds.), 141-162.

Mindt, D. (1986). Corpus, grammar and teaching English as a foreign language. In, Leitner, G. (ed.), **The English reference grammar.** 125-139. Tubingen: Niemeyer.

Noss, R.B., ed. (1983) **Varieties of English in Southeast Asia.** Singapore: SEAMO Regional Language Center.

O'Donnell, R.C. (1974). **Syntactic differences between speech and writing.** American speech, 59, 102-110.

Ochs, E. (1979). Planned and unplanned discourse. In Givon, T. (ed.), **Discourse and syntax**, 51-80. New York: Academic Press.

Olson, D.R. (1977). From utterance to text: the bias of language in speech and writing. **Harvard educational review, 47**, 257-281.

Pedersen, H. (1931). **The discovery of language: Linguistic science in the 19th century.** Bloomington: Indiana University Press.

Quirk, R. (1960) Towards a description of English usage. **Transactions of the philological society**, 40-61.

Renouf, A. (1987). Corpus development. In J. Sinclair, (ed.), **Looking up. An account of the COBUILD project in lexical computing.** London, Glasgow: Collins. 1-40.

Rosch, W.L. (1987). The modern modem: Bridge to the on-line world. **PC Magazine, 6**, 100-235.

Rubin, A. (1980). A theoretical taxonomy of the differences between oral and written language. In Rand, J.C., Bruce, B.C. & Brewer, F.W. (eds.), **Theoretical issues in reading comprehension:perspectives from cognitive psychology, linguistics, artificial intelligence, and education.** 411-438. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Sapir. E. (1921). **Language, and introduction to the study of speech.** New York: Harcourt, Brace and World.

Sey, K.A. (1973). **Ghanaian English: An exploratory survey.** London: Macmillan and Co.

Shastri, S.V. 1988. The Kolhapur corpus of Indian English and work done on its basis so far. **ICAME News, 12**: 15-23.

Sinclair, J. McH., & Coulthard, R.M. (1975). **Towards an analysis of discourse.** Oxford: Oxford University Press.

Sinclair, J. H. (1982). Reflections on computer corpora in English Language research. In Johansson (1982).

Smitherman, G. (1977). **Talkin and testifyin: The language of Black America.** Boston: Houghton Mifflin.

Spencer, J., ed. (1971) **The English language in West Africa.** London: Longman

Stubbs, M. (1980). **Language and literacy: The sociolinguistics of reading and writing.** London: Routledge and Kegan Paul.

Sutcliffe, D. (1982). **British Black English.** Oxford: Basil Blackwell.

Svartvik, J. (1990). **The London-Lund corpus of spoken English: Description and research.** Lund: Lund University Press.

Svartvik, J. & Quirk, R. (eds.). (1980). **A corpus of English conversation.** Lund: Lund studies in English 56.

Svartvik, J. , Eeg-Olofsson, M., Forsheden, O., Orestrom, N. & Thavenius, C. (1982). **Survey of spoken English: Report on research, 1975-1981.** Lund: Lund studies in English 63.

Taylor, L. & Leech, G. (1989). **Lancaster preliminary survey of machine-readable language corpora.** Bergen: Norwegian Computing Centre for the Humanities.

Tongue, R.K. (1974). **The English of Singapore and Malaysia.** Singapore: Eastern Universities Press.

Torode, B. (1976). Teachers' talk in classroom discussion. In Stubbs, M., and Delamont, S. (eds.), **Exploration in classroom observation.** London: Wiley.

Tottie, G. (1982). The co-occurrence of negation, modality and 'mental verbs' in spoken and written English. In Jacobson, S. (ed.), **Papers from the second Scandinavian symposium on syntactic variation,** 67-74. Stockholm: Almqvist & Wiksell International.

Tottie, G. (1986) The importance of being adverbial: Adverbials of focusing and contingency in spoken and written English. In Tottie, G. & Backlund (eds.).

Tottie, G. & Backlund, I. (1986). **English in speech and writing: A symposium.** Uppsala: Acta Universitatis Upsaliensis.

Valdman, A, ed. (1977). **Pidgin and Creole linguistics.** Bloomington & London: Indiana University Press.

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

Wilkins, D. A. (1976). **Notional Syllabuses.** London: Oxford University Press.

Winer, L. (1989). **Dictionary of Trinbagonian.** English today, 5, 17-22.

APPENDIX A: EXCERPT FROM THE CHAT CONFERENCE

[004 TEXT HMC]
|^Such a warm feeling pervades when one recognizes true appreciation of one's poetic prowess! ^Bless you my son! ^Jeeesh<tm> I'm sorry, Gary is supposed to do that for me!

[005 TEXT HMC]
|^Usen's here ain't got no reel 'spect fer no dern queen imposin' her english on usen's. ^You'ens speak yer wey an us'ens will all kick the queen en her royal tushie!
|^It's more nicer to not critisize no body than it is to talk yer way. ^Joel, does you have other compulsiveations in yer life? ^Bet ya eat off clean platters and put on church socks and starch yer undergarmets!
|^Hang out here an us'ens will teach you to talk right, write rite and think rite. ^Yer'er in heaven now, usen's is smert folks!
|^Sea ya Joel, I'm a goin' ta colage class now; studeyin Shakesarrow!

[006 TEXT GHA]
|^Always there to lend a helpin' hand, huh Mark-y?

[012 TEXT GHA]
|^Wrong conference...

[016 TEXT MMA]
|^Any other parts of your been damaged that we should know about?
|^Maisel

[017 TEXT DRO]
|^Why does that irritate you?
|^dave

[020 TEXT RMO]
|^(Looking around frantically) None that I'm aware of.
|^Rick

[022 TEXT HMC]
|^Frankly, I think I sense a strong neurological implication. ^Severe personality defects can start this way and major pschotic breaks can eventually be traced to obcessive manifestations that have to do with **correctness****.
|^This gentleman may be in need of help from VERBANON, a group that tries to help compulsive victims live in a world of imperfect verbs!

[023 TEXT HMC]
|^Now we have established that you are rather pompous but this statement makes it quite clear that Logic 101 may be a major goal you need to consider.
|^If you are, in fact, a sensitive scholar of english prose you certainly would have know better than to post statements that might discourage others from expressing themselves. ^Your choice of words reak with

insensitivity and I personally feel you should reconsider your entire approach to **helping*** others overcome their *"supposed***" deficiencies.
|^Many outstanding linguists completely disagree with your dogmatic conclusions. ^The written word can be considered completely effective if it correctly coveys the thoughts at issue. ^Whether the clauses are set out according to ancient tradition is probably totally moot!
|^To anyone out their who may **worry*** about their use of grammer; don't **worry*** about it. ^Write what you feel the way you feel it when you feel it and punctuate it the way you like to see it or don't use any punctuation at all as I am doing here. ^Enjoy expressing yourself in writtend form and don't let petty pedagogues discourage your efforts!
^If you decide to attend Oxford, you might consider brushing up on English concepts of correct grammatical expression!

[024 TEXT JFR]
|^Or just a salt lick sized valium. ^I think most folks that go off the deep end over little stuff need to have one sitting on the computer desk.

[027 TEXT KBU]
|^Heck, E-Mail is not a pretty sight!

[035 TEXT RHI]
|^That is a misspelling, you twit.
|^(grin)

[036 TEXT RHI]
|^Not in these parts it isn't.

[038 TEXT JCO]
|^Well, some of the response to this topic has begun to get ugly. ^Apparently, many people on these boards are pretty unwilling to listen to constructive feedback. ^My appreciation to those who feel like I do and who, at times, whispered their agreement without wanting to look un-cool. ^To those who really got nasty about this, I wish you luck (you're probably going to need it).
|^Funny how many of you are programmers who get upset if one character is out of place on a screen and yet have no problem making mistakes in communicating that a third grader would correct. ^No one has a problem if people criticize or make suggestions about their software. ^Why does this topic cause such resentment? ^Why are people so unwilling to make an effort to correct some basic mistakes in their English?
|^Some have told me to lighten up, but read some of the warboard material coming this way. ^I thought my message was pretty low key. ^Other people are allowed to vent their pet peeves, and talk freely, but apparently not about this topic.
|^Well, 'nuff said. ^I'm out of here. ^Keep making embarrassing spelling and grammar mistakes in your software. ^Let's keep it mediocre, folks!

[039 TEXT DMA]
|^What do you expect from a man with multiple wives?

|^-Dave

[040 TEXT DMA]
|^Oh come now, surely the modern world can do better than this.
|^-Dave

[043 TEXT RHI]
|^A singular event to say the least; it's not often that I show self-restraint.

[045 TEXT CMO]
|^Pal ol boy, as I said to another in a similair predicament to your own, check the batteries in your sarcasm detector. ^If anyone was really upset you'd know it. ^You have the unfortunate condition of being bothered by something relatively trivial, and many folks (myself inculded) are more than happy to give those with trivial concerns a bit of ribbing.
|^Probably because there is a marked difference between textbook English and colloquial English. ^You have made the mistake of assuming that since these messages are written that they should adhere to proper grammatical form. ^Such might be the case where there was formal debate on a particular topic (science, politics, theology). ^As it is, this conference is a catch-all blow-off zone, where most of the posts are idle chitchat, jokes, and informal discussion. ^Thus, colloquialism reigns.
|^Don't take it so hard, Joelerini.

[052 TEXT DRO]
|^You'd have a hard time enjoying reading THE COLOR PURPLE by Alice Walker. ^It's a masterful piece of writing, but it is not technically *"correct***. ^And you'd have trouble enjoying bluegrass music from Appalachia because although it has great feeling and meaning to many, it's not technically *"correct***. ^And you probably don't like rap, either. ^;-)
|^dave

[053 TEXT DRO]
|^:-)
|^dave

[054 TEXT HMC]
|^You're absolutely correct, this is not a *"WAR***" board! ^So, the next tim
you decide to openly criticize another person's use of grammatical expression you need to keep your critical thoughts to yourself.
|^Referring to other people as *"slim***" is certainly not the epitomy of genteel expression, either!
|^You knew when you made the posting your comments were not going to be well received. ^You potentially embarrassed others with a public posting of their expressions and I sincerely believe you should consider a public apology to any you might have offended.

|^As far as I am concerned this thread is terminated. ^I have no desire to do to you what you are apparently willing to consider doing to others. ^I'm sure you're a nice person, be tolerant of the weaknesses of others, you will be stronger for it!

[058 TEXT HMC]
|^Ulcers, severe despondency, chronic constipation, involuntary **tics*** about the eyes and arms and blood pressure problems. ^These would be the LEAST serious problems a man with several wives would confront!

[059 TEXT HMC]
|^I hear the Primary God of the ROASF lives as an amorphous mass in dismal depths of the Okeefenokee! ^Could it be Gary?

[060 TEXT WBR]
|^Maybe I'll try to make a meeting, I no longer WORK nights, so just maybe I can get up and try to find this place. ^Your the one with the crutchs right?

[063 TEXT WBR]
|^Hay we are not all good spelars, like urself. ^And who gives a rats ass anyway, we are all tring to have fun cumunicating. ^So enjoy or get out.
|^<WJB3)

[064 TEXT JZA]
|^It did. ^He, he, he, he.

[065 TEXT JZA]
|^I thought I saw his initials in the Penthouse letters Mag. ^:-)

[066 TEXT JZA]
|^Nice to be able to copy the same massage over and over.
|^Did you pick up on that one too?
|^PS. ^I know how ↄ spell. ^Its my typing that really is bad.
|^(did you also pick up on the missing punctuation mark in the above?)
|^See...it AIN'T that easy, is it?
:-)

[072 TEXT HMC]
|^I concur, Jeeesh<tm> it bothers me to see someone make another person feel bad about something they might not be able to help. ^OTOH, maybe we can start a new conference on southern grammer, northern grammer, continental grammer, colloquial grammer, Turabian style, Denealian style, MLA form and maybe, for the heck of it, plain ole' locally accepted expression.

[074 TEXT HMC]
|^Hey dude,
You are getting really aggressive, been eating the **beaties*** lately!?

[082 TEXT RHI]
|^Come on Hugh, no one that is easily offended hangs out in here.

[107 TEXT RHI]
|^Gawd! ^That makes me want to run right out and kiss a redneck!
|^(Actually, I thought it was kinda cute. ^Good shot.)

[108 TEXT RHI]
|^Don't fret Kevin; we know you're a Georgia Tech student and we will make special allowances for you.

APPENDIX B: TAGSET USED BY CLAWS 1 ON THE ELC

```
'!     ',PUN04;      (*punctuation tag - exclamation mark*)
'&FO   ',AFO  ;      (*formula*)
'&FW   ',AFW  ;      (*foreign word*)
'(     ',PUN03;      (*punctuation tag - left bracket*)
')     ',PUN08;      (*punctuation tag - right bracket*)
'*''    ',PUN07;     (*punctuation tag - open single quotes*)
'**''  ',PUN05;      (*punctuation tag - close single quotes*)
'*-    ',PUN06;      (*punctuation tag - dash*)
',     ',PUN11;      (*punctuation tag - comma*)
'-----',PUN10;       (*punctuation tag - new sentence marker*)
'.     ',PUN01;      (*punctuation tag - full stop*)
'...   ',PUN02;      (*punctuation tag - ellipsis*)
':     ',PUN13;      (*punctuation tag - colon*)
';     ',PUN09;      (*punctuation tag - semicolon*)
'?     ',PUN12;      (*punctuation tag - question mark*)
'ABL   ',ABL  ;      (*pre-qualifier*)
'ABN   ',ABN  ;      (*pre-quantifier*)
'ABX   ',ABX  ;      (*pre-quantifier/double conjunction*)
'AP    ',AP   ;      (*post-determiner*)
'AP$   ',APX  ;      (*post-determiner + genitive*)
'APS   ',APS  ;      (*OTHERS*)
'APS$  ',APSX ;      (*OTHERS'*)
'AT    ',AT   ;      (*singular article*)
'ATI   ',ATI  ;      (*singular or plural article*)
'BE    ',BE   ;      (*BE*)
'BED   ',BED  ;      (*WERE*)
'BEDZ  ',BEDZ ;      (*WAS*)
'BEG   ',BEG  ;      (*BEING*)
'BEM   ',BEM  ;      (*AM*)
'BEN   ',BEN  ;      (*BEEN*)
'BER   ',BER  ;      (*ARE*)
'BEZ   ',BEZ  ;      (*IS*)
'CC    ',CC   ;      (*coordinating conjunction*)
'CD    ',CD   ;      (*cardinal*)
'CD$   ',CDX  ;      (*cardinal + genitive*)
'CD-CD',CDXCD;       (*hyphenated pair of cardinals*)
'CD1   ',CD1  ;      (*ONE*)
'CD1$  ',CD1X;       (*ONE'S*)
'CD1S  ',CD1S ;      (*ONES*)
'CDS   ',CDS  ;      (*plural cardinal*)
'CS    ',CS   ;      (*subordinating conjunction*)
'DO    ',DO1  ;      (*DO*)
'DOD   ',DOD  ;      (*DID*)
'DOZ   ',DOZ  ;      (*DOES*)
'DT    ',DT   ;      (*singular determiner*)
'DT$   ',DTX1 ;      (*singular determiner + genitive*)
'DTI   ',DTI  ;      (*singular or plural determiner*)
'DTS   ',DTS  ;      (*plural determiner*)
'DTX   ',DTX  ;      (*determiner/double conjunction*)
'EX    ',EX   ;      (*existential THERE*)
'HV    ',HV   ;      (*HAVE*)
```

```
'HVD  ',HVD  ;      (*HAD past tense*)
'HVG  ',HVG  ;      (*HAVING*)
'HVN  ',HVN  ;      (*HAD past participle*)
'HVZ  ',HVZ  ;      (*HAS*)
'IN   ',IN1  ;      (*preposition*)
'JJ   ',JJ   ;      (*adjective*)
'JJB  ',JJB  ;      (*semantically superlative adjective*)
'JJR  ',JJR  ;      (*comparative adjective*)
'JJT  ',JJT  ;      (*superlative adjective*)
'JNP  ',JNP  ;      (*adjective with word-initial capital*)
'MD   ',MD   ;      (*modal*)
'NC   ',NC   ;      (*cited word*)
'NN   ',NN   ;      (*singular common noun*)
'NN$  ',NNX  ;      (*singular common noun + genitive*)
'NNP  ',NNP  ;      (*common noun with word-initial capital*)
'NNP$ ',NNPX ;      (*common noun with word-initial capital + genitive*)
'NNPS ',NNPS ;      (*plural common noun with w.i.c.*)
'NNPS$',NNPSX;      (*plural common noun with w.i.c. + genitive*)
'NNS  ',NNS  ;      (*plural common noun*)
'NNS$ ',NNSX ;      (*plural common noun + genitive*)
'NNU  ',NNU  ;      (*unit of measurement unmarked for number*)
'NNU$ ',NNUX ;      (*unit of measurement unmarked for number + genitive*)
'NNUS ',NNUS ;      (*unit of measurement marked for number*)
'NNUS$',NNUSX;      (*unit of measurement marked for number + genitive*)
'NP   ',NP   ;      (*proper noun*)
'NP$  ',NPX  ;      (*proper noun + genitive*)
'NPL  ',NPL  ;      (*locative noun with w.i.c.*)
'NPL$ ',NPLX ;      (*locative noun with w.i.c. + genitive*)
'NPLS ',NPLS ;      (*plural locative noun with w.i.c.*)
'NPLS$',NPLSX;      (*plural locative noun with w.i.c. +  genitive*)
'NPS  ',NPS  ;      (*plural proper noun*)
'NPS$ ',NPSX ;      (*plural proper noun + genitive*)
'NPT  ',NPT  ;      (*titular noun with w.i.c.*)
'NPT$ ',NPTX ;      (*titular noun with w.i.c. + genitive*)
'NPTS ',NPTS ;      (*plural titular noun with w.i.c.*)
'NPTS$',NPTSX;      (*plural titular noun with w.i.c. + genitive*)
'NR   ',NR   ;      (*adverbial noun*)
'NR$  ',NRX  ;      (*adverbial noun + genitive*)
'NRS  ',NRS  ;      (*plural adverbial noun*)
'NRS$ ',NRSX ;      (*plural adverbial noun + genitive*)
'OD   ',OD   ;      (*ordinal*)
'OD$  ',ODX  ;      (*ordinal + genitive*)
'PN   ',PN   ;      (*nominal pronoun*)
'PN$  ',PNX  ;      (*nominal pronoun + genitive*)
'PP$  ',PPX  ;      (*first possessive personal pronoun*)
'PP$$ ',PPXX ;      (*second possessive personal pronoun*)
'PP1A ',PP1A ;      (*I*)
'PP1AS',PP1AS;      (*WE*)
'PP1O ',PP1O ;      (*ME*)
'PP1OS',PP1OS;      (*US*)
'PP2  ',PP2  ;      (*YOU*)
```

```
'PP3  ',PP3  ;      (*IT*)
'PP3A ',PP3A ;      (*HE,SHE*)
'PP3AS',PP3AS;      (*THEY*)
'PP3O ',PP3O ;      (*HIM,HER*)
'PP3OS',PP3OS;      (*THEM*)
'PPL  ',PPL  ;      (*singular reflexive personal pronoun*)
'PPLS ',PPLS ;      (*plural reflexive personal pronoun*)
'QL   ',QL   ;      (*qualifier*)
'QLP  ',QLP  ;      (*post-qualifier*)
'RB   ',RB   ;      (*adverb*)
'RB$  ',RBX  ;      (*adverb + genitive*)
'RBR  ',RBR  ;      (*comparative adverb*)
'RBT  ',RBT  ;      (*superlative adverb*)
'RI   ',RI   ;      (*adverb which can also be a preposition*)
'RN   ',RN   ;      (*nominal adverb*)
'RP   ',RP   ;      (*adverb which can also be a particle*)
'TO   ',TO1  ;      (*infinitival TO*)
'UH   ',UH   ;      (*interjection*)
'VB   ',VB   ;      (*verb*)
'VBD  ',VBD  ;      (*verb past tense*)
'VBG  ',VBG  ;      (*present participle*)
'VBN  ',VBN  ;      (*past participle*)
'VBZ  ',VBZ  ;      (*verb 3rd person singular*)
'WDT  ',WDT  ;      (*wh-determiner*)
'WP   ',WP   ;      (*wh-pronoun, neutral between nomin. & obj.*)
'WP$  ',WPX  ;      (*possessive wh-pronoun*)
'WPA  ',WPA  ;      (*nominative wh-pronoun*)
'WPO  ',WPO  ;      (*objective wh-pronoun*)
'WQL  ',WQL  ;      (*wh-qualifier*)
'WRB  ',WRB  ;      (*wh-adverb*)
'XNOT ',XNOT ;      (*NOT or N'T*)
'ZZ   ',ZZ   ;      (*letter of the alphabet*)
```

Three additional tags, for *private*, *public* and *suasive* verbs, were appended to the tagged corpus, with the aid of a program written specially for that purpose.