A Comparative Analysis of Computer-mediated Communication (CMC) Versus Non-CMC Texts Along the Dimension of Abstract vs. Non-Abstract Information

電腦網路傳訊(CMC)與非電腦網路傳訊(non-CMC)之文本在抽象程度上之比較分析

蔡素薰

Abstract

The similarities and differences between written and spoken forms of language have been a focus of interest of many scholars. Recent studies usually establish one or two dimensions along which to measure the difference between spoken and written forms (Ure, 1971; Stubbs, 1986; Halliday, 1989; Ljung, 1991; Smeltzer, 1992; Botta, 1993; Kress, 1994). In addition to the many possible dimensions along which language use may be depicted as being more oral or written, different genres and media also have direct impact on its features. The coming into existence of Computer-mediated Communication (CMC) has made the line of distinction even less obvious. It is technically a writing (key-pressing) behaviour but may be used to carry out spontaneous communication.

This paper is intended to be a pilot study in text analysis to investigate the special linguistic features of CMC versus non-CMC texts. The model of analysis is based on that of Biber (1988, 1989), who, through statistical factor analysis, presents seven dimensions on which texts may be measured as being more spoken or written. Limited by the scope of research, this study focuses only on the features underlying the fifth factor: abstract versus non-abstract information.

Another limitation is on the objects of analysis. The CMC texts used for analysis are limited to the asynchronous mode only, and do not include texts in the synchronous mode. The texts selected for analysis are two Internet archives, Neteach-L and TESL-L. The theme of these texts is on Teaching English as a Second or Foreign Language, and almost all writers are TESL teachers. The texts are first tagged by the University of Birmingham tagging program, then analyzed using the concordancing program CLAN (MacWhinney, 1996a), and finally computed using SPSS statistical program (SPSS, 1993).

The features found in those CMC archives are compared with those in the non-CMC texts in Biber's corpus. Findings from this study are that the two CMC asynchronous text corpora stand somewhat in the middle of the dimension between the more abstract and technical genres of academic proses and the more concrete leisure genres of broadcasts and conversations.

It is hoped that on completion this pilot study will pave the way for a more comprehensive analysis of CMC (both asynchronous and synchronous modes) and non-CMC texts covering all the seven dimensions in Biber's model. It is thus hoped that the findings of this series of studies will provide EFL professionals with an extended understanding of the features of language in the computer age.

摘要

本研究係針對電腦網路通訊(CMC)中使用之語文之口說或書寫傾向所作的分析研究。研究之模式係採用Biber (1988, 1989)對口說語文及書寫語文間異同之各種層面所作之因素分析結果。Biber共歸納出七項有效之層面，因本研究係一全面性研究前之預備性研究，本文僅就其中第五項:「抽象與非抽象」作為分析之標準。

本研究分析之語料來自兩個非即時性(asynchronous)討論群

(Neteach-L及TESL-L)的內容，此兩討論群之討論主題均以英語教學為限。本研究在進行時，先將語料原文以電腦加註其詞類標籤，再以CLAN程式就符合各項語文結構之語詞挑出，其結果再由SPSS軟體加以統計。

本研究對CMC語料分析所得之結果經與Biber原作中對non-CMC語料之分析結果比對，發現非即時性網路討論群的語料，在所謂「抽象與非抽象」層面上表現的性質，大致介於non-CMC語料中口說及書寫兩種傾向之間。至於即時性(synchronou)的CMC語料所表現的特性如何，CMC語料在其它各項層面上可能顯示的性質如何，皆有待下一階段之繼續研究。

1. Introduction and Literature Review

The similarities and differences between written and spoken forms of language have been a focus of interest of many scholars (Chafe & Danielewicz, 1987; Halliday, 1989; Hasan, 1968; Kress, 1994; Sinclair et al., 1993; Tannen, 1982). In early years dichotomous differences were sought between orality and literacy (Ong, 1982; Finnegan, 1988). Spoken language, uttered by speech organs, was considered spontaneous and fragmented; while written language, produced with a pen or a typewriter, was viewed as prepared and highly organized (Close, 1994).

More recent studies in this field take into account the fact that language uses are in different genres (Bhatia, 1993; Littlefair, 1991). Instantaneous dinner-table talk between family members can be very different from a formal open speech by a high-ranking official. Personal letters between friends are usually different from research papers submitted to a professor.

In recent years, it is widely accepted that, instead of being dichotomous forms, the spoken mode and written mode may differ only in degree along some continuum (Bakhtin, 1986; Bhatia, 1993; Derewianka, 1996; Littlefair, 1991). Comparative studies between the two are usually done along one or several chosen dimensions. Halliday (1989), Stubbs (1986), and Ure (1971) all use lexical density as a measure to distinguish written from spoken messages. Kress (1994) puts more weight on structure, arguing that written language has more clausal complexity. Researchers may also use word choice (Halliday, 1989; Ljung, 1991; Vande-Kopple, 1995), readability (Botta, 1993; van Hout-Wolters & Schnotz, 1992), and message structure complexity (Laina, 1992; Smeltzer, 1992) to measure the differences. It can be seen that the difference between spoken and written forms of language lies on more than one dimension. The different findings in these studies can be attributed to the different genres of the texts being compared and the different dimensions chosen as the basis of comparison.

However, most of the earlier researchers analyze linguistic variation only from a single parameter, and many of them (Bernstein, 1970; Blankenship, 1974; DeVito, 1966; Ferguson, 1959; O'Donnell et al., 1967) tend to treat the linguistic variation in terms of dichotomous distinctions rather than continuous scales.

Some of the earlier researchers are even in disagreement with each other. Blankenship (1962, 1974) found that sentence length is nearly the same in speech and writing, while O'Donnell (1974) claimed that sentence length is considerably longer in writing. Besides, Kroll (1977) found that writing has more subordination than speech while Blass and Siegman (1975) found little differences on that. Viewing these disagreements, Ervin-Tripp (1972), Hymes (1974), Brown and Fraser (1979) give warnings that it is misleading if linguistic variation have been analyzed only with some specific, isolated linguistic markers without taking into account the sets of co-occurring features in

texts.

Seeing these disagreements, Biber (1985, 1986, 1989, 1995) argues that no any single dimension or function could adequately account for the linguistic differences among written and spoken modes, instead, a multi-feature/multi-dimension approach would be appropriate for this task.

Through an empirical study on the corpora LOB (Johansson et al., 1978; Johansson, 1982) and London-Lund (Svartvik & Quirk, 1980; Johansson, 1982), Biber examined the frequency counts of particular linguistic features (See Appendix 1) existed in texts across genres. He found that some features tended to occur strongly together across a range of texts. This gave the evidence that texts actually comprised several dimensions. For example, passives co-occurred with nominalizations in scientific texts with abstract and informational focus. First and second person pronouns co-occurred with contractions in face-to-face conversation underlying interactive situations. And past tense verbs co-occurred with third person personal pronouns in fiction being features in narrative focus. These patterns of feature co-occurrence appeared across different genres were the ones that defined the basic linguistic dimensions of English. Through the factor analysis, he established the multi-feature/multi-dimension approach from clusters of co-occurred feature patterns.

Unlike other researchers' studies based on the assumption of dichotomous distinction, Biber's approach takes the oral/written distinction as on a continuous scale of variation. Different types of texts in various styles, registers, genres are not the same or dichotomously different; rather they are "similar, or different", to differing extents with respect to each dimension (Biber, 1988; p.22). Based on the 67 linguistic features found in a corpus of 960,000 words and 23 written and spoken genres (Biber, 1988, p. 67), he finds a total of seven factors, which may serve as dimensions on which texts may be measured as being more speaking-like or writing-like. These include:

Factor 1: informational versus involved production,
Factor 2: narrative versus non-narrative concerns,
Factor 3: explicit versus situation-dependent reference,
Factor 4: overt expression of persuasion,
Factor 5: abstract versus non-abstract information,
Factor 6: on-line informational elaboration, and,
Factor 7: academic hedging.

Each dimension is characterized by some of these 67 features.

In addition to the above, the constant development of new communication technologies has also made the distinction between spoken and written modes less obvious (Ferrara et al., 1991; Maynor, 1994). Radio broadcast and tape recording, for instance, have broken the barriers of space and time that are usually associated with the spoken mode of language. The fax machine has caused certain changes in the style of written messages. The development of a new medium, it seems then, inevitably results in some new styles of language use.

The coming into existence of Computer-mediated Communication (CMC) has brought a new medium to human interactions (Adkins & Brashers, 1995; Baron, 1984; Hardy et al., 1994). It is technically a writing behaviour as messages are entered by key-pressing and are transmitted through visual symbols. The electronic transmission of the messages, however, has made possible a variety of modes of message exchange. Users may be engaged in spontaneous on-line talk and/or conference, and asynchronous modes such as e-mail or electronic journal.

This new mode of communication is certainly unconventional and comprises the

features found both in the spoken and the written modes. Its existence has made the distinction between spoken and written forms even more vague (Close, 1994; Maynor, 1994; Murry, 1988; Warschauer, 1996).

As computer-mediated communication is certainly going to be more popular in the years to come, it can be anticipated that this new medium will result in greater impact on the language styles of future generations.

This research is intended to be a pilot study of a text analysis project to investigate the linguistic characteristics of CMC texts versus non-CMC texts. The methods, the procedures, and the findings of this study are presented in the next sections.

2. Methods of Text Analysis

2.1 Model of Analysis

The model of analysis in this study is mainly adopted from the series of studies by Biber (1986, 1988, 1989, 1995) and others (Biber & Hared, 1992; Biber et al., 1994). A selection of 67 linguistic features (See Appendix 1) are collected as reflecting spoken versus written features. Further statistical factor analysis of these 67 features has produced a total of seven factors.

The main project will try to measure the CMC and non-CMC texts on all these seven dimensions. In the present pilot study, however, the researcher intends to focus on the fifth factor, i.e. abstract versus non-abstract information.

Among the 67 language features examined by Biber, eight are found to relate to the factor of abstract versus non-abstract information. Table 2.1 is the list of their serial numbers and their loadings on this factor.

Table 2.1 Language Features Underlying the Factor of Abstract and Non-Abstract Information (Factor 5)

Serial No.	Features	Loadings
F45	conjuncts	.48
F17	agentless passives	.43
F26	past participle clauses	.42
F18	BY passives	.41
F27	past participle WHIZ deletions	.40
F38	other adverbial subordinators	39
F41	predicative adjectives	.31
F43	type/token ratio	-.31

Following Biber's model, only the features with a loading larger than .35 in absolute value are considered valid in the computation of the factor score. Therefore, only the first six of the above eight features are searched and analyzed for this factor.

2.2 Objects of Study

While only Factor 5 is being studies in this pilot study out of a total of seven factors, the objects of the study are also limited. Data taken for analysis in the main project will be from several selected archives in Internet sources (CMC texts) as well as the traditional paper-format journals (non-CMC texts). In terms of the CMC texts, there will be synchronous texts (such as on-line conference logs from Neteach-L) and asynchronous

texts (such as postings on discussion lists like Neteach-L and TESL-L). It is hoped that a comparison can be made of the relative standings of these different types of CMC on each of the seven dimensions.

In the present pilot study, only the CMC asynchronous texts are collected as samples. They are postings from the discussion lists Neteach-L (Neteach-L, 1996) and TESL-L (TESL-L, 1996).

The theme of these archived texts mainly concerns issues relating to teaching English as a second or foreign language, and almost all the writers are TESL teachers, who are either native or non-native speakers teaching in primary, secondary, tertiary or adult levels in different countries

The selected archives are downloaded in electronic form from the related remote sites. To achieve a correct and precise analysis, all irrelevant lines in the files, like mail headers, are removed.

In short, while the main project will cover all the seven dimensions for comparison and use a variety of CMC and non-CMC texts as samples, the present study only deals with the fifth dimension, i.e. abstract versus non-abstract information; and only takes asynchronous CMC texts for the empirical analysis.

2.3 Tools for Data Preparation and Analysis

The sampled files from the above archives are processed using the computer programs: 1) the concordancing program CLAN, developed at Carnegie Mellon University (MacWhinney, 1995, 1996a, 1996b); 2) the part-of-speech tagging program TAGGER developed in University of Birmingham (1991); and 3) SPSS commercial quantitative program (SPSS, 1993).

2.3.1 CLAN

CLAN (Child Language ANalysis) is a set of programs written by Leonid Spektor at Carnegie Mellon University with design assistance from Brian MacWhinney. These programs are designed to allow users to perform a large number of automatic analyzes of transcript data formatted according to the CHAT system of Child Language Data Exchange System. However, many of the programs can run on ASCII files of any type. They include programs for doing frequency counts, Boolean searches, keyword in context searches, cooccurrence analyzes, mean length utterance counts, interactional analyzes, text changes, and so on. The two programs used in the present study are COMBO (for key-word-in-context search) and FREQ (for frequency count).

The programs have been written in the C language and can be compiled for a variety of operating systems, including MS-DOS, UNIX, and MVS. The one used in this study is MS-DOS system.

2.3.2 TAGGER

The program TAGGER is used to affix tags of the linguistic features to each word so that the features can be identified and counted by the CLAN program.

TAGGER is an automatic POS (Part-Of-Speech) program developed at University of Birmingham when John Sinclair edited the COBUILD Dictionary in 1988. It was originally used to facilitate the dictionary compilation, and was improved as a tagging program later on. Part-of-speech tagging is a linguistic procedure which attaches word-class information to the words in a text. This information is useful for further linguistic study, either for analyzing the syntactic structure of the texts' sentences or for statistical work such as counting the distribution of the different word classes in text corpora. A list

of part-of-speech tags is presented in Appendix 2.

TAGGER calculates the most probable word class in case of ambiguities (for example, a word can belong to several word classes like *light*, which can either be a noun, a verb or an adjective, depending on its actual use). Both the probability of the word belonging to a certain word-class and the probability of the word-class occurring at the specified position in the text are taken into account. Since it is probabilistic, there is no guaranteed correctness, but currently that is the limit of automatic tagging without human intervention. The correctness of this program is quite high, though it is not evaluated with any exactness as yet.

The program is now publicly accessible by means of an Experimental E-mail Tagging Service (TAGGER, 1996). The text can be sent to the TAGGER in University of Birmingham, and the output files would be automatically sent back. To avoid the difficulty of getting a long text tagged, it is advised that texts not exceed 50 KB in each file.

2.3.3 SPSS

SPSS is a commercial program dealing with numerical data for a statistical purpose. The program analyzes data in forms of Mean, Standard Deviation, Range, Sum, as well as Chi-square, T-test, Anova, Correlation etc. In this research, only Mean, Standard Deviation, Range, Minimum Value, Maximum Value are computed for the scores of Factor 5 abstract versus non-abstract.

2.3.4 Use of the CLAN and TAGGER

The researcher first develops notations of the language features to be searched for by the CLAN. For the six valid features relating to Factor 5, their notations are:

F17 Agentless Passives

@verb_be^*^(/VBN)^*^!by

This notation represents the occurrence of a BE verb followed by a word tagged as "/VBN" (i.e. a past participle), which is NOT followed by the word "by." The occurrence of the ^*^ means there may or may not be any number of words occurring at that position.

F18 BY Passives

@verb_be^*^(/VBN)^*^by

This notations is similar to the one above with the exception that the word "by" must follow the past participle in the sentence.

F26 Past Participle Clauses

@sym.txt^(/VBN)

The @sym.txt is a file of several punctuation marks (,.?!) that serve as delimiters. When a past participle occurs after one of these punctuations, it is counted as an occurrence of this feature.

This notation is supposed to locate sentences like:

He sat there with a smiling face, satisfied with what he had earned.

He couldn't say a word, astonished at what he saw.

Of course, manual proofreading and revising are necessary to distinguish this feature from other sentences like:

The diamond was found, bought, and finally sent to the queen as a present.

F27 Past Participle WHIZE Deletions

/VBN

This notation simply tells the CLAN to look for any occurrence of words with the tag "/VBN." However, here it is the intention of the researcher to locate texts like:

The solution produced by the process ...

A situation resulted from this decision ...

To distinguish texts with this feature with other texts simply containing a word with "/VBN" tagging, there seems no choice but to resort to manual editing.

F38 Other Adverbial Subordinators

@ad_sub

The @ad_sub means a pre-edited file containing all words that are considered adverbial subordinators. These include: *since, while, whilst, whereupon, whereas, whereby,* etc.

F45 Conjuncts

@sym.txt^(@conj)

The @conj is a file of all conjuncts, such as *alternatively, altogether, consequently, else, furthermore,* etc. This notation indicates the occurrence of such conjuncts after a delimiter.

For the CLAN to locate and count the occurrence of the specified features, the texts are usually first tagged with the symbols like /VBD and /VBN attached to all words. The researcher takes advantages of the automatic TAGGER service at University of Birmingham and has the CMC files tagged.

Although the jobs of tagging and concordancing are done by the computer programs, there is still the need for manual proofreading and correction. The TAGGER program still has limitations in its discrimination power and the notations designed with CLAN looking for the language features can not always elicit precisely the features needed. This is due to the clear limitations of the computer capability in linguistic analysis at the present stage.

2.4 Procedures of Analysis

Procedures of analysis cover several stages of data preparation, CLAN concordancer implementation, and results for statistical computation. They are illustrated in the following sections.

2.4.1 Data Preparation

Data preparation was divided into several steps:

First, the researcher ordered from TESL-L archives, by e-mail, five weeks' postings of the daily logs of November 1996. Neteach-L postings, however, were ordered for the 61 daily logs for October and November 1996 as their configuration system is different from TESL-L. It was then necessary to merge the five weeks' files, and the 61 days' files into two big corpora for later processing.

Next, the mail headers of each posting were removed as those were generated by the computer system and not written by the computer users. Each posting was then assigned serial numbers with the prefix of "no", "nn" to stand for the October/November postings of Neteach-L, and "ln" for November ones of TESL-L. All words including citations, and signatures were kept for data analysis as they were all written or arranged by the posters. At this stage, the data were still in two large corpora.

Further checking was done to remove any possibly weird symbols that would affect the computing process or results. Some modification was done to avoid these problems.

The revised postings were then sent to TAGGER through the e-mail system for automatic tagging. After the tagged texts were sent back, individual postings were divided into separate files. All of them were saved under different directories. At the same time, the untagged source data were also separated into individual files. This procedure resulted in two sets (tagged and untagged) of files for Neteach-L postings, with 265 files in each set, and also two sets of files for TESL-L postings, with 357 files in each set.

Before the data could be searched for linguistic features, a special directory was necessary for working purposes. All related files were stored in this working directory, including both sets of files of Neteach-L and TESL-L postings, the files of group words which needed to be searched for specific linguistic features, as well as the COMBO and FREQ commands from CLAN. The former two kinds of files has to be in the ASCII format.

2.4.2 Implementing the CLAN Concordancing

The COMBO and FREQ commands of CLAN were then used to locate and count the occurrence of features in the subject files by referring to the notations mentioned in 2.3.4.

While working in this way, it was necessary to leave enough memory in the computer for computing. Otherwise, the system might crash. Both COMBO and FREQ would report the stages of execution during the working process. Once it stopped unexpectedly, the search needed to be started again, or required some repair. Due to the limitations of the notations in fully representing the language features being analyzed, data generated form COMBO still needed to be carefully proofread for the precise entry discrimination.

Sometimes, it was necessary to refer back to the source data to make decisions as to whether the generated data were really the correct entries for specific features. An alternative way was to allow more context for this decision making. The "-w/+w" switch in the COMBO command could be used to attach more lines around the key pattern of specified features being searched. This "-w/+w" was a must while computing the tagged data as the tagging process added its tag-codes to the words and hence reduced the word numbers in a line.

2.4.3 Results for Statistic Computation

After proofreading, the correct frequencies of each of the six linguistic features in every posting were depicted in two tables compatible with the SPSS package, one for Neteach-L, and the other for TESL-L. These two tables then served as input files for computing by the SPSS program and gradually became converted into the two tables of factor scores of these two CMC discussion lists. The statistical procedures are discussed in the next section.

3. Statistical Procedures

3.1 Analysis of Frequencies and Scores on Individual Postings

3.1.1 Frequencies of Features in Individual Postings

The procedure in Section 2.4 generated a table of frequencies of occurrence of each linguistic feature relating to the factor. That is to say, each individual posting had a number showing the number of times one of the six features occurred in that posting. For

instance, for postings in the TESL-L, the following table was generated:

Table 3.1 A Sample Data Sheet of Frequencies of Features in Individual Postings of TESL-L

TSL96N	TTLWDS	F17FREQ	F18FREQ	F26FREQ	F27FREQ	F38FREQ	F45FREQ
ln0001	204	0	0	0	0	0	0
ln0002	236	3	0	0	0	1	1
.							
.							
ln0355	422	3	2	0	0	0	3
ln0356	183	0	0	0	0	0	1
ln0357	210	4	1	0	0	0	2

There are 357 postings adopted for analysis from TESL-L discussion list. The abridged table shows that Posting ln0001 has a total of 204 words, and none of the six features occurs in this posting; that Posting ln0002 has a total of 236 words and Feature 17 (agentless passives) occurs 3 times, Feature 38 (other adverbial subordinators) occurs once, and Feature 45 (conjuncts) occurs once in this posting; and so on.

3.1.2 Raw Feature Scores in Individual Postings

The frequencies found in Table 3.1 have to be converted into feature scores calculated on the same basis. As each posting is of different length, the frequency can not be used for any comparison unless the numbers are converted for calculation on the same basis. The "raw feature score" is an adjustment of the raw frequency to show the occurrence of a feature as if each posting were of the same length. The formula to be used for this conversion is:

*Raw Score = Frequencies of every feature in one posting / Total words of this posting * 1,000*

The raw feature scores received from this conversion are depicted in Table 3.2. The table shows that, among the 357 postings in TESL-L, the raw feature scores in ln0001 are all zero for the six features; that in ln0002, the raw scores are 12.71 for Feature 17 (agentless passives), 4.24 for Feature 38 (other adverbial subordinators), 4.24 for Feature 45 (conjuncts); and so on.

Table 3.2 A Sample Data Sheet of Raw Feature Scores in Individual Postings of TESL-L

TSL96N	TTLWDS	F17RAW	F18RAW	F26RAW	F27RAW	F38RAW	F45RAW
ln0001	204	.00	.00	.00	.00	.00	.00
ln0002	236	12.71	.00	.00	.00	4.24	4.24
.							
.							
ln0355	422	7.11	4.74	.00	.00	.00	7.11
ln0356	183	.00	.00	.00	.00	.00	5.46
ln0357	210	19.05	4.76	.00	.00	.00	9.52

3.1.3 Standard Feature Scores in Individual Postings

As these raw feature scores for individual postings are calculated on the same basis of posting length, they can be justifiably compared with each other in this pool of data. However, it is the intention of the researcher to compare the features with a larger pool of data. Besides, the comparison would not be of too much significance if the standard deviations of the features scores were not considered.

As the data collected for this study are limited in scope, only 123,986 words in total, the researcher has decided to take the statistical data of a larger pool of texts as a comparison. The most convenient source of data would be the corpus Biber (1988, p. 67) gathers for his study, approximately 960,000 words. The statistic figures of the six linguistic features based on his corpus are listed in Appendix 3 and serve as the basis of calculating the standard feature scores in the present postings.

The standard feature scores are calculated using the formula:

Feature Score = (Raw Scores - M (Biber)) / SD (Biber)

M (Biber): the mean score of a feature in Biber's corpus

SD (Biber): the standard deviation of a feature in Biber's corpus

With this formula, the feature score of each posting in Neteach-L and TESL-L is calculated as is depicted in the Table 3.3.

3.1.4 Factor Score of Each Posting

As the six features are believed to belong to Factor 5 "Abstract vs Non-Abstract Information" and they are all considered positive (see Appendix 3) in their loading in this factor, the standard feature scores of all the six are now added up to become the factor score of Factor 5. A table of these factor scores of TESL-L postings would be something like in Table 3.4.

Table 3.3 A Sample Data Sheet of Standard Features Scores in Individual Postings of TESL-L

TSL96N	TTLWDS	F17FTURE	F18FTURE	F26FTURE	F27FTURE	F38FTURE	F45FTURE
ln0001	204	-1.45	-.62	-.25	-.81	-.91	-.75
ln0002	236	.47	-.62	-.25	3.29	2.94	1.90
.							
.							
ln0355	422	-.38	3.03	-.25	1.49	-.91	3.69
ln0356	183	-1.45	-.62	-.25	-.81	-.91	2.67
ln0357	210	1.43	3.05	-.25	5.34	-.91	5.20

Table 3.4 A Sample Data Sheet of Factor-5 Scores of Each Posting in TESL-L

TSL96N	TTLWDS	FACTOR5
ln0001	204	-4.79
ln0002	236	7.73
.		
.		
ln0355	422	6.67
ln0356	183	-1.37
ln0357	210	13.86

3.1.5 Sum of Factor Scores of All the Postings in a Discussion List

Finally, the factor scores of Factor 5 of all the postings in Neteach-L and TESL-L are added up and examined against the counterparts in different genres in Biber's corpus.

3.2 Analysis of Abstract vs Non-Abstract Information

The figures gained from the tables in Section 3.1 above mostly relate to individual postings. To get a whole picture of the styles of the discussion lists, they are further computed to show the mean, the standard deviation, and other basic statistics of the features in the whole of each discussion list.

3.2.1 Frequencies of Features Found in the Two Lists

Tables 3.5 and 3.6 show the statistical data relating to the frequency with which the six features occur in the two discussion lists.

Table 3.5 Frequencies of F17, F18, F26, F27, F38, & F45 for Neteach-L Postings

Variable	Mean	Std Dev	Range	Min.	Max.	N = 265 Sum
TTLWDS	205.46	180.47	1245.00	4	1249	54446.00
F17FREQ	1.17	1.83	9.00	0	9	311.00
F18FREQ	.13	.42	2.00	0	2	34.00
F26FREQ	.09	.36	2.00	0	2	23.00
F27FREQ	.24	.66	5.00	0	5	64.00
F38FREQ	.26	.61	3.00	0	3	68.00
F45FREQ	.38	.80	5.00	0	5	100.00

Table 3.6 Frequencies of F17, F18, F26, F27, F38, & F45 for TESL-L Postings

Variable	Mean	Std Dev	Range	Min.	Max.	N = 357 Sum
TTLWDS	194.79	112.80	624.00	25	649	69540.00
F17FREQ	1.42	1.67	11.00	0	11	506.00
F18FREQ	.18	.55	5.00	0	5	65.00
F26FREQ	.01	.12	2.00	0	2	3.00
F27FREQ	.20	.57	4.00	0	4	71.00
F38FREQ	.25	.56	3.00	0	3	88.00
F45FREQ	.59	.88	5.00	0	5	212.00

3.2.2 Raw Feature Scores

Tables 3.7 and 3.8 show the statistic data of the raw feature scores generated from Table 3.2 and those of the Neteach-L as if the length of each posting is set as 1,000 words.

Table 3.7 Raw Feature Scores of F17, F18, F26, F27, F38, & F45 for Neteach-L Postings

Variable	Mean	Std Dev	Range	Min.	Max.	N = 265 Sum
TTLWDS	205.46	180.47	1245.00	4	1249	54446.00
F17RAW	5.36	8.49	52.63	.00	52.63	1419.32
F18RAW	.66	3.01	25.00	.00	25.00	176.12
F26RAW	.30	1.61	15.75	.00	15.75	80.00
F27RAW	.92	2.58	16.13	.00	16.13	243.45
F38RAW	1.17	2.96	16.67	.00	16.67	308.86
F45RAW	1.67	3.43	20.10	.00	20.10	441.39

Table 3.8 Raw Feature Scores of F17, F18, F26, F27, F38, & F45 for TESL-L Postings

Variable	Mean	Std Dev	Range	Min.	Max.	N = 357 Sum
TTLWDS	194.79	112.80	624.00	25	649	69540.00
F17RAW	7.21	8.16	45.05	.00	45.05	2575.41
F18RAW	.94	2.97	25.00	.00	25.00	336.41
F26RAW	.05	.81	14.81	.00	14.81	19.00
F27RAW	.87	2.74	27.78	.00	27.78	311.21
F38RAW	1.18	2.83	15.63	.00	15.63	422.56
F45RAW	2.91	4.41	25.86	.00	25.86	1039.21

3.2.3 Standard Feature Scores

Standard feature scores in Table 3.3 and those of the Neteach-L postings are then converted into Tables 3.9 and 3.10 to show the statistic data of the standard feature scores of the two discussion lists.

Table 3.9 Standard Feature Scores of F17, F18, F26, F27, F38, & F45 for Neteach-L Postings

Variable	Mean	Std Dev	Range	Min.	Max.	N = 265 Sum
TTLWDS	205.46	180.47	1245.00	4	1249	54446.00
F17FTURE	-.64	1.29	7.97	-1.45	6.52	-170.41
F18FTURE	-.10	2.32	19.23	-.62	18.62	-27.60
F26FTURE	.50	4.03	39.37	-.25	39.12	133.75
F27FTURE	-.51	.83	5.20	-.81	4.40	-135.18
F38FTURE	.15	2.69	15.15	-.91	14.24	39.87
F45FTURE	.29	2.15	12.56	-.75	11.81	77.12

Table 3.10 Standard Feature Scores of F17, F18, F26, F27, F38, & F45 for TESL-L Postings

Variable	Mean	Std Dev	Range	Min.	Max.	N = 357 Sum
TTLWDS	194.79	112.80	624.00	25	649	69540.00
F17FTURE	-.36	1.24	6.83	-1.45	5.37	-129.06
F18FTURE	.11	2.28	19.23	-.62	18.62	39.09
F26FTURE	-.12	2.04	37.04	-.25	36.79	-41.75
F27FTURE	1.52	2.63	14.53	-.81	13.72	542.87
F38FTURE	.17	2.57	14.20	-.91	13.30	59.60
F45FTURE	1.07	2.76	16.16	-.75	15.41	381.76

3.2.4 Factor Scores

All the standard feature scores of Factor 5 in Tables 3.9 and 3.10 are added up to be the factor score of Factor 5 for a discussion list, as shown in Tables 3.11 and 3.12.

Table 3.11 Factor Scores of F17, F18, F26, F27, F38, & F45 for Neteach-L Postings

Variable	Mean	Std Dev	Range	Min.	Max.	N = 265 Sum
TTLWDS	205.46	180.47	1245.00	4	1249	54446.00
FACTOR5	-.31	5.94	39.37	-4.79	34.58	-82.44

Table 3.12 Factor Scores of F17, F18, F26, F27, F38, & F45 for TESL-L Postings

Variable	Mean	Std Dev	Range	Min.	Max.	N = 357 Sum
TTLWDS	194.79	112.80	624.00	25	649	69540.00
FACTOR5	2.39	6.38	43.77	-4.79	38.99	852.50

4. Findings and Discussions

4.1 Middle-point Standing of the Two Lists on This Dimension

The findings shown in Tables 3.11 and 3.12 are to be compared with the factor scores of different non-CMC genres in Biber's corpus (1988, pp. 181-184). Table 4.1 shows the means of the factor scores of all the genres (including the two CMC lists in this study) along the dimension of Factor 5.

It can be seen from Table 4.1 that, for Factor 5: the abstract vs. non-abstract information, both TESL-L and Neteach-L stand near the middle position between the two poles. The higher the value of the factor score, the more abstract and technical the style is.

The academic prose of all types is certainly the more formal and abstract in style and the broadcast and telephone conversation are more informal in style. The texts in the two CMC discussion lists are somewhat in between and are similar in the standings to most press reportage genres. This certainly is not a surprise to us.

An interesting phenomenon to notice is that the Standard deviation (SD) for the texts from these two CMC lists are remarkably higher than those in other genres. It may be that the large number of postings in these CMC lists (265 and 357 respectively) represents a greater variety of writing styles while Biber's corpora are composed of only six or at most 80 different texts for each genre (Biber, 1988, p. 67).

It may also be speculated that writers posting on the CMC discussion lists are more free to exhibit their individual styles and this causes such variety.

Another observation that can be made here is the slight difference between the standings

Table 4.1 A Comparison of the Standings on Factor 5 of Each of the Genres Used in Biber (1988, pp. 181-184) and the Two Discussion Lists in This Study

Genres	Mean	Std Dev	Range	Min.	Max.
Technology/engineering academic prose	9.70	4.00	12.80	2.70	15.50
natural science academic prose	8.80	4.50	13.80	3.00	16.80
mathematics academic prose	7.60	2.60	6.30	5.00	11.30
medical academic prose	7.30	3.90	9.20	2.30	11.50
politics/education academic prose	3.70	3.10	13.00	-2.40	10.60
social science academic prose	3.40	4.70	14.10	-1.40	12.60
humanities academic prose	2.80	4.10	16.90	-1.60	15.20

financial press reportage	2.70	3.10	7.00	-1.50	5.50
TESL-L	2.39	6.38	43.77	-4.79	38.99
spot news reportage	1.60	2.40	6.50	-1.40	5.00
political press reportage	0.60	1.70	4.50	-1.60	2.80
personal editorials	0.60	2.20	6.80	-2.20	4.50
letters to the editor	0.40	2.10	5.60	-2.40	3.20
sports press reportage	0.10	2.20	6.90	-3.30	3.60
institutional editorials	0.10	1.80	6.10	-2.20	3.90
Neteach-L	-0.31	5.94	39.37	-4.79	34.58
cultural press reportage	-0.60	2.90	9.10	-4.40	4.80
society press reportage	-0.90	1.10	2.00	-1.60	0.30
sports broadcasts	-1.50	3.60	10.00	-4.70	5.40
non-sports broadcasts	-2.00	1.20	3.50	-3.40	0.10
telephone conversations/ business associates	-3.10	1.10	2.80	-4.20	-1.40
telephone conversations/ personal friends	-3.80	1.40	4.90	-4.80	0.10
telephone conversations/ disparates	-4.20	0.50	1.20	-4.70	-3.50

of the Factor-5 scores of TESL-L and Neteach-L. The score of TESL-L (2.39) is a little higher than that of Neteach-L (-0.31). It is hard to decide if the difference is really significant, and if it is, what might be the cause of the difference.

While both TESL-L and Neteach-L are of the CMC asynchronous type, they are two independent discussion lists and each has a certain group of members. The difference in their Factor-5 scores may be due to the composition of their member groups. It may also be due to the fact that, in TESL-L, there is a moderator and length limit for each posting (55 lines including all computer-generated mail headers) but there is no moderator controlling the content nor any length limit in Neteach-L.

It is hoped that the findings of this research will provide more concrete empirical evidence on the linguistic features of the new medium: Computer-mediated Communication. It is hoped to demonstrate that the new medium is characterized by some new linguistic features. It is also expected how these new features will differ from those of the more traditional printed mode of communication.

4.2 Limitations of the Study and Directions for Further Work

As mentioned at the beginning of this paper, this study is intended to be the pilot study of a text analysis project on the difference between CMC and non-CMC texts in terms of the speech/writing variations.

While the main project intends to use both CMC (asynchronous and synchronous) and non-CMC (asynchronous) text files as objects of analysis and to analyze the difference along the seven dimensions as proposed by Biber (1988), the present study only collects data from CMC asynchronous texts. The scores of these texts are compared with those found in genres in a large non-CMC corpus (960,000 words). Also, analysis is made only

along the fifth dimension of difference, i.e. abstract vs. non-abstract information.

The major finding in this study, viz., that the two CMC asynchronous text corpora stand somewhat in the middle of the dimension between the more abstract and technical genres of academic proses and the more concrete informal genres of broadcasts and conversations, is not surprising.

Despite the limitations of its scope, this study has established a feasible framework of research for the future main project and it is expected that more interesting findings can be obtained in the main project.

References

Adkins, M., & Brashers, D. E. (1995). The power of language in computer-mediated groups. *Management Communication Quarterly, 8*, 289-322.

Bakhtin, M. M. (1986). *Speech genres and other late essays.* Translated by McGee, V. W., edited by Emerson, C., & Holquist, M. Austin: University of Texas Press.

Baron, N. (1984). Computer-mediated communication as a force in language change. *Visible Language, 18*, 118-141.

Bernstein, B. (1970). *Class, codes, and control. Vol. 1: Theoretical studies towards a sociology of language.* London: Routledge & Kegan Paul.

Bhatia, V. K. (1993). *Analyzing genre: language use in professional settings.* London: Longman.

Biber, D. (1986). *Spoken and written textual dimensions in English.* Language 62, 384-414.

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

Biber, D. (1989). A typology of English texts. *Linguistics 27.*

Biber, D. (1995). *Dimensions of register variation, across linguistic comparison.* New York, NY: Cambridge University Press.

Biber, D., & Hared, M. (1992). Dimensions of register variation in Somali. *Language Variation and Change, 4*(1), 41-75.

Biber, D. et al. (1994). Corpus-based approaches to issues in applied linguistics. *Applied Linguistics, 15*(2), 169-189.

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

Blankenship, J. (1974). The influence of mode, submode, and speaker predilection on style. *Speech Monographs, 41*, 85-118.

Blass, T., & Siegman, A. W. (1975). A psycholinguistic comparison of speech, dictation, and writing. *Language and Speech, 18*, 20-33.

Botta, R. (1993). Does shorter mean easier to understand? A study of comprehension of USA Today information stories. Paper presented at the annual meeting of the Association for Education in Journalism and Mass Communication, 76th, Kansas City, MO, August 11-14, 1993.

Brown, P., & Fraser, C. (1979). Speech as a marker of situation. In Scherer, K. R., & Giles, H. (Eds.), *Social markers in speech.* Cambridge: Cambridge University Press.

Chafe, W. L., & Danielewicz, J. (1987). *Properties of spoken and written language.* Berkeley, Calif. : Center for the Study of Writing.

Close, E. O. (1994). Recollect orality. *Visible Language, 28*(2), 100-109.

DeVito, J. A. (1966). Psychogrammatical factors in oral and written discourse by skilled communicators. *Speech Monographs, 33*, 73-76

Derewianka, B. (1996). *Exploring the writing of genres.* UKRA.

Erwin-Tripp, S. (1972). On sociolinguistic rules: Alternation and co-occurrence. In Gumperz, J. J., & Hymes, D. (Eds.), *Directions in sociolinguistics.* New York: Holt, Rinehart and Winston.

Ferguson, C. A. (1959). Diglossia. *Word, 15*, 325-340.

Ferrara, K., Brunner, H., & Whittemore, G. (1991). Interactive written discourse as an emergent register. *Written Communication, 8*, 8-34.

Finnegan, R. (1988). *Literacy and orality: studies in the technology of communication.*

Oxford: Basil Blackwell.
Halliday, M.A.K. (1989). *Spoken and written language.* Oxford: Oxford University Press.
Hardy, V., Hodgson, V., & McConnell, D. (1994). Computer conferencing: A new medium for investigating issues in gender and learning. *Higher Education, 28*, 403-418.
Hasan, R. (1968). *Grammatical cohesion in spoken and written English,* Part 1. London: University College, London.
Hymes, D. H. (1974). *Foundations in sociolinguistics.* Philadelphia: University of Pennsylvania Press.
Johansson, S. (Ed.) (1982). *Computer corpora in English language research.* Bergen: Norwegian Computing Centre for the Humanities.
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: University of Oslo, Department of English.
Kress, G. (1994). *Learning to write.* London: Routledge & Kegan Paul Ltd.
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 (SCOPIL no. 5).* Los Angeles: University of Southern California.
Laina, H. (1992). American teen books easier than British ones. *Journal of Reading, 35*(4), 324-327.
Littlefair, A. B. (1991). *Reading all types of writing: the importance of genre and register for reading development.* Milton Keynes: Open University Press.
Ljung, M. (1991). Swedish TEFL meets reality. Eric Document ED343133.
MacWhinney, B. (1995). *The CHILDES project: Computational tools for analyzing talk.* Hillsdale, N. J.: Lawrence Erlbaum.
MacWhinney, B. (1996a). *CD-ROM: The CHILDES database.* Pittsburgh, PA.: Carnegie Mellon University
MacWhinney, B. (1996b). *Readme manual for CD-ROM: The CHILDES database.* Pittsburgh, PA.: Carnegie Mellon University.
Maynor, N. (1994). The language of electronic mail: written speech? In Little, G.D., & Montgomery, M. (Eds.), *Centennial usage studies.* Tuscaloosa, AL: Alabama UP.
Murray, D. (1988). The context of oral and written language: A framework for mode and medium switching. *Language in Society, 17*, 351-373.
Neteach-L. (1996). Daily logfiles available on line <mailto listserv@thecity.sfsu.edu>.
O'Donnell, R. C. (1974). Syntactic differences between speech and writing. *American Speech, 49*, 102-110.
O'Donnell, R. C., Griffin, W. J., & Norris. R. C. (1967). A Transformational analysis of oral and written grammatical structures in the language of children in grades three, five, and seven. *Journal of Educational Research, 61*, 36-39.
Ong, W. J. (1982). *Orality and literacy: The technologizing of the word.* London: Methuen.
Sinclair, J. M., Hoey, M., & Fox, G. (eds). (1993). *Techniques of description: Spoken and written discourse.* London: Routledge.
Smeltzer, D. K. (1992). Computer-mediated communication: An analysis of the relationship of message structure and message intent. *Educational Technology*, June 1992.

SPSS Inc. (1993). SPSS for windows. Chicago, IL: SPSS Inc.

Stubbs, M. (1986). Lexical density: A technique and some findings. In Coulthard, M. (ed). *Talking about text*. Birmingham: Birmingham Instant Print Ltd.

Svartvik, J., & Quirk, R., (Eds.) (1980). *A corpus of English conversation*. Lund: CWK Gleerup.

TAGGER. (1996). Experimental email tagging on line <mailto tagger@clg.bham.ac.uk>

Tannen, D. (1982). Spoken and written language: Exploring orality and literacy. Norwood, NJ: Ablex.

TESL-L. (1996). Weekly logfiles available on line <mailto listserv@cunyvm.cuny.edu>.

University of Birmingham. (1991). COBUILD's TAGGER. Birmingham, UK: University of Birmingham.

Ure, J. (1971). Lexical density and register differentiation. In Perren, G. E. & Trim, J. L. M. (Eds). *Applications of linguistics*. Cambridge: Cambridge University Press.

van Hout-Wolters, B., & Schnotz, W. (Eds.). (1992). *Text comprehension and learning from text*. Amsterdam: Swets & Zeitlinger.

Vande-Kopple, W. J. (1995). Some explorations of the synoptic and dynamic styles. Paper presented at the 46th annual meeting of the Conference on College Composition and Communication, Washington, DC, March 23-25, 1995.

Warschauer, M. (1996). Comparing face-to-face and electronic discussion in the second language classroom. *CALICO Journal 13*(2/3), 7-26.

Appendix 1

The 67 linguistic features that may be related to the variation of speech/writing styles (Biber, 1988, pp. 77-78).

past tense
perfect aspect verbs
present tense
place adverbials
time adverbials
first person pronouns
second person pronouns
third person pronoun
pronoun IT
demonstrative pronouns
indefinite pronouns
DO as pro-verb
WH questions
nominalizations
gerunds
nouns
agentless passives
BY passives
BE as main verb
existential THERE
THAT verb complements
THAT adj. complements
WH clauses
infinitives
present participial clauses
past participial clauses
past prt. WHIZ deletions
present prt. WHIZ deletions
THAT relatives: subj. position
THAT relatives: obj. position
WH relatives: subj. position
WH relatives: obj. position
WH relatives: pied pipes
sentence relatives
adv. subordinator - cause
adv. sub. - concession
adv. sub. -condition
adv. sub. - other
prepositions
attributive adjectives
predicative adjectives
adverbs
type/token ratio
word length
conjuncts
downtoners
hedges
amplifiers
emphatics
discourse particles
demonstratives
possibility modals
necessity modals
predictive modals
public verbs
private verbs
suasive verbs
SEEM/APPEAR
contractions
THAT deletion
stranded prepositions
split infinitives
split auxiliaries
phrasal coordination
non-phrasal coordination
synthetic negation
analytic negation

Appendix 2

The list of tags used by TAGGER hosted in University of Birmingham:

??? -- no tag assigned
CC -- coordinating conjunction
CD -- cardinal number
DT -- determiner
EX -- existential `there'
FW -- foreign word
IN -- preposition
or subordinating conjunction
JJ -- adjective
JJR -- adjective, comparative
JJS -- adjective, superlative
LS -- list item marker
MD -- modal
NN -- noun, singular or mass
NNS -- noun, plural
NP -- proper noun, singular
NPS -- proper noun, plural
PDT -- predeterminer
POS -- possessive ending
PP -- personal pronoun
PP$ -- possessive pronoun
RB -- adverb
RBR -- adverb, comparative
RBS -- adverb, superlative
RP -- particle
SYM -- symbol
TO -- infinitive marker `to'
UH -- interjection
VB -- verb, base form
VBD -- verb, past tense
VBG -- verb, gerund or present participle
VBN -- verb, past participle
VBP -- verb, non-3rd person singular present
VBZ -- verb, 3rd person singular present
WDT -- wh-determiner
WP -- wh-pronoun
WP$ -- possessive wh-pronoun
WRB -- wh-adverb
" -- simple double quote
$ -- dollar sign
# -- pound sign
` -- left single quote
' -- right single quote
`` -- left double quote
'' -- right double quote
(-- left parenthesis
(round, square, curly or angle)
) -- right parenthesis
(round, square, curly or angle)
, -- comma
. -- sentence-final punctuation
: -- mid-sentence punctuation

Appendix 3

The statistical Data of Six Features in Biber's Corpus
(adopted from Biber, 1988, pp.77-78)

Linguistic Features	Mean	Std Dev	Range	Min.	Max.
Feature 17 agentless passives	9.60	6.60	38.00	0.00	38.00
Feature 18 BY passives	0.80	1.30	8.00	0.00	8.00
Feature 26 past participial clauses	0.10	0.40	3.00	0.00	3.00
Feature 27 past prt. WHIZ deletion	2.50	3.10	21.00	0.00	21.00
Feature 38 adv. subordinator - other	1.00	1.10	6.00	0.00	6.00
Feature 45 conjuncts	1.20	1.60	12.00	0.00	12.00