CHAPTER THREE

Researching Internet Interaction: Sociolinguistics and Corpus Analysis

Simeon J. Yates

Introduction

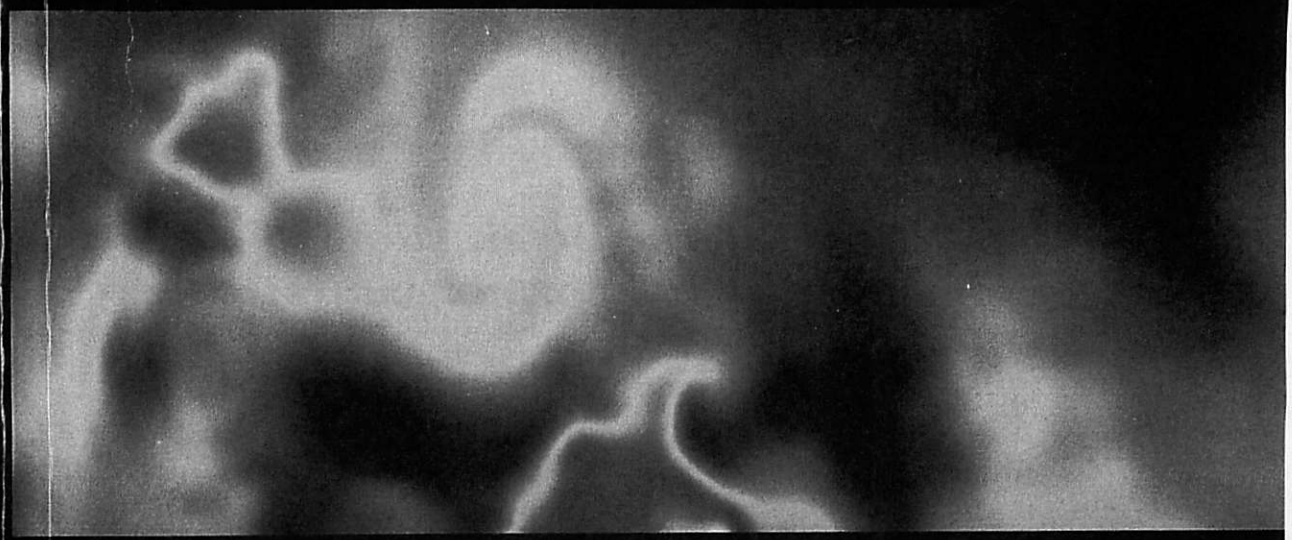
Extract 1: Electronic mail message

```
Date: Mon, 21 Jul 1997 12:38:48 +0100
To: t.sumner@open.ac.uk
From: s.j.yates@open.ac.uk
Subject: Interviews
Cc:
Bcc:
X-Attachments:

Tried to get you all day but your phone is always
engaged! I have the examples of written work. I'll look
at them tonight and then pass them on. Are you free for
a drink tonight? We could do an evening drink in the
sun and swap papers.

S.
```

Extract 2: Internet relay chat

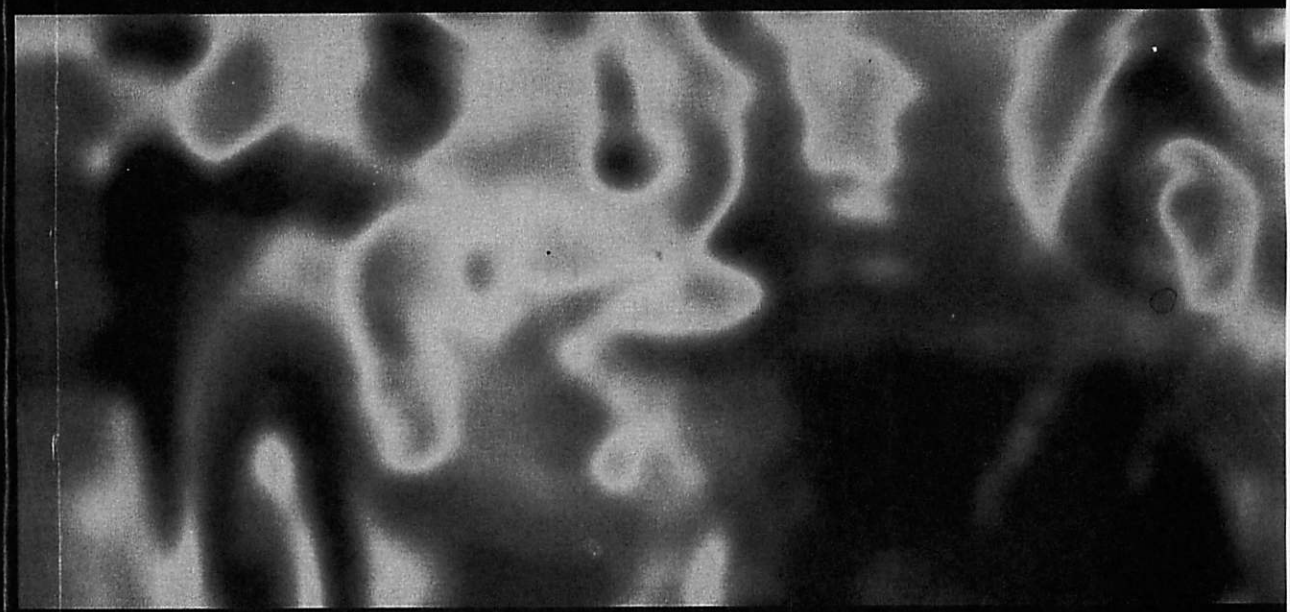
```
<zed> i like this channel
<mouse> Why are all us American's in the Ireland group
<mike> likes GA
<frd3> Really you came here, why!? :)
<GA> I lived in North Carolina!
<zed> we are drawn toghrther
```

The above examples come from one of the fastest growing new communications media – namely the internet. (Other examples of electronic mail messages given in the chapter are anonymous contributions to Open University on-line conferences.) The first being an electronic mail (e-mail) message, the second being a 'transcript' from an internet-relay-chat (IRC)



Discourse as Data

A GUIDE FOR ANALYSIS



MARGARET WETHERELL • STEPHANIE TAYLOR
SIMEON J. YATES

interaction. In fact they are both examples of a wider range of media that go under the title of '**computer-mediated communication**' (CMC). In this chapter CMC will be used in two ways. First, CMC technologies raise a number of questions and problems for researchers interested in discourse. Second, CMC provides data that can readily be employed by those discourse researchers using more 'linguistic' or '**corpus**'-based methods. Both of these points will be discussed in greater depth later. The broad aims of this chapter are therefore:

- to introduce you to the 'linguistic' or 'corpus' methods employed by some discourse analysts
- to introduce you to the types and forms of data that can be explored using such methods
- to introduce you to some of the issues raised by these kinds of data and methods
- to make use of CMC data and research as an example of this work.

The methods presented in this chapter do not follow a specific methodological framework, such as that employed by conversation analysis, or a specific theoretical position, for example that employed in Foucauldian analysis. This chapter brings together a set of methods that can be broadly described as linguistic and reflect some of the approaches to discourse taken by those working within an '**interactional linguistics**' framework (see Wetherell *et al.*, 2001). The methods employed range from corpus linguistics (see McEnery and Wilson, 1996), sociolinguistics (see Stubbs, 1996), as well as linguistic approaches to discourse structure and development (see Coulthard, 1992; 1994). There are two major differences between the analyses presented here and those in some of the other chapters in this volume. First, corpus work tends to be more **quantitative** and general. It is based on large samples of language use that the researchers hope are representative of general **language practices** across a group, culture or even a society. It therefore often involves the counting or **measuring** of linguistic features. Having said this, not all work in this tradition is purely quantitative and more qualitative methods, akin to those used in the rest of this volume, are also employed. Second, the analyses tend to be comparative across groups, cultures and contexts, etc. The hope being that the relationship between language practices and other variables such as context or culture can be identified. In the examples given in this chapter, comparisons are made across different modes of communication (e.g. speech, writing and CMC), as well as across cultural variables (e.g. gender). The assessment of variations across different examples of discourse can involve the use of statistics.

The chapter begins in Section 1 with an exploration of one of the questions that CMC researchers have focused upon – is CMC spoken-like or written-like? The section also introduces the basic ideas and methods to be

explored further in the rest of the chapter. Section 2 discusses the broad range of methods used in corpus work. The section also explores some of the reasons why CMC data has been approached in this way. Section 3 explores the question of identity on-line. Section 4 provides you with the opportunity to conduct your own analysis of a CMC interaction based on data provided in the Appendix to the chapter. The focus of the activity in this section is the exploration of gender and CMC issues. Before jumping straight into these discussions, let's begin by exploring the two examples of CMC that were given at the start of the chapter.

Activity 1

Look again at Extracts 1 and 2 and try to answer the following questions:

1 Do they look like any other forms of communication or records of that communication?

2 Looking at Extract 1, can you work out any of the following:

- Was it written by a man or a woman?
- Where was it written?
- Is it happy, sad, ironic, factual, friendly, or aggressive?

3 Looking at Extract 2, can you work out any of the following:

- Are the speakers men or women?
- Do the turns make sense?
- Is this one conversation or several?

Discussion

I'm not sure there are clear answers to these questions. Taking Question 1 first, you could argue that Extract 1 looks like a letter or a memo. There has been quite a bit of research conducted on the history of the e-mail message and how it connects to the history of the letter and the memo (Yates and Orlikowski, 1992; Yates, 2000). Extract 2 looks like the transcript of spoken interaction. In fact it is the record of a typed exchange.

You might be able to make a guess as to the gender of the author of Extract 1, but it would have to be a guess. I actually know that the person's gender is male but it is hard to tell from the message itself. There are some indicators as to where the message was written. The '@open.ac.uk' element of the writer's e-mail address lets you know that they are part of the 'Open University' (the 'open' part), which is part of the UK's (the 'uk' part) academic (the 'ac' part) internet network. You still can't tell (though you might be able to guess from the content of the message) whether this was written at work, at home, even abroad, as this address only tells you where the 'mailbox' of the user is, not the actual location where they read and write their e-mails. As to the tone of the e-mail message, it seems quite friendly to me but all I have to go on is the text. The answers to Question 3

are even less clear. Extract 2 looks like a transcript but the turns do not seem in order. There is, in fact, an order but this can only be seen if we view this as a set of multiple conversation threads involving several overlapping sets of speakers. As for the gender of the speakers, all we have are the usernames to go on, and anything they might say in the interaction.

The issues highlighted in Activity 1 have formed the basis of a large amount of CMC research. These issues can be expressed in three simple questions:

1 How is the medium of CMC related to other communications media?

2 How do users interact on-line?

3 How do users construct identities on-line?

These three general questions will follow us through this chapter. Section 1 will present an analysis that explores specific aspects of the questions 'How is the medium of CMC related to other communications media?' and 'How do users interact on-line?' Section 3 will present an analysis of the question 'How do users construct identities on-line?' Section 4 takes a closer look at how gender functions on-line.

1 What is CMC?

CMC technologies come in a wide range of forms. Each of these technologies supports a range of interactions. In order to make sense of this range, CMC researchers tend to categorize CMC systems across two dimensions. First, there is the extent to which the interactions are **synchronous**. Synchronous communication, as exemplified by speech, requires both parties to be present for the interaction to take place. An example of synchronous CMC is IRC. Other CMC systems such as e-mail do not require participants to both be present. E-mail messages are sent asynchronously, that is, in delayed time, in the same manner as letters. We expect a delay between sending a letter and receiving a reply. Second, CMC can be categorized by the nature of the interaction. This can be one-to-one, one-to-many or group communication.

By using these criteria we can categorize the various types of CMC. At the synchronous end of the scale are 'chat' systems where the interacting parties must be '**co-present**' at their respective computers at the same time. Given this direct virtual co-presence, the interactions tend to consist of short single-line turns. At the other end of the scale are technologies such as e-mail and computer-conferencing. In this case the medium is designed to be asynchronous with users sending and receiving messages under the assumption that the others are not virtually co-present. In between these extremes are CMC media that support varying degrees of synchronicity/asynchronicity, these include such things as Multi-User-Domains/Multi-User-Domains-Object-Orientated (MUDs/MOOs) and Usenet. Figure 1 is an idealized representation of the types of synchronicity/asynchronicity that CMC systems support. Each technology supports a mixture of one-to-one,

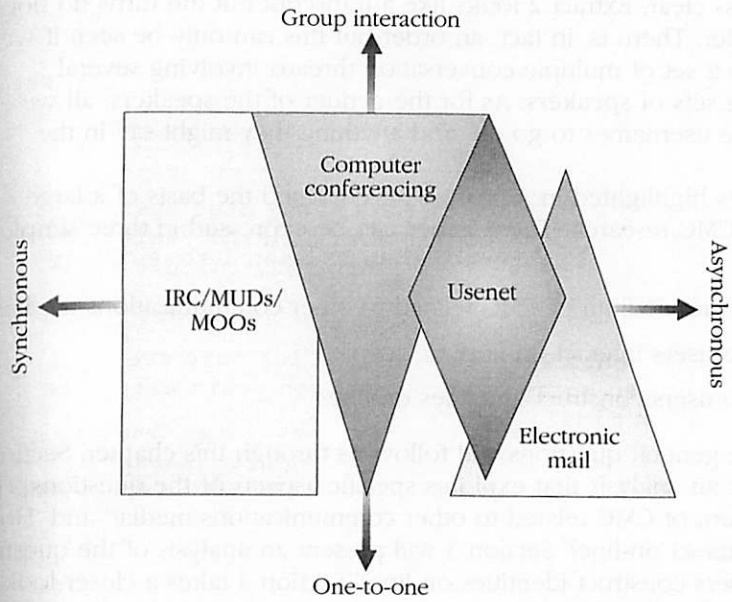


Figure 1 CMC and synchronicity

one-to-many and group communication (see Yates, 2000, for a full discussion).

One of the first questions CMC researchers asked was: 'Is CMC written-like or spoken-like?' This question arises from the fact that CMC consists of typed texts, but, as Figure 1 implies, CMC is used in contexts such as synchronous group communication where we would normally expect a spoken interaction. This is, of course, a specific version of one of the three questions highlighted above – namely 'How is the medium of CMC related to other communications media?'

Activity 2

The following are some more examples of CMC texts. This time the elements of the text that mark them out as being from an e-mail or 'chat' interaction have been removed and they are presented as though they are 'turns' as they occurred 'next' to each other in an interaction. Read these two extracts and try to decide which of the 'turns' are 'more spoken-like' or 'more written-like' – note down why you decided they were more written-like or spoken-like.

Extract 3

```
It is fairly obvious that managing information
technology change is an area which can be expected to
be of importance to more or less every organisation.
Every organisation has an information system and every
information system is susceptible to the use of new
technology.
```

Sorry, not been around for a couple of days. What I meant was, we should examine all the processes involved in our operation, not just the way in which we write or deliver our courses.

Extract 4

A: 13 was very special for me, B. But it went downhill a bit at 18, then brightened up at 19 when I threw my ex out.

B: Your teen years were obviously a lot more eventful than mine. Then again everybodys teen years were more eventful than mine!

A: Life is what you make it, kid ... though I could have done without some of my events!

B: I know that now. It just took me a long time to work that out.

Discussion

Extract 4 seems obviously spoken-like to me. Not only does it look like the transcript of a spoken interaction but also the topic and tone of the content reminds me of a quite personal face-to-face interaction. Having said this we must be careful. In Chapter One, Taylor pointed out that the way in which language data is transcribed is directly linked to the analysis. Transcription is not a neutral activity. In the case of Extract 4, the fact that I presented this as a transcript can be seen to have had two possible effects. First, it may have made it easier to see the spoken-like nature of the interaction. Second, it may have hidden some of the written-like aspects of the interaction. What do you think? Here is how the interaction would have appeared on screen:

Extract 5: Computer-conferencing chat

==========

dt200-chat #209, [A]

This is a comment to message 208

There is/are comment(s) on this message

There are additional comments to message 208

---

13 was very special for me, B. But it went downhill a bit at 18, then brightened up at 19 when I threw my ex out.

A

```
==========================
dt200-chat #210, [B]
This is a comment to message 209
There is/are comment(s) on this message
--------------------------
Your teen years were obviously a lot more eventful than
mine. Then again everybodys teen years were more
eventful than mine!


B


==========================
dt200-chat #211, [A]
This is a comment to message 210
There is/are comment(s) on this message
--------------------------
Life is what you make it, kid ... though I could have
done without some of my events!


A


==========================
dt200-chat #212, [B]
This is a comment to message 211
There is/are comment(s) on this message
--------------------------
I know that now. It just took me a long time to work
that out.
```

Looking at the version of the interaction in Extract 5, I find that the 'spokenness' is less clear. The message headers provide a 'memo'-like feel to the interaction.

Extract 3 in Activity 2 is less clear. First, though it is laid out as a transcript, the turns don't seem to follow from one another. Once again it is an example of the ways in which many CMC interactions involve multiple threads running concurrently and where the specific 'temporal' sequence of messages does not fit exactly the turns of the interaction – as it does in speech. Turn one seems more written-like not only because of the content and tone but also because of the grammatical structure (I come back to this point later when we analyse data in Example Analysis 1). You could argue that turn two is also written-like – though more like a chatty letter than the turns in Extract 4. I might argue that turn two of Extract 3 is, in fact, spoken-like because of its the grammatical structure.

Once again we could ask if the form of transcription is getting in the way of a fuller understanding/analysis? Here is the 'on-screen' form.

Extract 6: Computer-conferencing discussion

```
==========================
dt200-forum/IT Management #40, [1]
This is a comment to message 25
There is/are comment(s) on this message
There are additional comments to message 25
--------------------------
It is fairly obvious that managing information
technology change is an area which can be expected to
be of importance to more or less every organisation.
Every organisation has an information system and every
information system is susceptible to the use of new
technology.

[1]

==========================
dt200-forum/IT Management #41, [1]
This is a comment to message 28
There is/are comment(s) on this message
--------------------------
Sorry, not been around for a couple of days. What I
meant was, we should examine all the processes
involved in our operation, not just the way in which we
write or deliver our courses.
```

You might have noticed that in both cases the turns are marked in the headers of the messages 'This is a comment to message XX'. Users have to follow these threads rather than the temporal sequence of messages in order to understand the turns. But also note that in both Extract 5 and Extract 6, some messages have multiple responses so even the threads don't follow 'normal' face-to-face turn-taking conventions (such as 'There are additional comments to message XX'). The problems of defining CMC as being spoken-like or written-like, and the problems of understanding turn-taking in CMC, have quite important implications for how CMC researchers have gone about their work.

1.1 Example Analysis 1: Exploring the medium of CMC

In the rest of this section I will walk you though an analysis designed to answer the question 'Is CMC spoken-like or written-like'. First, I will discuss some of the reasons for choosing the data and the methods. Second, I will describe the collection and analysis of the data. Third, I will present and discuss the results. The purpose of this walk through is to explore the

implications of Activity 2 and to introduce you to some of the ideas and concepts that will be examined in more depth later in the chapter.

Taylor, in Chapter One, outlined four approaches to discourse work. In many respects the analyses that follow can be placed in the first two of these approaches. Wetherell *et al.* (2001) separate discourse work into the study of three broad areas.

- social interaction
- minds, selves and sense-making
- culture and social relations.

In this typology, the following analyses fall into the area of social interaction and the 'interactional sociolinguistic' method of analysing such interactions. It therefore draws upon ideas and methods from linguistics more than from conversation analysis, psychology or sociology.

Activity 3

As you read through Example Analysis 1, make some notes on the following issues:

- How was the data collected?
- How was the data analysed?
- How did the analysis link data, theory and methods?

I'll discuss my responses to these questions once you have finished reading this Example Analysis.

Background

In general, the analysis that follows is based in a 'cultural' or 'critical linguistic' understanding of the difference between speaking and writing (see Hodge and Kress, 1988, reproduced as Reading Twenty-one in Wetherell *et al.*, 2001). It is based on the assertion that CMC, as a typed 'text-based' medium, must draw upon users understandings of 'doing writing'. It assumes that CMC communication is a new kind of '**literacy practice**'. Barton (1991) claims that the definition of literacy practices comes from various sources, the main two being Scribner and Cole (1981) and Street (1984). Scribner and Cole define practice in the following manner:

> By a practice we mean a recurrent, goal-directed sequence of activities using a particular technology and particular systems of knowledge. ... Practice always refers to socially developed and patterned ways of using technology and knowledge to accomplish tasks. Conversely, tasks that individuals engage in constitute a social practice when they are directed to socially recognized goals and make use of a shared technology and knowledge system.
>
> (Scribner and Cole, 1981: 236)

So we can begin the analysis by assuming that the way people 'write' CMC messages is probably defined by the ways in which 'doing speaking', 'doing writing' and 'doing CMC' have been socially and culturally organized. This in turn will leave its mark in the actual discourse produced through CMC interactions. In order to explore this we will need CMC data as well as spoken and written data for comparison.

Collecting the data

As the first part of this section made clear, there are very many kinds of CMC. What we want to know is the range of spoken-like to written-like forms in CMC. One form of CMC that tends to display a whole range of message types, levels of synchronicity (virtual co-presence) and one-to-one to group interaction is computer conferencing. This analysis therefore collected data from one computer conferencing system. There are a number of technical reasons (to do with how messages are stored) and ethical reasons (to do with access to 'private' messages) why computer conferencing is an easier option than other media. Cherny (1999) provides a full discussion of the ethical concerns raised by the study of CMC.

The data for this analysis were collected from an Open University computer conferencing system (CoSy). The examples of computer conferencing interactions you have seen so far (Extracts 1, 5 and 6) come from this system. Given that corpus work tries to be general in scope, the CoSy interactions provided useful data for a number of reasons. First the user base was spread across the whole of the United Kingdom and there were, on average, 2,000 registered members of the CoSy system at any one time. It therefore provided a large social sample, including both novice and experienced users. Second the system supported a range of **genres** of interaction including both formal educational interactions and more informal 'chat' interactions. The data can therefore be seen as representative of a range of users and types of interaction. As Taylor noted in Chapter One of this volume, selecting the respondents and data for discourse studies is a key part of the process. Here we have gone for a very 'quantitative' approach – collecting a lot of data from a range of sources to produce a 'generally representative' sample.

While one obvious method for collecting the CMC material would be to download all that was available, this would produce a very large set of data! One also needs to consider the corpora against which comparisons will be made, as well as taking seriously the ethical issues about access. The first selection therefore consisted of a sample of 50 messages from 152 publicly open conference discussions. This provided a total of some 648,550 words and an average of 4,267 words per conference discussion. This set of data will be referred to as the CoSy:50 corpus.

Given that the above selection consisted of only partial discussions, and given some of the possible methodological problems this raises, a corpus of full conferencing interactions was also collected. Rather than

randomly sample a set of conferences, it seemed more useful to purposefully sample from specific sources. The sources once again included a range of 'chat' and formal interactions. In total 66 conference discussions were collected totalling some 1,573,499 words. I will refer to this as the CoSy:Full corpus. In the end, most analyses used a combination of the CoSy:50 and the CoSy:Full corpus. This provided a total of 218 texts and 2,222,049 words.

The two sets of data have been described as '**corpora**'. So what is a corpus? A corpus is collection of linguistic data, nowadays stored on computer, which is seen to be representative of a certain type of text, interaction or discourse. It is interesting to note that the development of interest in corpus work has coincided with the availability of personal computers capable of storing and manipulating very large amounts of text-based data. There are now a large number of corpora available for use by language and discourse researchers. In the UK, the British National Corpus contains about 100 million words. 20 million of these are transcribed from 'naturally occurring speech' recorded on 'Walkman' tape recorders by a representative sample of the UK population. The other 80 million words were sampled from printed and written materials.

Another possible reason why CMC researchers have tended to use corpus or content analysis methods derives directly from the fact that CMC is **digital text**. It is, in fact, far easier to build a CMC corpus than any other. As the data is already digital it requires no transcription or copying from another form such as tape recording or the printed page. As one of the earlier corpus researchers noted:

> The computer revolution has brought with it new forms of discourse which also deserve systematic study. One of these is electronic mail ... Electronic mail reveals features of both speech and writing. Like other forms of discourse, new as well as old, it deserves the attention of future corpus workers.
>
> (Johansson, 1991: 307–8)

The main idea behind collecting a corpus is the exploring of general features of a language or a specific kind of text or interaction.

In order to conduct our analysis we will need to compare our CMC corpus with spoken and written corpora. The analysis presented here used the Lancaster-Oslo/Bergen corpus of written British English and the London-Lund corpus of spoken British English. The Lancaster-Oslo/Bergen (LOB) corpus of written texts consists of 500 individual texts of approximately 2,000 words each. These texts are distributed across 15 separate text categories.

Table 1 Number of texts per category in the LOB corpus

Code	*Category*	*Number of texts*
LOB:A	Press: Reportage	44
LOB:B	Press: Editorial	27
LOB:C	Press: Reviews	17
LOB:D	Religion	17
LOB:E	Skills, Trades and Hobbies	38
LOB:F	Popular Lore	44
LOB:G	Belles lettres, Biography, Essays	77
LOB:H	Miscellaneous (government documents, foundation reports, industry reports, college catalogue, industry house organ)	30
LOB:J	Learned and Scientific Writing	80
LOB:K	General Fiction	29
LOB:L	Mystery and Detective Fiction	24
LOB:M	Science Fiction	6
LOB:N	Adventure and Western Fiction	29
LOB:P	Romance and Love Story	29
LOB:R	Humor	9
	Total Texts:	500

An example of the data in this corpus is presented in Corpus Example 1. This is a 'transcription' of a newspaper report. As you can see it contains markers designed to indicate such things as headings and sentence breaks. This is in fact from the 'untagged' version of the corpus. The 'tagged' version gives information on the 'part-of-speech' or grammatical function of each of the words.

Corpus Example 1: Extract from 'untagged' LOB corpus text

```
A01 2 *<*'*7STOP ELECTING LIFE PEERS**'*>
A01 3 *<*4By TREVOR WILLIAMS*>
A01 4 |^A *OMOVE to stop \OMr. Gaitskell from nominating any more Labour
A01 5 life Peers is to be made at a meeting of Labour {OM P}s tomorrow.
A01 6 |^\OMr. Michael Foot has put down a resolution on the subject and
A01 7 he is to be backed by \OMr. Will Griffiths, {OM P} for Manchester
A01 8 Exchange.
```

```
A01 9 |^Though they may gather some Left-wing support, a large majority
A01 10 of Labour {OM P}s are likely to turn down the Foot-Griffiths
A01 11 resolution.
A01 12 *<*7*'ABOLISH LORDS**'*>
A01 13 |^*0\OMr. Foot's line will be that as Labour {OM P}s opposed the
A01 14 Government Bill which brought life peers into existence, they should
A01 15 not now put forward nominees.
A01 16 |^He believes that the House of Lords should be abolished and that
A01 17 Labour should not take any steps which would appear to *''prop up**'' an
A01 18 out-dated institution.
A01 19 |^Since 1958, 13 Labour life Peers and Peeresses have been created.
A01 20 |^Most Labour sentiment would still favour the abolition of the
A01 21 House of Lords, but while it remains Labour has to have an adequate
A01 22 number of members.
```

The London-Lund corpus of spoken English consists of 100 spoken texts. The full corpus consists of some 500,000 words. Each individual text therefore contains some 5000 words (see Greenbaum and Svartvic, 1990) and is therefore comparable to our CoSy:50 corpus.

Table 2 Categories of spoken texts in the London-Lund corpus

Code	*Category*	*Number of texts*
LLc:1	Conversations between equals	14
LLc:2	Conversations between equals	14
LLc:3	Conversations between disparates	7
LLc:4	Conversations or discussions between equals	7
LLc:5	Radio discussions and conversations between equals	13
LLc:6	Interviews and discussions between disparates	9
LLc:7	Telephone conversations between equals	3
LLc:8	Telephone conversations between equals	4
LLc:9:1–2	Telephone conversations between disparates	2
LLc:9:3–5	Telephone, dictaphone and answerphone	3
LLc:10	Spontaneous commentary, mainly radio	11
LLc:11	Spontaneous oration	6
LLc:12	Prepared oration	7

Much of the data in the London-Lund corpus consists of academics in discussion, indicating that the main source for much of the data was the researchers' own social environment.

Corpus Example 2: Extract from London-Lund corpus

```
1 1 1   20 1 1 A  11  ^w=ell# . /
1 1 1   30 1 1 A  11 ((if)) did ^y/ou _set _that# /
1 1 1   40 1 1 B  11 ^well !J\oe and _I# /
1 1 1   50 1 1 B  11 ^set it betw\een _us# /
1 1 1   60 1 1 B  11 ^actually !Joe 'set the :p\aper# /
1 1 1   70 1 1 B  20 and *((3 to 4 sylls))* /
1 1 1   80 1 1 A  11 *^w=ell# . /
1 1 1   90 1 1 A  11 ''^m/\ay* I _ask# /
1 1 1  100 1 1 A  11 ^what goes !\into that paper n/ow# /
1 1 1  110 1 1 A  11 be^cause I !have to adv=ise# . /
1 1 1  120 1 1 A  21 ((a)) ^couple of people who are !d\oing [dhi: @] /
1 1 1  130 1 1 B  11 well ^what you :d\/o# /
1 1 1  140 1 2 B  12 ^is to - - ^this is sort of be:tween the :tw\/o of /
1 1 1  140 1 1 B  12 _us# /
1 1 1  150 1 1 B  11 ^what *you* :d\/o# /
1 1 1  160 2 1 B  23 is to ^make sure that your 'own . !c\andidate /
1 1 1  170 1 1 A  11 *^[\m]#* /
1 1 1  160 1 2 (B  13 is . *.* ^that your . theres ^something that your /
1 1 1  160 1 1 (B  13 :own candidate can :h\/andle# - - /
1 1 1  180 2 1 B  21 ((I ^wont)) /
1 1 2  190 1 1 A  11 *((^y\eah#))* /
1 1 2  180 1 1 (B  11 ((be a m/inute# - - /
1 1 2  200 1 1 B  20 3 to 4 sylls)) - - - /
```

Analysing the data

One difference between speech and writing that many researchers have commented upon is that of the differing modes of production and **consumption**. Speech is produced 'on the fly' and is intended to be consumed – heard – in the same rapid and dynamic manner. Writing on the other hand is static; it is produced at the pace set by the writer alone and can be consumed at any speed that the reader chooses. The effects of such differences in production are likely, it is claimed, to generate differences in the language used. One aspect of this concerns vocabulary use. Chafe and Danielewicz claim that:

> As a consequence of these differences, speakers tend to operate with a narrower range of **lexical choices** than writers. Producing language on the fly, they hardly have time to sift through all of the possible choices they might make, and may typically settle on the first words that occur to them. The result is that the vocabulary of spoken language is more limited in variety.
>
> (Chafe and Danielewicz, 1987: 88 my emphasis)

In order to examine the differing use of vocabulary more empirically Chafe and Danielewicz make use of the **type/token ratio**. The type/token ratio used by Chafe and Danielewicz is that of the number of different words (**types**) divided by the total number of words (**tokens**). In this calculation all words of all types were counted (Chafe and Danielewicz, 1987: 88).

Activity 4

What is the type/token ratio of this sentence?

The answer is 1.0 or 100%. There are 9 words, all different.

What about the following utterance? 'I would like to agree, but I would not be able to take that position, I would not agree.'

The type/token ration here is 0.58 or 58% as there are 11 different words (types) and a total of 19 words (tokens). Now try to calculate the type/token ratio for Extract 1 at the start of this chapter – ignore the 'header' part of the message.

I calculate the type/token ratio to be 0.875. You probably found this quite a boring and possibly tedious process. You will not be surprised to hear, then, that computers normally carry out such analyses. In fact most corpus work involves the use of computers, as even simple analyses such as this would take excessive amounts of time if conducted 'by hand'.

There are some problems with the above method that are too detailed to discuss at this point. A more rigorous measure of vocabulary use is to remove the words that are essentially **grammatical** (e.g. and, if, it, the, etc.) and to focus on lexical words (content words). One then calculates the ratio of the number of different lexical items (**lexical types**) to the total number of lexical items (**lexical tokens**). Computer software did this analysis across the three corpora. The general result is presented in Table 3. If a statistical analysis is conducted on the data in this table the differences prove to be statistically significant. In other words, the differences are not the product of simple chance.

Table 3 Mean type/token ratios for three corpora

Corpus	*CoSy (CMC)*	*LOB (writing)*	*London-Lund (speech)*
Mean type/token ratio	0.590	0.624	0.395

This result would seem to indicate that CMC is more akin to writing than speech in terms of vocabulary use. The most obvious conclusion is to follow Chafe and Danielewicz and see this as a product of the medium itself, and the opportunity it allows for longer gestation over the content of utterances. Careful thought about this point, though, raises a complication. The texts contained within the LOB corpus were mostly produced prior to the advent of the word processor or the electronic text. The CMC corpus contains only word-processed electronic texts. What the electronic text

brings is an even greater set of opportunities to correct, change, restructure and review utterances. Following Chafe and Danielewicz, one would expect to see even higher levels of vocabulary use – therefore higher type/token ratios. Even more interestingly, it is writing that demonstrates the greatest variation in vocabulary use (type/token ratios) as compared to speech and CMC which have comparable variances. The implication of this result is that factors other than simply the mechanical aspects of the medium are at work in producing utterances. Vocabulary use covers a large number of social and cultural issues and, despite having a clear relationship to the production and consumption of a text, it is not the only measure of the textuality of an utterance.

One measure that attempts to incorporate these social and cultural issues is '**lexical density**'. Halliday (1985) begins his explanation of lexical density with the following examples:

- If you invest in a rail facility, this implies that you are going to be committed for a long time. (*Lexical density = 0.35*)
- Investment in a rail facility implies a long-term commitment. (*Lexical density = 0.7*)

Halliday argues that the first of these sentences appears much more like the record of spoken communication than written. The first of these sentences has 7 lexical items and 13 grammatical ones. The lexical items are: *invest; rail; facility; implies; committed; long* and *time*. In the second sentence there are 7 lexical items and 3 grammatical ones. The lexical items are: *investment; rail; facility; implies; long; term* and *commitment*. Halliday attributes this difference to the different cultural and contextual uses to which we put speech and writing and the social roles our linguistic practices have to support. It is not the medium but society and culture that motivates differences in speech and writing.

In the above examples the ratio of lexical items to grammatical ones (the lexical density) is lower for spoken language than it is for written. Lexical density therefore provides a quantitative difference between spoken and written utterances. In the above example the first sentence has a ratio of 7 to 13 whilst the second sentence has a ratio of 7 to 3. These can be better represented as a ratio or percentage of the number of lexical items to the number of total items within an utterance. On this measure the first sentence would score 7 out of 20 or 0.35 (35%). The second sentence would score 7 out of 10 or 0.7 (70%). These types of score are *lexical density* scores (Halliday, 1985). Such a test was conducted using our three corpora and the results are given in Table 4. Once again the differences in the results were statistically significant.

Table 4 Mean unweighted lexical densities for three corpora

Corpus	*CoSy (CMC)*	*LOB (writing)*	*London-Lund (speech)*
Mean lexical density (unweighted) (%)	49.258	50.316	42.292

The interpretation of these results is not straightforward. On one level it implies that some texts in the corpus are more written-like and some more spoken-like – thus making an average result in between. On the other hand a more detailed study indicates that there are styles – genres – of CMC that are themselves between the oral and the literate. In fact these genres are often used by experienced users who have developed their own 'CMC' practices (in contrast to oral and literate practices) and their use is very dependent on social context (see Yates, 1996).

1.2 Comment on Example Analysis 1

Example Analysis 1 is very quantitative and very different from nearly all the analyses you will find elsewhere in this volume. It does, though, highlight some of the key issues faced by corpus-based research. Let's review the three questions that were set out in Activity 3.

How was the data collected?

The only data that was collected by the researcher was the CMC data. This was simply 'downloaded' from a CMC system (CoSy) and used as given. Unlike both the written and spoken corpora, this data did not need to be transcribed. On one level this might seem to be a benefit of CMC data. The practical problem it raises is that of having too much data. In fact one could argue that the ease with which large amounts of CMC data can be collected leads researchers towards corpus methods. Being pushed into a methodology by the nature of your data is something that all discourse analysis workers need to take note of. It is also something that should not be done 'blindly'.

The spoken and written data were gained from existing corpora. In Chapter One, Taylor notes that large amounts of qualitative data are now being archived for use by future researchers. She notes that for qualitative researchers, re-using others' data has a number of problems that derive from the separation between the data collection and the analysis processes. In the case of corpora, most linguists do not see this problem, all they are concerned about is the representativeness of the corpus. As you will see in the analyses that follow in this chapter, this may, in fact, be a difficult position to maintain.

How was the data analysed?

It could be argued that the data was 'analysed' by computer. The three corpora were processed by software to provide the relevant type/token ratios and lexical density scores. In fact the setting up of these measures was theory driven and they represent the 'operationalization' of these theories into a specific form of quantitative analysis. These results were then subjected to statistical analyses. This represents a very quantitative approach informed by essentially positivist models of the research process. In the analyses to follow you will see how CMC researchers have moved away from such methods as they begin to address more social and cultural questions, though they continue to make use of the opportunities provided by computer software tools.

How did the analysis link data, theory and methods?

We have already begun to answer this question above. The link between theory and methods is based on the following logic. Oral and literate practices leave their mark in the structure and content of utterances. This must also be true of CMC. We can theorize what these marks might be – for instance vocabulary variations. We can use these ideas to construct empirical quantitative measures – for instance type/token ratios or lexical density. By applying these measures we can make statements about the nature of CMC as a medium. This model of the process is based in a 'positivist' framework in that it follows the model of the natural sciences. In this model, research follows the stages of: theory; hypothesis; operationalization; measurement; results; analysis and theory.

Summary

What have we learnt from Section 1? It introduced a number of important methodological issues and the analysis made use of corpora. Corpora are large sets of linguistic data, often in the order of several million words. For linguists interested in discourse analysis, corpora provide readily available data. Section 1 also introduced the idea that a key aspect of discourse – the very manner in which the interaction is grammatically and linguistically constructed – can be studied in a very quantitative manner. This approach drew upon a range of linguistic ideas and methods and made the assumption that the record of linguistic interactions can be taken as evidence of discursive processes.

2 Methods

Section 1 briefly mentioned the use of computer software to conduct data analyses. Computer software allows the researcher to access or analyse the millions of words in a corpus which would simply be extraordinarily difficult and time-consuming otherwise (though not impossible, see Zipf, 1935). Before looking at another example analysis let's reconsider these three issues:

- the development of a corpus of data
- the range of corpus methods
- using computers.

2.1 The selecting corpora and corpus methods

As McEnery and Wilson (1996) note, corpus methods have been used in everything from formal linguistic studies of lexis and grammar through to cultural studies, social psychology and discourse analysis. At the same time, those linguists interested in the sociological, cultural and social-psychological aspects of language use often employ much more qualitative methods than those presented in Section 1. To say that this range of methods follow a well-defined set of rules, or that methods are coherent across disciplines would be untrue. Corpus methods, especially in the area of discourse analysis, are still 'under development' and there are few full and detailed discussions of the practical and methodological implications of such an approach to the study

of language and communication. One of the few discussions of the use of corpus-type methods in the study of discourse and culture is provided by Stubbs (1996). In this work, Stubbs makes use of varying methods and corpora. Table 5 lists the data and the methods Stubbs employs.

Table 5 Corpus methods employed by Stubbs

Topic/Data	*Methods*	*Analysis*
Institutional language use: spoken and written texts.	Lexical density	Differences in language use in context
Sexist language: speeches by Baden-Powell to Boy Scouts and Girl Guides	Content analysis – word frequency and use	Exploration of lexical choices in a text
Language in courtrooms: transcript of a judge's summing up	Content analysis – word frequency and use, discourse markers, modal verbs and syntactic complexity, concordance	Exploration of interpretation of 'complex' language and the production of 'facts'
Language in teaching texts: geography text books	Content analysis – word frequency and use; Key-words-in-context	Exploration of the role of agency in pedagogic explanations
Cultural connotations of words: 130 million words of written English from the *Cobuild* corpus	Collocations of words	Exploring the connotations and contexts of word use
Positioning and framing of utterances: various selections from corpora	Content analysis – word frequency and use; collocations	Exploration of the use of modal auxiliaries to position speakers in relation to utterances

Elements of all of these methods will be employed in the examples in the following sections. Word frequency lists allow you to see the terms and topics that are most frequently used in a text. Some software also provides graphical displays of the position and frequency of target words throughout texts. **Key-words-in-context (KWIC)** analyses provide lists of all the occurrences of a word and the context (e.g. sentence etc.) around it. Concordancers do a similar job. **Collocations** are frequency lists of the words that occur in the context around a key word. For example if one were to study the use of racist language in political debates, one could explore the words that occur near to a key word such as 'immigration'. Other software provides 'part-of-speech' analyses. These allow you to explore the forms of grammar and syntax in use. The use of word frequencies, key-words-in-context, and qualitative content analyses will be used in this chapter.

2.2 Using computers

All of these methods can and often have to employ computers. In order to search corpora for specific key words, or to count and display occurrences of words and phrases in a corpus of several million words, requires the use

of computers and specialist software. In the following sections two types of computer software were employed. The first can best be described as 'linguistic tools'. The second type are **computer assisted qualitative data analysis software** (CAQDAS) designed for use by qualitative researchers. Both of these types of software are usable by anyone working with 'linguistic' data. In fact they represent a spectrum of tools from quantitative word counters to tools specifically designed to help those working within some form of qualitative data analysis framework.

Corpus linguistic tools include concordances, key-word-in-context (KWIC) systems, and word counting and sorting tools. Various forms of these are available on nearly all major computing platforms, though classic examples include the Oxford Concordance Program for UNIX, or WordSmith™ for the PC which brings a suite of such tools together in one piece of software. Though such tools do not allow coding of the text they can easily make considerable use of 'part-of-speech' as well as any other annotations and transcription markers included within the text. Some of these systems are designed to be interactive, allowing the researcher to dynamically explore large bodies of text at a detailed linguistic level. Figures 2, 3 and 4 show a corpus linguistic tool in action on some CMC data.

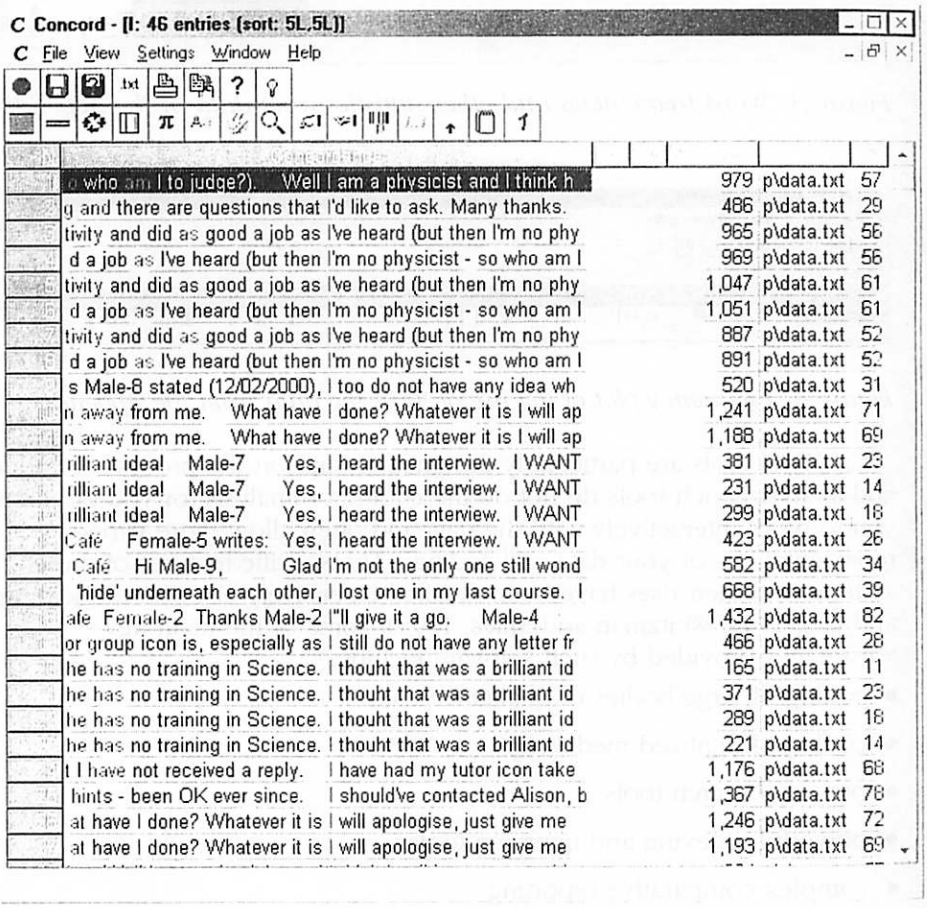


Figure 2 Concordance (KWIC) of the use of 'I' in the data from the Appendix

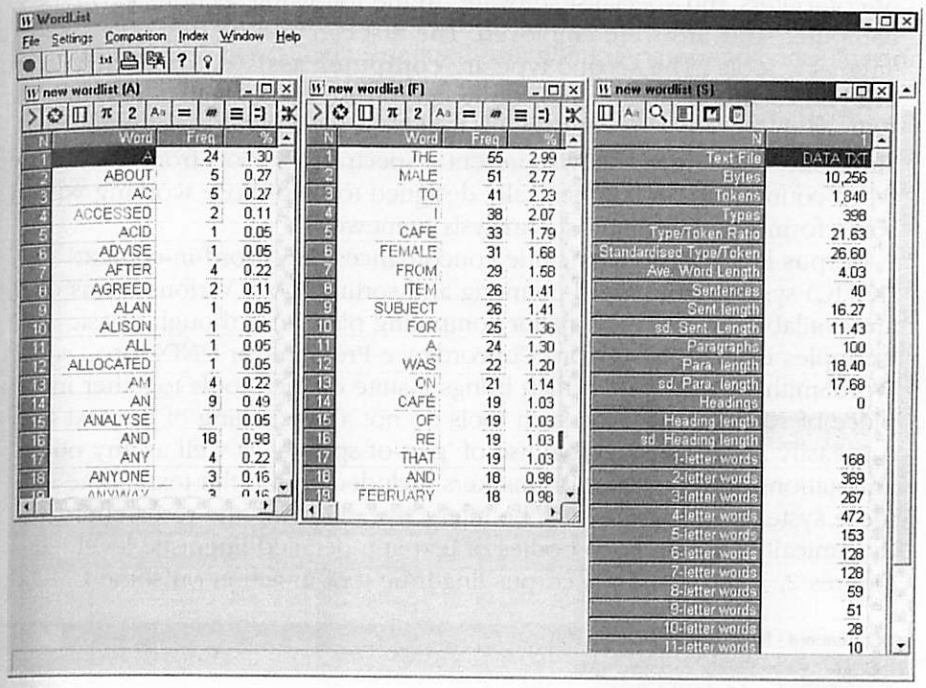


Figure 3 Word frequencies and other statistics from data in the Appendix

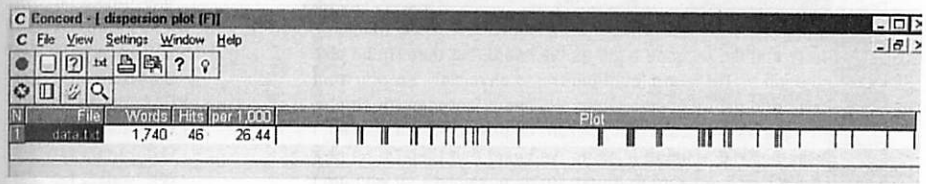


Figure 4 Frequency plot of the use of 'I' in the data from the Appendix.

CAQDAS tools are particularly designed to support the process of coding and analysis. Such tools do not do the coding or analysis for you but allow you to work interactively with the data and often allow more varied representations of your data and codings. The specific features of these systems and their uses have been detailed extensively elsewhere (Fielding and Lee, 1994; Weitzman and Miles, 1995). Some of the main key advantages provided by such software include:

- coding of large bodies of qualitative data
- coding of digitized media clips
- powerful search tools
- powerful indexing and querying tools
- complex comparative reporting
- visualization of relations between data and codes.

One way of viewing such systems is as tools for adding **metadata** – to borrow a phrase from computing – to the transcribed data. Metadata is a structured set of comments, annotations and mark-ups that provide information about the data to which it is linked. The addition of the metadata makes it possible to find, sort and restructure your data at a more abstracted level. The addition (coding) and manipulation (analysis) of meta-level data is, of course, part and parcel of written/printed text-based qualitative data analysis. Figure 5 shows the Atlas/ti™ CAQDAS tool in use. This system allows the researcher to see their data and the coding of that data side-by-side on the screen. Such systems provide a large range of output, including reorganizations of data by codes, or graphical representations of inputted relationships between codes, as well as numerical tables of types and amounts of coding.

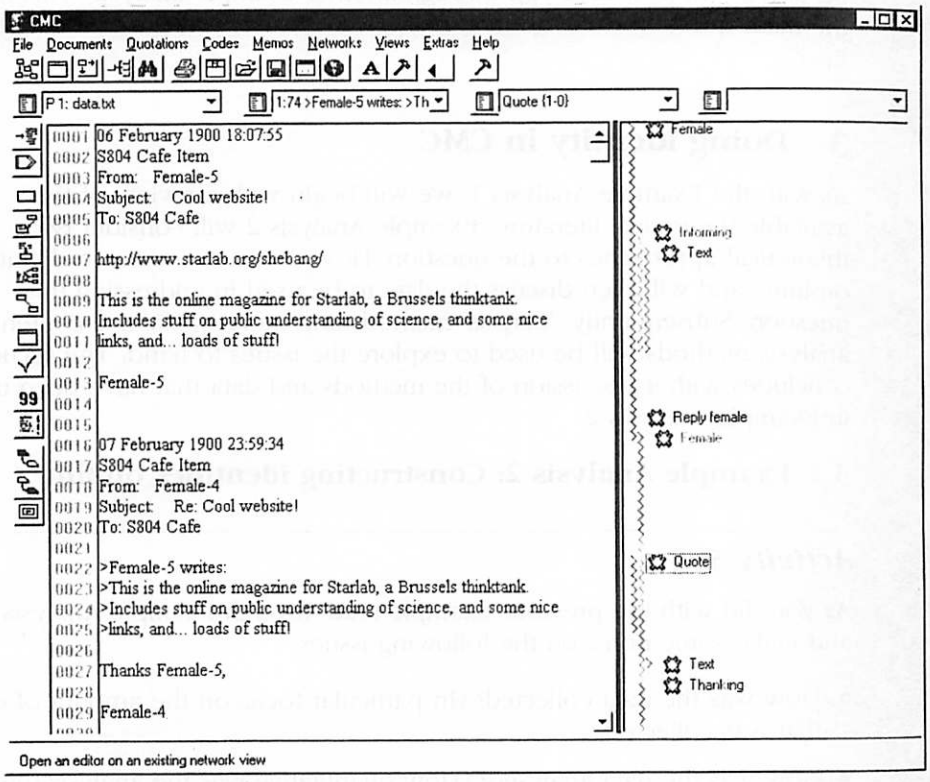


Figure 5 Initial coding of data from the Appendix using Atlas/ti™

Summary

What can we conclude from this short section on methods? First, it is clear that corpora and corpus methods provide and opportunity for those linguists interested in discourse analysis issues. Second, that there are a number of 'costs' that are incurred in taking up this opportunity: Some of these are:

- There is no 'text-book' that lays out the use of corpus methods in discourse analysis. Corpus methods employ a range of activities from quantitative word counting to more complex and qualitative content analyses.
- Corpus methods take on board some epistemological baggage. In particular they take linguistic data as direct evidence of discursive processes and practices. This contrasts with the other approaches in this book, especially Chapters Two, Four and Five.
- Corpus methods rely upon the use of computer software that ranges from 'linguistic' tools, such as word counters and concordances, to CAQDAS tools.

Section 3 will now take you through another example analysis. This time the main question to be addressed is 'How do users construct identities on-line?'.

3 Doing identity in CMC

As with the Example Analysis 1, we will begin with a review of the available theoretical literature. Example Analysis 2 will consider two theoretical approaches to the question 'How do users construct identities on-line?' and will then discuss the data to be used in addressing this question. Subsequently, 'corpus' methods and more qualitative content analysis methods will be used to explore the issues to hand. The section concludes with a discussion of the methods and data that have been used in Example Analysis 2.

3.1 Example Analysis 2: Constructing identities on-line

Activity 5

As you did with the previous example read through Example Analysis 2 and make some notes on the following issues

- How was the data collected? (In particular focus on the amount of data that was collected)
- How was the data analysed? (How quantitative was the analysis this time?)
- How did the analysis link data, theory and methods? (Was there such a clear link between theory and methods in this case?)

I'll discuss my responses to these questions once you have finished reading Example Analysis 2.

Background

To date there have been several models of 'identity on-line'. This analysis will focus on two of these. These can be described as the '**limited cues**' and the '**social presence**' models. The first model predominated early work on CMC and developed from social-psychological experiments designed to elicit the 'nature' of CMC interactions. These experiments were based on the 'social presence' model derived from the work of Short *et al.* (1976). Short compared face-to-face, video and telephone/intercom interactions and explored how users rated telecommunications media across a number of factors such as:

- unsociable–sociable
- insensitive–sensitive
- cold–warm
- impersonal–personal.

The crucial issue which Short found to influence use and perception of the communication was the level of 'social presence'. Media with less 'social presence' were viewed less favourably. The explanations of why users perceive such media as being 'asocial' or lacking in social presence focus on the lack of face-to-face cues. Interestingly, many analyses based on such 'cue-less' models have tended to set up dichotomies between task-orientated (un-emotional) and 'socio-emotional' forms of communication and interaction. Such approaches have been used in the study of CMC (see Rice and Love, 1987). One outcome of this research was a model of CMC interaction put forward by Kiesler (1986) and Kiesler *et al.* (1984). In this model the interplay of various factors (especially the lack of face-to-face cues) leads, according to the results of Kiesler *et al.*, to group polarization in CMC interactions. This in turn leads to aggressive forms of interaction (called 'flaming' in the CMC jargon). Interestingly, despite the focus on 'group polarization' these models leave out any discussion of how group membership and identities, both individual and group, are negotiated or produced on-line.

The 'social presence' model develops from but criticizes the 'limited cues model'. Spears and Lea (1992) offer an alternative to that of Kiesler *et al.* which starts from the physical separation of CMC users. They claim that two important factors need to be reconsidered if a more rounded explanation of social processes in CMC interaction is to be developed. First, one must distinguish between *in situ* interpersonal cues normally derived visually in face-to-face interaction from general social context cues and markers of self-identity. The general social context can be defined by the subject matter of the interaction, the location of the interaction within a specific 'CyberSpace', the power relationships between members of the interaction, and so forth. Self-identity can be expressed in many ways from the use of 'signatures' at the end of e-mail messages to the choice of usernames in IRC or MUD interactions. A lack of *in situ* visual cues does not imply the lack of social context and self-identity markers. Nor does a lack of visual communication prevent the complex expression of identity.

Spears and Lea also distinguish between personal identity and social identity. Personal identity is a person's complex understanding of himself or herself as an individual. Social identity derives from people's presentation of identity as part of group membership or the taking on of a social role within the interaction. Using these factors, Spears and Lea argue firstly that the lack of an active visual feedback channel leads to greater, rather than less, self-awareness. Such extra self-awareness produces differing results dependent upon the social context. Coherence to group norms and the strong expression of group identity is only likely to take place in contexts where the maintenance of the group is more salient than individual personal expression. Visual anonymity in CMC does not directly lead to group polarization but, rather, sets up structures that under different contextual conditions lead to greater or lesser reliance on the expression of group identity.

We therefore have two models of CMC identity. The first implies that the lack of other channels than the typed text leads to simplified and polarized presentations of self and group identity. The second argues that CMC identities can be very complex and that CMC does provide a lot of contextual cues. How then can a 'interactional sociolinguistic' approach combined with 'corpus methods' help to explore these models? In the analysis that follows, 'linguistic markers', seen as indicators of identity-related behaviours, are used to measure and explore differences between media.

Methods and data

The analyses of CMC are based in corpus-based techniques and have relied upon several computer-based analysis tools. Much of the 'quantitative' work was conducted through the use of 'key-word-in-context' (KWIC) and 'concordance' software. The other system used in the analysis was a CAQDAS system. The analysis follows the advice of Stubbs (1996) who argues that useful and informative analyses need to be comparative. Stubbs claims that such comparisons provide the researcher with a handle on what are in most cases relative measures. Not only are many linguistic measures relative functions of the media, more importantly, they are also functions of social context. With this issue in mind, several corpora were used in the analyses.

The two corpora of spoken material used were the British National Corpus (BNC) and the London-Lund corpus. Two samples of face-to-face interaction material were used. In the case of the BNC this consisted of 350,000 words taken from the 20 million words of conversational data available in the British National Corpus. The second sample of conversational interaction was taken from the London-Lund corpus and consisted of 196,028 words collected from interactions between equals. Telephone data (46,084 words) was also taken from the London-Lund corpus.

Four types of CMC data were also collected. These included 'chat' logged from three publicly accessible internet relay chat-rooms (IRC) on the eu.undernet server. This data included a little more than 25,000 words representing 10 hours of interaction (2,500 words per hour). In all three

rooms several nationalities met to discuss a very broad range of topics. This allowed comparison with data from computer-based video conferencing (CU-SeeMe) that comprised just over 14,000 words from 8 hours of interaction (1,750 words per hour). The CU-SeeMe data was collected from a European site where a broad range of nationalities met to discuss numerous issues. The CoSy computer conferencing data consists of over 300,000 words posted on the Open University's computer conference system – this is a sub set of the CoSy:50 corpus described in Example Analysis 1. The Usenet postings comprised over 2,000 postings made to alt.usage.english during May 1996. The software used for the analysis automatically removed all headers, signatures and – where conventionally marked – material quoted from earlier posts which formed a significant portion of most postings. This editing left just over 200,000 words of data.

Quantitative measures of identity

One of the first ways in which the importance of identity within each of the media can be measured is through an examination of the relative occurrence of **identity markers** within the overall data set. It is often claimed that the word 'the' is the highest frequency word in the English language but this is true only of the written genres that were conventionally included in such word counts. In the data being used here, the personal pronouns 'I' and 'you' appear with higher frequency in interactive genres (whether face-to-face or mediated) than the overall means for the full BNC. In other words, direct reference to self and others is higher in the data than in a very large sample of general spoken and written English. We can compare the CMC data with the conversational and telephone interaction data from our various corpora. We can collect data on the most frequent words (say the top 10) and use these frequency figures to establish an 'egocentric' index for the materials according to the frequency of self-referring pronouns. We can measure this by counting self-referential pronouns per thousand words. Table 6 shows the results.

Table 6 Egocentricity of various media

Media	*Pronoun reference to self per thousand words*
CU-SeeMe	52
Telephone	44
Conversation	42
IRC	40
CoSy	28
Usenet	21

On the other hand, the counts for 'you' perhaps indicate the degree of interactivity and concern for other. The results of this analysis are presented in Table 7.

Table 7 Reference to others in various media

Media	*Pronoun reference to other per thousand words*
Cu-SeeMe	42
Conversation	33
Telephone	33
IRC	28
CoSy	12
Usenet	11

It is interesting that CU-SeeMe scores highest on both counts, ahead of either IRC or face-to-face conversation. Computer conferencing and Usenet interactions score lowest and have the word 'the' as the highest ranking item. This can be seen as a marker of 'literate' communications practices being employed. An attempt might be made to explain such results using a 'limited social/visual cues' approach with the presence of cues leading to greater interactivity. Yet such an argument would ignore nearly all the complex aspects of the data.

There are three important findings in this data. First, all of these media have much higher levels of the use of self-reference and interpersonal reference than the mean scores derived from the full BNC. This makes it very hard to claim that CMC is a de-personalized medium. Though 'I' is not the highest ranked item in the CoSy and Usenet data, it occurs almost twice as often in these data sets as it does in the full BNC. On these measures, CMC is a highly personalized medium where individuals are making considerable use of identity markers. Second, the result must be viewed not in technological (media) terms but in relation to the genres of communication. The choice of textual features is as much (in most cases more) a product of perceived social context and the social function of the interaction, as it is of the technology of communication.

In order to look more closely at the ways in which individuals are expressing and defining their identity we need to look at specific points in the interaction where direct reference to self takes place. An obvious starting point is the use of the phrase 'I am' (or 'I'm') in the various interactions. Taking three sets of sub-corpora data matched for a size of around 14,000 words (the size of CU-SeeMe data) from the CU-SeeMe, BNC Spoken and IRC data sets, a key-word-in-context (KWIC) search was conducted. The results of the search are far too large to present here, though Table 8 overleaf provides an extract.

Table 8 An example of a KWIC result

IRC addict is what	I AM	
says the same things	I AM	.
sharOn>: well	I AM	a nice wee lassie
	I'M	a working girl
HUGO_BOSS>:	I'M	from Lisbon, POrtugal
whizz -- you're an op,	I'M	not, I'm just an ordinary user
Druidd>: yah	I AM	not really from Scotlands
turkydog>:	I AM	not sure Moo?
chaos>: loesie:	I AM	not talkin about things
	I AM	not mixing my academic
Beaker>:	I AM	not breaker I am a dude
alexarose>: Yes and	I'M	not a dude I'm cool
The rest of the time	I'M	there.
Beaker>:	I AM	tired.
	I'M	trying it now and I'll get back
Sheena>: yeah,	I'M	trying to get tby online
Beaker>:	I AM	trying to get geocities acess
ceili>:	I AM	trying Dixon
maybe it's cos	I'M	usually on in the morning

Overall, the IRC and CU-SeeMe data, despite being multilingual and therefore containing less English data, contain more 'I am' sequences than the conversation data. It is also interesting that 'not' is found to be the most frequent word following 'I am' in conversational interaction. One might want to argue that the 'too much information' of face-to-face interaction has to be controlled and mediated by the negative definition of an individual's identity. In this case possible interpretations of contextual or social cues are bracketed off or controlled through statements containing the phrase 'I am not'.

Qualitative measures of identity

A further analysis of the use of 'I am' within the corpora was conducted using a CAQDAS system. Following from the findings above, individual occurrences of 'I am' in four corpora were examined. These corpora were the CU-SeeMe data, the IRC data, the London-Lund telephone interactions between equals and the London-Lund spoken interactions between equals. One of the important initial findings concerned the repetition of the phrase 'I am' within spoken and telephone discourse. The following examples are taken from spoken and telephone corpora respectively:

A: I see ((I'm I'M I'm it's)) very kind of you Sam

A: I AM ((I'm)) going out too

This type of repetition does not occur in the CMC data. Nor does the following case which occurs often in spoken discourse where 'I am' forms part of an incomplete utterance:

A: but I put it on the wall quite often and [m] and then I did some portraits and that I find absolutely fascinating for a stupid reason which is that I'M [m] my eyes focus differently you know and I see quite differently from a distance

Such a finding indicates that the numbers of occurrences of 'I am' in the spoken and telephone data is not directly comparable with that of the CMC data. Table 9 presents the data differently pointing out the number of 'turns' in all four data sets where an 'I am' phrase was coded in the analysis that follows. It is important to note that this is also an over-count as several turns were double coded (see below) though this double coding occurred across all media to a similar degree. This provides an even more striking result with speech having the lowest occurrence of turns containing the phrase 'I am/I'm' than all the other media measured as the number of occurrences per 1000 words.

Table 9 Occurrences of 'I'm/I am' in six corpora

Text	*Total words*	*Total occurrences of 'I am/I'm'*	*Mean per 1000 words*	*Coded turns containing 'I am'*	*Mean per 1000 words*
CU-SeeMe	16717	106	6.34	126	7.54
IRC	37489	171	4.56	226	6.03
Telephone	46084	187	4.06	122	2.65
Speech	196028	498	2.54	302	1.54
CoSy	450000	1411	3.14	Not categorized	
BNC Speech	345000	1107	3.21	Not categorized	

Such a result cannot be simply fitted into the model of reduced social cues leading to limited self-presence. On the contrary, a large amount of self-expression is taking place in the CMC media – with or without the visual channel. It does, though, fit the argument of Spears and Lea that the structure and content of the interaction is tied to the social context and perceived social relations. Both the CU-SeeMe and IRC data were taken from open forums where people from a large number of nations and societies met. In both cases the social context was defined by this meeting of strangers. One would, therefore, expect a greater discussion of self. In order to assess how this is being done we need to explore the aspects of self that are referenced by the use of the phrase 'I am'.

Each of the turns in which 'I am/I'm' occurred in all the texts were qualitatively categorized. The categories were derived from a close reading of the interactions.

Coding Scheme

The complete set of categories were as follows:

- Being in action – statements for the general form 'I am doing X'
- Gender – statements of the general form 'I am female/male'
- Being in physical/geographic space – statements of the general forms 'I am in Scotland/I am in the office'
- Knowing and feeling about self and others – statements that express knowledge of self and others both objective and subjective (i.e. their emotional/personal state)
- Nationality – statements of the general form 'I am Scottish'
- Being in social relations (including social status) – statements that express social relations or status such as 'I am her tutor'
- Age – statements of the general form 'I am X years old'
- Appearance – statements about appearance such as 'I am thin/tall'
- Occupation – statements about occupation such as 'I am a programmer'
- Negativity – negative versions of the coded statements such as 'I am not an economist/I am not sure how he feels', etc.
- Futurity – statements about future events such as 'I am off to the shops later'
- Being in CyberSpace – statements about place in CyberSpace such as 'I'm off to another IRC room/I am usually at the Cornell reflector'
- Contextual and undefined expressions – statements of 'I am' that are difficult to disambiguate from context, often used as answers to questions, also statements of identity that do not readily fit above categories.

In some cases, turns were double coded where either two occurrences of 'I am/I'm' were present or where two categories were involved. In the case of the negativity category all cases here were examples of this double coding.

Table 10 gives the numerical results of this analysis and Table 11 ranks the categories for each medium.

Activity 6

Using the occurrences of 'I am' in Table 8, categorize these uses of 'I am/I'm' using the coding scheme described above. You will have to work out some criteria for each category – what is often described as 'operationalizing' your concepts. You might find that some statements fit into more than one category. Did they easily fit into these categories? Did you want to be able to explore the surrounding text to be sure of their role or purpose?

Table 10 Categorization of 'I am/I'm' in four media

	CU-SeeMe			*IRC*			*Telephone*			*Speech*		
Category	*Number*	*Normalized per 1000 words*	*Percent of turns*	*Number*	*Normalized per 1000 words*	*Percent of turns*	*Number*	*Normalized per 1000 words*	*Percent of turns*	*Number*	*Normalized per 1000 words*	*Percent of turns*
Being in action	22.00	1.32	17.46	38.00	1.01	16.81	29.00	0.63	23.77	71.00	0.36	23.51
Gender	0.00	0.00	0.00	5.00	0.13	2.21	0.00	0.00	0.00	0.00	0.00	0.00
Being in physical/ geographic space	22.00	1.32	17.46	51.00	1.36	22.57	10.00	0.22	8.20	16.00	0.08	5.30
Knowing and feeling about self and others	42.00	2.51	33.33	25.00	0.67	11.06	34.00	0.74	27.87	92.00	0.47	30.46
Nationality	4.00	0.24	3.17	14.00	0.37	6.19	0.00	0.00	0.00	1.00	0.01	0.33
Being in social relations	1.00	0.06	0.79	3.00	0.08	1.33	5.00	0.11	4.10	18.00	0.09	5.96
Age	3.00	0.18	2.38	6.00	0.16	2.65	0.00	0.00	0.00	4.00	0.02	1.32
Appearance	10.00	0.60	7.94	4.00	0.11	1.77	1.00	0.02	0.82	1.00	0.01	0.33
Occupation	8.00	0.48	6.35	21.00	0.56	9.29	2.00	0.04	1.64	8.00	0.04	2.65
Negativity	2.00	0.12	1.59	2.00	0.05	0.88	21.00	0.46	17.21	41.00	0.21	13.58
Futurity	0.00	0.00	0.00	0.00	0.00	0.00	11.00	0.24	9.02	13.00	0.07	4.30
Being in CyberSpace	9.00	0.54	7.14	43.00	1.15	19.03	0.00	0.00	0.00	0.00	0.00	0.00
Contextual & undefined expressions	3.00	0.18	2.38	14.00	0.37	6.19	9.00	0.20	7.38	37.00	0.19	12.25
	126.00	7.54	100.00	226.00	6.03	100.00	122.00	2.65	100.00	302.00	1.54	100.00

Table 11 Ranking of top five categories for four media (excluding undefined category)

	CU-SeeMe	*IRC*	*Telephone*	*Speech*
Rank position				
1	Knowing and feeling	Being in physical/ geographic space	Knowing and feelings	Knowing and feelings
2	Being in action *and* Being in physical/ geographic space	Being in action	Being in action	Being in action
3	Appearance	Being in CyberSpace	Negativity	Negativity
4	Being in CyberSpace	Knowing and feeling	Futurity	Being in social relations (inc. social status)
5	Occupation	Appearance	Being in physical/ geographic space	Being in physical/ geographic space

Discussion

Table 12 shows how I coded the occurrences of 'I am/I'm'.

Table 12 Coding of I AM/I'M

Occurrence			coding
IRC addict is what	I AM		Knowing and feeling
She says the same things	I AM	.	Knowing and feeling
shar0n>: well	I AM	a nice wee lassie	Gender
	I'M	a working girl	Gender
HUGO_BOSS>:	I'M	from Lisbon, Portugal	Being in space/ Nationality
whizz – you're an op,	I'M	not, I'm just an ordinary user	Occupation
Druidd>: yah	I AM	not really from Scotlands	Negativity/ Nationality
turkydog>:	I AM	not sure Moo?	Knowing and feeling
chaos>: loesie:	I AM	not talkin about things	Contextual and undefined
	I AM	not mixing my academic ideas	Knowing and feeling
Beaker>:	I AM	not breaker I am a dude	Negativity/Knowing and feeling
alexarose>: Yes and	I'M	not a dude I'm cool	Negativity/Knowing and feeling
The rest of the time	I'M	there.	Being in space
Beaker>:	I AM	tired.	Knowing and feeling
	I'M	trying it now and I'll get back	Being in action
Sheena>: yeah,	I'M	trying to get tby online	Being in action/ Being in Cyberspace
Beaker>:	I AM	trying to get geocities acess	Being in action/ Being in Cyberspace
ceilî>:	I AM	trying Dixon	Being in action
maybe it's cos	I'M	usually on in the morning	Being in action/ Being in Cyberspace

Discussion

From the results in Table 10 and Table 11 it is clear that important differences exist between the four media. In the case of the CMC media, it is hard to accept the claim that reduced social/visual cues lead to a loss of self-awareness. On the contrary, as Spears and Lea argue, it seems to lead to heightened self-awareness and self-reference. Knowledge of self and others is high for all media. This implies that in all the media, from video conferencing to speech, interpersonal knowledge and understanding is central. Also the extensive self-reference in relation to action implies that even in virtual space individuals perceive themselves as social and material actors. For the CMC media, self reference is direct, positive, concerned with location – be that physical or virtual – and with appearance. This contrasts with the spoken and telephone data that are more concerned with the social structural factors (social relations and social time) and the negotiation/control of contextual factors (negativity and use of negative forms).

Accepting that CMC identities can only be constructed via the communications medium would imply, as has been argued time and again, that it allows only a limited number of methods for their expression and interpretation. Yet the work of many CMC researchers has noted the importance of gender identity markers (see Graddol and Swann, 1989; Herring, 1992; 1993a; 1993b; Spender, 1996; Turkle, 1996). The results above also make clear that CMC users spend considerably more time presenting aspects of identity through the communicative text than do people interacting in other media. The construction of CMC identities is only limited by the user's ability to provide coherent textual descriptions. By exploring the use of phrases such as 'I am/I'm', we are examining one of the direct methods through which such textual identities are constructed.

3.2 Comment on Example Analysis 2

How was the data collected?

Once again two existing corpora were used: the London-Lund and the BNC. In each case, though, the analysis focused on only part of these corpora. The CMC data were once again downloaded although some of these were then 'processed', for example, the Usenet data had the 'quoted text' removed. As Usenet does not have information on a messages position in an interaction, as conferencing does, users often include material (quotes) from the original posting to which they are replying. The corpora, then, were not 'exact copies' of the original interaction as was the case of the IRC and CU-SeeMe data. The size of the data sets was much smaller as well. There are a number of reasons for this. First, the IRC and CU-SeeMe data represented a lot of hours of interaction but not so much text. Second, and this is an important point, the data were to be subjected to qualitative methods. Even though these were supported by the use of a CAQDAS system they still required the researcher to painstakingly go though the transcripts coding up the relevant utterances. This is something that would not be possible for a corpus of several million words.

How was the data analysed?

The analysis this time included both quantitative and qualitative methods. It began with the creation of some basic measures of the levels of reference to self and to others in the data. The analysis then moved on to conduct qualitative content analyses of the data. In this case the occurrences of self-reference were put into categories that seemed to reflect their role or purpose in the interaction. This involved the kind of close reading that is more indicative of other discourse analysis approaches in this volume.

How did the analysis link data, theory and methods?

Unlike the previous example, the analysis (which followed an essentially 'positivist' model of developing a hypotheses and then measuring this analysis) proceeded from two possible explanations to an analysis that was more exploratory. Though some quantitative measures were put forward and the results of the qualitative content analysis were tabulated, the overall approach was one of understanding identity construction in CMC. No statistical methods were used and no hypotheses were proposed. The results were used to indicate the communicative practices of CMC users and to argue against the simplistic 'cue-lessness' explanation of CMC behaviour.

4 Gender on-line

This section is designed to give you an opportunity to analyse a CMC interaction for yourself. The data, which comes from an Open University computer conferencing interaction, are provided in the Appendix at the end of this chapter. The interaction took place in a 'chat' conference attached to a course focused on 'science communication' issues. The identities of the posters has been removed and replaced by markers indicating their gender combined with a number to differentiate posters (e.g. Male-1; Female-3). The topic under scrutiny is that of gender differences in language use.

Before setting you off on this task there are some background issues to be covered. It is now commonly accepted that there are differences in men and women's use of language – though the source of these differences is hotly debated. A large amount of psychological, sociological and linguistic literature on this issue now exists (e.g. Coates, 1997; 1996; 1993; Coates and Cameron, 1989; Coates *et al.*, 1989; Tannen, 1994; 1993). This research raises three main interrelated issues. First, there are the inequalities in the structure of male/female interactions. Second, there are differences in the linguistic practices and strategies that the sexes use in interactions. These can include differences in turn taking conventions, means of gaining the conversational floor, and means of directing the flow of the interaction. Third, there are differences in the purposes for which people engage in linguistic interaction.

Susan Herring (1992; 1993a; 1993b) has written extensively upon the issue of gender, language and CMC. This section will follow her main arguments and make use of her research methods. Herring used data collected from a study of two educational/academic discussion lists to

clearly demonstrate differences in gender use and access to CMC. The two e-mail lists which Herring explored were the LINGUIST electronic mail list which is devoted to the discussion of all aspects of language and linguistics, and Megabyte University (MBU) which is a list concerned with computers and writing. Both lists represent university-level interactions between academics and students; the LINGUIST list having over 1,000 members and the MBU list over 250 members at the time of Herring's research. Over a period of one year, Herring conducted an ethnographic observation of the interactions taking place. This involved collecting a full transcript of the interaction and from this collecting data on the participants, the issues raised and discussed, and other relevant information (for a full discussion of a linguistic ethnography of CMC see Cherny, 1999). During this time she subjected two extended discussions of specific subjects from each discussion list to detailed linguistic and sociolinguistic analysis (Herring, 1993a).

The first important finding that Herring uncovered was the disparity in participation. Both lists had reasonably high numbers of women members (36% on LINGUIST; 42% on MBU). Despite this, women contributed far less than the men and the levels of participation varied according to the topic under discussion. For example 30% of women on both lists participated in 'sexism'-related discussions compared to 16% of women who participated in 'theory'-related discussions. Herring also found that messages from women were shorter on average with only men posting messages ten screens or more in length. Finally, Herring notes that women's messages gain fewer replies in these mixed gendered interactions.

Herring also notes differences in the CMC practices of men and women. The first of these is topic selection. Looking at her data from the LINGUIST list, Herring notes that men were more likely to post messages on specific issues or to provide specific information whereas women were more likely to post on personal aspects of the discussion or to post queries to other list members (see Table 13, based on Herring, 1993a).

Table 13 Topic selection and gender in CMC interactions

Gender	*Most postings*	←	→	*Least posting*
Women	personal	queries	information	issues
Men	issues	information	queries	personal

Herring next noted differences in language style and content. Using a set of features defined in terms of their attribution to different genderlects – gender-based language styles – Herring discovered large differences in style (see Table 14 below, taken from Herring, 1993a). Of the 261 LINGUIST messages she analysed, 68% of women's postings contained women's genderlect compared to only 31% of male messages. On the other hand, 48% of male messages contained *only* male features. This contrasted with only 18% of women's messages solely containing male genderlects. Lastly

whilst 46% of women's messages *combined* both male and female genderlects, only 14% of male messages did so (Herring, 1993a). This result indicates that women are engaging in 'male' strategies in order to remain in the interaction.

Table 14 Features of women's language and men's language

Women's Language	*Men's Language*
attenuated assertions	strong assertions
apologies	self-promotion
explicit justifications	presuppositions
questions	rhetorical questions
personal orientation	authoritative orientation
supports others	challenges others
	humour/sarcasm

There is, of course, much more to Herring's work but for now we have a number of measures of gender differences in CMC interactions. These are:

- disparity in postings – measured by the different number of messages posted, with men posting more messages
- disparity in message length – measured by the longer postings by men
- disparity in replies – measured by the lower response rate to women's messages – they get fewer replies
- differences in message topics – as measured by the types of topics in men and women's postings (see Table 13)
- differences in language styles (see Table 14).

Activity 7

Using the five measures developed by Herring that are listed above, conduct an analysis of the data in the Appendix at the end of this chapter. Data on the proportions of men and women involved in the conference was not available so you will not be able to tell definitively if men posted proportionally more messages to the conference as a whole. You will also have to decide your own criteria for the categories in Tables 13 and 14 – you will have to operationalize your concepts. It might be easier to do this if you read the data through first. This should be fairly straightforward for Table 13. Is the message personal, querying, information-providing or issue-based (or possibly a combination)? Table 14 may be more difficult and you will find that you have to make some subjective choices. To help you along, Table 15 overleaf provides some suggestions.

Table 15 Examples of gender-based differences in language styles

Women's Language		*Men's Language*	
Attenuated assertions	Assertions on a topic that are prefaced with some form of distanced modality. E.g. 'It might be true that...'	Strong assertions	Assertions made without or with involved modalities. E.g. 'It is true that...', 'I am very sure that...'
Apologies	Retractions of position or statements employing apologetic language	Self-promotion	Statements that stress the rightness, importance or social standing of the individual
Explicit justifications	Statements that provide the justification or basis for opinions	Presuppositions	Statements that assume certain facts or opinions
Questions	Genuine open questions seeking a response	Rhetorical questions	Questions set up for an answer from the speaker themselves
Personal orientation	Presentation of statements and ideas from a personal position	Authoritative orientation	Presentation of statements and ideas from a claimed, assumed or asserted position of authority
Supports others	Statements in support of others' ideas or opinions	Challenges others	Statements that challenge others' ideas or opinions
		Humour/ sarcasm	Jokes or statements used to belittle others or others opinions.

You may wish to present your data as tables that allow comparisons between the genders. Do you find the same patterns as Herring in this small data set?

You might also, if you have the time and inclination, conduct analyses of the types described in Sections 1 and 3, for example type/token or lexical density measures, or the categorization of all the occurrences of 'I' or 'I am'.

Discussion

I will not provide you with all the 'results' but there are a few things you might wish to consider. First, defining criteria by which to categorize data is not as straightforward as it seems and requires you to make use of your understandings of both language (in this case English) and possibly of British culture as well. You might also have noticed how posters copy bits from previous messages – this is one way that the 'thread' of a conversation is maintained. Second, once you have worked out some categorization scheme, the actual coding is a relatively easy, if repetitive, task. The use of computer tools might make the process simpler and may allow you to deal with a greater body of data but they do not get rid of the intellectual effort required to select criteria.

To give you some indication of the types of results that you may expect, I looked at two criteria: message length and number of replies. I counted the actual number of new lines in each message (I ignored quotes) and compared this across gender and got the results shown in Table 16. This analysis took me a couple of hours. I will admit that I used a CAQDAS tool and statistical software. Even at this rate, to look at several hundred messages as Herring did, or to look at thousands as in Example Analysis 1, would have taken about 20 to 40 hours.

Table 16 Analysis by gender of message length and number of replies

Gender	*Number of messages*	*Mean number of messages*	*Maximum length in lines*	*Mean number of lines*	*Median number of lines*
Female	11	1.57	11.00	3.45	2.00
Male	15	1.86	12.00	5.07	5.00

This table seems to show that, on average, the men posted more and posted longer messages; the mean and median lengths of messages, as measured by the number of lines, are longer. Having said this, the sample is very small and the result is not statistically significant. Is the use of 'lines' the best measure of message length? I might be tempted to re-do this analysis by counting words per message as a more accurate measure of message length. I also looked at the number of replies to men and women and found that there were eight replies to women's postings and 13 replies to men. This implies that women had a 73% chance of being replied to compared to 87% for men, though this difference cannot be easily statistically tested.

Appendix

Please note that many of the messages contain 'quoted material' from other messages, even quotes within quotes. This part of the messages is marked with the usual '>' convention at the start of lines.

```
06 February 1900 18:07:55
N900 Cafe Item
From: Female-5
Subject: Cool website!
To: N900 Cafe

http://www.starlab.org/shebang/

This is the online magazine for Starlab, a Brussels
thinktank.
Includes stuff on public understanding of science, and
some nice
links, and... loads of stuff!

Female-5


07 February 1900 23:59:34
N900 Cafe Item
From: Female-4
Subject: Re: Cool website!
To: N900 Cafe

>Female-5 writes:
>This is the online magazine for Starlab, a Brussels
thinktank.
>Includes stuff on public understanding of science,
and some nice
>links, and... loads of stuff!

Thanks Female-5,

Female-4
```

09 February 1900 15:46:15
N900 Cafe Item
From: Male-3
Subject: Re: Cool website!
To: S804 Cafe

>http://www.starlab.org/shebang/

Brilliant web site Female-5!

Regards

Male-3

09 February 1900 20:02:09
N900 Cafe Item
From: Male-7
Subject: Re: Cool website!
To: N900 Café

Thanks for posting that website.
The chairman of Starlab was on the radio the other day. He was offer
job because he has no training in Science. I thouht that was a
brilliant idea!

Male-7

10 February 1900 18:30:38
N900 Cafe Item
From: Female-5
Subject: Re(2): Cool website!
To: N900 Café

>Male-7,oufcnt2.open.ac.uk writes:
>Thanks for posting that website.
>The chairman of Starlab was on the radio the other day. He was offer
>job because he has no training in Science. I thouht that was a

>brilliant idea!
>
>Male-7
Yes, I heard the interview. I WANT HIS JOB!!!!!

11 February 1900 16:42:15
N900 Cafe Item
From: Male-7
Subject: Re(3): Cool website!
To: N900 Café

>Female-5 writes:
>Male-7,oufcnt2.open.ac.uk writes:
>Thanks for posting that website.
>The chairman of Starlab was on the radio the other day. He was offer
>job because he has no training in Science. I thouht that was a
>brilliant idea!
>
>Male-7
>
>
>Yes, I heard the interview. I WANT HIS JOB!!!!!

Sorry, you can't have it because it's going to be mine!

12 February 1900 10:51:40
N900 Cafe Item
From: Female-5
Subject: Re(4): Cool website!
To: N900 Café

>Male-7,oufcnt2.open.ac.uk writes:
>Female-5 writes:
>Male-7,oufcnt2.open.ac.uk writes:
>Thanks for posting that website.
>The chairman of Starlab was on the radio the other day. He was offer
>job because he has no training in Science. I thouht that was a

>brilliant idea!
>
>Male-7
>
>
>Yes, I heard the interview. I WANT HIS JOB!!!!!
>
>Sorry, you can't have it because it's going to be mine!

Job share??

12 February 1900 12:12:03
N900 Cafe Item
From: Female-4
Subject: Re(3): Cool website!
To: N900 Café

>Female-5 writes:
>Yes, I heard the interview. I WANT HIS JOB!!!!!

Me too!

Female-4

12 February 1900 15:35:22
N900 Cafe Item
From: Male-8
Subject: TUTOR ICON
To: N900 Café

Hi,
Can someone fill me in on what this tutor group icon is,
especially as I still do not have any letter from my tutor and the
TMA is looming and there are questions that I'd like to ask. Many
thanks.

Male-8

13 February 1900 11:53:42
N900 Cafe Item
From: Male-9
Subject: tutor icon
To: N900 Café

G'day to all fellow communicators,

As Male-8 stated (12/02/2000), I too do not have any idea what this
tutor icon is or whom is allocated as my tutor for this module. I
would appreciate any help (via this medium) or do I need to contact
the helpdesk?

please advise asap

13 February 1900 15:22:46
N900 Cafe Item
From: Male-8
Subject: Re: tutor icon
To: N900 Café

Hi Male-9,
Glad I'm not the only one still wondering what is going on.
Thanks for that

Male-8

13 February 1900 17:15:29
N900 Cafe Item
From: Female-6
Subject: Re(2): tutor icon
To: N900 Café

Don't know if its any help, but mine was 'hiding'.
It was underneath the OU icon.....have a look.....it could
be there!
Female-6

13 February 1900 19:50:17
N900 Cafe Item
From: Female-7
Subject: Hidden icons
To: N900 Cafe
Male-9/Male-8,

As Female-6 said, the icons can 'hide' underneath each other, I lost
one in my last course. If you drag and drop the existng icons around
for a bit you may find the hidden ones. Good luck!

Female-7

13 February 1900 19:53:52
N900 Cafe Item
From: Female-3
Subject: Missing Tutor icons
To: N900 Café, Male-8, Male-9

If you look at my message on the Noticeboard about N900 Tutors for
2000, you will see it lists which tutors are looking after students
in particular regions.
Thus Male-8, you should be looking for an icon on your desktop that
says N900 Sue's Group, which is the icon for Sue Barker's group, and
Male-9, you should be looking for an icon for John Forrester's group,
N900 John's Group.
If you can't see them immediately, try viewing by name.
Female-3

Female-3
N900 Course Manager
MSc in Science

14 February 1900 18:09:04
N900 Cafe Item
From: Male-8
Subject: Re: Missing Tutor icons
To: N900 Café

Thanks Female-3,

it magically appeared after my message but definietly wasn't there
before, X-files stuff!! Thanks.

14 February 1900 19:58:54
N900 Cafe Item
From: Male-7
Subject: TV Prog worth watching
To: N900 Café

Did anyone see ''6 Experiments that Changed the World'' last night?
Ken Campbell was explaining General Relativity and did as good a job
as I've heard (but then I'm no physicist -- so who am I to judge?).

He's reproducing a different experiment every week on Sunday nights
at 7:00 on Channel 4 (just after Time Team).

16 February 1900 00:14:46
N900 Cafe Item
From: Female-4
Subject: Re: TV Prog worth watching
To: N900 Café

>Male-7 writes:
>Did anyone see ''6 Experiments that Changed the World'' last night?
>Ken Campbell was explaining General Relativity and did as good a job

>as I've heard (but then I'm no physicist -- so who am I to judge?).
>
Well I am a physicist and I think he did a pretty good job too. Excellent stuff!

Female-4

16 February 1900 17:04:50
N900 Cafe Item
From: Male-5
Subject: Re: TV Prog worth watching
To: N900 Café, N900NoticeBoard, N900Male-5's Group

>Male-7 writes:
>Did anyone see ''6 Experiments that Changed the World'' last night?
>Ken Campbell was explaining General Relativity and did as good a job
>as I've heard (but then I'm no physicist -- so who am I to judge?).
>
>He's reproducing a different experiment every week on Sunday nights >at 7:00 on Channel 4 (just after Time Team).

Well spotted Nick (and Jacqui). This was indeed an enjoyable programme and one which could provide an example for those searching for an example of science communication to analyse for question 1 of TMA 01.

The series ''6 Experiments that Changed the World'' continues next Sunday evening (7pm on Channel 4) with an examination of the work of Marie Curie.

best wishes
Male-5

04 March 1900 18:13:00
N900 Cafe Item
From: Male-6
Subject: Tutor Icon
To: N900 Café

Hello,

I have tried a mail to the help desk but I have not received a reply.

I have had my tutor icon taken away from me.

What have I done? Whatever it is I will apologise, just give me my
tutor icon back!!!

Thanks
Male-6

04 March 1900 18:31:44
N900 Cafe Item
From: Female-3
Subject: Re: Tutor Icon
To: N900 Cafe
Cc: Male-5

>Male-6 writes:
>I have had my tutor icon taken away from me.

>What have I done? Whatever it is I will apologise, just give me my
>tutor icon back!!!
Don't know what happened Stuart -- one of these FC 'glitches'!
Anyway, it should back now -- ok?
Female-3

Female-3
Programme Course Manager
MSc in Science

04 March 1900 19:06:09
N900 Cafe Item
From: Male-2
Subject: Re(2): Fwd(2): Re: Problems with Athens and passwords
To: MSc Sci Library conf, Male-4, N900 Café, Female-2

Alan has had probs with sciencedirect logging on -- I did too, [made
my initial reg --> accessed sd OK once, then failed] -- contacted John
Doe -- their helpline on e-mail John.Doe@somewhare.net which is on the
sciencedirect initial page -- he sent back an e-mail with user ID and
password + hints -- been OK ever since.

I should've contacted Alison, but confess I forgot 'til now -- sorry!

Does this help -- I find it works well, 'tho not been there for a week
or so.....

Regards

Male-2

05 March 1900 10:56:34
N900 Cafe Item
From: Male-4
Subject: Re(3): Fwd(2): Re: Problems with Athens and passwords
To: Male-2
Cc: MSc Sci Library conf, N900 Café, Female-2

Thanks Male-2 I''ll give it a go.

Male-4

09 March 1900 23:05:10
N900 Cafe Item
From: Female-1
Subject: GM Foods
To: N900 Cafe
Tonight's Horizon was about the GM food debate and can be accessed
on: www.bbc.co.uk/horizon.

10 March 1900 08:49:26
N900 Cafe Item
From: Male-3
Subject: Greenpeace = Starvation
To: N900 Café

Last nights Horizon programme brought out the true colours of
Greenpeace. That they were willing to see people in the Third Worlds
die of starvation.

That may have been a slip by their UK rep but it does reveal their
non-public attitude.

I hope their support dries up!

Male-3

10 March 1900 20:49:41
N900 Cafe Item
From: Male-2
Subject: Re: Greenpeace = Starvation
To: Male-3, N900 Café

Come on, Greenpeace don't make the situation in these countries -
that's a political situation -- like the Iraq one that I was talking
about yesterday [to an Iraqi consultant]. If the regime there is

maintained, either for balance of power, or for the desires of a
dictator, the way around it is not to keep punishing the people,
agreed -- but then why keep the food stocks, medicines etc. for the
elite anyway?

Male-2

10 March 1900 22:46:30
N900 Cafe Item
From: Male-1
Subject: Re(2): Greenpeace = Starvation
To: N900 Café

>Male-2,oufcnt2.open.ac.uk writes:
>Come on, Greenpeace don't make the situation in these countries -
>that's a political situation -- like the Iraq one that I was talking
>about yesterday [to an Iraqi consultant]. If the regime there is
>maintained, either for balance of power, or for the desires of a
>dictator, the way around it is not to keep punishing the people,
>agreed -- but then why keep the food stocks, medicines etc. for the
>elite anyway?
Perhaps Male-3 was talking about Mexico where the ground is too acid for normal corn.

Male-1

Further Reading

CMC

Cherny, L. (1999) *Conversation and Community: Chat in a Virtual World*, Stanford, CSLI Publications.

Cherny, L. and Weise, B. (eds) (1996) *Wired Women: Gender and New Realities in Cyberspace*, Seattle, Seal.

Herring, S. (ed.) (1996) *Computer-Mediated Communication: Linguistic, Social and Cross-cultural Perspectives*, Amsterdam, John Benjamins.

Hiltz, S.R. and Turoff, M. (1978) *Network Nation: Human Communication via Computer*, Reading, MA.: Addison-Wesley.

Jones, S. (1998) *Cybersociety 2.0: Revisiting Computer-Mediated Communication and Community*, Thousand Oaks, CA: Sage.

Lea, M. (ed.) (1992) *Contexts of Computer-Mediated Communication*, New York, Harvester Wheatsheaf.

Rheingold, H. (1993) *The Virtual Community*, Reading, MA.: Addison-Wesley.

Turkle, S. (1996) *Life on the Screen: Identity in the Age of the Internet*, London, Orion.

Analysis

McEnery, T. and Wilson, A. (1996) *Corpus Linguistics*, Edinburgh, Edinburgh University Press.

Stubbs, M. (1996) *Text and Corpus Analysis: Computer Assisted Studies of Language and Culture*, Oxford, Blackwell.

Tannen, D. (1993) *Gender and Conversational Interaction*, New York, Oxford University Press.

References

Barton, D. (1991) 'The contribution of a social view of literacy to promoting literacy among adults', *Centre for Language in Social Life: Working Paper Series, No.30*, Lancaster University.

Chafe, W. and Danielewicz, J. (1987) 'Properties of spoken and written language' in Horowitz, R. and Samuels, S.J. (eds) *Comprehending Oral and Written Language*, San Diego, CA, Academic Press.

Cherny, L. (1999) *Conversation and Community: Chat in a Virtual World*, Stanford, CSLI Publications.

Coates, J. (ed.) (1997) *Language and Gender*, London, Blackwell Publishers.

Coates, J. (1993) *Women, Men and Language*, London, Longman Higher Education.

Coates, J. (1996) *Women Talk*, London, Blackwell Publishers.

Coates, J. and Cameron, D. (eds) (1989) *Women in Their Speech Communities*, London, Addison Wesley.

Coulthard, M. (ed.) (1992) *Advances in Spoken Discourse Analysis*, London, Routledge.

Coulthard, M. (ed.) (1994) *Advances in Written Text Analysis*, London, Routledge.

Fielding, R. and Lee, R. (1991) *Using Computers in Qualitative Analysis*, Thousand Oaks, Sage.

Graddol, D.J. and Swann, J. (1989) *Gender Voices*, London, Blackwell.

Greenbaum, S. and Svartvic, J. (1990) 'The London-Lund corpus of spoken English' in Svartvic, J. (ed.) *The London-Lund Corpus of Spoken English: Description and Research*, Lund, Lund University Press.

Halliday, M.A.K. (1985) *Spoken and Written Language*, Deakin, Deakin University.

Herring, S. (1992) 'Gender and participation in computer-mediated linguistic discourse' in *ERIC Clearing House on Languages and Linguistics*, October.

Herring, S. (1993a) 'Gender and democracy in computer-mediated communication', *Electronic Journal of Communication*, vol.3, file Herring V3N293.

Herring, S. (1993b) 'Culture vs. Technology, or Why Computer-Mediated Communication Is Still Sexist', Paper presented to the International Pragmatics Conference.

Hodge, R. and Kress, G. (1988) *Social Semiotics*, Cambridge, Polity, extracts reproduced in Wetherell, M., Taylor, S. and Yates, S.J. (eds) (2001) *Discourse Theory and Practice: A Reader*, London, Sage in association with The Open University.

Johansson, S. (1991) 'Times change, and so do corpora' in Aijmer, K. and Altenberg, B. (eds) *English Corpus Linguistics*, London, Longman.

Kiesler, S. (1986) 'The hidden messages in computer networks', *Harvard Business Review*, Jan-Feb, pp.46–58.

Kiesler, S., Siegel, J. and McGuire, T. (1984) 'Social psychological aspects of computer-mediated communications', *American Psychologist*, no.39, pp.1123–34

McEnery, T. and Wilson, A. (1996) *Corpus Linguistics*, Edinburgh, Edinburgh University Press.

Rice, R.E. and Love, G. (1987) 'Electric-emotion: socio-emotional content in a computer mediated communication network', *Communication Research*, no.14, pp.85–108

Scribner, S. and Cole, M. (1981) *The Psychology of Literacy*, Cambridge, MA, Harvard University Press.

Short, J., Williams, E. and Christie, B. (1976) *The Social Psychology of Telecommunications*, London, Wiley & Sons.

Spears, R. and Lea, M. (1992) 'Social influence and the influence of the "social" in computer-mediated communication' in Lea, M. (ed.) *Contexts of Computer-Mediated Communication*, Hemel Hempstead, Harvester Wheatsheaf.

Spender, D. (1996) *Nattering on the Net*, North Melbourne, Spinifex Press.

Street, B.V. (1984) *Literacy in Theory and Practice*, Cambridge, Cambridge University Press.

Stubbs, M. (1996) *Text and Corpus Analysis: Computer Assisted Studies of Language and Culture*, Oxford, Blackwell.

Tannen, D. (1993) *Gender and Conversational Interaction*, New York, Oxford University Press.

Tannen, D. (1994) *Gender and Discourse*, New York, Oxford University Press.

Turkle, S. (1996) *Life on the Screen: Identity in the Age of the Internet*, London, Orion

Weitzman, E.A. and Miles, M. B. (1995) *Computer Programs for Qualitative Data Analysis: A Software Source Book*, Thousand Oaks, Sage.

Wetherell, M., Taylor, S. and Yates, S.J. (eds) (2001) *Discourse Theory and Practice: A Reader*, London, Sage in association with The Open University.

Yates, J. and W. Orlikowski, (1992) 'Genres of organizational communication: A structurational approach to studying communication and media', *Academy of Management Review*, vol. 17, pp.299–326.

Yates, S.J. (1996) 'Oral and literate linguistic aspects of CMC discourse: A corpus based study' in Herring, S. (ed.) *Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives*, Amsterdam, John Benjamins.

Yates, S.J. (2000) 'Computer-Mediated Communication – The future of the letter?' in Barton, D. and Hall, N. (eds) *Letter Writing as a Social Practice*, Amsterdam, John Benjamins.

Zipf, G.K. (1935) *The Pscho-biology of Language: An Introduction to Dynamic Philology*, Boston, MIT Press.