INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

**Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600**

UMI[®]

Creation and Analysis of a Corpus of Internet Relay Chat

by

Wernfrid Doell, B.A.

A thesis submitted to:

the Faculty of Graduate Studies and Research

in partial fulfilment of

the requirements for the degree of

Master of Arts

School of Linguistics and Applied Language Studies

Carleton University

Ottawa, Ontario

15 May 2000

© 2000

2000, Wernfrid Doell

**National Library
of Canada**

**Acquisitions and
Bibliographic Services**

**395 Wellington Street
Ottawa ON K1A 0N4
Canada**

**Bibliothèque nationale
du Canada**

**Acquisitions et
services bibliographiques**

**395, rue Wellington
Ottawa ON K1A 0N4
Canada**

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-52347-0

Canada

***The undersigned recommend to the Faculty of Graduate Studies
acceptance of the thesis***

"Creation and Analysis of a Corpus of Internet Relay Chat"

submitted by Wernfrid Doell

***in partial fulfilment of the requirements for
the degree of Master of Arts***

Thesis Supervisor

Director

School of Linguistics and Applied Language Studies

***Carleton University
May 2000***

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION	1
--------------------	---

CHAPTER 2

RESEARCH REVIEW	4
Technical introduction	4
Ferrara et al. (1991)	9
Murray, D. E. (1991a)	11
Lundstrom, P. (1995)	12
Collot, M. & Belmore, N. (1996)	14
Wilkins, H. (1991)	15
Bechar-Israeli, H. (1995)	17
Bays, H. (1999)	18
Lunsford, W. (1996)	19
Werry, C.C. (1996):	19
Paolillo, J. (1999)	21
Doell, W. (1997)	23
Doell, W. (1999)	24
Summary	25

CHAPTER 3

METHODOLOGY	27
Data collection and corpus organization	27

Channel selection	29
Privacy	31
Population	32
Study organization	33

CHAPTER 4

DESCRIPTIVE STATISTICS	35
Basics	35
Correlations between variables	39
100 most frequent words	40
Correlations between the 100 most frequent words	47
Frequency distributions in the 100 most frequent words	49
Type/token ratios	50
ANOVA and post hoc tests	51
Summary	55

CHAPTER 5

ANALYSIS OF CIRC FEATURES COMPARABLE TO SPEECH AND TO WRITING	56
Deictics	56
Personal Pronouns	56
Demonstrative pronouns and determiners	71
‘Here’	73
‘There’	74
‘This’	76

‘That’	77
‘These’ and ‘those’	78
‘Then’	79
Summary	80
Questions	80
Wh-questions	81
Yes/no questions and response items	84
Summary	88
Hedging	88
Contractions	90
Contractions with ‘not’	90
Contractions with ‘have’	94
Contractions with ‘is’ and ‘has’	95
Contractions with ‘will’	96
Dialects	98
Ellipsis	101
Colloquial Usage	102
Syntactical complexity	104
Subordination and coordination	104
Passives	106
Greeting and Leave-taking	107
Greeting	107

Leave-taking	110
CHAPTER 6	
ERSATZ - FEATURES OF IRC TALK WHICH ATTEMPT TO SIMULATE SPEECH ...	111
Emoticons	111
Action messages	115
Backchanneling	117
Phonetic spelling	119
Paralinguistic and suprasegmental cues	121
CHAPTER 7	
INNOVATION AND CREATIVITY	123
Nickname morphism	123
ASCII art	126
Neologisms	127
Innovative spellings	129
Abbreviation	130
CHAPTER 8	
CONCLUSION	136
Results of the study	136
Applications	138
Suggestions for further research	140
CHAPTER 9	
APPENDIX	147

Appendix A:	147
Appendix B:	151
Appendix C:	153
Appendix D:	155
Appendix E:	156
Appendix F:	157
Appendix G:	160
Appendix H:	160
Appendix I:	171
Appendix J:	172

ACKNOWLEDGEMENTS

I would like to thank Kassia Balian for her unwavering support for this project and the many hours she spent proofreading the different draft versions of this paper. Likewise, I would like to thank my supervisor Professor Helmut Zobl for his invaluable advice and directions. Further thanks go to Hillary Bays for sharing her enthusiasm for IRC research with me, Professor Leslie Siegrist for his technical trouble shooting of the wconcord software, Lynne Young and Richard Darville for their contributions in the early stages of the project, Janna Fox and Martin Kessner for their help with the statistical analysis, and Lola and Dude for their moral support.

ABSTRACT

This thesis reports the process of creation and the subsequent computer-aided analysis of a large-scale corpus of Internet Relay Chat (henceforth IRC). The corpus created from log files of several IRC sessions contains one million running words collected in ten IRC chat rooms or channels. For the analysis of the corpus, concordancing software was used to extract words and their frequencies.

A statistical analysis of some basic measures obtained with the concordancer yielded insights into the internal structure of the corpus. For example, the data from one channel, i.e. a subcorpus of 100,000 words, was found to be significantly different from the others. Influences from dirty data could also be pinpointed due to the statistical analysis.

The analysis of the corpus concentrated on features of IRC which have been identified by prior research in the field, including our own pilot studies. A special focus was placed on linguistic characteristics which are typically associated with casual spoken conversation. The analysis revealed that our corpus data generally supported claims made by prior research, however, using our corpus the patterns examined can now be quantified. Moreover, the analysis yielded massive evidence in support of the hypothesis that IRC interaction is more similar to casual spoken conversation than any traditional written genres. Furthermore, patterns of distinctive systematic linguistic variation were observed between groups of IRC channels. The applications for these findings in the corpus-based study of spoken interaction and variation studies are discussed in the concluding sections of this paper.

CHAPTER 1

INTRODUCTION

The explosive rise of the Internet during the last decade has brought with it the emergence of a novel kind of communication that is made possible by the use of networked computers. Certainly the most popular variant of this computer-mediated communication, or CMC for short, is email. In just a few years it has evolved from an occupation of computer specialists and academics to a means of mass communication which is clearly illustrated by the eight million messages which are sent daily. A more recent development is the various so-called chat systems which have begun to appear throughout the Internet. Starting with the creation of the first popular chat software called Internet Relay Chat or IRC¹ in 1988, their use has seen a tremendous proliferation in recent years.

There is virtually no major website which does not host its own website chat, almost no Internet service provider who does not make on-site chat programs available to its customers. Moreover, the technology has been taken another step forward with the creation of instant messaging systems. First introduced by a small Israeli company called ICQ², these programs have taken the Internet by storm and now all the big players offer their own versions such as

¹ See appendix F for some samples of Internet Relay Chat.

² ICQ is an innovative spelling for 'I seek you'.

yahoo! Messenger, MSN messenger, or AOL instant messenger. The newest development using this technology is a program which allows Internet users who surf the same website to exchange comments while they browse the same pages.

However, the original program, IRC, has not been replaced by its more recent relatives. On the contrary, its popularity continues to burgeon. This is proved by IRC user statistics, some of which are available online. For example, IRC-Net, one of several big IRC networks, counts between 40,000 and 60,000³ participants every day. In comparison, in an article published just five years ago, Bechar-Israeli (1995) reports an average number of 8,000 users on the IRC network.

Shortly after its conception, researchers began exploring this new type of communication that was evolving on the Internet. Researchers from a plethora of disciplines have been studying phenomena related to CMC ever since. Much work has been done in applied linguistics in order to assess the value of CMC for delivering distance education and enhancing language learning. Work in other branches of linguistics is more sporadic, especially in the area of synchronous CMC, i.e. simultaneous interactive CMC of which chat systems are a part. Studies in this area are mostly qualitative and based on the analysis of small databases of individual log files.

The present study is designed to address this imbalance in linguistic research on synchronous CMC. The project encompasses the creation and analysis of a large scale corpus of Internet Relay Chat, which continues to be the most popular chat system. The creation of the corpus can serve as a model for later studies on synchronous CMC which choose a quantitative

³ Source: <http://statistix.xs4all.nl/public/irc/ircnet.html>

approach based on a large linguistic database. The analysis of the corpus will focus on linguistic characteristics of IRC which have been identified by prior research. This research comes from a variety of disciplines and includes work on other types of synchronous CMC, as well as my own pilot study (Doell 1997). The goal of the present study is to verify the claims made by this prior research about lexical, syntactic and pragmatic conventions of IRC and arrive at a more detailed description thereof. Special emphasis in the analysis will be placed on the features of spoken language which many researchers have found for this type of written interaction.

The thesis is structured as follows: the chapter following this introductory section includes a short technical introduction to computer-mediated communication systems and a review of the research pertinent to the study at hand. Chapter three describes the study's methodology and the creation of the corpus of Internet Relay Chat. This is followed by a chapter describing the corpus' basic statistics and their analysis. Chapter five focusses on the analysis of linguistic features in IRC which also occur in speech and writing. Chapter six reports the results of the analysis of IRC features which are associated exclusively with speech and chapter seven treats IRC-specific innovations. Finally, chapter eight contains implications for future research as well as concluding remarks.

CHAPTER 2

RESEARCH REVIEW

Technical introduction

The following overview of research on CMC serves the purpose of establishing the traditional paradigms found in the field. A special focus will be placed on characteristics of CMC which have already been identified in the research and which will be investigated further in this study. The technologies used for CMC are as varied as can be expected in the non-standardized world of computers. Therefore an introduction in the terminology and technology of CMC in general and IRC in particular is imperative at this point.

Usually, CMC is grouped into synchronous and asynchronous forms. Asynchronous CMC is characterized by the provision to store messages. Thus the communicating parties do not have to be on-line simultaneously in order to receive each other's messages and communicate. In fact, considerable amounts of time can pass in-between communicative turns. Typical examples of this type of CMC are e-mail and newsgroups.

Synchronous CMC, in contrast, is ephemeral and requires simultaneous on-line presence of all participants. IRC and Unix "talk" are examples of this type. However, as in other areas of networking technology, there is a strong trend towards convergence and that trend also applies to these two forms. More and more synchronous CMC technologies have integrated

features for asynchronous communication. MUDs (i.e. Multi-User Dungeons) and MOOs (i.e. MUD Object Oriented), for example, are usually used for synchronous interaction but they also have MUDmail which is an e-mail feature for all users of a particular MUD. Similarly, ICQ and other messaging systems of its kind were developed for synchronous communication but offer the opportunity to leave messages for users who are off-line.

Another distinction can be drawn between one-way and two-way CMC. The former refers to systems which transmit the message whenever the user hits enter or a send button. The latter transmits whatever is typed on the participants' terminals one keystroke at a time. All asynchronous forms of CMC naturally fall into the one-way category whereas synchronous CMC programs can use either technology. IRC and MOOs are one-way forms, UNIX "talk" and VAX "phone", on the other hand, are two-way systems (Herring 1999).

CMC as an object of research is a fairly recent development starting in the mid-1980s. Since then, the field, like the technology of CMC, has experienced considerable expansion. Researchers from a variety of disciplines such as philosophy (Cicognani 1998), sociology (Bays 1998, 1999), ethnography (Murray), engineering (Wilkins 1991), history (Reid 1991), communication (Bechar-Israeli 1999), and linguistics (Lundstrom 1995, Werry 1996) have examined this new form of communication. Some researchers have used experimental methodologies (Ferrara et al. 1991, Condon & Cech 1996) and e-mail questionnaires (Lundstrom 1995), whereas others have chosen naturalistic observation (Bays 1999, Herring 1999, Werry 1996).

The bulk of the research thus far has focused on asynchronous CMC, probably because it is much more common and more easily accessible for analysis than synchronous CMC. Due to

the wide range of CMC technologies available for study, research on IRC is both relatively rare and comes from a multitude of disciplines with a large variety of methodological approaches. In order to overcome these drawbacks the research overview will also include relevant work on similar types of synchronous CMC and work on IRC that was done in disciplines other than linguistics, such as sociology and ethnography. The linguistic studies of IRC, including my own pilot projects, will be covered last in the overview.

In order to fully understand the issues covered in these studies it is necessary to introduce some of the technological aspects of IRC first. IRC is a computer program originally written in 1988 by Jarrko Oikarinen as an enhanced version of existing Chat-programs like the Multi-User-Talk or MUT for short. The program allows the creation of hundreds of channels, which are a kind of virtual meeting room, on a server or a network of servers. In these channels people from all over the world who are logged onto the server with a so-called client program can meet and communicate in groups or privately.

Figure 1: A schematic drawing of an IRC network

Newer versions of the program also allow participants to exchange data-files over the Internet. Some users of a channel, the operators or ops for short, have access to special

commands. They can “kick”, that is remove, and ban participants from the channel and modify the channel properties, for example make it impossible for non-ops to change the channel topic. In IRC, it is a convention not to use one’s real name. Instead, every participant has one or more nicknames. Operators have an “@” symbol in front of their nicknames to signal their special status. When one joins a channel, two windows are displayed on the computer screen: a small one listing the nicknames of all the people who are currently in that channel, and a bigger one which displays the messages that the participants send to the IRC-server. This is where the group communication takes place. By clicking on a nickname in the smaller window one can open a new channel with that person. This is where private communications are held. The process of communication goes as follows: the participant types up the statement she wants to make on her computer. As soon as she hits enter, the message is sent to the IRC-server which then multiplies it and sends it to all the people currently present on the channel. This usually takes only a few seconds. Thus everybody is able to read what everybody else is saying on the channel. A communicative turn can extend over several messages because people hit enter frequently, thus splitting long sentences or thoughts. This allows the other participants can follow the conversation, and also the composition process more easily and respond faster. My turn/message distinction follows the terminology of Denise Murray (1991a).

A phenomenon that is widespread in IRC yet non-existent in traditional genres are the so-called emoticons. According to the Hypertext Webster’s Dictionary’s 1996 jargon file, an emoticon is “[a]n ASCII glyph used to indicate an emotional state in e-mail or news.” It appears that emoticons were first introduced by Scott Fahlman around 1980. They have since come into common use in CMC when one wants to convey information such as irony, surprise, or doubt

which, in speech, would be signaled by kinesic cues or suprasegmental features. In order to infer their meaning, one must tilt one's head 90 degrees (usually to the left) when looking at them and imagine that they symbolize a small face. Here are some examples: :-) ;-) >:(The first of the emoticons above would thus be read as 'a smiling face' the second as 'a winking face' the third as a frowning face and so forth. New kinds of emoticons are constantly created, even though their meanings sometimes remain opaque to anyone but the inventor.

Another particularity in IRC communication is the so-called "action command". When a participant wants to communicate some activity such as "I am laughing my head off" to the channel, she can use this command. She would type: "/me is laughing her head off." Her message would then be preceded by an asterisk followed by the her nickname, and, depending on the software parameters, might be displayed in a different colour. The example above would be converted into an action message which would look somewhat like this: *nickname is laughing her head off.

When a message is preceded by three asterisks it is a system message, which is automatically generated by the IRC program. These messages report, among other things, people changing their nicknames, people joining, leaving or being kicked off the channel. An example would be: *** nickname(email@address.de) has joined #cyberchat.

Any IRC user can record all the conversations they participate in or watch by using the IRC client's log function. This function saves everything that scrolls across the participant's screen in an ASCII text file on the computer's hard drive. Such log files serve as observation data for many of the studies on CMC and IRC which are summarized in the following research overview.

Ferrara et al. (1991)

In their groundbreaking article Ferrara et al. describe an early study on synchronous CMC. The technology they examine is different from IRC in that it does not allow for one-to-many conversations nor does it have the IRC-client programs extended command functions. Furthermore, Ferrara et al. do not observe naturally occurring interaction but use an experimental setup in which a controlled group of subjects engage in what they call Interactive Written Discourse or IWD with a single recipient on a single topic. The subjects are 23 computer specialists and their spouses, the topic is travel arrangements and the audience is a travel advisor. During the observed 46 hours of computer conversation Ferrara et al. gather a corpus of about 18,800 running words on which they base their analysis.

They report three main results of this analysis. Firstly, they conclude that IWD is a reduced register because their corpus contains forms of linguistic reduction such as:

- omission of (unstressed) pronouns, especially subject pronouns
- omission of articles, both the definite and the indefinite
- omission of finite forms of the copula
- shortening of words through use of abbreviations and symbols

Ferrara et al. suggest this reduction might be due to real-time production constraints or to an impression that the recipient of the message (like a baby or a foreigner) is somehow not fully competent in English .

Ferrara et al.'s second finding is the hybrid status of the IWD register between spoken and written language. Using results from Biber's (1988) work⁴ they identify several features typical of spoken language in IWD. These are the presence of first and second person pronouns, wh-questions and yes/no questions. Additionally, they note the occurrence of informal spelling and colloquial usage. However, characteristic structures of written language are also present in Ferrara et al.'s corpus. They observe "frequent use of relative clauses, adverbial clauses, and subordination." (Ferrara et al. 1991, p. 24) and items of cataphoric reference besides. They conclude that IWD is a hybrid register which is at the same time interactive and edited. This conclusion poses a challenge to Biber's approach because interactive and edited are the opposite ends of one of his dimensions of register variation.

The third finding Ferrara et al. report is the state of flux in IWD conventions. They notice that their subjects are not consistent in their use of capitalization at the beginning of messages and for proper nouns. Additionally, they find a high degree of individual variation, for example in the mean length of messages. Ferrara et al. attribute these features of their corpus to the status of IWD as an emergent register, i.e. the norms of IWD have not been conventionalized to a degree where all members of the discourse community display more homogenous linguistic behaviour.

⁴ In Biber's approach a large number of linguistic features are extracted from a grammatically tagged corpus and then counted for each genre or subcorpus in question. Then a factorial analysis is used to determine which features tend to co-occur or are more or less mutually exclusive. The resulting feature clusters are called dimensions and given labels such as 'involved vs. informational production', and 'narrative vs. non-narrative concerns' depending on the functions of the linguistic features included. The texts which are being analysed can then be localized along these dimensions according to their scores for that particular factor. Thus, face-to-face conversations, for example, find themselves at the 'involved' end of Biber's first dimension whereas "official" documents are at the 'informational' end.

Based on their findings, Ferrara et al. come to the conclusion that IWD is a rewarding subject of study for researchers interested in language variation and change. As an emergent register with unstable conventions, the features of IWD could provide insight into the process of register formation and acquisition. Ferrara et al. advance the hypothesis that IWD inherits most of its characteristics from registers such as postcardese, headlinese, telegraphese, and note taking. Also, according to Ferrara et al., IWD's position between typical written and spoken genres in combination with the claim that "this type of written communication will proliferate in the future" (p. 30) makes it likely that it will exert a strong influence on traditional writing styles.

Murray, D. E. (1991a)

Murray's article is an ethnographic account of the linguistic behaviour of a senior employee at a large computer company. Murray employs a participant-observer methodology to describe the patterns of communication in a computerized workplace. Her work focuses predominantly on her subject's choice of medium for specific communicative purposes.

However, it also includes a description of linguistic features found in the participants e-messages, i.e. when they employed the medium of synchronous CMC. The spoken characteristics Murray mentions as occurring in CMC are: active voice, personal pronouns, informal diction, hedging, vagueness, paralinguistic cues, quotations, contractions, ellipsis, openings, closings, and turn-taking. The last three items pertain to the organization of the discourse and are, according to Murray, realized differently than in spoken language. Murray states that turn-taking is more complex, adjacency pairs are missing, and openings and closings

are often completely absent. She also mentions features typical of written language such as technical terminology, the pronoun 'one', nominalisation, attributive adjectives, complement clauses, and relative clauses. However, Murray fails to provide the statistics on which she bases these conclusions.

Lundstrom, P. (1995)

Lundstrom's paper evaluates the possible use of Internet Talkers, a MUD-like environment used for synchronous CMC, in teaching English as a foreign or second language. He comes to the conclusion that integration of this particular text-only CMC program into an EFL/ESL curriculum is not advisable because the nature of the language encountered is nothing like standard English. In this context, his use of the term communicative competence (Hymes 1971), which he interprets as being proficient in the standard English variety that is the alleged target language in ESL/EFL courses, is problematic. Contrary to Lundstrom's interpretation of it, Hymes' communicative competence is an umbrella term which covers all aspects of language knowledge and use including non-standard usage. Lundstrom's conclusions are based on the analysis of an e-mail questionnaire and of a log taken during a 24 hour period on a talker. The corpus text contains around 80,000 words, roughly two thirds of which are computer messages and the other third is user generated.

Lundstrom groups the linguistic characteristics he finds under the following headings: reconstruction of affect, disambiguation, and speed. Under the first heading Lundstrom lists the

devices used to convey emotional content. He finds that action commands, emoticons⁵, creative capitalization, multiple exclamation and question marks, reduplicated letters, asterisks, brackets, and vocalizations all serve this purpose.

Under the second heading, Lundstrom claims that the poverty of context in the Talker causes participants to use a relatively narrow range of vocabulary in order to avoid ambiguity. Unfortunately, Lundstrom does not elaborate his argument and it is not clear how a reduction in vocabulary would disambiguate the communicative intent of the speakers. If anything, one would expect an increase in ambiguity when a limited vocabulary is used to express semantically complex ideas. In the context of this disambiguation argument, Lundstrom reports that a small number of verbs accounted for a large number of messages. This finding is somewhat problematic because the actual frequencies are not reported and such a distribution is quite normal for any but the most technical texts. Lundstrom also mentions high frequencies of the following items:

- first person singular, second person singular and plural pronouns
- the indefinite pronouns <anyone> and <anybody>
- the demonstrative pronouns <that> and <this>
- the possessives <me>, <my>, and <your>
- the articles <a> and <the>
- wh-questions with <what>, <why>, <how>, <where>
- the modal <can>
- the conjunctions <and>, <but>, and <or>

⁵ See the technical introduction for a description of emoticons

- the prepositions <to>, <in>, and <from>

Again Lundstrom does not supply the frequencies for the words listed.

Under the third heading of speed Lundstrom treats several features that he attributes to the need to communicate with as few keystrokes as possible. Firstly, there are numerous abbreviations in his corpus. Secondly, he observes the use of macros, short computer programs which display a preprogrammed sentence or expression when a special combination of keys is hit. Thirdly, he finds a tendency for simplification on a syntactic level through omission of auxiliary, determiner, and subject. He also finds this trend at the level of spelling and punctuation exemplified by frequent typos, sentence initial lower case, omission of sentence final period and generally a disregard for conventional spelling. Finally, a very useful part of Lundstrom's study are the appendices. They provide the detailed results of Lundstrom's e-mail questionnaire and additionally lists of nicknames, abbreviations, and emoticons.

Collot, M. & Belmore, N. (1996)

In their study Collot & Belmore apply Douglas Biber's Multi-Feature/Multi-Dimensional or MF/MD⁶ approach to their 100,000 word corpus of "electronic language" as it is used in an Internet usegroup. In the field of CMC this study is unique because of the large range of linguistic features that it includes in the analysis. However, the statistical data for these features are then collapsed into six dimensions by the factorial analysis of Biber's approach so that they cannot serve as a basis of comparison between my corpus of synchronous CMC and

⁶ See footnote 3 for details of Biber's MD/MF approach.

Collot & Belmore's corpus of asynchronous CMC. They mention that their corpus contains an extraordinarily high frequency of 1st and 2nd person pronouns, hedging, contractions, and amplifiers, features which have also been found by other researchers for other forms of CMC (Lundstrom 1995, Ferrara et al. 1991).

Wilkins, H. (1991)

Wilkins studies 377 entries in a publicly accessible 'topic', which is an asynchronous computer messaging system similar to a newsgroup but with the capability for synchronous conversations. Thus her data contains a mixture of synchronous and asynchronous material. She addresses the question of whether or not "this traditionally oral activity - interactive discourse - now in graphic form" (Wilkins 1991:57) would inherit its conventions from the written or spoken end of the discourse spectrum.

She finds that turn-taking mechanisms that have been identified in oral conversation were not applied in her data. Instead, the means for signaling a response to a previous topic and the person it is addressed to are "taken from the conventions of written language in informal letters and memos." (Wilkins 1991:63) Examples of these means are the abbreviation *re:*, the identification of the addressee by full name, and the number of the message which is responded to.

Next, Wilkins discusses the means for topic maintenance she finds in her corpus. She states she "had little difficulty following, reading and correctly interpreting a log of computer conversation in which [she] did not participate." (Wilkins 1991:64) She attributes this in part to

frequent lexical repetition as a means of creating textual cohesion. In contrast, oral discourse relies more heavily on pronouns to the same effect. As a second factor she mentions the large degree of shared cultural knowledge between the discussion's participants who were all North-American and mostly associated with the Presbyterian Church.

Wilkins finds several other features typically associated with spoken discourse in her corpus. In 11% of the entries participants explicitly stated paralinguistic features. These were phrased as direct statements of physical actions such as grinning or ducking, and put in brackets like stage directions. An exemplary quote would be: "we consider trying to make Unison a church for the tax advantages but (sighs) we didn't think the IRS would buy it" (Wilkins 1991:66).

Furthermore Wilkins notes a high degree of involvement in her corpus which had been found characteristic of oral texts by Biber. Ego involvement is represented by the high incidence of first person pronouns in 63% of the entries in the corpus. Involvement with the audience was expressed by the use of second person pronouns or names in 56% and the occurrence of questions in 21% of the entries. Additionally, devices signaling lively interest such as exaggeration, exclamation, expressive vocabulary and vivid particles were also present in Wilkins' corpus.

Finally, Wilkins mentions two more features associated with oral discourse that occur in her data: disfluencies and innovative use of language. The first refers to phenomena such as hesitations, false starts, and the adding of afterthoughts. Incidentally, Wilkins' examples for these also display signaling of prosodic features through use of capitalization and punctuation. As an example of innovation in language she cites a play on words in her corpus.

Wilkins concludes that her sample shares more characteristic features with traditional spoken/unplanned/involved discourse than with written/edited/informational texts. She advances the hypothesis that this oral character contributes to the strong feeling of community that the participants reported in her study. Possibly, this is enhanced by the fact that the informal, oral features occur in the written medium.

Bechar-Israeli, H. (1995)

Bechar-Israeli's study examines the use of nicknames or 'nicks' on IRC. She raises several issues about the importance of nicknames for the construction of identity in the text-only medium of IRC and establishes a typology of nicknames. The database for her categorization is 231 nicknames she collected from her logs of IRC conversations on two channels. First she introduces 16 groups of nicks, which she then refines to the following seven:

- people using their real name
- self-related names
- names related to medium, technology and their nature
- names of flora, fauna, and objects
- play on words and sounds
- names related to figures in film, literature, fairy tales and famous people
- names related to sex and provocation

The vast majority of nicknames she examines, namely 45%, fall into the category of self-related names. Names of flora, fauna, and objects come second with only 16% and sex &

provocation, with 3%, is the smallest category. Bechar-Israeli argues that nicknames in IRC constitute a worthy object of study because they represent the online personality of users condensed into one word. Therefore users apply considerable care, creativity and innovation when they establish their nicknames.

Bays, H. (1999)

Bays uses a socio-cognitive approach in her examination of framing and face on IRC. She finds that due to the lack of a situational frame based on physical copresence, people on IRC must construct a frame of their own. The framing on IRC is modeled after the community metaphor and the concept of a physical room. Bays observes an occasional overcompensation for the lack of physicality. This is expressed in exaggerated descriptions of actions and events that do not occur in real life.

Face in Bays' study is defined as the positive social value a person places on herself. Bays finds that people in IRC, as in spoken conversations, do a lot of facework. That is, participants of IRC carefully create and shape their online identities. Also, a set of social conventions has been established and violations of these norms are reprimanded. As examples of the above, Bays elaborates on the creativity of IRC nicknames, the consistency of the online personas' talk, the repair or reproof of face threatening acts and the distribution of virtual gifts.

Lunsford, W. (1996)

In his study of turn-taking organization in IRC, Lunsford utilizes the model proposed by Sacks, Schegloff and Jefferson in 1974 and applies it to IRC. His conclusions are based on the analysis of the logs of four 15 minute IRC conversations. Lunsford finds that although turn-taking occurs, i.e. speakers alternate, IRC technology imposes specific adaptations on the mechanisms at work. For example, due to the fact that the IRC-server automatically displays the messages in the order it receives them, there is neither speaker selection nor overlap in the conversation. Therefore repair mechanisms do not exist since turn-taking errors cannot occur and all forms overlap are manifested as separate turns. Furthermore, the technology enables the user to participate in multiple simultaneous conversations. Lunsford notes that messages tend to be short, that screen names are used to direct utterances to their addressee, and that gaps regularly occur between turns and between messages and responses. Lunsford identifies miscommunication due to physical distance and backchanneling as issues for further research.

Werry, C.C. (1996):

Werry's is the only study which addresses linguistic and interactive features of IRC specifically. It is based on logs of two ten minute conversations in two channels. After a description of some general characteristics of IRC conversation, such as the absence of turn-taking mechanisms and overlap, Werry examines conventions of addressivity. He finds that the intended addressee's name is often put at the beginning of a message. He interprets this convention as an attempt to avoid ambiguity. However, he also notes that the convention is

dropped when addressing the entire population of the room or when the intended addressee is made sufficiently clear by the context.

Next, Werry treats abbreviation phenomena which according to him result from factors such as screen size, average typing speed, minimal response times, competition for attention, channel population, and the pace of channel conversations. All these in conjunction lead to the observed abbreviations. For example, the average length of message in Werry's corpus is around six words. Furthermore, Werry mentions a frequent use of syntactically reduced forms, acronyms, symbols, and clipping of words. As evidence for the first he reports a high incidence of subject pronoun deletion and reduction, particularly for first and second person pronouns. For the three others he quotes examples of commonly used Internet acronyms such as 'lol'⁷, emoticons, and the abbreviation of participants' names and words to the least number of letters possible. Yet another way of compensating for the medium's restrictions is the use of macros, that is, phrases that are stored and can be displayed instantly by pressing a succession of certain preprogrammed keys.

Furthermore, Werry examines the imitation of intonation and paralinguistic cues through creative use of capitalization, punctuation, and spelling. Capitalization most often signals stress and volume, periods and hyphens signal pauses, and reduplication of letters signals drawn-out or expressive intonation. In this context Werry also notes that IRC tends to be highly colloquial. He states: "[c]olloquial verbalizations and non-standard spelling appear to be self-consciously selected in preference to 'standard' linguistic expressions." (Werry 1996:57)

⁷Lol = Laughing out loud

Lastly, Werry explores the IRC-conventions by which participants use words and images to symbolize gestural aspects of face-to-face interaction. Such descriptions of physical activities are usually preceded by an asterisk in IRC. Werry notes that users have different preferences in the two channels he observes as to which activities are thus enacted. Werry also mentions special programs that allow users to display graphical images consisting of ASCII characters and other software, the so called 'bots', these pose as IRC participants and can be prompted to participate in the conversation by, for example, quoting famous writers or entries from a dictionary.

Werry concludes that the linguistic and interactive features he encountered function to compensate for and adapt to the constraints of the medium and that "communication on IRC is shaped at many different levels by the drive to reproduce or simulate the discursive style of face-to-face spoken language." (Werry 1996:61)

Paolillo, J. (1999)

Paolillo studies the speech community of an IRC-channel called #india. Based on the log of a 24 hour period on that channel he conducts a social network analysis. This analysis was developed by sociolinguists and it correlates linguistic variables with network ties. Network ties is a concept which designates the relative closeness of people to each other. Frequency of contact is one way to differentiate between strong and weak social network ties. Paolillo points out that, unlike in real life communities, IRC offers the researcher the unique possibility to exactly quantify and measure the contacts the participants have with each other.

The speaker and the addressee can usually be identified very easily in IRC. This was done for Paolillo's corpus and over 6,000 lines of text were then entered into a database accordingly. A statistical program was used to establish the social network ties. Finally, the following five linguistic variables were assessed and correlated with the social network database:

- use of Hindi
- use of 'u' for 'you'
- use of 'r' for 'are'
- use of 'z' for 's', especially in word final position
- use of obscenities

Paolillo finds that use of Hindi is the in-group variable for the most central group of participants with the strongest network ties. The use of obscenity seems more popular in groups that contain many operators, probably due to the fact that this practice would result in the removal from the channel for normal participants. The other linguistic variables are distributed in other locales of the social network and do not correlate with any particular group of speakers. Paolillo concludes that linguistic variables in IRC do not map onto social networks in a simple and straightforward way. A major shortcoming of Paolillo's study is the failure to provide a reason for the choice of the variables which are examined. Three of them are IRC-specific spellings, one is a code switch, and no working definition of what the researchers count as obscenity is given in the article.

Doell, W. (1997)

My 1997 examination of German and English IRC channels is one of the pilot studies for the present work. The bases for my analysis were logs from a German speaking, an English speaking and a German/English bilingual channel. The corpus contained approximately 140,000 words. A detailed analysis of three 50 message passages from each channel was combined with a quantitative analysis of the entire log files using the 'wconcord' concordancing software.

For the examination of the passages there were four levels of analysis: lexis, orthography, grammar, and discourse. The following table summarizes the findings in these four realms:

Lexis	emoticons, onomatopoeia, abbreviations
Orthography	non-standard spelling, capitalization & punctuation signal prosody, phonological spelling
Grammar	omissions, contractions, no subordination or passives
Discourse	no turn-taking rules, no interruptions, addressivity, greeting routines

Onomatopoeia in the above table refer to attempts at making paralinguistic sounds graphically visible. An example would be the use of 'hehe' to symbolize laughter. Phonological spelling is attempts to write things the way they are pronounced especially in a dialect variety. The mean length of the messages in the analyzed sample ranged between 3.2 and 4.3 words. These observations largely concur with similar studies on synchronous CMC. Some differences between the German and English channels were observed but they were mostly limited to the lexical level. I concluded that IRC is a register that is quite similar to spontaneous conversation

and that the use of IRC as data is cheap and easy to implement. Moreover, the study of IRC-talk offers a unique possibility to observe communicative acts in their entirety.

Doell, W. (1999)

In the preliminary phase of the present study, an initial pilot corpus of approximately 100,000 running words was created from 16 IRC log files. These logs were recorded from six channels simultaneously over a period of four days. Several exemplary wconcord word searches were run on the pilot corpus. It was determined that a corpus of that size would be too small to meaningfully study linguistic phenomena, especially at the syntactic level of analysis. Furthermore, in order to shorten the period of data collection and to diversify the data sources, it was decided to include four more IRC-channels.

Then, using the same procedure, a second pilot corpus of about 600,000 running words from ten different channels was established. However, the differences in file size for the channel logs proved to be a major hindrance in the statistical analysis of the corpus data. Crucial measurements such as type/token ratios are dependent on the sample length. To render the comparison of subcorpora more practical, a standardized sample text length had to be instituted. Up to that point the collection of data from all the channels had always been conducted simultaneously because it was thought that the time of recording might be a significant variable. However, since the amount of data that can be collected on the different channels through a certain time varies greatly, this approach had to be altered in order to arrive at a standardized sample length. Prior to the change in collection method, it was ascertained that the time when

the data is logged does not affect the linguistic makeup of the sample. This was done by running a series of t-tests on the pilot corpus files using a size-independent variable, namely the number of words per message. The results of these tests showed no significant differences for the data not recorded at the same time and on the same day. Thus it was clear that the simultaneous data collection could be safely abandoned. The final method of data collection will be discussed in detail in the methodology section after the following brief summary of the research discussed.

Summary

Research on IRC is rather scarce. This is in part due to the relative novelty of the technology and in part to a general tendency in the field to study CMC in its asynchronous forms. Moreover, much of the research on CMC and IRC is done in disciplines other than linguistics. The bulk of studies adopt sociological and ethnographical approaches and study aspects of community building, construction of on-line identity and other processes of social interaction (for example: Reid 1991, Bays 1998). Of the papers discussed above, only three beside my own pilot studies specifically examine linguistic aspects of IRC. Lunsford (1996) studies turn-taking mechanisms, Paolillo (1999) focuses on language variation and Werry (1996) gives an overview of the distinguishing linguistic and interactional features of IRC supported by examples from his personal logs. Their studies, combined with research on other forms of synchronous CMC (Murray 1991a, Ferrara et al. 1991, Wilkins 1991, Lundstrom 1995, Collot & Belmore 1996), make it possible to identify several common fields of interest: use of pronouns or, more generally put, use of deictic items; use of features that occur frequently in casual speech; creative innovations such as abbreviations or non-standard spelling; features inherent in the technology

such as special turn-taking conventions, nicknames or the use of action messages. It is these areas identified in prior research which will constitute the main focus of the analysis of the CIRC.

CHAPTER 3

METHODOLOGY

Data collection and corpus organization

The log function of the client software mIRC version 5.41 was used for the collection of language data for the corpus of IRC. The researcher joined the selected IRC-channels and recorded the ongoing interactions without participating himself. In IRC, being present without taking part in the chat is called 'lurking'. Therefore I dubbed this method of data collection "lurking & logging". Conversations were logged in 10 different channels on 5 different IRC-networks. Somewhat less than half the log files were taken from the second pilot corpus which was recorded in May 1999. The rest, especially for the channels with less traffic overall and those with peak hours after midnight, was logged during September 1999. As described above and in the research review, the CIRC is the product of a process involving several pilot corpora and studies. This procedure of gathering a corpus, then analysing it, then collecting a more refined corpus, subsequent analysis of this second corpus and so on and so forth is called cyclical corpus design. It is strongly advocated by Biber (1993) in order to ensure exact sampling and a high degree of representativeness.

Using the word count feature in the 'concordancer for windows v.3.0' software, the log files were then split into the corpus files containing approximately 5,000 words each. Note that

for the computer software “word” was defined mechanistically as any string of letters between two spaces. Due to the creative use of punctuation on IRC it was necessary to disable punctuation as a word boundary indicator in the Wconcord program . The size of 5,000 words for each corpus file was chosen for several reasons. The designs of the three classic English language corpora, the London-Oslo-Bergen (LOB), the Brown, and the London Lund Corpus (LLC) all adopted sample text size limits of several thousand words. The first two are corpora of written English and have a 2,000 word limit, and the third is a corpus of spoken English and has a 5,000 word limit. It has been shown that larger sample texts do not yield significantly more accurate results (Biber 1993). Since the CIRC would be inflated by some amount of dirty data, e.g. computer generated messages, and because previous research indicated that IRC was quite similar to speech, it was decided to follow the example of the London Lund Corpus.

Consequently, the CIRC contains twenty 5,000-word files for each of the ten logged channels. This yields a total corpus size of slightly more than one million running words. Again, this follows the tradition of the three classic corpora mentioned above which all adopt the exact same size. Each of the 200 CIRC files was assigned a coded name which contains the IRC-network, the channel name, the original log file number, and the passage’s location in the original log file. Following a standard procedure in corpus linguistics the data was stored exactly in the form it was collected, i.e.. as ASCII text files. The researcher did not add, exclude or alter any of the linguistic material in the data. This clean-policy (Sinclair 1991:21) is adopted for two reasons. Firstly, this is done in order to facilitate accessibility to corpora which might have been designed for a single research purpose but can be used in other projects with different priorities and perspectives if the original data remains untouched. Secondly, there is no standardized way

of coding corpus texts yet. Therefore it is vital that any alterations made to the original data are reversible should the need arise to transfer the corpus into standard code.

In order to minimize the collection of dirty, i.e. computer generated data, most system messages were suppressed during the logging process. Exceptions were system messages which are instigated directly by the users and which contribute essential clues for the study of IRC interaction. These are: nicknames changes, kicks, and the changing of the channel topic. Incidentally, the first two of these three are impossible to remove with the mIRC client software anyway. This highlights their importance in helping participants interpret IRC discourse.

Channel selection

There is a large variety of IRC-channels, often more than ten thousand per server on the larger networks, and choosing a sample of ten which was to be representative required the development of some general criteria. First of all, the channels which were logged were all large by IRC-standards with at least 25 participants at any time of day. This was done in order to ensure a speedy data collection and also to avoid getting too much notice as an outsider. Although the large majority of IRC-channels is publicly accessible and strangers come and go at all times, the smaller the group of participants, the more likely they are to notice and negatively sanction a non-participating intruder. Being kicked off the channel constantly would have rendered the data collection process, especially the goal of at least 5,000 words per session, extremely difficult.

Another difficulty in the selection process was the decision of which kind of channel to log. Out of the multitude of channels, two main categories can be distinguished. There are

channels which focus on social interaction and others which are used mainly for the exchange of data or information. In the former category, one would find channels named after social activities such as #cyberchat, #party, #friends or #cybercafe, channel names which have a national, regional or local significance, such as #uk or #london, channel names which indicate that they are frequented by a certain audience such as #over30 or #ircnewbies and a plethora of others named after their founder, a hobby, stars, TV series etc.

In the latter category you would find channels where people trade software, mp3 music files, or pictures and also channels in which certain information can be obtained, such as channel #help, #linux, or #mIRCScripts. It is important to keep in mind the fact that an absolute control of the kind of interaction which is being recorded is not possible. Although the permanent channel names declare a channel's general topic or characteristic, nothing can stop participants from doing something entirely different. A channel with a local focus such as #berlin might very well be used to trade software for example. Nevertheless, the channel name is usually a fairly good indication of the kind of interaction that is occurring in the chat-room.

Channels which were obviously used for trading were not included in the corpus because the linguistic interaction in these channels is minimal. Furthermore, logging and analyzing these sometimes illegal activities would be unethical. Since the main focus of the present study is casual conversation, four channels with general chat names were selected. These were: #cheers, #cyberchat, #beginners, and #chat-world. #beginners is a channel for newcomers to IRC where they can try out the new medium in a friendly and helpful atmosphere. It was included as a chat-channel but it is interesting in its own right to examine how newcomers respond and adapt to the new communicative context they enter.

Since one of the pilot studies (Doell 1997) and some prior research (Werry 1996) had investigated channels with geographic channel names and found significant variation among them, it was decided to further pursue this research interest. Naturally, the channels selected for a corpus of IRC-English had to be English speaking. What is more, in order to fully explore the extent of English language variation in IRC, the two national channels selected were neither British nor US-American but channels #canada, and #australia. Unfortunately, the South African channels all failed to reach the critical mass of participants necessary for the study. Furthermore, two channels named after urban centers in Australia were chosen in order to also capture linguistic variation at a more local level. These channels were #melbourne and #perth. Finally, in order to achieve a well-rounded corpus, two more channels whose topics indicated more of a focus on the exchange of information and ideas were selected. These were #help and #philosophy. The former is a channel where a group of regulars offers advice to anyone who cares to ask. Naturally, most of the information exchanges in this channel center around computer technology in general and IRC in particular. The latter is a virtual meeting place for people inclined to discuss philosophical ideas in a more or less serious manner. However, it is not an academic forum. Both channels are also used, albeit to a lesser degree than the others, for social interaction.

Privacy

It has been pointed out (Reid 1991) that although IRC-users use nicknames, that does not lead to real anonymity on the chat network. Regulars are known in their respective communities and tend to keep their nickname/s. This situation has been termed pseudonymy.

People are well known under their pseudonym and thus can be identified by their online communication partners. In order to protect IRC-users' privacy, the policy in this study is to alter all pseudonyms and not to disclose on which IRC-networks the channels were recorded. The channel names themselves will be used since they provide important clues as to the conversational content and the group makeup of the respective IRC-communities.

Population

A detailed description of the IRC-user population is somewhat difficult to establish due to the fact that users usually remain anonymous on the network. However, since IRC-users form a subgroup of the Internet population, it can safely be assumed that they share a lot of characteristics with the latter. Despite falling computer and connection prices, the Internet is still largely an affair of the well-to-do, educated upper and middle class⁸ white male and dominated by Anglo-American culture and the English language⁹. That the same holds true for IRC is supported by the few available user surveys for IRC and similar synchronous CMC media. For example, an online survey on IRC use conducted by Mary Lynn Hyman found that all the respondents had at least some college education¹⁰. An article by Kai Seidler on 'Relay Parties' in Germany mentions that only an estimated 20% of IRC-users are women¹¹ and in Phoenix

⁸ Source: American government survey at: <http://www.ntia.doc.gov/ntiahome/fttn99/>

⁹ 57% of people accessing the Internet are native speakers of English
(<http://www.euromktg.com/globstats/>)

¹⁰ See survey at: <http://home1.gte.net/fyre/NewPage7.htm>

¹¹ See article at: <http://oswald.pages.de/chat/rps/>

Lundstrom's (1995) survey on talkers 37% of respondents were female. More recent consumer surveys show that the males' lead in this area is certainly shrinking. The American Internet User Survey¹² puts the gender ratio of online adults for 1999 at 53% men vs. 47% women. The average age of users, according to Lundstrom and Hyman, lies around 25 years. The range of ages is probably considerable. One of the users in my corpus claims to have been on IRC since he was 12 and as there are also channels that are named 40+, 50+ etc., a significantly older population is indicated.

Study organization

The goal of this study is to evaluate existing claims about the nature of IRC talk and to achieve a more detailed description thereof. To this end the study will reexamine characteristics of IRC which have been identified in prior research using the CIRC. It will also focus on problem areas in the existing description of IRC talk and will explore some newly identified phenomena. The overall perspective will be shaped by the hypothesis that IRC language shares characteristics with the language of spontaneous, casual conversation. Therefore the features that are to be studied will be assigned to one of three categories. These will also be reflected in the division into chapters, namely chapters five, six and seven correspond to these three categories. The following table gives an overview of these categories and some examples.

¹² Source: http://www.cyberdialogue.com/free_data/index.html

<u>Comparison</u>	<u>Ersatz</u>	<u>Innovation/Creativity</u>
ellipsis	emoticons	neologisms
leave-taking	phonetic spelling	nickname formation
deictics	backchanneling	innovative spelling

The 'comparison' category will include the features for which a direct comparison between spoken and written language is possible. For example, the occurrence of ellipsis, deictics, and conventions of greeting. Due to the large variety of these features this part will make up the bulk of the detailed analysis. Features of IRC- talk which do not occur in traditional written genres but clearly mimic characteristics of casual conversation will be treated in the 'ersatz' category. The most important item in this category will be the use of emoticons in IRC which replace the kinesic clues of face-to-face interaction. The IRC-features which occur neither in traditional spoken nor written genres but which reflect the IRC-users' linguistic innovation and creativity will be included in the last category. Such playful alterations and innovations have been found to be characteristic for spoken language rather than written genres where they occur only rarely, for example in poetry or in informal letters between close friends. The distinctions made above are a heuristic tool and not absolute categories. This can be illustrated by a borderline case. Attempts at phonetic spelling can be seen as an 'ersatz' phenomenon when they are used to express in writing what would be suprasegmental information in speech for example. At the same time they may be interpreted as expressions of linguistic creativity. Where these borderline cases occur their special status will be mentioned.

CHAPTER 4

DESCRIPTIVE STATISTICS

Before a detailed analysis of the language material in the CIRC can be undertaken, it is vital to describe some fundamental characteristics of the data found in the corpus. Furthermore, a set of statistical tests will be run on some basic measurements extracted from the CIRC. This will yield more insight into the internal makeup of the corpus and provide some working hypotheses as to the amount and distribution of linguistic variation that will be encountered in the data.

Basics

As explained above, the CIRC consists of 200 text files - twenty for each of the ten channels selected for study. The files have a standardized length of approximately 5000 words (margin of error <1%). The average length of the files lies at 5004 words. This margin of error is the result of a standardization process which preserved the integrity of messages.

Six basic measures were taken for each file. These are:

- the number of words or tokens per file
- the number of types (multiple occurrences of a word or token are only counted as one) per file

- the number of standard messages (composed by IRC participants and preceded by their nickname in brackets) per file
- the number of action messages (composed by participants and preceded by an asterisk) per file
- the number of system messages (generated by the IRC client software) per file
- the number of hapactes legomena (words that occur only once) per file

From these six basic measures, four derived measures were calculated. Adding the figures for standard, action, and system messages yielded the total number of messages. This allowed for the calculation of the number of words per message which is the second derived measure. Additionally, the type/token ratio was calculated. The number of hapax legomena was divided by the number of words per file and then multiplied by one thousand in order to calculate the frequency of hapax legomena per thousand words. The following table shows the channel averages for the six basic and four derived measures¹³. Figure 2 displays the channels means for standard, action and system messages as a stacked 100% bar graph.

Table: averages for the ten channels:

Name	Words	Types	<*>	*	***	Hapax	Total M	W/M	T/T	Hap ‰
mel	5003.55	1329.75	641.4	64.65	106.15	759.15	812.2	6.1872	0.2657	151.72
per	5004.3	1448.25	732.35	59.85	66.85	871.7	859.05	5.9124	0.2894	174.19
can	5005.25	1399.3	694.45	40.35	61.85	829.15	796.65	6.3248	0.2795	165.66
che	5004.85	1360.1	639.8	56.45	38.8	790.3	735.05	6.8499	0.2717	157.91
cyb	5003.35	1365.6	787.5	73.15	28.5	798.4	889.15	5.7454	0.2729	159.57
hel	5004.15	1395.15	634.35	13.85	17.35	819	665.55	7.5489	0.2787	163.67
aus	5003.75	1433.35	861	40.55	28.7	883.5	930.25	5.4223	0.2864	176.57
phi	5004.15	1511.25	616.15	18.95	12.75	970.35	647.85	7.9751	0.302	193.91
beg	5005.1	1460.15	544.8	77.65	31.75	852.75	654.2	7.6696	0.2917	170.38

¹³See appendix A for the table of variables for the 200 CIRC files.

cha	5003.9	1404.5	749.55	73.1	44.1	823.25	866.75	5.7958	0.2806	164.52
-----	--------	--------	--------	------	------	--------	--------	--------	--------	--------

Legend: <*> = standard messages; * = action messages; *** = system messages; T/T = type/token ratio;

Total M = total number of messages; W/M = average number of words per message; Hap/M = Hapax/Mille =

Hapax Legomena per thousand words. Channels: mel = #melbourne; per = #perth; can = #canada; che = #cheers;

cyb = #cyberchat; hel = #help; aus = #australia; phi = #philosophy; beg = #beginners; cha = #chat-world

Figure 2: Channel means for standard, action and system messages

A brief look at these very basic statistics can already reveal several things. For example, both #help and #philosophy, the two informational channels, score very low for action and system messages and very high for the number of words per message. This might indicate linguistic similarities in the talk in these channels despite their distinctly different subject matters. The channel #beginners has a high average for the number of words per message, which might mean that newcomers to IRC compose longer messages than regulars. Lastly, clearly observable in figure two, the channel #melbourne has the highest number of system generated messages, which might be the cause for its low scores on the type/token and hapax/mille measures.

At this point, an explanation of these two measures is in order. The type/token ratio is the classical measure of lexical density in corpus linguistics. The rationale behind this concept is that words which occur frequently are either grammatical items such as 'the', and 'to', or they are words with very broad, general meanings such as 'thing' and 'way'. Words with very specific or even specialized meanings, such as 'prototype' and 'corpus', on the other hand, occur much more rarely in a large non-specialized collection of texts. These lexical items carry a bigger, i.e., more closely specified semantic load, which is why they are said to be more lexically dense. Similarly, a text which contains relatively more of these rare lexical items is lexically more dense. The type/token ratio is the tool with which corpus linguists measure the relative rarity of the words in their corpora. A ratio of one indicates that every word in the text occurs only once. For example, the sentence 'The car crashed' has a type/token ratio of one. In contrast, the more detailed sentence 'The car crashed into the tree' has a type/token ratio of 5/6 or 0.83.

The hapax/mille ratio works very similarly to and measures the same phenomenon as the type/token ratio. A sentence such as 'The horse kicked the stall' would have a hapax/mille ratio of $3/5 \times 1000 = 600$ whereas 'The black horse kicked the old stall' would rate higher with $5/7 \times 1000 = 714$. Of the two measures the type/token ratio is considered to be the more reliable one. I include the hapax/mille ratio in my study in order to double-check the results obtained using the type/token ratio. Naturally, both ratios are dependent on the length of the text to which they are applied. This is one of the reasons why the CIRC files were trimmed to a standard size of 5000 words. Otherwise, type/token ratios could not be compared meaningfully. For example, the type/token ratio for the entire one million words of the corpus is 0.08293. The corresponding

hapax/mille ratio is 44.8520. Both these figures are considerably lower than the averages obtained for the 5000 word files.

Correlations between variables

Pearson Correlation tests were run on some of the basic and derived measures for the CIRC's 200 files. Beside the obviously strong correlations between derived measures and the basic measures they were derived from, some interesting results can be reported. Predictably, since they both measure lexical density, a strong (0.678; $p < 0.01$) correlation was found to exist between the type/token ratio and the hapax/mille ratio. A channel by channel correlation test for the same two variables revealed that this strong correlation holds true within each of the channels as well. The tests also showed that there is no correlation between the number of words per message and the measures of lexical density for each corpus file. This demonstrates that longer messages do not necessarily mean that their content is lexically more complex.

The correlations between the number of system messages and the type/token ratio (-0.308; $p < 0.01$) and the hapax/mille ratio (-0.349; $p < 0.01$) attest to the distortion effect the computer generated messages have on these ratios. This is due to the fact that there exist only a small number of different computer messages. In the CIRC the bulk of system messages report either a person being kicked off the channel or a someone changing their nickname. A large number of these messages lead to frequent repetition of preprogrammed computer phrases such as: "<booh> has changed his nickname to <baah>" where only the nicknames in brackets have been entered by the IRC user. Naturally, the repeated occurrence of these phrases lowers the two

ratios which measure lexical density. Note that the correlation figure for type/token ratio and system messages is lower than the corresponding figure for the hapax/mille ratio. This corroborates the choice of type/token as the more reliable measure of the two.

100 most frequent words

One of the basic tools of corpus linguistics is word lists extracted from the corpus. They can be sorted alphabetically in order to enable quick retrieval of a word and its frequency or by order of occurrence if the word's position in the text is of importance for the researcher. Of special interest, however, are word lists in which the words are ranked by frequency. They are not only necessary for the determination of how many hapax legomena exist in the text but also provide important clues about the overall structure of the corpus and the language contained in it. For this purpose, researchers usually study the list of the 100 most frequent words in the corpus. This list is especially useful because it contains most of the high frequency items which form the lexico-grammatical core of the language and are common to all genres and media. Despite the limited range and relative crudeness of this tool, the list of the 100 most frequent words allows us nevertheless to identify several eminent tendencies in the CIRC, especially when compared to other corpora. Our analysis will begin with the rank list of the hundred most frequent words and then proceed to the frequency list of the hundred most frequent words.

Below is the rank list of the hundred most frequent words in the CIRC as extracted with the wconcord software. For purposes of comparison, the ranks for the listed words in the 100

most frequent words lists of the London Lund Corpus (LLC) of spoken language and the Brown and London Oslo Bergen (LOB) corpora of written English are included as well.

	CIRC	LLC	LOB	Brown
i	1	3	17	20
is	2	11	8	8
to	3	4	4	4
the	4	1	1	1
a	5	6	5	5
you	6	7	32	33
*	7	n/a	n/a	n/a
***	8	n/a	n/a	n/a
and	9	2	3	3
it	10	10	10	12
now	11	51	72	76
me	12	66	66	82
in	13	9	6	6
as	14	29	13	14
that	15	8	7	7
hi	16	n/a	n/a	n/a
lol	17	n/a	n/a	n/a
was	18	13	9	9
known	19	n/a	n/a	n/a
of	20	5	2	2
:)	21	n/a	n/a	n/a
u	22	n/a	n/a	n/a
for	23	20	11	11
are	24	42	27	24
on	25	16	16	16
what	26	37	55	54
have	27	19	26	28
hey	28	n/a	n/a	n/a
my	29	58	58	75
all	30	33	39	36
not	31	35	23	23
by	32	65	20	19
with	33	32	14	13
do	34	40	73	72
)	35	n/a	n/a	n/a
here	36	n/a	n/a	n/a
how	37	n/a	n/a	n/a
no	38	31	47	49
from	39	53	25	26
kicked	40	n/a	n/a	n/a
so	41	30	46	52
be	42	21	15	17
hello	43	n/a	n/a	n/a
chat	44	n/a	n/a	n/a

but	45	15	24	25
just	46	52	n/a	n/a
any	47	n/a	75	74
like	48	71	83	78
there	49	38	36	38
know	50	22	n/a	n/a
your	51	93	n/a	98
can	52	70	61	61
get	53	74	n/a	n/a
if	54	41	45	50
anyone	55	n/a	n/a	n/a
or	56	44	31	27
one	57	36	38	32
good	58	n/a	n/a	n/a
up	59	61	56	55
wanna	60	n/a	n/a	n/a
at	61	26	19	18
this	62	14	22	21
ya	63	n/a	n/a	n/a
ok	64	n/a	n/a	n/a
I'm	65	57	n/a	n/a
go	66	86	n/a	n/a
am	67	n/a	n/a	n/a
we	68	23	40	41
msg	69	n/a	n/a	n/a
well	70	17	95	100
too	71	n/a	n/a	n/a
about	72	39	54	57
they	73	24	33	30
:	74	n/a	n/a	n/a
im	75	n/a	n/a	n/a
out	76	69	53	51
he	77	18	12	10
hehe	78	n/a	n/a	n/a
back	79	n/a	n/a	94
want	80	n/a	n/a	n/a
oh	81	47	n/a	n/a
<kinoko>	82	n/a	n/a	n/a
see	83	50	n/a	n/a
an	84	81	34	29
thanks	85	n/a	n/a	n/a
<chmod>	86	n/a	n/a	n/a
think	87	28	n/a	n/a
then	88	63	68	71
some	89	90	57	65
got	90	49	n/a	n/a
bye	91	n/a	n/a	n/a
who	92	83	50	46
its	93	n/a	n/a	56
where	94	n/a	n/a	96
yeah	95	46	n/a	n/a

don't	96	n/a	n/a	n/a
will	97	n/a	n/a	47
has	98	94	n/a	44
why	99	n/a	n/a	n/a
2	100	n/a	n/a	n/a
Sources:	CIRC	Svartvik 1982	Svartvik 1990	Francis & Kucera 1967

Table 1: List of the hundred most frequent words in CIRC compared with LLC, LOB, and Brown.

A first glance at the table makes it abundantly clear that the language in the CIRC has several unique features that are not present in the other corpora. Almost half of the items listed for CIRC are not listed in all of the other corpora's hundred most frequent word lists. Partly, these are items which would not count as words but would fall under the category of punctuation in other corpora. Examples would be asterisks or the emoticon :). For the CIRC, however, the inclusion of such characters in the analysis is crucial because of their special functions, for example distinguishing action and system messages from standard messages.

Another category of words for which we do not find equivalents in the comparison corpora are those which arise out of unique CMC conventions and technological limitations. These include abbreviations, such as 'u' for 'you', 'msg' for 'message' or 'lol' for 'laughing out loud'. Another part of these are unusual spellings, such as the omission of the apostrophe in 'dont' and 'im'. Finally, there are two IRC user names which of course would not occur in a corpus of standard spoken or written English, because the speaker's/writer's name is not automatically be added to every utterance. Both speakers' contributions are below seven percent of all messages in the channels they participate in which is equivalent to less than one percent of the whole corpus.

Lastly, and most prominently, there are words which are listed for the CIRC but not the comparison corpora simply because their frequency is too low for inclusion in the top 100 in these corpora. Some examples of this category are ‘why’, ‘thanks’, ‘too’, ‘want’ and the digit ‘2’. There are also a number of words in this category which are listed in some of the comparison corpora. More often than not they appear in the list for the LLC, the corpus of spoken English, but not in the lists of the two written corpora. For example, ‘oh’, ‘see’, ‘think’, ‘yeah’, ‘get’ and ‘got’ are some of the words that are high frequency items in CIRC and LLC but not the LOB and Brown corpora. This indicates that the CIRC is more similar to the LLC than to the two written corpora. Another fact that supports this hypothesis is the nature of the words in the above category. The vast majority of these items are typical for casual speech. Some striking examples would be ‘ok’, ‘oh’, ‘thanks’, and ‘wanna’.

When looking at the ranking in the word lists some items can be noticed that are considerably (more than five ranks) higher or lower than in all the other corpora. The ones with lower ranks are: ‘and’, ‘of’, ‘be’, ‘there’, ‘but’, ‘or’, ‘one’, ‘at’, ‘this’, ‘we’, ‘he’, ‘then’, ‘that’, ‘about’, ‘out’, and ‘who’. It is interesting to note the high incidence of deictic items as well as coordinating conjunctions in this list. Abnormal distributions for these items have been described in previous research (Lundstrom 1995). These phenomena will be analysed in more detail in later sections dealing with deixis and syntax.

The analysis of the items with higher ranks and frequencies is complicated by the distortion effect exerted by the system messages contained in the corpus. Most frequent¹⁴ among

¹⁴ The ratio messages announcing nickname changes and kicks vs. ones announcing a change in channel topic is about 30 to 1. None of the words in the latter system messages appear in the list of the hundred most frequent words.

these are the messages announcing a change of nickname and the ones announcing someone being kicked off the channel. The respective phrases displayed are: “X is now known as Y” and “X was kicked by Y”. The lexical items in these phrases either rank considerably higher than the respective words in the comparison corpora or they rank closer to the written corpora than the spoken one. The latter would run counter to the general tendency of the CIRC to be more similar to the LLC than to Brown and LOB but can be easily explained by the influence of system generated messages.

In order to better gauge the extent of the system messages’ distortion effect the total number of occurrences for the two automatically generated phrases mentioned above was extracted from the corpus. Then these frequencies were subtracted from wconcord’s figures for the word count and two additional lists of the hundred most frequent words corrected for the system message effect were created¹⁵. Figure 3 below shows the words in question and their original data compared to the data corrected for the system messages’ influence.

words	frequencies				ranks				
	CIRC	CIRC cor.	LLC	Brown	CIRC	CIRC cor.	LLC	LOB	Brown
is	1.551	1.016	0.475	1.009	2	7	11	8	8
now	0.727	0.192	0.141	0.131	11	60	51	72	76
known	0.540	0.005	n/a	n/a	19	n/a	n/a	n/a	n/a
as	0.660	0.125	0.227	0.725	14	93	29	13	14
was	0.543	0.263	0.412	0.982	18	40	13	9	9
kicked	0.294	0.014	n/a	n/a	40	n/a	n/a	n/a	n/a
by	0.368	0.088	0.118	0.53	32	n/a	65	20	19

Figure 3: Comparison of the words contained in system messages. Unaltered and corrected frequencies and ranks and the corresponding figures for the comparison corpora are included.

¹⁵ See appendix C for the lists of the hundred most frequent words corrected for the system messages’ influence.

The fact that three of the seven words affected by the system messages' distortion effect, namely 'kicked', 'known', and 'as', fall out of the list of the hundred most frequent words when the effect is statistically corrected, attest to the strong influence of these messages. For the others the removal of the distortion results in lower ranks and lower frequencies. In four out of five applicable cases the comparison corpora have a pattern where the LLC has lower ranks and lower frequencies than the written corpora. Therefore the corrected data supports our hypothesis that the CIRC is more similar to LLC than the LOB or Brown corpora even more strongly than the original data. The words whose frequencies are considerably higher (more than five ranks higher) without being inflated by system messages are: 'me', 'like', 'what', 'my', and 'can'. Again, with the exception of 'like', which has recently become a filler in spoken English, similar tendencies have been noted by Lundstrom (1995).

The analysis of the 100 most frequent words above was based on a rank list. In order to verify its results a list of the 100 most frequent words displaying the absolute frequencies of these words was also generated¹⁶. Then the frequencies for the corresponding words in the LLC and the Brown corpora were added. Unfortunately, the same figures for LOB were not readily available. The results from this second 100 most frequent words list correspond to those of the first except that there were no items in the CIRC list with significantly lower frequencies than their counterparts in the comparison corpora. This is due to the fact that the CIRC shows a slightly different frequency distribution than the other corpora. Overall, the CIRC's frequencies for the 100 most frequent words are higher than those of LLC and Brown. This phenomenon will

¹⁶See appendix B for the 100 most frequent words list with frequencies

be described in more detail below. As for items with significantly higher frequencies, the same results as with the rank list were found and in addition 'your', 'no', and 'do' had above average frequencies. Again, the first item has been mentioned by previous studies (Lundstrom 1995) the latter two are typically found in spoken discourse.

Correlations between the 100 most frequent words

So far, the similarity between the CIRC and the LLC has only been demonstrated at a relatively intuitive level through presence of typically spoken items in the 100 most frequent words lists of both corpora. However, the similarity hypothesis can also be supported by evidence from statistical tests. Correlation analyses were performed for the two versions of the 100 most frequent words lists. Pearson's correlation was used with the list of absolute frequencies whereas a Spearman's Rho test was utilized to determine the correlations within the rank lists.

For the frequency list, significant correlations were found between all three corpora. However, the CIRC's correlation coefficient with the LLC was higher than that with the Brown corpus (0.793 vs 0.557; $p < .001$ for both). The degree of similarity becomes even clearer when we look at the coefficients for the Brown corpus. The correlation figure for the LLC is much higher (0.845; $p < .001$) than the same figure for the CIRC (0.557; $p < .001$). Finally, comparing the LLC to the other corpora, the figures suggest that it is slightly more highly correlated with the Brown corpus than with the CIRC (0.845 vs. 0.793; $p < .001$).¹⁷

¹⁷See appendix D for more details of the Pearson test

The results of the correlation test for the 100 most frequent words rank lists further corroborate the hypothesis that the CIRC is more similar to the LLC than to the two corpora of written English. When correlating the CIRC with the other corpora the Spearman's correlation coefficient for the LLC is higher (0.596; $p < .001$) than for the two written ones (0.545 and 0.543 for LOB and Brown respectively; $p < .001$). The correlations obtained for the written corpora show the same trend but in inverse order. For example, the Brown corpus is highly correlated, in fact it is almost identical to the LOB corpus (correlation coefficient 0.993; $p < .001$). Its correlation with the LLC is weaker (0.740; $p < .001$) and the coefficient for the CIRC is lower still (0.543). The numbers calculated for the LLC show that in terms of similarity it is almost equidistant to both the corpora of written English and the CIRC. However, the correlation coefficient for LOB and Brown (0.762 and 0.740 respectively; $p < .001$) is higher than the one for CIRC (0.596; $p < .001$).¹⁸ Illustration 3 shows a schematic drawing of the relations between the CIRC and the comparison corpora based on the results of the correlation tests described above.

Figure 4: Correlation schema for CIRC and comparison corpora

The exact same patterns which are illustrated in figure 4 above also emerge when the correlation tests are run on the lists of the hundred most frequent words which were corrected for the system message distortion effect¹⁹. The only difference is an even more pronounced

¹⁸See appendix D for more details of the Spearman's Rho test

¹⁹ See appendix E for the detailed results of these test on the corrected rank and frequency lists.

correlation between the CIRC and LLC. When displayed in a schema such as the one above the LLC would therefore be slightly closer to the CIRC.

Frequency distributions in the 100 most frequent words

Figure 4 below displays a frequency plot of the twenty most frequent items in the CIRC, the LLC and the Brown corpus. The graphs of the spoken corpus and the CIRC run parallel, thus clearly showing the great similarity between these two. Both start with frequencies approaching two percent and then gradually decline. In contrast, the written corpus starts with high frequencies near seven and then drops sharply until it falls into a line slightly above the other two corpora around the seventh most frequent item.

Figure 5: Frequency decline, high end.

This frequency distribution of the written corpus has an impact in the area of corpus description that is usually referred to as type/token ratios as well.

Type/token ratios

Type token ratios can not only be used to measure the relative lexical complexity in different parts of the corpus. They are also a tool that describes the corpus as a whole. For this purpose the wconcord software includes the cumulative frequency in its word lists. The cumulative frequency simply adds up the frequencies of all the items or types above the current one. Thus, when looking at the list of word types ordered by frequency, one can easily determine how many tokens are accounted for by a specific number of types. These type/token ratios are used in corpus linguistics to describe the internal makeup and level of technicality in a corpus.

The following quote by Soyama (1999²⁰) illustrates their use:

It is generally known that, in a fairly large corpus of non-genre specific written English, the first 50 to 100 word-types typically account for about a half of the total word tokens of that corpus, and 600 to 800 word-types cover approximately 70% of the total running words. Beyond this level, figures vary according to researchers and the types of text/corpus under consideration, but a very “conservative” estimate given in Kennedy (1998, p. 96) is that we need up to about 1,500, 3,000 and 15,000 word-types to achieve 80%, 90% and 95% coverage of the total word tokens of a given corpus respectively (See Table 4-3, p. 92). The *COBUILD Dictionary* claims that 15,000 word-types cover approximately 95% of the running words of their corpus, which partly confirms the estimate given by Kennedy. In a corpus as diverse as the Brown and LOB Corpora, however, we need a slightly larger vocabulary to achieve the same percentages of text coverage than this estimate, particularly above the 80% range. My calculation indicates that approximately 1,700 word-types are needed to cover 80% of the total tokens in the Brown Corpus, and about 61% more word-types as that of Kennedy's estimate are required to achieve 90% coverage thereof. These differences, however, are not at all surprising considering the “catchall” nature of the Brown Corpus, which is made up of 500 texts from as many as 15 different genres.

For the CIRC these type/token measurements are summarized in the table below.

²⁰ The quote is taken from an unpaginated version of the thesis downloadable from the Internet. It is from chapter one.

percentage of tokens accounted for	50%	60%	70%	80%	90%	95%
by the number of types	316	778	1911	5050	16447	35513

These figures are clearly well above, and in most cases the triple of, those that Soyama mentions as typical for written corpora of English and the ones he reports for the Brown corpus. The reasons for these high type/token ratios are manifold and stem from the nature of the data contained in the CIRC. First of all, we have seen that the CIRC frequencies are much more similar to a spoken corpus than to a written one. The frequency distribution table for the twenty most frequent words already indicated that the LLC and the CIRC had overall lower frequencies than the Brown corpus.

Secondly, as explained in the methodology section, punctuation was not disabled in the word count. Therefore, words which have punctuation characters adjacent to themselves, e.g. 'stop.', are counted as a new type instead of being subsumed under the lemma 'stop'. Thirdly, due to the rather quick pace of typing and lack of time for editing in an interactive conversation there are more typing mistakes in CIRC than in a typical corpus of written English. Again, these are recorded as separate types in wconcord's word counts. Finally, the fact that IRC-nicknames are automatically included in every message, leads to a disproportionately high number of proper names in the CIRC which in turn results in higher type/token ratios.

ANOVA and post hoc tests

In order to achieve a more detailed description of the data that goes beyond simple correlations between variables, a series of ANOVA tests was run on the measures available for

the 200 CIRC files. This procedure is used to find statistically significant differences between the variation within groups and the variation across groups. Significant effects were found for the words/message, the type/token ratio and the hapax/mille ratio variables. The figures are $F(9-190) = 32.140$, $p < .001$; $F(9-190) = 11.549$, $p < .001$; and $F(9-190) = 6.987$, $p < .001$ for these three variables respectively²¹. This signifies that the data of one or more channels for the variables mentioned above is significantly different from the other/s.

In order to pinpoint the source of these effects more closely a set two of post hoc tests was also applied to the data from the 200 corpus files. The tests used were Scheffe and Bonferroni²². Again two tests were chosen so that results could be double checked. Both tests compare each channel with the nine others and indicate where significant differences in means are found. The following tables²³ summarize the post hoc tests' results.

Table2: ANOVA post hoc tests (Scheffe and Bonferroni) for the words/message variable. For every square marked with a letter the post hoc tests found a significant difference in means between the respective two channels.

1 = mel						zZ	Z	zZ	zZ	
2 = per				Z		zZ		zZ	zZ	
3 = can						zZ	Z	zZ	zZ	
4 = che		Z				zZ		zZ	zZ	Z
5 = cyb				zZ		zZ		zZ	zZ	

²¹ See appendices G, and H for details on the ANOVA analysis.

²² For details see SPSS 9.0 for Windows.

²³ Each table matrix is halved by a vertical line because both halves display the same data. In order to facilitate table interpretation, however, both halves have been included.

6 = hel	zZ	zZ	zZ		zZ		zZ			zZ
7 = aus	Z		Z	zZ		zZ		zZ	zZ	
8 = phi	zZ	zZ	zZ	zZ	zZ		zZ			zZ
9 = beg	zZ	zZ	zZ	Z	zZ		zZ			zZ
10 =cha				zZ		zZ		zZ	zZ	
	1	2	3	4	5	6	7	8	9	10

Legend: words/ message with Scheffe = z; words/ message with Bonferroni = Z; for all results the mean difference is significant at $p < .05$.

Table 3: ANOVA post hoc tests (Scheffe and Bonferroni) for the type/token ratio variable. For every square marked with a letter the post hoc tests found a significant difference in means between the respective two channels.

1 = mel		X					X	xX	xX	
2 = per	X									
3 = can								X		
4 = che								xX	X	
5 = cyb								xX		
6 = hel								X		
7 = aus	X									
8 = phi	xX		X	xX	xX	X				X
9 = beg	xX			X						
10 =cha								X		
	1	2	3	4	5	6	7	8	9	10

Legend: type/token ratio with Scheffe = x; type/token ratio with Bonferroni = X; for all results the mean difference is significant at $p < .05$.

Table 4: ANOVA post hoc tests (Scheffe and Bonferroni) for the hapax/mille ratio variable. For every square marked with a letter the post hoc tests found a significant difference in means between the respective two channels.

1 = mel		oO					oO	oO	O	
2 = per	oO							O		
3 = can								oO		
4 = che							O	oO		
5 = cyb							O	oO		
6 = hel								oO		
7 = aus	oO			O	O			O		
8 = phi	oO	O	oO	oO	oO	oO	O		oO	oO
9 = beg	O							oO		
10 = cha								oO		
	1	2	3	4	5	6	7	8	9	10

Legend: hapax/mille ratio with Scheffe = o; hapax/mille ratio with Bonferroni = O; for all results the mean difference is significant at $p < .05$.

From the analysis of the above tables a number of conclusions can be drawn. Firstly, the Scheffe test is slightly more conservative in its results than the Bonferroni test. In every instance where Scheffe finds a significant difference Bonferroni yields the same result. However, there are a few comparisons where only Bonferroni finds a significant difference in means. Secondly, the two measurements for lexical density have much less variation than the words/message measure. Thirdly, channel eight, #philosophy, has the highest number of significantly different means of all channels and across each of the three variables. Therefore, it is likely that in this channel distinct linguistic patterns will be found. As a result, examples taken from this channel should not be considered typical for the whole of the CIRC.

Summary

The descriptive statistics have shown us that the CIRC is structurally more similar to a corpus of spoken English than to a corpus of written English. They have also allowed us to identify several areas of interest for our study. These are the deictic items, the IRC specific expressions, and the words typically associated with spoken conversation, all of which featured prominently in the list of the 100 most frequent words.

Finally, through the study of the descriptive statistics several relevant characteristics of the CIRC could be ascertained. Channel #philosophy was found to be a special case within the corpus data. In terms of word frequency counts distortion effects were found due to the inclusion of computer generated messages on the one hand and due to the inclusion of punctuation characters on the other.

CHAPTER 5

ANALYSIS OF CIRC FEATURES COMPARABLE TO SPEECH AND TO WRITING

Deictics

The present study concerns itself only with the most basic form of deixis, i.e. pronouns. These lend themselves most easily to the computer-aided analysis using a corpus, which is the methodology of this work. The analysis of all deictic devices in IRC, while certainly worthwhile, lies outside the scope of this study. The treatment of deictics in the following section will be informed by one overarching hypothesis. It is expected that the use of deictics in IRC is heavily influenced by the special nature of IRC communication technology. Namely, the fact that IRC-conversations do not take place in the real world but in an imagined, virtual space will impact the usage of deictics whose essence it is to refer to entities other than themselves. In IRC there are two spaces of reference for deixis, the physical world and the virtual world, thus creating a whole new dimension which IRC-users can refer to. It will be investigated how this communicative challenge is met.

Personal Pronouns

Among the deictic items examined in this study, personal pronouns are probably the

least affected by the dual referential background described above. In fact, people, manifesting their online presence through nicknames and the messages they send are the most tangible part of the virtual space of an IRC-chat room. However, all second and third person pronouns in IRC will have an added sense of ambiguity due to the lack of a physical situational context. For example, 'you' in a real life conversation can only be ambiguous if more than one person is present. In contrast, the chat rooms observed for the CIRC always have a co-presence of at least several other participants.

Due to the factors outlined above, it is not surprising that personal pronouns have been identified in prior research (Ferrara et al. 1991, Murray 1991a, Collot & Belmore 1996) as an area where CMC discourse conventions deviate from other language varieties. In particular, it has been observed that first and second person pronouns occur more frequently in CMC than is normally the case. This observation is also borne out by our data. In the lists of the hundred most frequent words the pronouns 'I', 'you', 'me', 'my', and 'your' are have considerably higher ranks and frequencies in the CIRC than the comparison corpora. This trend is even more pronounced if the variants of 'I' and 'you', 'i'm', 'im', 'u', and 'ya' are included in the word count.

The other personal pronouns contained in the list of the hundred most frequent words are 'we', 'they', 'he', and 'its'. All of them are less frequent in the CIRC than in Brown and the LLC. It appears that in IRC the aforementioned trend only applies to the first person pronoun in the singular. Unfortunately, it is not possible to ascertain if the same applies to the pronoun 'you' because it functions as both singular and plural. The differentiation of the two in a large scale corpus such as the CIRC would require grammatical tagging which lies beyond the scope of

the present study. However, there is reason to believe that the second person plural pronoun ‘you’ follows the same pattern as the other plural pronouns and is therefore less frequent in the CIRC than the comparison corpora. Firstly, a sample of 80 instances of ‘you’ taken at random from the CIRC contained only five occurrences in the plural. Moreover, in all five instances the plural pronoun was accompanied by quantifiers such as “both of you”, “you guys” or “you all”. Due to the fact that IRC discourse is taking place in a virtual locale where multiple conversations with multiple participants can take place at the same time, ‘you’ alone is too ambiguous to serve efficiently as a plural pronoun. The same reason is probably behind the low frequencies of the other plural pronouns as well.

The table below summarizes the occurrences of personal pronouns in the ten channels. The most frequent pronoun is ‘I’, followed by ‘you’. The same relation somewhat less pronounced holds still true when we take the variations of ‘you’ and ‘I’ in the CIRC, namely ‘u’, ‘ya’, ‘yew’, ‘i’m’, and ‘im’ as displayed in table 7 below, into account.

	i	you	he	she	it	we	they
acmel	1460	631	137	103	489	109	107
acper	1504	697	172	75	605	146	159
dacan	2089	1123	195	133	723	148	188
dache	1543	1014	163	124	631	207	148
escyb	1812	1172	148	112	618	129	116
eshel	2707	1380	107	18	1627	108	225
itaus	2371	1407	185	175	834	170	171
itphi	2214	1427	247	73	1084	371	290
utbeg	1229	1272	90	49	470	184	123
utcha	1536	1138	105	126	576	123	94
total	18465	11261	1549	988	7657	1695	1621

Figure 6: Distribution of personal pronouns (subject case) in the ten channels.

	i	i'm	im	you	u	ya	yew
acmel	1460	137	198	631	761	207	0
acper	1504	114	289	697	682	276	148
dacan	2089	251	179	1123	545	197	1
dache	1543	159	159	1014	534	317	0
escyb	1812	177	241	1172	405	293	1
eshel	2707	187	111	1380	271	50	1
itaus	2371	252	67	1407	589	230	2
itphi	2214	230	76	1427	292	36	0
utbeg	1229	144	107	1272	385	198	1
utcha	1536	234	167	1138	539	173	0
total	18465	1885	1594	11261	5003	1977	154

Figure 7: Distribution of the spelling variants of 'i' and 'you' in the ten channels.

The absolute frequencies of 'I' vs. 'you' 18,465 vs. 11,261 change to 21,944 vs. 18,395 with the inclusion of variations. While this does not reverse the two items' rank order it does give 'you' a higher frequency proportionally. Higher in fact, than in both of the two comparison corpora where 'I' and 'you' occur 5170 vs. 3290 (Brown) and 14890 vs. 8010 times (LLC). The ratios corresponding to these figures are 1.19 (CIRC), 1.57 (Brown), and 1.86 (LLC), i.e., in the CIRC there are about 20% more instances of 'I' than of 'you' compared to 57% and 70% more in the Brown corpus and the LLC.

These figures are interesting in two aspects. Firstly, it is surprising that they are so high in absolute terms even when compared to the corpus of spoken English. This is due to the fact that the language material contained in the CIRC consists entirely of highly interactional multi-party conversations whereas the LLC is compiled from a variety of genres including ones with little or no interaction, such as radio broadcasts, public speeches, sermons and the like. These types of discourse, where there is no immediate feedback and no need to address the audience frequently, naturally rely to a much lower degree on personal pronouns to create the rapport between speaker and addressee. I suspect that the casual conversation subcorpus of the LLC

would yield high numbers of the first and second person singular pronouns similar to the CIRC.

The second point of interest is the low number of occurrences of 'I' in proportion to 'you' in the CIRC. In fact, the numbers even suggest that the CIRC is closer to the Brown corpus than the LLC in this aspect, which seems to contradict our similarity hypothesis. However, a closer look at IRC-technology can easily disprove that argument. Recall that every IRC-message is preceded by the nickname of the sender in brackets. Thus, the identity of the speaker is never in question in IRC. The computer generated nickname provides a gratuitous subject, so to speak, for every sentence where the actor and the speaker are identical. The frequent occurrence of messages with a pattern such as: "<participantX> thinks thats sweet" proves that IRC users are aware of this relationship and utilize it in their speech. This omission of the subject pronoun I has also been identified in prior research on IRC (Werry 1996). In messages such as "<participantY> i think thats right" the 'I' is essentially redundant because participantY is obviously the 'I' who is speaking.

The question we need to ask then is not why does 'I' occur relatively infrequently but why does it occur at all? One possible reason is the awkward shift to the third person singular when talking about oneself. A solution to this problem would be to simply drop the redundant pronoun 'I'. Even though such messages with 'I' omission, e.g.: "<participantG>get to look at hallucinating things...weird trickery on my sences." occur they are extremely rare. This rarity is probably due to the possible confusion of this construction with the imperative, which incidentally is much more frequent than the above construction in the CIRC. Another possible explanation is that IRC-users simply assume that the nickname in brackets implies something along the lines of: "<participantX> says/asks/remarks/etc." An IRC conversation would

therefore be read like an extended succession of direct quotes.

Figure 8: Distribution of personal pronouns (subject case) in the ten channels. Numbers stacked to 100%.

A channel by channel comparison of the use of personal pronouns reveals several interesting tendencies. Including alternative spellings there are twelve different pronoun forms which enter the analysis as illustrated by figure 8 above. The channel #philosophy scores highest on four of these, namely: 'you', 'he', 'we', and 'they'. The latter three fall into the category of pronouns which we have found to be much more rare in the CIRC than in the comparison corpora. This supports the earlier finding of the ANOVA analysis which revealed that #philosophy was distinctly different from the rest of the CIRC. Now, it becomes clear that this difference very likely lies in a closer similarity of this channel to language typically associated with written genres.

The channel #help has the highest score of all channels for the pronouns 'i', and 'it'. This reflects the kind of interaction that takes place in this channel. The channel regulars in

#help provide free online support for computer related questions. Therefore, participants more often have need to refer to things, usually computer parts or software, than in channels with more social interaction. Similarly, social interaction which is the main objective in channels such as #cyberchat or #canada requires more involvement with the audience than giving and requesting problem solving information. These are usually concerned with the speaker's personal experience. Thus, there is a higher incidence of 'i' in #help than in the other channels. Below is a short example of a typical #help interaction.

"<ParticipantB> does anybody know if I can view a .art file without having aol?"

<ParticipantD>participantb : Internet explorer handles it"

When analyzing the four pronouns with non-standard spellings #melbourne and #cheers both get one highscore and #perth has two. One of these variants, namely 'yew' is highly localized in #perth and is probably one of the characteristic features of that channel. Therefore it will be excluded, for the moment, from the further analysis of personal pronoun variants. The distribution for the non-standard spelling becomes more complex when we examine the second and third ranking channels. As with the first ranks we find #melbourne and #perth, but #cyberchat and #australia are also found in the top ranking places. Later analyses of non-standard spelling variants will verify whether these channels, and especially the two local Australian ones which appear more often than the others, have indeed much looser spelling conventions than the other seven. Finally, the last two personal pronoun forms, 'I'm' and 'she' are headed by the channel #australia. The reasons for this remain unclear.

Looking at low frequencies it is not surprising that #philosophy proves to be a special case yet again. It ranks very low on all forms of non-standard spelling. This is in accordance

with our earlier hypothesis which stipulated that #philosophy adopts more conservative linguistic conventions than the other channels. #help also disfavours non-standard spellings. It might be that the exchange of highly technical information results in a more formal conversational climate in this channel. It remains to be seen if other linguistic material supports this hypothesis. The channel #beginners ranks low on some non-standard items, namely ‘im’ and ‘u’ but not ‘ya’. It seems to be a hybrid in this respect. Further investigation is needed to ascertain which non-standard features are utilized by newcomers to IRC and why some seem to be acquired later.

As discovered above the channels #melbourne, #perth, #australia and #cyberchat are at the lower end of the level of formality. An examination of the low frequencies for the standard forms of ‘you’, and ‘I’m’, however, finds that only the first two channels have low ranks corresponding to their high ranks in the non-standard variants. It seems that #melbourne and #perth are more at the cutting edge of the linguistic development in IRC which is moving IRC discourse away from written norms than #cyberchat. Incidentally, these are the two channels in the CIRC with a local channel name. It will take further investigation to determine whether these ‘progressive’ conventions are due to the channels’ locale, the channels’ in-group makeup, or other as of yet unidentified factors.

As mentioned earlier, the fact that grammatical tagging of the CIRC was not an option for the present study puts restrictions on the grammatical analysis of the pronouns examined. In order to partly compensate for this shortcoming, a brief analysis of the non-referential usage of ‘you’ was conducted on the basis of samples taken from the CIRC. The studied sample consisted of 80 instances of the personal pronoun ‘you’ taken at random from four channels. These were #help and #philosophy on the one hand, chosen because prior analysis had shown them to deviate

from the rest of the CIRC in pronoun usage, and #canada and #cyberchat to represent the pronoun usage typical for channels with primarily social interaction.

The results for ‘you’ showed that of twenty occurrences in the four channels, no instances of non-referential usage were found for the channel #cyberchat, one instance for #canada, four for #help and five for #philosophy. Although the sample size is certainly too small to make any statements about the statistical significance of these findings, they are nevertheless completely in accordance with the tendencies identified above. Logically, the distribution makes sense as well. In discussions about philosophy people often generalize about the human condition and universal patterns of behaviour and thinking. The non-referential ‘you’ is an ideal construction for that sort of statement e.g.:

“<ParticipantL> don't kill is universal moral law

<ParticipantL> u break it..u face judgement”

Similarly, the interaction in #help often involves asking or instructing how to do something. This is the other area where IRC-users show a preference for employing the non-referential ‘you’. This is plainly illustrated by the fact that all four of the instances in the sample from #help occur in questions.

Tables 9 and 10 below summarize the channel frequencies for the personal pronouns in the object case and the possessive pronouns and determiners²⁴. Due to their in general much lower frequencies generalizations and interpretations on the basis of this data are naturally more tentative than those made for the lists above. Nevertheless, some interesting observations can be

²⁴ The item ‘her’ which is both a personal pronoun in the object case and a possessive determiner was included only once in table 9. The word ‘his’ which is a possessive determiner and pronoun is included in both tables.

made.

	mine	yours	his	hers	its	ours	theirs
acmel	6	3	75	1	120	0	0
acper	15	7	106	0	212	0	1
dacan	33	10	88	1	109	1	1
dache	27	4	79	0	115	1	0
escyb	30	6	97	3	93	1	0
eshel	25	7	58	0	212	0	0
itaus	29	11	87	1	127	0	0
itphi	16	13	118	0	112	4	1
utbeg	22	3	77	1	86	1	0
utcha	24	14	79	1	138	1	0
total	227	78	864	8	1324	10	3

Figure 9: Distribution of possessive pronouns.

	me	him	her	us	them	em	my	your	yer	his	our
acmel	129	54	99	33	46	39	263	170	27	75	25
acper	93	76	12	49	43	65	310	207	88	10	20
dacan	66	71	16	39	73	23	444	231	58	88	42
dache	492	69	132	38	82	15	33	21	18	79	27
escyb	654	77	15	33	77	15	41	27	14	97	20
eshel	722	29	10	19	10	14	590	352	18	58	14
itaus	60	72	152	33	70	31	48	23	40	87	20
itphi	35	67	50	54	11	3	39	29	6	11	71
utbeg	437	36	84	74	60	8	29	261	11	77	23
utcha	747	84	17	35	45	10	391	19	11	79	11
total	691	63	115	407	717	22	392	244	291	864	273

Figure 10: Distribution of personal pronouns (object case) and possessive determiners.

First of all, the analysis of the distribution tables displayed in figures 9 and 10 above shows that the same pattern that was found with the frequencies of the personal pronouns in the subject case exists for the possessive pronouns and determiners and personal pronouns in the object case as well. The third person and plural pronouns are much less frequent than the first and second person pronouns. The same caveat about the second person pronoun 'your' holds true in this case. The proportion between first and second person pronouns in the singular does

not follow the pattern discovered above. That is, the first person is by far more frequent than the second. Obviously, the syntactic restrictions discussed above, which were posited to account for the lower frequency for 'i', do not apply to the other forms of the first person singular pronoun. If the frequencies for the possessive pronouns and determiners and the pronoun in the object case are an indication of what the proportion would look like without these restrictions, they must be quite powerful.

It is important to keep in mind that 'her' listed in the tables above is both a possessive determiner and personal pronoun in the object case. Otherwise one could reach the conclusion that 'her' is more frequent than 'his' or 'him'. However, if we add the frequencies for 'his' and 'him' we find that overall the masculine pronoun is used more often than the feminine pronoun. This trend is common to all large scale non-genre specific corpora of English. To find the contrary in the CIRC would be very unusual. The use of gender neutral pronouns such as 'hir', which has been reported (Schaap 2000) to be quite common in text based virtual realities or MUDs, is not attested for IRC. No instance of such pronouns was found in the CIRC.

The data for the object case forms of the personal pronoun and the possessive pronouns and determiners for the channel #philosophy follows the same pattern already identified above. #philosophy ranks first for 'his', 'our', 'ours', and 'them'. All of them are third person and/or plural pronouns, thus supporting the hypothesis that, unlike the rest of the CIRC, #philosophy's conventions are more similar to those of written genres than speech. This receives further backing from the extremely low frequencies of the non-standard forms 'yer' and 'em' in #philosophy.

It is important to note that 'its' listed in the distribution is not only a form of the

possessive third person pronouns but also an IRC spelling variant of ‘it is’ as in:

“<ParticipantZ> Its night.. but I cant sleep... “

We have already observed the IRC spelling convention of dropping the apostrophe in the alternate forms ‘I’m’ and ‘im’. An analysis of a sample 100 instances of ‘its’ indicates that the vast majority of occurrences is a non-standard spelling rather than the possessive pronoun. In fact, of the one hundred analyzed instances only one was a possessive pronoun. Seen from this perspective it makes sense that #help which had the largest number of ‘it’ ties for first rank with #perth for the variant ‘its’. Naturally, this also confirms #perth’s role as a channel with a high incidence of non-standard spellings.

More evidence for this comes from the analysis of the two non-standard forms ‘em’ and ‘yer’ which show a striking parallel to the other non-standard variants of the personal pronouns in the subject case. For both forms #perth ranks highest followed by #canada in the first case and #melbourne in the second, which reinforces the validity of our earlier finding that both #perth and #melbourne are prominent in the loosening of spelling conventions. We also find another trend which was not as apparent in the first analysis of non-standard spelling personal pronouns. Namely, for both forms, the four channels with geographical names account for more than two thirds of the total number of occurrences. See figure 11 below for a visual presentation of the phenomenon.

Pers. Pron. Distribution Cont.

Figure 11: Distribution of personal pronouns in the object case, possessive pronouns and possessive determiners. Stacked to 100%.

This finding could have far-reaching theoretical significance. If we take the use of non-standard forms in IRC to be a means of signaling in-group familiarity or strong group ties as Paolillo (1999) suggests, then the clustering of these forms in channels with geographical names would indicate a link between the channel name and group cohesion within the channel. The next question is why stronger group ties develop in channels with geographical names and not in others.

An argument can be made that membership in channels with general chat names is more in flux than it is in channels with geographical names. However, this argument has two serious flaws. Firstly, although it is plausible, it cannot be easily verified. Every public IRC-channel has a group of regular participants and a number of guests who come and go with variable frequency. Therefore it is not a question of whether or not fluctuation occurs but it is a question of degree which renders the measuring of the phenomenon impractical. Secondly, why does a channel

such as #philosophy, for which there is no reason to assume a high participant turnover, not display strong in-group ties signaled by non-standard spellings?

Another hypothetical answer to the observed phenomenon may lie in the geographical origin of the channels' participants. It can be assumed that most people gathering in channels with geographical names have some connection to the real world equivalent of the virtual locale. They might be born there, reside there or simply have a strong interest in the place. It is possible that this shared real world background facilitates comprehension in the virtual setting. This in turn allows participants to deviate more freely from real world norms of communication without an added risk of communication breakdown. In contrast, the population of a general chat channel lacks a common background apart from the desire to communicate and a certain affinity to computer technology. Moreover, the presence of non-native speakers of English adds to the level of linguistic clarity required of the channel's participants. Unfortunately, due to the difficulty of tracking IRC-users locations, origins and motivations this answer can only be speculative. It also remains to be seen if other data confirms or disconfirms the link between non-standard linguistic forms and channels with geographical names.

Last but not least, at this point, a mention should be made of what I would term a deictic sign. The following example illustrates the way this deictic arrow is employed by IRC participants:

“<ParticipantGl> <- loves participantf”

Use of this deictic sign to replace the first person pronoun has been first documented by Werry (1996). However, he attributes its emergence to a general tendency for abbreviation in IRC which is erroneous since it results in more keystrokes than would be required for the same

sentence using the first person pronoun. When we consider the fact discussed above that phrases such as '<participant> thinks this is cool' are perfectly acceptable in IRC it becomes clear that this arrow deictic is not an abbreviation but a creative play with the medium's possibilities. This interpretation receives further support from the fact that the use of this arrow deictic is not rare even though it requires more typing. In fact, in the CIRC there are more than 300 occurrences of the deictic arrow. Moreover, the deictic arrow does not consist of only two letters in all of these 300 instances. Sometimes, the IRC participants prolong it extensively by pressing the minus key repeatedly to yield something similar to "<-----" which is an actual instance taken from the CIRC.

Furthermore, the use of the deictic arrow is not restricted to anaphoric reference as seen in the example above. In the following example the arrow points at the window with the list of channel participants on the user's computer screen:

"<ParticipantF> where did participantm go [...]"

<ParticipantS> participantf - ParticipantM is still here . . .way down there---->"

ParticipantS tells the person looking for ParticipantM that he/she is still present in the channel but that his/her name has probably scrolled off the screen in the participant list window. It is not clear if this is a case of exophoric reference because technically the arrow deictic points to an integral part of the IRC text. Additionally, it cannot be interpreted as an abbreviation because in some cases it serves to indicate a pause or to emphasize a point. In the following example it seems more likely to indicate the latter.

"<ParticipantA> individualistic, the way of survival

<ParticipantH> overtly individualistic - the way of eternal conflict with others ->

survival? heh”

Demonstrative pronouns and determiners

Since the demonstrative pronouns and determiners are more affected by the presence of two frames of deictic reference, the physical and the virtual, an item by item analysis will be conducted after the general discussion of the demonstratives’ distribution. The following table gives an overview over the occurrence patterns of the demonstrative pronouns and determiners and their spelling variants in the ten channels.

	this	that	dat	these	those	here	there	then
acmel	136	406	11	6	16	392	230	125
acper	176	450	13	16	14	269	157	139
dacan	177	645	5	11	38	398	260	134
dache	219	546	0	13	34	358	229	115
escyb	216	536	1	17	28	339	207	98
eshel	270	899	2	19	52	212	319	201
itaus	215	650	9	24	38	249	201	171
itphi	206	1020	5	26	65	163	249	196
utbeg	252	607	0	14	34	471	348	82
utcha	151	470	0	19	31	383	307	99
total	2018	6229	46	165	350	3234	2507	1360

Figure 12: Distribution of demonstratives in the ten channels

First of all, a caveat is in order for the items ‘that’, ‘there’ and ‘then’. As with ‘you’ in the discussion of the personal pronouns and determiners, these forms of the demonstrative items can also have other functions. Without grammatical tagging it is impossible to ascertain how many of these forms are demonstratives and how many are instances of the conjunctions ‘that’ the adverb ‘then’ or instances of existential ‘there’. This fact also makes it impossible to

ascertain which of the demonstrative items are more frequent, the ones that express close proximity, 'this', 'these', and 'here', or the ones that express distance 'that', 'those', and 'there'. Taking into consideration that the other functions of 'that' and 'there' can only inflate the numbers for these items, it is clear that the only case where the balance between these items is in question is 'that'. However, whichever way the balance is tilted for the 'this'/'that' pair no clear pattern emerges.

Figure 13: Distribution of demonstratives; numbers stacked to 100%.

In the distribution of the demonstrative pronouns/adverbs and determiners illustrated in figure 13 above many patterns familiar from the analysis of the personal pronouns and determiners can be found. #philosophy and also #help once again prove to be special cases. They have most of the highest and second highest scores for the items 'that', 'these', 'those', 'this' and 'then'. The high incidence of 'that' in #philosophy can probably be attributed to the use of that item as a conjunction rather than a demonstrative. This is supported by an analysis of 40 sample

instances taken equally from #philosophy and #cheers which yielded six conjunctions in the #philosophy sample and only two for the #cheers sample. However, as outlined above a definite quantitative proof of that claim lies beyond the scope of this study. The leading role of #help in the use of ‘this’ and ‘then’ might be explained by the presence of instructions and computer problem descriptions in the form of “first do (I did) this, then try (I tried) this ...”. Of course the ‘then’ in these phrases is an instance of the conjunction not the demonstrative. The channel #beginners scores highest of all channels for the use of ‘here’ and ‘there’. The reasons for this remain unclear. At the low end of the distribution we find once again mostly the channels #melbourne and #perth except when it comes to non-standard spelling forms.

There is only one non-standard spelling variant for the demonstratives. However, it is interesting that ‘dat’ exhibits the exact pattern already discussed for the non-standard spelling personal pronouns above. The four channels with geographical names account for more than 75% of all instances.

‘Here’

Demonstratives are much more affected by the presence of two frames of reference than the personal pronouns discussed above. The following two examples of ‘here’ illustrate this problem of deictic duality:

(1) “<ParticipantC> does anyone here know much about html??? “

(2) “<ParticipantW> ooooooooo lots of snow there participantS [...]

<ParticipantG> where are you ParticipantP?

<ParticipantP> 'it's cold here pretty much straight now [...]

<ParticipantP> northern vermont"

The 'here' in the first example refers to the chat room located in virtual space, the 'here' in the second refers to the user's physical location in northern Vermont. What are the conventions that have developed around this sort of ambiguity? An analysis of 50 sample instances of 'here' taken at random yielded ten occurrences where it referred to the physical world and forty others where the reference was to the virtual space or the discourse taking place there. This suggests that IRC-users display a definite preference for the virtual frame of reference. This is also supported by a collocation analysis for 'here'. The two collocations which are by far the most frequent are 'in here' and 'anybody here' with 398 and 223 instances respectively. Both collocations are typical for questions used by IRC-participants to inquire after potential conversational partners such as : "<ParticipantS> i gots a question...is there ne1 in here named Anthony?" or "<ParticipantI> anyone here from Caulfield Grammar of Preshil?" The demonstrative item 'here' in both types of question clearly refers to the chat room in which the question is asked. Judging from these frequencies the virtual space takes preference over the physical world in the linguistic interaction of an IRC chat room.

'There'

As mentioned above, the analysis of the item 'there' is complicated by the fact that it occurs not only as a demonstrative but also as the existential 'there'. Without grammatical tagging it is not possible to separate these two functions in the automated concordancing process.

However, a collocation analysis for ‘there’ in the CIRC revealed 451 instances of ‘there’ collocating with ‘is’ in adjacent position before or after ‘there’ and 188 instances of collocation with ‘are’. This indicates that about one fifth or slightly more instances of ‘there’ are occurrences of the existential ‘there’.

Another major collocation with 424 instances altogether are greeting routines involving ‘there’. The most frequent is ‘hi there’, followed by ‘hello there’ and ‘hey there’. These will be dealt with in more detail in the section on greeting and leave-taking.

All these uses of ‘there’ were disregarded for the analysis at hand. Like ‘here’, ‘there’ is used by IRC participants to refer to both physical and virtual locations. The following two examples clearly illustrate this point:

“<ParticipantS> what time is it there?”

“<ParticipantC> i don't talk on there”

In both examples the IRC participants use there to express distance in their chosen frame of reference. In the first example, ‘there’ refers to a physical location that is clearly far away, even in a different timezone, from the speaker’s. In the second example, ‘there’ refers to a channel or chat room other than the one the user is currently interacting in. However, this keeping to one dimension of reference need not always be heeded. Consider the following example:

“<ParticipantF> ParticipantP,you there??

<ParticipantF> he must be away:(

<ParticipantK> he must be participantF , he hasn't said anything anyway

<ParticipantL> he must be busy, or just watching>”

The question “are you there” is ambiguous in that it can refer to ParticipantP’s presence in the chat room or to his physical presence in front of his computer screen. The comments and reactions to ParticipantF’s question are very revealing in this context. First of all, P might not be present at either location, i.e. “away”. Secondly, he might be present in the virtual room but he will not and has not participated because he is either unable, i.e. “busy”, or unwilling to do so, i.e. he is “just watching”. We see that it is possible for participants to use ‘there’ to express distance within a referential space but also between the two frames of reference. As for the question which referential space is preferred in combination with ‘there’ we can only hazard an educated guess. The analysis of 20 sample instances of ‘there’ which found that 50% of the items referred to physical space suggests that IRC users have no clear preference either way.

‘This’

Like the other demonstratives we have examined so far, ‘this’ is used in IRC conversations to refer to phenomena of the real and the virtual world. The following two examples illustrate this point:

“<ParticipantN> why is this room so quiet????”

“<ParticipantG> what is this new age thing abt anyway?”

When ‘this’, as in the above examples, is used as a determiner it accompanies the noun it refers to. In this instance it is therefore much less ambiguous in its deictic reference than ‘here’ and ‘there’. Its meaning remains clear no matter what referential space the accompanying noun belongs to. The usage of ‘this’ as a demonstrative pronoun as exemplified in the two examples

below also occurs in the CIRC. However, it is less frequent than the demonstrative determiner.

“<ParticipantC> this is a slow channel....hi everyone”

“<ParticipantT> participantd: false, they use the signs creatively. even birds do this”

The quotations above make it clear that the demonstrative pronoun, similar to the demonstrative determiner, can refer to both the physical and the virtual space. An analysis of a 20 instance sample of ‘this’ showed that it was used with real life referents 14 times. This suggests that ‘this’ more often than not refers to things not of the virtual world. This is not surprising since the virtual world of IRC does not provide much of a virtual locale. CMC talk in MUD environments with their detailed descriptions of the virtual worlds they create is very likely different in that aspect. Some MUD programs used for fantasy role play even provide a special communication system for users who want to talk about the physical world instead of the virtual environment.

‘That’

The problem of ‘that’ as a conjunction and a demonstrative determiner and pronoun has been touched upon already. In its use as a demonstrative determiner ‘that’ works very much like ‘this’ in that it usually accompanies a noun. These nouns can pertain to either physical or virtual environments as the two examples below illustrate.

“<ParticipantS> hey is there an easier way to switch between windows w/o having to click on that window in irc”

“<ParticipantT> participants: i know that movie that's comin up, somethin slat lake city”

Unlike ‘this’ the occurrence of ‘that’ as a demonstrative pronoun seems to be more frequent than that of the demonstrative determiners. Just like its counterpart, the pronoun can refer to both referential dimensions as the examples below demonstrate:

“* ParticipantK wiggles around the room [...]

<ParticipantK> gawdssssssss I live that ”

“<ParticipantM> but then his msg said he'd be at midnight too

<ParticipantM> :P and we all know that that didn't happen”

The first example consists of an action message and a comment on it. Since the event described in the action message takes place in the virtual chat room, the ‘that’ which refers to it can be said to have a virtual referent. In contrast, ‘that’ in the second example refers to being stood up in the physical world. Again the answer as to which referents, physical or virtual, are more has to remain very tentative. An analysis of a 20 instance random sample yielded 13 physical world referents, indicating yet more similarity between the conventions of ‘that’ and ‘this’.

‘These’ and ‘those’

‘these’ and ‘those’ follow the patterns of their singular equivalents. The usage of ‘these’ and ‘those’ as determiners or pronouns seems to be roughly equivalent. Again, both can refer to the virtual and physical frames of reference. However, an analysis of a 20 instance sample revealed only two virtual referents. It appears that the plural forms of the demonstrative determiners are even more restricted to collocation with real world referents than the singular

demonstrative determiners and pronouns. Therefore, ‘these’ and ‘those’ are used predominantly like they would be used in more standard texts and contexts. This aspect of the usage of ‘these’ and ‘those’ can also explain why they occur so frequently in #philosophy. As we have seen the interaction in this channel is atypical for CMC in that it adheres much more to the standardized conventions of written language and is thus the most linguistically conservative of the channels.

‘Then’

The use of ‘then’ as an demonstrative adverb is extremely rare in the CIRC. The analysis of a 60 instance sample yielded only one occurrence where ‘then’ was not a conjunction. This example is given below:

“<ParticipantB> This was back when NO ALCOHOL between lunch and dinner

<ParticipantL> haha - I thought that was WWI

<ParticipantL> lame-o-rama

<ParticipantD> no alocohol between lunch and dinner?

<ParticipantL> pm alcohol is my fave

<ParticipantD> Wooh Glad I wasn't around then”

Presumably, the virtual timeframe of an IRC conversation is not extended enough to produce statements with ‘then’ like the one above. Due to the rarity of this construction it is impossible, however, to rule out an occurrence of ‘then’ referring to virtual time. Clarification of this point will depend on a more detailed study of the phenomenon which would be extremely time consuming without the aid of an automated parser which can add grammatical tags to the

corpus.

Summary

The analysis of the personal and demonstrative pronouns and determiners uncovered several findings. Firstly, the observations of prior research that first and second person pronouns are more frequent in CMC than in other language varieties receive further support from our study. Secondly, the presence of two dimensions of deictic reference could be ascertained for all deictic pronouns and determiners which are likely to be affected by it. Thirdly, irregularities in distribution could be explained by properties of IRC technology and channel specific interaction patterns. Once again, #philosophy was identified as a special case with conventions which are more conservative and closer to written standards than the other nine channels. Fourthly, the four channels with geographical names were found to display an interesting cluster of similarities in the distribution of non-standard spelling items.

Questions

Prior research (Ferrara et al. 1991) has identified a high frequency of questions, and yes/no questions in particular as typical for CMC and spoken interactions. In the analysis of questions our options are limited by the capabilities of the wconcord software. It is not possible to detect the presence of yes/no questions directly since they are produced through an inversion of subject and verb. Naturally, it is not feasible to manually check every subject and verb in the CIRC for such an inversion. Moreover, not even the analysis of question marks allows us gauge

the presence of questions. This signal of questions in traditional discourse is not obligatory in IRC conversation. As exemplified by the following it can be omitted:

“<senna73> r u there”

And it can be multiplied as in the example below:

“<ParticipantP> what are you people doing... Why aren't you at work???”

Furthermore, the letter ‘?’ cannot be examined using the concordancer’s word search function, because, following standard programming conventions, it stands as a wildcard for any letter. The means of quantitatively analysing questions in the CIRC left to us is then the analysis of wh-question words and the analysis of traditional answer items such as ‘ok’, ‘yes’ and ‘no’. These can be easily counted using wconcords word count function. Of course a caveat is in order since these items can also be used in various other functions such as conjunctions, exclamations, and relative pronouns²⁵. Therefore, they are only an indirect measure for the presence of questions.

Wh-questions

Table 14 below displays the results of the analysis of the wh-question words and figure 15 further down illustrates the proportions of the distribution table.

²⁵ The analysis of syntactical properties of IRC talk described further down reveals a low occurrence of subordinating conjunctions. This fact lends further support to the rationale of using wh-words as indicators of questions.

	how	where	when	what	why	who	which	whose
acmel	225	91	65	287	121	150	13	0
acper	201	111	97	317	100	149	13	0
dacan	229	152	122	368	135	145	31	2
dache	351	152	115	323	94	79	8	1
escyb	320	106	90	397	105	128	12	1
eshel	592	155	223	717	127	83	57	0
itaus	302	121	135	535	171	170	35	4
itphi	241	108	163	511	169	165	79	1
utbeg	370	137	73	283	78	93	32	0
utcha	352	188	76	335	137	164	6	0
total	3183	1321	1159	4073	1237	1326	286	9

Figure 15: Distribution of wh-words.

Figure 14: Distribution of wh-words, figures stacked to 100%

A comparison of the total numbers of occurrence of the wh-words in the CIRC with those of the Brown corpus yielded the following results. The items 'which' and 'who', which are frequently used as relative pronouns, were more frequent in the Brown corpus. Additionally the very formal word 'whose' and also 'when' occurred more often in the Brown corpus than in the CIRC. All the other wh-words are more frequent in the CIRC than in the Brown corpus. Even

though these observations are based on only one comparison corpus it seems safe to say in light of these proportions that the data from the CIRC supports earlier results that have found wh-questions to be of high frequency in CMC discourse. Some more support comes from the data of the four items available for comparison from the list of hundred most frequent words in the LLC corpus, namely 'when', 'who', 'which' and 'what'. All of these follow the same pattern as the CIRC i.e. 'what' has a higher frequencies in LLC than in the corpus of written English but the other three have lower frequencies. With the exception of 'which' their frequencies are still lower than the respective frequencies in the CIRC. It is clear that the high numbers of wh-questions in CMC stem from the interactivity in the discourse. Without a participating audience to answer the speaker's questions there is not much point in the asking except for rhetorical reasons. It would be worthwhile to compare the figures for the CIRC with those of a corpus of spoken English that, unlike the LLC, does not include non-interactive genres.

As for the channel subcorpora of the CIRC the patterns observable for the wh-word distribution are less and less obvious than the ones found for the deictics. Also, recalling the caveat mentioned above the interpretation of these frequencies must needs be more tentative than that for less ambiguous items. The general pattern of the above tables shows that 'what' and 'how' are by far the most frequent wh-question words in the CIRC. All the others have frequencies of less than a third of these two.

Both 'which' and 'whose' have very low frequencies and are therefore not suitable to serve as a basis for generalizations. There are several possible reasons for these low numbers. First of all, 'whose' is a word with a very formal connotation and therefore tends to occur in formal communicative contexts which are for the most part restricted to written genres.

Secondly, if the lack of syntactic subordination that researchers have found for other kinds of CMC also holds true for IRC, the numbers for ‘which’ would scarcely be inflated by its use as a relative pronoun. Thirdly, the kind of question that is introduced by ‘which’ usually presupposes the presence of a group of things or people from which the appropriate answer is chosen. For example the question: ‘Which car did you take?’ requires the respondent to know which cars you are talking about, i.e. the red, the blue; or the Chevy, the Ford etc. This sort of knowledge is dependent on the situational context which is generally not a detailed one in an IRC chat room. Therefore it might be that ‘which’ in questions does not occur very often in IRC because the necessary elaborate situational context requires too much work to create verbally.

Channel #help scores highest for ‘how’, ‘when’, and ‘what’. This again is in accordance with the finding that the interaction in this channel centers around information requests and responses. Below are two typical examples taken from #help:

“<ParticipantC> mine friend's master boot records are corrupted. how can he fix it”

“<ParticipantD> ParticipantC : try fdisk /mbr [...]

<ParticipantC> ParticipantD: what will that do?”

It is important however not to overinterpret these numbers. #beginner, for example has very low frequencies for most of the wh-question words although one would expect that newcomers to IRC would have frequent questions about the medium and its conventions.

Yes/no questions and response items

The other venue open to our exploration of questions is the indirect detection of yes/no questions through the analysis of the common response items to these questions namely ‘yes’ and

'no' and their various variants. The tables 16 below show the distribution of these items.

	yes	yeah	ok	yep	okay	k	aye	no	nope	nah
acmel	71	183	379	31	20	17	2	269	26	59
acper	70	172	117	35	22	26	42	332	26	52
dacan	101	159	171	26	28	24	1	327	35	11
dache	105	87	278	48	19	172	1	292	38	5
escyb	124	124	196	46	16	86	0	299	40	14
eshel	115	96	253	27	26	69	3	309	35	8
itaus	151	248	143	41	51	23	13	395	40	30
itphi	144	66	86	23	19	12	5	366	18	15
utbeg	84	69	148	23	28	41	0	325	22	1
utcha	141	115	195	39	43	56	1	267	48	12
total	1106	1319	1966	339	272	526	68	3181	328	207

Figure 16: Distribution of positive and negative response items.

Figure 17: Distribution of response items. Figures stacked to 100%

Surprisingly, the proportion of 'yes' and 'no' seem to indicate that the latter, and therefore a negative response, is much more frequent in the CIRC than the former. However, if we take into account variations of the positive and negative response such as 'yeah', 'ok', 'yep',

and 'nope' the emerging picture is different. When added up, the frequencies of the positive responses outnumber the negative ones, a finding which is in accord with the patterns observed in spoken language. In the LLC, for example, 'yes' is twice as frequent as 'no'. It is interesting however, that proportionately, the CIRC still has more negative items than the LLC.

It is possible that this can be attributed to the fact reported by prior research (Murray 1991b, Lunsford 1996) that certain politeness conventions have been weakened or rendered impractical in IRC. Maybe the trend to avoid negative responses when possible has been affected by the same mechanisms that are said to cause the decline in politeness standards. These are things such as anonymity, the need for quick typing or unreliability of connections.

Furthermore, research has found that there is "twice as much negation overall in speech as in writing" (Biber 1988: 245). This would indicate that the language contained in the CIRC is in this respect even less like written English than the language of the LLC, a result that makes sense considering the representation of non-interactive genres in the LLC, which renders the statistics for the whole corpus less representative of spontaneous spoken conversation.

As in our earlier discussions of personal pronouns the importance of non-standard spellings in IRC discourse cannot be stressed enough. 'Ok', whose frequency in the LLC is so low that it is not even part of the 100 most frequent words, has a frequency of almost 2000/million in the CIRC. Moreover, not only the frequency but also the range of non-standard forms far exceeds that of the classic corpora. For example, 'nope' occurs only once in the Brown corpus yet the CIRC contains over 300 instances of it. Other items, with quite low frequencies in the tables above, do not appear in the Brown corpus at all. An example would be 'nah', which is a phonological spelling of negative response items peculiar to spoken language. Only the

inclusion of the plethora of non-standard forms of the positive and negative response allows us to analyze the phenomenon in its entirety.

It is interesting that some of the non-standard forms, namely ‘yeah’, ‘aye’, and ‘nah’ display the same pattern we have already found with non-standard spelling forms of the deictics. In all three cases the channels with geographical names account for the vast majority of instances. However, there are others, such as ‘yep’, ‘nope’ and ‘k’ which do not follow that pattern. In fact ‘k’ has a lopsided distribution with almost a third of all cases concentrated in the channel #cheers whereas ‘yep’ and ‘nope’ show a fairly normal distribution.

It is interesting to note that ‘yeah’ is actually more frequent than ‘yes’. This is the opposite of the distribution of these items in the LLC and can therefore not be attributed to a difference between spoken and written English. It might be due to the fact that the Internet population and Internet technology are still most heavily influenced by North American culture in general and US American culture in particular. This is also supported by the fact mentioned above that the channels with geographical names are at the forefront of the frequent usage of ‘yeah’. It would be worthwhile to study if this trend is reversed in channels with a clearly British focus.

When analysing the low end of the distribution #philosophy stands out. It scores very low for all items except ‘yes’ and ‘no’. In this case it reveals its special conventions not through the overuse of a certain form but rather through the avoidance of all non-standard variants. Another interesting case, but at the opposite end of the spectrum are the channels #melbourne and #perth in which ‘yes’ is used only rarely. This lends further support to the hypothesis that it is these two channels which have evolved the furthest away from standard written English usage.

The grammar of these response items is identical to their usage in spoken English. They often comprise a turn of their own. In IRC that means they are the only word in a message. Otherwise they are followed by a pause. In IRC this is signaled through the use of the three dots or the exclamation mark. The following examples should illustrate the syntax of ‘yeah’ as an example of a response item:

“<ParticipantT> yeah...all of them!!! “

”<ParticipantG> participantm: Can you open the bleem file up with a text editor and look at the file header? [...]

<ParticipantM> yeah”

Summary

With the means of the present study it is not possible to positively verify the claim made in prior research that questions are more frequent in IRC than in other more conventional genres. However, the indirect analysis of wh-question words and response items indicates that these are more frequent in the CIRC than in the comparison corpora. Further evidence was found for the role of #philosophy as a linguistically more conservative channel and for the role of #melbourne and #perth in the linguistic vanguard. Moreover, additional distribution patterns that identify the channels with geographical names were uncovered.

Hedging

The so-called hedges, ‘sort of’ and ‘kind of’, have been identified in prior research

(Murray 1991a, Collot & Belmore 1996) as items typical for spoken language which also occur frequently in CMC discourse. The analysis of the hedges was complicated by the fact that ‘kind of’ and ‘sort of’ can also be used in constructions such as: ‘what kind of music do you like?’ where they are not hedges at all. However, due to the relatively low frequencies of hedges, compared to personal pronouns for example, it was decided to manually remove all instances where ‘kind of’ and ‘sort of’ were used that way from the lists. The table below shows the distribution of occurrences of hedges in the CIRC.

	sort of	kind of	sorta	kinda
acmel	2	4	1	9
acper	1	3	5	9
dacan	1	6	5	17
dache	3	0	2	14
escyb	2	4	1	15
eshel	6	3	2	15
itaus	6	2	2	29
itphi	8	7	5	17
utbeg	1	3	2	18
utcha	1	2	2	17
total	31	34	27	160

Figure 18: Distribution of hedges in the CIRC.

Hedges are another case in point for the importance of considering the non-standard spelling forms in the analysis of the CIRC. In this case, an examination of the standard forms ‘sort of’ and ‘kind of’ would miss the vast majority of hedges in the CIRC. The non-standard spelling variants ‘sort’ and ‘kinda’ taken together are, in fact, three times more frequent than the standard forms. This is the first time that we have found a non-standard spelling form that outnumbers the standard form it is derived from. This phenomenon will be examined in more detail in the section on abbreviation.

When looking only at the standard forms both hedges seem to be equally popular when

in reality the two forms based on the word 'kind' are much more frequent. It was impossible to verify if this relationship between the two forms also exists in the Brown corpus because the corpus was not available to me in computerized form so that the collocation could not be checked. Due to the low frequency of the individual items a complete channel by channel analysis of the distribution was not undertaken, since the smallest of idiosyncracies would skew the results. It is interesting to note, however, that #melbourne and #perth are not leading the field in the non-standard spelling of hedges. In fact, the contrary is the case.

Contractions

Contractions with 'not', 'had', 'would', 'is', 'has' and 'have' are accepted only in informal writing and are typical for spoken discourse. Not surprisingly, since CMC is similar to both these varieties, they have been identified in prior research on CMC (Collot & Belmore 1996) as being of high frequency. In the standard corpora there are two forms taken into account when analysing contractions: the long form, e.g. 'have not', and the contracted form 'haven't'. In the CIRC there is a third variant, which has been mentioned briefly above in the discussion of the variants of the personal pronoun 'i'. This is the contracted form with the apostrophe omitted as in 'havent'.

Contractions with 'not'

Figure 19 below shows frequencies of the three contraction variants of 'not' with the respective verbs. Verbs such as 'might', 'need', and 'had' which also allow contraction with 'not' but whose frequencies were so low that there were no instances for all three variants are not

included in the table²⁶.

	is	are	was	were	has	have	can	will	would	should	could
not	317	107	20	2	4	21	12	25	14	10	11
n't	220	74	85	10	42	126	472	131	100	34	67
nt	173	65	69	7	16	98	405	159	68	21	34

Figure 19: Contractions with 'not'. Absolute frequencies.

	do	does	did
not	107	65	32
n't	1546	270	322
nt	1291	162	263

Figure 21: Contractions with 'not'. Absolute frequencies.

Figure 20: The three variants of contractions with 'not'. Numbers stacked to 100%.

²⁶ In the special case of 'can' the figures for both 'can not' and 'cannot' were collapsed into one.

Figure 22: The three variants of contractions with 'not'. Numbers stacked to 100%.

Illustrations 20 and 22 above clearly show that in the vast majority of cases the uncontracted form of 'not' is dispreferred in the CIRC. The only exceptions are the items 'is' and 'are' in connection with which 'not' occurs more often than its contracted variants. However, this does not alter the fact that the general conventions of 'not' contraction in the CIRC follow those of spoken language and very informal writing. It is also interesting to note that the IRC specific form, 'nt', unlike the uncontracted form, never ranks below 20%. This goes to show that the convention of dropping the apostrophe in the 'not' contraction is firmly established in IRC talk. Considering such widespread use, it is surprising that this trend has not been mentioned in prior IRC research.

Since the analysis of the non-standard spelling variants so far has been very consistent in

showing that certain channels are more progressive than others in this respect, a channel by channel comparison for the standard and IRC specific contraction variants was performed. To this end the frequency data for all 'n't' and 'nt' variants was extracted for the ten channels. Then these frequencies for the different 'not' contraction collocations such as 'didn't' etc. were totaled. The results are displayed in the pie charts 23 and 24 below.

"nt" distribution by channel

Figure 24: Distribution of the IRC specific 'not' contraction by channel.

"n't" distribution by channel

Figure 23: Distribution of the standard 'not' contraction by channel.

Similarly to the majority of standard vs. non-standard forms discussed so far, #philosophy proved to be the leading channels for the more conservative variant. For the IRC specific contraction with the omission of the apostrophe, the four channels with geographical names had four of the five biggest slices. #perth and #melbourne are again the highest scoring channels in this category.

Contractions with 'have'

The contractions of 'have' are less frequent than those with 'not' discussed above.

Similarly to the 'not' contraction, there is an IRC specific variant of the 'have' contraction which omits the apostrophe. The following figures display the frequencies for the three variants of contraction with 'have' in the CIRC.

	i	you	u	we	they
have	1073	454	132	118	92
*'ve	319	54	6	16	10
ve	129	10	4	3	3

Figure 26: Contraction with 'have'. Absolute frequencies.

Figure 25: Contractions with 'have'. Numbers stacked to 100%.

Surprisingly, the pattern of distribution for 'have' contraction is nothing like that for 'not' contraction. In all instances, the uncontracted form is by far more frequent than the contracted variants. The reasons for this remain unclear. It is interesting to note, however, that the IRC specific contraction variant is always present although the contracted form is strongly

disfavoured overall. It is interesting to note that the non-standard spelling variant ‘u’ seems to strongly discourage the contraction. However, the item’s frequency is too low to determine if this is a general trend or just a quirk of the data.

Contractions with ‘is’ and ‘has’

Since the contracted forms of ‘is’ and ‘has’ share the same graphematic representation alone and are therefore indistinguishable by the concordancer, they will be analyzed together. The tables 27 below display the results of the frequency analysis for the contractions with ‘is’ and ‘has’:

	he	she	it	that	who	what
has	34	48	67	23	31	2
is	104	83	319	210	76	233
's	177	161	1134	616	88	304
s	83	95	1324	816	58	484

Figure 27: Frequency distribution of contractions with ‘has’ and ‘is’.

First of all, it is clear that the long form of ‘has’ is less frequent than the uncontracted form of ‘is’. It can therefore be assumed that the contracted forms with and without the apostrophe have similar proportions and therefore are more often than not contractions of ‘is’. In all cases, the standard contracted form with the apostrophe is more frequent than the long forms. In fact, for all items except ‘who’, the contracted form is more frequent than the uncontracted forms taken together. Therefore, it is clear that overall the contraction of ‘has’ and ‘is’ follows the pattern established by ‘not’ contractions and not the one for contractions with ‘have’.

Another interesting aspect of table 27 above is the distribution of the non-standard

contracted form. The case of ‘it’ must be excluded here because ‘its’ collapses both the contracted variant of ‘it is’ and the possessive determiner. For the other cases we find that the contraction without the apostrophe is less frequent than the standard contraction except for the combinations with ‘what’ and ‘that’ where standard contractions are outnumbered by the IRC-specific ones. Incidentally, these two items are also the ones with the most letters. It is possible that the likelihood of abbreviation of the standard contraction by dropping the apostrophe is proportional to the length of the word that is created through the contraction. However, the same trend was not found for the other contractions examined thus far. It is possible that the length of the contracted item, i.e. two letters in ‘nt’ and ‘ve’ versus one letter ‘s’ plays a role in this too. This phenomenon will be analysed in more detail when we examine abbreviations in the CIRC in chapter 7.

Contractions with ‘will’

The contractions with ‘will’ are the least frequent of the several contraction types investigated so far. Figure 28 below shows their frequency distribution.

	I	you	u	he	she	it	we	they	that
will	250	95	38	17	11	97	36	24	39
*'ll	449	96	5	21	13	17	51	17	23
ll	270	15	1	n/a	n/a	3	n/a	4	6

Figure 28: Distribution of the three contraction variants with ‘will’

From the results of the table above there emerges no clear pattern of preference for either the standard contraction or the non-standard variant. This is probably due to the low overall frequencies of these contractions, which make it more likely for idiosyncratic usages to

significantly influence the distribution pattern. Moreover, three forms of the non-standard contraction yielded no applicable results because they are homograph with the words ‘hell’, ‘shell’, and ‘well’. Note that the non-standard spelling ‘u’ seems to disfavour the contraction with ‘will’. This is the same phenomenon we already noted with the contraction of ‘have’ and ‘u’. It could be that a double abbreviation within one word is too complex to be easily understood and therefore not frequently used. On the other hand, this would run counter to the frequently mentioned (Lundstrom 1995, Werry 1996, Wilkins 1991, Murray 1991a) trend for abbreviation.

Contractions with ‘would’ and ‘had’ were not examined in the present study because their frequency is even lower than that of the contractions with ‘will’. As was mentioned above, these frequencies were already too low to establish patterns of distribution.

When comparing the frequencies of the contractions discussed above with the comparison corpora, several interesting trends are revealed. Unsurprisingly, the CIRC’s frequencies of contraction are far higher than the Brown corpus’. However, when comparing the CIRC to the contracted items from the LLC, which are among the list of the 100 most frequent words, this pattern is reversed. The contractions entering the comparison are ‘it’s’, ‘don’t’, ‘I’ve’ and ‘that’s’. When we include the non-standard spelling form and add the frequencies for the CIRC the LLC items still score higher with the exception of the contraction ‘that’s’. From these results it is clear that the conventions for contractions in IRC are very similar to the ones for spoken language. This supports the findings of prior research. Moreover, the above analysis yields more evidence for the phenomenon discovered for contractions with ‘is’ and ‘has’, namely that the longer the word that the contraction is adjoined to, the more likely it is to trigger the

contraction.

Dialects

One of the linguistic characteristics that is much more salient in spoken than in written language is the fact that speakers' origins are often betrayed by the presence of dialectal features in their speech. As mentioned in the methodology section, it is close to impossible to identify the socioeconomic background of the participants in IRC. Therefore a study of social dialects in this medium is problematic. While linguistic markers of socioeconomic class such as 'yous' instead of 'you' are certainly present in IRC, there is no way of meaningfully correlating them with specific IRC communities, i.e. channels.

The existence of channels with geographical names, on the other hand, allows us to identify with some confidence gatherings of people who share some sort of common geographical background. As discussed earlier, the region after which the channel is named need not necessarily be the place of physical location of the participants in that channel. However, it can be assumed that there exists some connection between the people, the channel they join, and the name of that channel.

The existence of geographical dialects in IRC channels has been documented in the first of my pilot studies (Doell 1997) and in Werry (1996). Furthermore, in the course of the present analysis, a trend of the geographical channels to share certain speech-like qualities has been found as well. Moreover, some clearly geographical variation could also be detected at the lexical level. The table below shows the frequencies of some words which are characteristic of a

national or regional variety.

	eh	mate	wog/s	heap/s	bloke
acmel	10	14	25	15	2
acper	43	24	2	10	5
dacan	107	0	0	1	0
dache	13	3	0	0	2
escyb	8	12	0	0	0
eshel	11	0	0	1	0
itaus	30	18	0	4	1
itphi	12	8	0	1	1
utbeg	20	3	0	0	0
utcha	15	7	0	1	1
total	269	89	27	33	12

Figure 29: Distribution of words characteristic for a geographical dialect.

The Canadian ‘eh’ is the most frequent of the above lexical items and therefore its heavy concentration in the channel #canada provides the strongest proof for the existence of geographical dialects on IRC. The other four words are commonly associated with Australian English and their distribution shows that this is reflected in the linguistic communities of IRC channels. One item, namely ‘wog’ occurs almost exclusively in the channel #melbourne. The Concise Oxford Dictionary 8th edition has two entries for ‘wog’. The first describes it as a derogatory slang term for non-white foreigners. The second entry explains ‘wog’ as an Australian slang term for illness or disease. The online version of the Cambridge International Dictionary lists only the first meaning but additionally identifies it as an Australian slang term. This matches our data because the instances of ‘wog’ in the CIRC are localized in an Australian channel but seem to have the meaning of the first entry. Moreover, the derogatory quality seems to have been replaced by a positive one. The following examples illustrates this usage:

“<ParticipantM> any bombed wogs msg me”

“<ParticipantY> i just discovered that wogs hate americans”

What is more, 'wog' is also used in nickname formation for creations such as 'coolwog' or 'wildwog' which certainly does not indicate a derogatory meaning. It is not clear if this new meaning of the word is IRC specific, or Australian in origin.

Differences between geographical dialects in IRC are not only realized through usage of certain lexical items but also through spelling. Table 30 below shows the frequency distribution of some common words with regional spelling variants.

	favour	favor	colour	color	night	nite	night/nite	realise	realize
acmel			2	2	26	8	3.25	3	1
acper	2		4		59	25	2.36	3	
dacan			5	2	42	19	2.21		3
dache	1		1	4	105	109	0.96		3
escyb		2		3	71	87	0.82		1
eshel		4	1	22	15	16	0.94		2
itaus	2		7	3	62	14	4.43	1	
itphi			2	9	30	12	2.50	8	5
utbeg			3	17	62	42	1.48		2
utcha				6	70	46	1.52		
total	5	6	25	69	542	378	1.43	15	17

Figure 30: Distribution of lexical items with regional spelling variants.

With the exception of 'night', all the items in the table above have such low frequencies that one look suffices to reveal that the channels #melbourne, #perth, #australia and #canada generally favour the British spelling. The other channels whose population, similarly to the Internet as a whole, is presumably more influenced by US American linguistic standards and spelling conventions.

The strongest case for this can be made with the item 'night' which has high enough frequencies to warrant a more detailed analysis. For this purpose the ratio of the standard British

versus the non-standard spelling, which is generally more acceptable in the US²⁷, was calculated. On average the British spelling variant was 43% more likely than the non-standard variant. The five channels with a significantly higher incidence of the standard spelling were #melbourne, #perth, #canada, #australia, and #philosophy. It comes as no surprise that #philosophy, which we have found to be conservative when it comes to spelling conventions, appears in this group. The channels with geographical names, however, are another matter altogether. The analysis of personal and demonstrative pronouns had shown them to be very susceptible to non-standard spelling variants. It appears that in the case of 'night' the influence of regional linguistic norms overrides the usual preference for non-standard spelling in these channels.

Ellipsis

Various forms of ellipsis have been mentioned by researchers working on CMC (Ferrara et al. 1991, Murray 1991, Lundstrom 1995, Werry 1996, Doell 1997). Omission of subject pronouns, articles, auxiliaries, and finite forms of the copula have all been well documented by this research. Unfortunately, the present study, which uses computer software to search for strings of letters in a database, is not capable of detecting ellipsis which, by definition, is an absence of words. However, the presence of ellipsis in IRC discourse is so pervasive that the analysis of a brief passage from the corpus can yield examples for all of the omissions reported in prior research. Below are a few examples of ellipsis:

²⁷ The item 'nite' is listed as a variant of 'night' in the Merriam Webster's dictionary and the Cambridge Dictionary of American English but not in the Cobuild Student Dictionary, Oxford Dictionary, and Cambridge International Dictionary of English.

“<ParticipantM> anyone from washington state??”

“<ParticipantL> ParticipantH, heard from ParticipantS”

“<ParticipantK> what ya been up to ? [...]

<ParticipantL> not much, just haven't been on line and working two jobs”

In the following order the examples contain omissions of the copula ‘is’, auxiliary ‘have’ and subject pronoun ‘you’, auxiliary ‘have’ and subject pronoun ‘I’. The frequent omission of the pronoun ‘I’ has already been touched upon during the discussion of the personal pronouns. The three examples are all selected from a short 60 message sample chosen from the CIRC at random, yet they contain all but one of the omissions reported in prior research. This attests to the widespread use of ellipsis in IRC.

Colloquial Usage

Colloquial usage has been reported by prior research (Ferrara et al. 1991, Murray 1991a) to be a characteristic feature of CMC discourse. Unfortunately, colloquial is a rather fuzzy concept. Ferrara et al. (1991) quote expressions such as ‘let’s’ and ‘how about’ as examples. Murray does not elaborate on what she means by “informal diction”. It seems that what both authors are trying to capture under the heading colloquial are words, expressions and constructions which are considered typical of spoken rather than written language and which are appropriate to informal rather than formal speech situations. The number of constructions and vocabulary items in the CIRC which could be subsumed under this catch-all category is virtually

limitless. Therefore, I will describe just one example in some detail and briefly mention two others which could be added here but which fit more appropriately under another heading.

A prime example of colloquial, i.e. informal spoken language is the use of expletives. Research on CMC, especially from authors who explore the potential value of CMC technology for distance education (Kelm 1992), has often documented and deplored the extensive use of expletives, or ‘flaming’, in CMC discourse. A different approach was taken by Stenström (1991) who analyzed the use of expletives in the London Lund Corpus of spoken English. Her analysis focuses on the differences in usage between male and female speakers, syntactic characteristics and discourse functions of expletives. In her conclusion Stenström points out that the data from the LLC that her analysis was based on was collected in the 50s and 60s from educated speakers of British English. Therefore, she stresses the need for a comparison of her results with more recent data from other, for example younger, subjects. The present study cannot attempt a detailed examination of the usage of expletives in IRC along the lines of Stenström’s analysis. However, a frequency comparison of the most frequent expletives in the LLC with my data indicates the importance for comparative studies that incorporate more recent data. The table below shows the frequencies of some expletives in the LLC and the CIRC.

	god	gosh	goodness	bloody	damn	hell	fucking	ass	shit
LLC	75	30	18	21	4	9	9	2	2
CIRC	236	15	7	57	402	244	147	273	392

Figure 31: Frequencies of some expletives in the CIRC and the LLC.

It is clear that there is a huge disparity between the expletives used in educated British speakers’ conversations of the 50s and 60s and the computer mediated interactions of the 90s. Some of the disparity can surely be attributed to the fact that the LLC also contains non-

interactive speech, but the shift in usage is not only quantitative but also qualitative. Items with what Stenström calls strong expletive force, such as 'shit' and 'fucking' have registered big gains in popularity whereas words with weak expletive force, such as 'gosh' have stagnated or fallen into disuse. In part this apparent change in the strength of the expletives is probably due to regional variation. Since the North American influence is very strong on the Internet, some items like 'bloody' and 'gosh', which are more popular in British English, can be expected to have proportionately lower frequencies in the CIRC than in the LLC. The frequent use of expletives which are very rare in traditional written genres is another case in point which proves the similarity of IRC discourse to spoken conversation.

There are two more examples of colloquial usage which can be mentioned at this point. Firstly, use of the colloquial forms 'wanna' and 'gonna' instead of 'want to' and 'going to' is very widespread in IRC. In fact, the colloquial forms consistently outnumber the more formal variants. These items will be examined in more detail in chapter 7 in the section on abbreviation. Secondly, the occurrence of the get-passive which is widespread in spoken English but strongly stigmatized in written English (Miller & Weinert 1998) can definitely be regarded as a form of colloquialism in IRC. It has not been mentioned in prior research on CMC. This construction will be analyzed in more detail in the section on syntactical complexity.

Syntactical complexity

Subordination and coordination

Many studies have examined syntactical complexity and its relation to written and

spoken genres. Clausal subordination and coordination have been key concepts in this ongoing research effort. Mostly, it was claimed that subordination of clauses signaled greater syntactical complexity and was therefore typical for written genres. On the other hand, researchers such as Halliday (1985) found more subordination in spoken than in written texts. The results for CMC are also conflicting. Ferrara et al. (1991: 24) report “frequent use of relative clauses, adverbial clauses, and subordination” in their corpus but they do not supply the actual frequencies. My pilot study (1997) on the other hand found a low incidence of these same constructions. Biber (1988) stresses that “subordination is not a unified construct, [...] different types of structural dependency have different discourse functions, and [...] particular subordination features are therefore used to different degrees in different types of discourse” (Biber 1988: 229).

Unfortunately, it is impossible without grammatical tagging of the corpus to isolate all the different types of subordination. For example, infinitives, that-clauses, or participial clauses cannot be extracted from the corpus and quantified with the wconcord software. The one feature that can be analysed with the concordancer are conjunctions. Table 32 displays the frequencies for the most common coordination and subordination conjunctions in the CIRC, the LLC and the Brown corpus²⁸.

	and	or	because	although	though	tho	if	unless
CIRC	8365	2262	632	28	313	155	2368	70
LLC	12699	1658	1249	n/a	n/a	n/a	1764	n/a
Brown	28852	4207	883	319	442	1	2199	101

Figure 32: Frequencies for common coordinating and subordinating conjunctions in the CIRC and the comparison corpora.

Predictably, considering the complexity of the problem as explained above, no clear

²⁸ Some of the figures for the LLC were not available.

pattern of distribution emerges. When compared to the LLC and Brown corpus the CIRC seems to contain surprisingly high numbers of subordinating conjunctions, which runs counter to previous findings (Doell 1997). However, the results are in accord with the finding Biber (1988) reports, namely that adverbial subordinators such as 'because', 'if', etc., co-occur with features he has identified as typical for interpersonal and reduced-content type interactions. Such interactions are typically spontaneous casual conversations.

Passives

The frequent occurrence of the passive has often been used as an indicator of syntactical complexity. The detection of passives with the word search function of the wconcord software is impossible because it would require the identification of all verbs in the corpus. Therefore the analysis will have to rely on estimates. For this purpose all forms of the auxiliary 'be' and the word 'get' for respectively the be-passive and the get-passive were extracted from the corpus. Then 120 instances of each item were analysed to determine if they were part of a passive construction. Note that the system messages 'is now known as' and 'was kicked by' were not counted as passives in the analysis. The percentage of passives found in the sample was then applied to the total number of occurrences of 'be' and 'get' to arrive at the following estimates for the occurrence of passive constructions in the CIRC. Of course, these estimates are educated guesses rather than statistical proof. The be-passives outnumber the get-passives by approximately 3000 to 400. The get-passive is less frequent proportionately in the CIRC than in spontaneous spoken language for which Miller and Weinert (1998: 88) report more get-passives than be-passives. However, the get-passive is also much more frequent in the CIRC than it is in

traditional written genre where its use is strongly stigmatized and highly restricted.

Greeting and Leave-taking

Greeting

My first pilot study (Doell 1997) compared in detail the greeting and Leave-taking routines in several IRC channels. Some particularities of these routines were also briefly mentioned by Werry (1996). The present study follows up on this previous work, although the inter-channel differences are likely to be less pronounced in the CIRC than in the pilot study because the CIRC's channels are all English speaking. The table below shows the distribution of various words commonly used in greeting.

	hi	hello	hey	hiya	heya	rehi	re w/o hi	total
acmel	435	228	390	26	29	0	0	1108
acper	227	97	421	22	63	0	0	830
dacan	545	238	411	69	47	0	0	1310
dache	1126	439	480	285	15	1	2	2348
escyb	899	288	670	169	89	0	5	2120
eshel	190	101	100	9	3	0	5	408
itaus	281	172	605	34	71	0	5	1168
itphi	239	102	88	5	6	1	20	461
utbeg	981	582	330	337	16	0	1	2247
utcha	1276	585	446	51	53	1	2	2414
total	6199	2832	3941	1007	392	3	40	14414

Figure 33: Distribution of terms of greeting in the ten channels.

The most frequent greeting term is 'hi', which is another case in point to prove the strong influence of the North American variety of English on the Internet. What is particularly striking in the table above are the overall very high frequencies of these items in the CIRC. The

first three, 'hi', 'hello', and 'hey', with frequencies of 2800 and higher are all part of the 100 most frequent words in the CIRC. In the LLC they are not part of that list, i.e. they occur less than 740 times and in the Brown corpus they occur six, ten and fifteen times respectively. There are several factors which contribute to this important role of greeting routines. Firstly, the constant coming and going of participants leads to increased greeting compared to more the more static conversational groups which tend to form in traditional face-to-face interaction. Secondly, the conversational groups in the channels that were logged are larger than they are in real life. Twenty people attending a party could never participate in one single conversation. Thirdly, in IRC, greetings are the easiest and most accepted way for newcomers to establish contact with the rest of the group when they want to join the conversation. In real life conversations, the same usually requires an introduction of the newcomer by some third party. Fourthly, some people in IRC channels run so-called auto-greet scripts. These are small software programs which are integrated by the users into their IRC-client software and which send an automated greeting messages to every user who joins the channels.

The items 'rehi' and 're' are IRC specific items with the meaning 'hi again' which have been discussed by Werry (1996). He reports the 're' "was first used on various English-speaking channels. Its usage has since declined on English-speaking channels (although variant forms such as 'rehi' occur reasonably often)." The CIRC data does not support Werry's statement, since 're' is much more frequent in the CIRC than 'rehi'. However, the items' overall frequencies are too low to warrant a generalization of that usage.

When analysing the greeting items' total frequencies as displayed in column nine of table 33 above, one can distinguish three distinct groups of channels. The informational channels

#help and #philosophy have very low frequencies. The channels with geographical names have frequencies of around 1.000 greeting items per channel and the general chat channels show the highest incidence of greeting items with over 2.000 occurrences each. Thus the groupings whose existence has been established by the examination in the course of this study of other linguistic features, receive further support from the distribution of greeting items. It is logical that in channels where information exchange is the prior concern of the participants, greeting routines with their function of establishing social contacts are not as commonly used as in social channels. In the case of the channels with geographical names, the relatively low occurrence of greeting items lends support to the two hypotheses advanced during the discussion of the personal pronouns, namely that there exist stronger in-group ties in the geographically named IRC communities and that there is less participant turnover in these channels. Less turnover results in fewer newcomers who need to be welcomed, and stronger familiarity between group members might lead to the abandonment of formalities such as greeting.

It is interesting to note that the IRC community has developed greeting expressions that are slightly different from those which can be observed in face-to-face interactions. Collocation analyses of the three most frequent greeting items reveal that they often combine with words that refer to an indefinite number of other persons such as 'all', 'everyone', 'people', and its abbreviated form 'ppl'. Another item, the informal 'guys', only collocates with 'hi' and 'hey'. Typical collocations for greetings are therefore expressions such as 'hello all', 'hi ppl', 'hey guys'. This probably reflects the participants' need to clarify that their greeting extends to the whole multitude of users present in the channel.

Leave-taking

Leave-taking routines in IRC are both less vital and less frequent than greeting. The main reason for this is probably the instability of network connections which results in people been disconnected from the conversation without prior warning. When these connection problems occur between IRC servers the result is a so-called 'net split'. When it happens, it looks as if half the people in an IRC channel have just left simultaneously. Due to these aspects of IRC technology, a failure to engage in leave-taking routines is not regarded as a breach of netiquette. This is reflected in the frequencies of leave-taking items displayed in table 34 below.

items	cya	c u	c you	cu	c ya	see you	see u	see ya	bye
frequency	297	8	0	7	33	74	21	166	1332

Figure 34: Frequencies of several leave-taking expressions.

None of these items comes close to the frequencies of the greeting routines. Even the figures for the most frequent item 'bye' are slightly inflated due to the fact that wconcord counts both occurrences of 'bye' in the expression 'bye bye'. As with other linguistic features, it is vital to include spelling variants in the analysis. Among the leave-taking items, the non-standard spellings with 'ya' are consistently more popular than standard spellings e.g. 'See ya' is more frequent than 'see you', 'c ya' is more frequent than 'c you' etc. Furthermore, a channel by channel analysis of the most frequent item 'bye'²⁹ revealed that the pattern found in the distribution of greeting routines also applies to leave-taking, namely, the informational and geographical channels have less leave-taking routines than the others.

²⁹ See appendix I for the distribution figures for 'bye'

CHAPTER 6

ERSATZ - FEATURES OF IRC TALK WHICH ATTEMPT TO SIMULATE SPEECH

Emoticons

Emoticons have already been mentioned briefly in the methodology section. At this point a brief restatement of the definition of emoticons is useful. Emoticons are short ASCII glyphs which symbolize a human face and which can be read by tilting one's head 90 degrees to the left or right. They are used to convey information such as irony or doubt which is mostly transmitted through kinesic and paralinguistic cues in face-to-face conversation. Since these features of speech do not exist in the context-reduced situation of an IRC channel, emoticons are used to replace them and compensate for this lack. Emoticons are a feature that is unique to the language variety that has emerged through the use of computers or, as Collot & Belmore (1996) call it, electronic English. Their use in synchronous CMC has been widely documented (Lundstrom 1995, Werry 1996, Doell 1997). Emoticons come in such an astounding variety that a complete inventory of the ones used in the CIRC could not be undertaken here. Instead the analysis will focus on establishing a typology of emoticons and the patterns of usage and distribution of some of the most frequent ones.

Emoticons come in two varieties: the ones for whose interpretation one has to tilt one's

head to the right e.g. (-: , and the ones such as :-) where you tilt your head to the left in order to read them. Accordingly, emoticons can be divided into a left bend and a right bend category.

The next distinction can be made on the basis of the two basic ingredients of an emoticon. These are a symbol for the eyes and one for the mouth of the face which is to be represented by the glyph. The schematic drawing below illustrates the different categories in this typology and gives a few examples.

Figure 35: Schema of the typology of emoticons.

Using the schema above, emoticons can be grouped into definite categories such as all left bend emoticons using the standard bracket as a mouthpiece. Examples of that category would include the icons: :) ;) =) :-) ;-) 8-) 8) and so forth. The combinatory variations of the emoticon schema are only limited by the IRC users' imagination and the number of keys on the computer keyboard. In addition to mouth and eye piece variation various optional features can be added to the face. These can be features such as nose, chin, mustache, hair, headdress etc. To

illustrate some of the existing variety, a list of some of the emoticons encountered in the CIRC is provided in appendix J. Although the list includes almost exclusively left bend emoticons with colons and semicolons as eye pieces, it contains nearly 100 items. For a more extensive list of emoticons with explanations of their meanings refer to Lundstrom's appendix (1995).

An aspect of emoticons that has been neglected by research on CMC is their pattern of usage, the grammar of these lexical items so to speak. Unlike the kinesic and paralinguistic cues they replace, emoticons cannot occur parallel to the message they accompany. The linearity of writing in IRC technology prevents it. The question is: at what point then does the emoticon occur? Does it precede the message whose meaning it modifies? Is it inserted in the middle or does it postmodify the items it refers to? Our data show that IRC users strongly prefer to use emoticons at the end of their messages. For example, :), the most frequent emoticon in the CIRC occurs 5300 times. In 4998 instances or 94% of cases its position is at the end of message. The second most frequent emoticon, ;), shows the same positional distribution. Of 1052 occurrences 1014 or 96% are at the end of a message. This pattern has also been identified in prior research (Lundstrom 1995). In those few instances where the emoticon does not occur at the end of the message, it is inserted at various syntactical boundaries. Usually it is inserted at the end of a sentence or phrase. The examples below illustrate this usage pattern of the emoticon :) :

"<ParticipantD> oh but it does in my dreams :) trust me "

"<ParticipantP> hows ParticipantT today? :)

<ParticipantT> well, I've just come back from having a lovely meal thanks ParticipantP
:) u?"

There are very few cases of the emoticon occurring at the beginning of a message. For

example, for :) only 83 instances or 1.5% of the total follow this pattern. Usually these messages consist only of one or two words beside the emoticon. Below are some examples:

“<ParticipantZ> :) yummy”

“<ParticipantT> :) tn timer sasha “

As with other linguistic features discussed already, a channel by channel analysis of emoticons' distribution patterns was undertaken. Only the most frequent emoticons were included in the analysis. Where possible, both the left bend and the right bend variant were analysed. Table 36 below summarizes the results of this analysis.

	:)	same rb	;) same rb	:(	:o)	:P	:-)	same rb	
acmel	533	1	44	44	12	202	5		
acper	241	1	54	33	10	231	2		
dacan	378	1	59	21	52	66	23		
dache	468	1	104	42	4	12	9		
escyb	508	5	83	37	16	92	6		
eshel	293		60	28	7	41	22		
itais	557	4	95	42	41	229	5		
itphi	667	44	290	6	14	33	15	1	
utbeg	821	1	122		29	16	8	13	
utcha	834	1	141	1	63	149	57	2	
total	5300	59	1052	7	353	344	971	102	1

Figure 36: Distribution of the most commonly used emoticons and their right bend (rb) variants.

The most striking feature of the table above is the distribution discrepancy between left bend and right bend emoticons. The data shows that right bend icons are largely dispreferred in IRC. The few instances of these emoticons are concentrated in the channel #philosophy, once more proving that channel's special status within the corpus.

Another notable fact is the predominance of smileys i.e. emoticons with a smiling mouth piece, versus frownies. This follows the same pattern that we found for the positive and negative response items. Moreover, it attests to the generally friendly atmosphere prevalent in IRC

communities.

Lastly and most importantly, the table above shows that the distribution of emoticons in the ten channels does not display a distinctive pattern like greeting items. The frequencies of emoticons in the ten channels are distributed over large ranges between 4 and 149 for :-o, between 8 and 231 for :p, between 44 and 290 for ;). This failure to find an emergent pattern can be attributed to the emoticons' special characteristics. As explained above, they are a recent linguistic phenomenon which is largely restricted to CMC. Therefore, they have not yet acquired distinctive connotations of formality or geographical background along which lines we have found the distribution patterns in the CIRC to emerge. However, distributional variation for emoticons between English and German-speaking channels has been reported (Doell 1997).

Action messages

Action messages which have been addressed briefly in the methodology section serve to convey contextual information which is given by the environment in a face-to-face interaction. Such information can be the description of complex actions undertaken by the user or it can simply provide paralinguistic cues such as grinning or nodding. Action messages are clearly a feature of IRC talk which replaces a characteristic of speech which is unavailable in the interactive written medium. They have also been widely recognized as a characteristic feature of IRC talk in prior research (Reid 1991, Bays 1998, Werry 1996, Doell 1997).

Originally, users manually created action messages and inserted asterisks in front and sometimes behind them. Soon they became so popular among IRC users that the IRC client software now provides a special command for their creation which has been explained in the

methodology section. However, some users still prefer to manually insert the action message into their discourse. Below are examples of one action message created with using the software command and one which was created manually:

“* ParticipantS sticks her tongue out at you”

“<ParticipantC> *jumps on ParticipantR”

Later on, the ever growing popularity of action messages resulted in the creation of preprogrammed action messages which can be accessed in the IRC client program with a simple mouse click. Naturally, since the interaction takes place in a virtual locale, all the activities described in action messages need not have a real world equivalent. In fact, some action messages describe action which would be impossible or highly improbably in a real world setting. This is illustrated by the action message below which, also, is the most frequent preprogrammed action message.

“* ParticipantF slaps ParticipantM around a bit with a large trout”

Note that action messages similar to the ones shown in the examples above are almost exclusively composed in the third person singular. This creates an illusion of detachment between the subject of the message and the author of the message which renders the description more vivid and believable. This process is similar to the creation of a narrator in literature which serves the same purpose.

Unfortunately, it is not possible to accurately analyse the use of action messages in the CIRC quantitatively channel by channel. This is due to the fact that the appropriate wconcord word search function interprets the asterisk as a wildcard for any number and kind of letter. However, the fact that the single asterisk is the eighth most frequent item on the list of the 100

most frequent words attests to the popularity of action messages in the CIRC. Of course this figure is not identical with the number of action messages in the corpus. It is inflated by asterisks used in emoticons and by manually created action messages which are often not only preceded but also followed by an asterisk. The list of frequencies for asterisks in the ten channels, which is shown below, was created with this caveat in mind:

	acmel	acper	dacan	dache	escyb	eshel	itaus	itphi	utbeg	utcha
*	129	119	807	112	146	277	811	37	155	146

Figure 37: Frequencies of asterisks in the ten channels.

The informational channel #help and #philosophy have by far the lowest number of asterisks and therefore action messages. Obviously there is less need for vivid, life-like description of a situational context in these channels. Moreover, since the actions contained in action messages happen only in virtual space they are of not much use in these more fact-oriented channels. The action messages that we find in #help and #philosophy probably occur during those stretches of conversation which focus on social interaction and not information exchange.

Backchanneling

Backchannel items occur exclusively in spoken discourse. They are generally very short, sometimes not even lexicalized utterances such as ‘oh’, or ‘mhm’ which signal to the speaker that their message is being received. Backchannel items’ meanings are hard to define apart from the fact that they generally express agreement (Stenström 1990). Their presence in IRC has not been mentioned by prior research but was noted in the course of the pilot studies for

this work. The table below shows the distribution of some common backchannel items in the CIRC.

	hmm	hrm	ah	oh	yeah
acmel	40	4	16	84	183
acper	27	3	15	100	172
dacan	22	0	21	204	159
dache	11	0	7	168	87
escyb	30	4	7	150	124
eshel	35	88	24	113	96
itaus	66	5	43	255	248
itphi	66	4	33	152	66
utbeg	12	0	9	101	69
utcha	30	1	13	193	115
total	339	109	188	1520	1319

Figure 38: Distribution of some common backchannel items in the CIRC

Of course, not all instances of these items function as backchannels. Due to their high frequencies however, we can safely assume that a considerable amount of backchanneling takes place in IRC conversations. The following example illustrates how participants make use of backchannels in IRC:

“<ParticipantP> actually yes.. ;) Larry loves his boats... but he truly did need a rest... [...]

<ParticipantP> had a little scare with his heart....ealier this year.. :/ [...]

<ParticipantD> oh...

<ParticipantP> i wanted him to take the trip..”

Naturally, backchanneling in IRC cannot occur almost simultaneously as it does in spoken discourse, because IRC technology only allows for linear display of messages. Therefore, backchanneling is more obtrusive in synchronous CMC than it is in spoken conversation. This slightly disruptive quality probably results in a lower incidence of backchanneling in IRC than in spoken discourse. To further support this hypothesis with quantitative data from both spoken and

written interactive discourse is a task for future research.

Phonetic spelling

Frequent attempts at phonetic spelling in synchronous CMC has been widely documented in prior research (Werry 1996, Lundstrom 1995, Doell 1997). They are generally attributed to IRC participants' desire to make their written interactive discourse more speech-like. As mentioned in the methodology section, phonetic spelling is a fuzzy category. Many words which are spelled phonetically are also shorter than their standard variants and could therefore also be examined under the heading of abbreviation. Moreover, many phonetic spellings also attest to their creators' linguistic creativity. Therefore, some items will be mentioned both here and in the following chapter in which these phenomena will be examined.

Due to the great variety of phonetically spelled words and corresponding low frequencies of the individual items this phenomenon does not lend itself well to the automatic analysis with the concordancer. However, phonetic spelling is so prevalent that an analysis of any sample taken from the corpus will contain examples of the phenomenon. Throughout the analysis of CIRC, items which are particularly interesting in this respect were noted. Furthermore, all the discussions of non-standard spelling variants throughout chapter five contained examples for phonetic spelling. For example, the variants of the personal pronouns 'ya' and 'u' for 'you', 'em' for 'them', 'dat' for 'that' can all be regarded as instances of phonetic spelling.

The last item in the list is probably also a case of phonetic spelling imitating a geographical dialect. The figures in table 13 in chapter five illustrate that this spelling is most

frequent in the Australian channels. This phenomenon has already been noted by Werry (1996).

An analysis of the items ‘what’ and ‘because’ and their spelling variants yields further evidence for this use of phonetic spelling. Table 39 below displays the distribution of the two items and their spelling variants.

	what	wut	wot	whut	because	becos	becoz	cuz	coz	cus
acmel	287	1	33		20		1	6	36	
acper	317		25		11	1	3	8	42	8
dacan	368	3			40			18	7	1
dache	323	2			19			8	1	
escyb	397	3	1	1	15			18	1	1
eshel	717	3	1	1	47			20	3	
itaus	535		12		32			5	14	
itphi	511	1		1	101				1	
utbeg	283	2			8		1	2	2	
utcha	335	1	1	1	21			8	4	1
total	4073	16	73	4	314	1	5	93	111	11

Figure 39: Frequencies of ‘what’ and ‘because’ and their spelling variants.

The variants ‘whot’, ‘becuz’, and ‘becus’ are not included in the table because no instances of them were found. The distribution clearly shows a preference for the spelling variants using ‘o’ instead of ‘u’ in the Australian channels. It is not clear which linguistic feature of Australian English is imitated thus but the distribution which concentrates these forms in the channels #melbourne, #perth, and #australia strongly suggests the spelling reflects a geographical dialect.

Phonetic spelling is also used to imitate social dialect features. The occurrence of ‘yous’ instead of ‘you’ was mentioned earlier. Other examples are the spellings ‘dood’ for ‘dude’, ‘gawd’ for ‘god’, ‘dawg’ instead of ‘dog’, and the reduced form ‘in’ of the ‘ing’ ending. The latter is obviously the most frequent of these items. For example for ‘doing’ and ‘going’

approximately one in seven instances is a reduced form.

Phonetic spelling in IRC is not only used to imitate dialects but also other language variations. The spellings 'otay' for 'ok', 'fink' instead of 'think' and 'wif' instead of 'with' for example, clearly imitate the babbling of babies. Phonetic features that are especially prone to having their spelling changed in the CIRC are the 'th' sound and vowels. This tendency is probably due to the several pronunciation variants for the first and the many spelling variants and phonetic reductions for the second.

Paralinguistic and suprasegmental cues

As we have seen in the discussion above, IRC users utilize phonetic spelling to imitate the pronunciation of individual words. However, their ingenuity at transposing features of speech into their interactive written discourse does not end there. They also imitate various paralinguistic sounds and suprasegmental features of speech using the limited resources offered by a computer keyboard. These tendencies have been documented under various names by previous research. Lundstrom (1995) calls it reconstruction of affect, Werry (1996) refers to it as paralinguistic and prosodic cues and Doell (1997) used the term onomatopoeia.

The most frequent paralinguistic feature which is transposed into writing is laughter. Numerous variants of 'hehe', 'haha', 'hihi' etc. are in use on IRC. The most frequent in the CIRC are the variants with 'e' followed by those with 'a'. Reduplication of letters is used to indicate the length of the laugh. Some users make extensive use of this option until the laughs are twenty or more letters long. The reduplication of letters to indicate drawn out pronunciation

is also used in many other expressions such as the one displayed in the example below.

"<ParticipantA> hi, anyone wanna coooool chat, msg me now"

Creative phonetic spelling is another means employed by IRC users to express vowel length. The example below illustrates this use of spelling.

"<ParticipantF> !ping my plumbers-crack-of-an-ass puh-leaze"

Other paralinguistic sounds that are imitated by members of the IRC community include expressions of surprise, agreement, hesitation, apology, despair and many more. The corresponding items in the same order are 'ooh', 'uh huh', 'er', 'oops', and 'awww'. It lies outside the scope of this study to attempt a complete enumeration and quantitative evaluation of these paralinguistic expressions because they are immensely popular on IRC. An analysis of a any short sample of text from the CIRC will yield numerous examples of this phenomenon.

The most common and well known suprasegmental feature incorporated into written discourse by participants of IRC is volume. Capitalization indicates that something was 'spoken' in a loud voice. Some channels have strict conventions regarding the use of yelling and penalize users who overapply this principle. The conversation passage given below is a prime example of this mechanism at work.

**"<ParticipantK> TO HELL WITH THE CORPORATIONS WHO DO NOT HIRE
FULL TIME ANYMORE [...]"**

<ParticipantM> STOP YELLING YOU TARD!"

Sometimes capitalization is also used to indicate stress as in the example below.

"<ParticipantD> wooHOO"

CHAPTER 7

INNOVATION AND CREATIVITY

The features of IRC talk discussed in this chapter occur neither in spoken nor in traditional written genres. They attest to the innovative and creative alteration of language by the IRC participants. This in turn is a quality which is considered more characteristic of speech than writing. Written language, with its centuries old standards and high levels of formality, is much more conservative and resistant to change than spoken language and IRC talk.

Nickname morphism

Bechar-Israeli (1995) described in detail the great importance that nicknames have in IRC culture. They are the first and foremost component of a user's online presence. An analysis of the CIRC's nicknames and the establishment of a typology along the lines of Bechar-Israeli's study is not the goal of this work, although, there is certainly plenty of material contained in the CIRC for such an undertaking. Based on a count of the number of different nicknames in three log files in the course of the second pilot study, a conservative estimate of the number of different nicknames per channel is 300. This means that there are at least 3000 nicknames available for

study in the CIRC compared to the 260 which formed the basis of Bechar-Israeli's study.

It is important to keep in mind that the number of nicknames is not the same as the number of participants because nicknames can be changed at will and are changed frequently by the IRC participants. Proof of this comes from the number of system messages which announce a change of nicknames. This measure tells us that IRC users changed their nicknames 5349 times during the conversations contained in the CIRC. Considering these high numbers, it is surprising that this phenomenon has not been examined in more detail in prior research. While many researchers (Lundstrom 1995, Werry 1996, Bays 1998) mention the use of nicknames and the fact that they change, none of them explore the mechanisms at work. It is this changing of nicknames that our analysis will focus on here. We propose to term this phenomenon nickname morphism.

To better understand the workings of nickname morphism it is vital to know one basic principle of IRC technology. On an IRC network, nicknames must be unique and there is no way of permanently registering a nickname for yourself. Once a participant quits the network her nickname becomes available for use. Bechar-Israeli describes this technical setup and the resulting quarrels over nickname ownership in detail. For the discussion of nickname morphism this fact is important because it explains why so many nicks contain special characters and numbers. For example, if you decided to name yourself 'Bill' on the IRC network but this name is taken you can try combining 'Bill' with special characters to yield names such as '_Bill', 'B+I+L+L', 'Bill01', and so forth. Many name changes in the CIRC result from this quest for a unique combination of letters. The two examples from the CIRC below illustrate this kind of nickname morphism.

“*** Bad^Boyz is now known as Bad^Boy01”³⁰

“*** F`L`I`C`K is now known as Flick”

Another result of the frequent nickname conflicts stemming from this particularity of IRC technology is the development of dual nicknames. Many IRC users have a second nickname to fall back on in case their first one is taken. There is even a provision for this in the IRC client software which will automatically try to log on with your alternate nickname if your first one is taken. Often the second nickname is completely different from the first. It is like a second aspect of the user's online persona. This use of multiple nicknames accounts for a large number of nickname changes on IRC. Below is an example which illustrates this process.

“*** _blah_ is now known as EFFEM”

However, the most interesting phenomenon in nickname morphism is not due to the technological constraint which requires the use of unique nicknames. Some users change their nickname to indicate that they are not physically present at the computer screen and therefore cannot participate in the conversation at this time. This enables other users to realize that participant's online status with a glance on the participant roster rather than having to verify it through the IRC client software's status check feature. Below are some example of this phenomenon.

“*** MiK|Sleepz is now known as MiK”

“*** ch0isE^sh0WerZ is now known as lilmisschoisey”

“*** _Duff_ is now known as DuffBRB”

³⁰ In the nickname analysis the quoted nicknames were not changed. The fact that there is no conversational context was deemed sufficient protection of IRC users' privacy.

“*** ^Angel_kitty^ is now known as Angel|Away”

In the first two examples the users provide information on their current activities e.g. sleeping and showering. The last two, on the other hand, merely indicate that they are not physically present at their computer screens. ‘Away’ is the standard IRC expression for this fact while ‘BRB’ stands for ‘be right back’ which is a common abbreviation of IRC talk. Morphologically speaking, the authors of all four examples signal the add-on quality of the status information. This is done through the use of special characters such as ‘|’ and ‘ ^ ’ or capitalization in ‘BRB’. However, there are also instances where the distinctness of the online status morpheme is not expressed. An example is provided below.

“*** slezy is now known as slezygonebyebye”

Over time the most common of these status indicators, namely ‘away’ and ‘brb’ may even evolve from preposition and abbreviation to true suffixes.

ASCII art

The occurrence of ASCII art in synchronous CMC, that is, pictures that are made up entirely of ASCII characters, has been mentioned by several researchers (Werry 1996, Lundstrom 1995, Wilkins 1991). However, its use in IRC conversation, with the exception of emoticons which could be classified as ASCII art, is quite marginal. Since most ASCII art pictures require more than one line and often several lines of text, their use is discouraged by the anti-flood conventions which are enforced in most channels. A flood in IRC terminology is a massive influx of usually identical messages coming from the same user. This is very disruptive to the

discourse and therefore most channels have programs in place that automatically kick users who flood off the channel. Incidentally, this practice also prohibits the spread of multi-line ASCII art on IRC. Some one-line pictures, however, are used. The classic example for these one liners is the picture of a rose which can also be found in the CIRC as illustrated below.

“* ParticipantB Hands ParticipantR a Rose! @ -}--}— “

These examples of ASCII art are often given as gifts or to show off the creators' creativity. The conventions of gift giving on IRC have been examined in detail by Bays (1999) and will not be discussed here.

Despite the flood convention, one space remains for ASCII art to flower on IRC. This is the so-called message of the day or MOTD for short. This message is displayed at the very beginning of an IRC session, i.e. immediately after the logging on to the network but before any channels are joined. The message usually contains administrative information such as who to contact with questions, which norms of behaviour are enforced on this particular IRC-server, and sometimes the message also displays large scale ASCII art. We do not include any examples here because this usage of ASCII art does not impact on the linguistic nature of IRC talk.

Neologisms

Akin to phonetic spelling, neologism is a fuzzy category. Firstly, it is left to the personal judgement of the researcher to determine how recent a word must be in order to be recognized as a neologism, i.e. a new word. Moreover, some overlap exists with this category and the abbreviation category because the coinage of a new abbreviation can also be regarded as a

neologism. The analysis of neologisms in the CIRC is further complicated by the fact that they must be identified manually before they can be sought out with the concordancer. If the corpus were to be tagged grammatically, the parser would identify the words which are not in its language database, thus possibly uncovering large numbers of neologisms.

Unsurprisingly, a large number of neologisms used in the CIRC pertain to computer and IRC technology. Several of the latter have been mentioned already. Some examples are: ‘flood’ with the meaning of sending multiple messages in quick succession, ‘kick’ with the meaning of removing a participant temporarily from a channel, ‘nick’ as a short form of ‘nickname’, ‘operator’ and its abbreviation ‘op’ meaning a privileged IRC user who moderates the interaction in a channel.

Naturally, IRC talk also incorporates and sometimes adapts many general computing terms. For example a ‘ghost’ on IRC is not a spirit from the netherworld but a participant whose connection to the server is broken but who, due to a computer glitch, is still registered as being online. Often, IRC users extend the grammatical functions of such computer terms. For example, the term ‘dcc’ is originally a short form of ‘direct client-to-client’ and as a noun describes a feature of IRC client software which allows two users to bypass the IRC server and establish direct connections between their computers. Now, ‘dcc’ is equally if not more often used as a verb which can mean send via a dcc connection or establish a dcc connection. This usage is illustrated in the examples below.

“<ParticipantC> anyone care to creat a good friendship please dcc or msg me!!!”

“<ParticipantS> well.....ParticipantT...I’ll dcc you the list...have about 5 gig left?”

Moreover, IRC participants use such computer terms as bases for new word formations.

The term ‘ping’, for example, is defined in the Hypertext Webster jargon file³¹ as 1. “n. Slang term for a small network message (ICMP ECHO) sent by a computer to check for the presence and alertness of another” and 2. “/vt./ To send a message to all members of a {mailing list} requesting an {ACK} (in order to verify that everybody's addresses are reachable).” In IRC the derivational noun ‘pinger’ and adjective ‘pingy’ are in use as well.

Naturally, due to the limitations outlined at the beginning of this section, the list of neologisms mentioned above is not exhaustive. This must remain a project for further research. As a source of 1999 IRC talk, the CIRC will enable later studies of neologisms on IRC to verify just how recent the words they find are.

Innovative spellings

As in the neologism and the phonetic spelling category, the innovative spellings category contains borderline cases which could also be assigned to other categories. The prime example of this is the word ‘peeps’ which was found in the CIRC as a variant of ‘people’ or ‘peoples’. It is clearly an innovative spelling, but it is also shorter than ‘people’ and could therefore be classified as an abbreviation. Furthermore, it might be an attempt at a phonetic spelling of some idiosyncratic pronunciation. Finally, it could be regarded as a new word and would therefore fall into the neologism category. In order to limit the inclusion of such problematic borderline cases, the words that will be examined in this category are all of comparable length compared to their standard spelling. However, some might still represent

³¹ Available online at: http://work.ucsd.edu:5141/cgi-bin/http_webster

attempts at phonetic spelling which defy clear identification.

One of the innovative spellings has already been mentioned in the discussion of personal pronouns. It is the form 'yew' instead of 'you'. Another item with a similar spelling is 'tew' which is a variant of both 'too' and 'to'. An innovative spelling which follows the pattern of 'too' is the spelling 'yoo' instead of 'you'. This spelling might also indicate a long vowel and therefore is possibly an attempt at a phonetic spelling.

There are several innovative spellings for the sound [ʌ] in the CIRC. The spelling 'lurve' or 'luv' instead of 'love', the spelling 'forking' instead of 'fucking' are other examples of this formation. There is an alternative spelling 'skool' instead of 'school' which might be an abbreviation since it is shorter than the original. That interpretation is unlikely, however, because of the replacement of the consonants 'ch' by 'k', which is a letter that does not occur in the original spelling, and not 'c'.

As with the phonetic spellings and neologisms, the examination of the phenomenon in this study cannot be exhaustive. Moreover, most of the innovative spellings have such low frequencies that a channel by channel analysis would not yield reliable results. There is, however, a noticeably high incidence of these spellings in the channels #melbourne and #perth, which strengthens the evidence for their especially progressive and innovative character.

Abbreviation

Abbreviations are perceived by many people as one of the most central features of IRC

talk. Introductions on IRC available online³² usually provide interested newcomers to the medium with whole lists of abbreviations which are common on IRC. Researchers (Ferrara et al. 1991, Reid 1991, Lundstrom 1995, Werry 1996) agree with this general assessment of abbreviations in synchronous CMC and argue that their use is largely motivated by the IRC users' need for speed. When addressing the issue of abbreviation, most researchers follow the example of the IRC 'how to' websites and focus on acronyms. Lundstrom (1995), for example, publishes a list of about 130 abbreviations in his appendix which contains only five expressions that are not acronyms. On the one hand, this focus makes sense because acronyms are usually more difficult to interpret than other abbreviations. On the other hand, the data in the CIRC suggests that users are much more likely to encounter abbreviations which are the result of clipping, innovative spelling, and phonetic spelling than acronyms. Moreover, the former are not only numerically more frequent than acronyms, they also come in a larger variety. Of Lundstrom's list of abbreviations, i.e. acronyms, only a handful occur in the CIRC but four out of five of the non-acronym items occur in CIRC with the lowest frequency being 46. The table below displays the distribution of the most frequent acronyms in the CIRC³³.

³² See <http://www.htmltips.com/patrick/irc.html> or <http://www.rzuser.uni-heidelberg.de/~mseuffer/irc.html> for examples.

³³ Individual channel frequencies are not included for the long forms. The long forms are: 'laughing out loud', 'be right back', 'be back later', 'by the way', 'rolling on the floor laughing'.

	lol	lol/long	brb	brb/long	bbl	bbl/long	btw	btw/long	rofl	rofl/long
acmel	562		26		5		1		7	
acper	636		36		5		3		34	
dacan	817		52		12		1		6	
dache	1083		133		19		4		51	
escyb	822		103		34		5		19	
eshel	72		30		3		3		3	
itaus	611		52		11		15		63	
itphi	55		12		10		12		2	
utbeg	611		84		16		5		110	
utcha	665		64		17		5		38	
total	5934	5	592	20	132	25	54	12	333	9

Figure 40: Distribution of common acronyms in IRC.

Overall the acronyms outnumber the long versions by far. Even the most frequent long version does not reach a quarter of its acronym's frequency. However, the most interesting feature of the table above is the proportion of the long forms versus their acronyms. The relationship between the two is an inverse correlation: the more frequent the acronym, the less frequent the extended version. This is probably due to the fact that the less frequent acronyms must be explained more often than the more frequent ones.

The distribution of the abbreviations created through clippings and innovative spelling is quite different from that of the acronyms. The tables below display the distribution of several of the most frequent examples of this kind of abbreviation and their long forms.

	please	pls	plz	thanks	thanx	thx	you	u	ya
acmel	167	44	99	48	23	0	631	761	207
acper	55	9	155	33	29	0	697	682	276
dacan	65	21	20	102	28	8	1123	545	197
dache	147	35	44	234	40	47	1014	534	317
escyb	122	30	32	206	52	19	1172	405	293
eshel	133	14	19	170	54	23	1380	271	50
itaus	83	13	5	70	37	8	1407	589	230
itphi	35	1	4	22	2	0	1427	292	36
utbeg	279	76	3	326	58	39	1272	385	198
utcha	111	93	25	191	75	51	1138	539	173
total	1197	336	406	1402	398	195	11261	5003	1977

Figure 41: Distribution of some frequent abbreviations through clipping and innovative spelling.

	going to	gonna	want to	wanna	message	msg	suppose	spose	about	bout +'
acmel		34		698	73	633	11	2	110	12
acper		52		489	50	472	4	4	136	26
dacan		75		190	17	127	15	0	159	16
dache		39		68	13	40	9	4	128	9
escyb		61		169	13	102	15	0	110	21
eshel		40		42	27	35	8	2	231	9
itaus		46		83	4	19	29	3	232	23
itphi		25		17	3	0	24	3	267	7
utbeg		30		118	41	70	11	2	140	14
utcha		30		228	16	176	8	2	118	9
total	337	432	924	2102	257	1674	134	22	1631	146

Figure 42: Distribution of some frequent abbreviations through clipping and innovative spelling.

It is interesting to note that the proportion of long versus short form in the tables above is quite the opposite of the pattern observed for acronyms. Here, the long form outnumbers the short form in the majority of cases and even in the few cases where it does not, i.e. 'want to' and 'going to', the short form is only slightly more than twice as frequent as the long one. The reason why exactly these two forms display this deviant distribution probably lies in the fact that both of them are so common in speech that they had already found their way into written language before the arrival of CMC. They are still not acceptable in formal writing but they occur on billboards, in advertisements and other forms of popular writing. Abbreviations such as 'u' for 'you', 'r' for

'are', or 'spose' for 'suppose', on the other hand, are still not found outside the realm of electronic English.

At this point a short correction of a claim made by Werry (1996) needs to be made. He reports the use of the abbreviation 'yr' on IRC, which, according to him, stands for 'your'. Data from the CIRC shows that it is used with that meaning only in 10 out of 75 instances and in 65 cases it is in fact an abbreviation for 'year'. It might be that Werry's database was not large enough to detect the second and more popular meaning, but it is also possible that this change is due to diachronic variation.

Finally, mention should be made of a kind of abbreviation which is very characteristic for IRC talk, namely abbreviation through use of numbers. There are many examples of this in the CIRC. The most frequently used numbers are '4' instead of 'for', '2' instead of 'to' or 'too', and '1' instead of 'one'. The most productive of these is probably '1'. The table below shows the distribution of some popular abbreviations created with the digit '1'.

	no1	every1	some1	any1	ne1	noone	none	everyone	someone	anyone
acmel	4	6	6	36	24	2	12	117	49	581
acper		5	1	36	55	3	5	90	70	368
dacan	3	1		4	1	2	13	145	70	277
dache		10		3		4	7	139	54	68
escyb		3	2	7	4	1	5	149	47	151
eshel		3	3	6	1	7	11	17	183	339
itaus	1	4	5	8		3	6	69	53	101
itphi				4		3	3	41	37	42
utbeg		2	1	7		3	6	150	124	174
utcha		4	2	6	1	1	3	170	65	197
total	8	38	20	117	86	29	71	1087	752	2298

Figure 43: Distribution of abbreviations created with the digit '1'.

From the figures of the table above it is clear that the abbreviations using numbers, while certainly productive, are not as widespread as the abbreviations using clipping and innovative spelling. None of their frequencies is more than 5% of those of the long forms. Probably, this type of abbreviation is too recent a development to have gained general acceptance, even in a medium which is as open to change as IRC. Centuries of keeping numbers and letters separate cannot be forgotten in the barely 20 years that CMC has existed. Evidence to support this hypothesis comes from the two most frequent abbreviated items in the table above, namely 'anyl' and 'nel'. Their distribution shows a distinctive concentration in the channels #melbourne and #perth which we have found to be the most linguistically experimental of the ten channels contained in the CIRC.

CHAPTER 8

CONCLUSION

Results of the study

The creation of the CIRC revealed several factors which will be important for the further development of corpora of synchronous CMC. Firstly, the date of data collection, i.e. the day of the week or time of day, does not significantly influence the nature of the data collected. Secondly, a standardized length of text passages is vital for comparisons within the corpus. Thirdly, the distortion effect created by system messages must be taken into account and carefully gauged during the statistical analysis of the corpus data. Fourthly, the nature of the data in the corpus creates important structural differences, for example in the amount of dirty data, the type/token ratio, and the frequencies of the individual words, compared to corpora of traditional written and spoken genres.

Overall, the analysis of the CIRC supported the findings of earlier studies on synchronous CMC. Where differences were found, for example in the usage of abbreviations or spelling, they were usually limited to questions of degree rather than principle. Throughout the analysis, features were found, e.g. in the discussion of the personal pronouns and innovative spelling, which rendered expressions in IRC longer than necessary. This finding runs counter to the widely proposed theory that IRC participants use linguistic reduction wherever possible in

order to simulate the speed of speech.

Statements of trends in the linguistic behaviour of IRC participants which had been based largely on educated guesses could be replaced by observations based on frequency data. IRC communities or channels were found to be the locale of systematic linguistic variation. Especially important is the discovery of channels with geographical names as the most linguistically progressive and innovative communities on IRC. Equally significant is that the influence of geographical dialects from the real world can overrule the tendency for experimentation in these channels in the virtual world. At the other end of the data spectrum in the CIRC, the channels with a more informational focus were found to have linguistic conventions which were closer to those of traditional written genres. It is important to note that although variation across the ten channels was found, this variation was never so fundamental as to result in non-occurrence of IRC-specific features. For example, even #philosophy had its share of non-standard spellings, abbreviations, emoticons, etc. Several phenomena that had been omitted or mentioned only in passing in prior research were examined in more detail. These are, for example, the identification of dual deictic reference, conventions of greeting and leave-taking, the categorization of emoticons, the mechanisms of nickname morphism, and the relative frequency of different types of abbreviation.

Perhaps the most important finding of this study is the abundance of evidence for the similarity of interactive written discourse as exemplified in IRC to spontaneous casual conversation. These similarities were found through statistical comparison, and through detailed analysis at both the lexical and the syntactic level. Of course this is not to say that IRC is identical to speech, but that it displays characteristics of it at every level of discourse right down

to the imitation of phonetic, paralinguistic and suprasegmental features of speech.

Applications

The possible applications for this study are many. First of all, the corpus itself is adaptable to a variety of research interests. For example, during the creation of the CIRC, I was approached by a professor of language engineering whose students wished to use the corpus data to train a conversation machine. The CIRC could also be an excellent source of authentic examples to complement researchers' personal log files. For example, there are 30 instances of the so-called cookie convention³⁴ in the CIRC which is a topic that Hillary Bays (1998,1999) is currently working on.

Moreover, the CIRC can be easily and greatly expanded at low cost so that, for example, variation across channels could be studied in more detail. The ease of expansion is demonstrated by the HAL corpus (Burgess, K. & Livesay, K. 1998) which is a corpus of asynchronous CMC containing more than 200 million words. It has been used in psycholinguistic research and has been found to yield significantly better results than the Brown corpus for word recognition tasks. It is possible that a similarly large corpus of synchronous CMC would improve these results even more.

Possibly the most far reaching application of the CIRC would be its use as a complement to corpora of spoken English. Countless researchers have strongly urged for the creation of more corpora of spoken English. John Sinclair (1991:15) for example writes: "Most

³⁴ The giving of virtual cookies is the prime example of gift giving on the Internet.

corpora keep well away from the problems of spoken language - with some honourable exceptions - and, for a corpus which in any way purports to reflect a 'state of the language', this is most unfortunate." Geoffrey Leech (1997:17) puts it even more strongly: "This prioritizing of the written language has persisted through the history of corpus development. For example, the BNC consists of c. 90 million words of written English, and only c. 10 million words of spoken language. Why this discrepancy? The reason, simply, is that spoken language is much more expensive to collect, at the present stage of technological progress, than written language. It costs roughly as much to collect 10 million words of spoken language (requiring, as this does, manual transcription) as to collect 90 million words of written English - and even then, the result is little more than a basic orthographic transcription."

Given the high costs of spoken corpora on the one hand and the lack of quality mentioned by Leech on the other, it would make sense to complement existing spoken corpora with low cost corpora of interactive written discourse which has so many features of speech. Moreover, the influence of CMC on other varieties of English will continue to grow since the increases in computer use and the economic importance of the Internet show no sign of abating. This trend may be illustrated by a story I was told recently by a friend. He confessed to using the acronym 'brb' which means 'be right back' in a conversation with other friends and was confronted with the blank, uncomprehending faces of his friends. However, considering the growing popularity of CMC, this state of affairs might not remain for long. As the following quote from Paollilo (1999³⁵) illustrates, I am not alone with this prediction: "Many scholars

³⁵ This quote is from an unpaginated version of the article published on the Internet. It is taken from the paper's abstract.

anticipate that online interaction will have a long-term effect on the evolution of language”

Suggestions for further research

As mentioned throughout the study, the analysis of the CIRC could be greatly enhanced and facilitated by grammatical tagging. It is likely, however, that the use of a standard parser for this task would encounter serious problems. These would surface especially in the areas of innovative and phonetic spelling, abbreviation, and typing mistakes. In order to address these problems, major alterations to the parsing software seem inevitable.

Once grammatical tagging is achieved, a more ambitious project would be the creation of a monitor corpus. John Sinclair coined this term for a corpus which no longer has a given size or number of included texts. Instead, language data would be constantly fed into the corpus as through a filter and be discarded after a while. Researchers would submit their queries and the corpus software would retrieve the appropriate data from the flow of linguistic information. For traditional written and spoken genres, the creation of such a monitor corpus is a very costly and complicated undertaking. However, the same is not true for CMC in general because it is already available in electronic form and for IRC in particular because it naturally occurs as a stream of language data.

The setup of a research IRC server could be achieved at minimal cost and effort. The method of lurking & logging would be replaced by simply logging. Users would be informed in the message of the day that their contributions may be used for research purposes which would eliminate the copyright issue that has plagued corpus development since the beginning.

Moreover, if users were required to register with the server, administering user surveys or interviewing individual participants could be implemented with ease. The establishment of the necessary software for a monitor corpus or IRC would be the most difficult part of the task. It is possible that suitable software already exists and is used by spy agencies to monitor the electronic traffic over the Internet. Whether similar software could be obtained easily for research purposes is another question. A monitor corpus of this kind would be immensely useful for linguistic study and especially variation studies. Researchers would be able to observe linguistic change as it happens. It would be possible to pinpoint exactly the time, author, and spread of a change. This could help uncover mechanisms which have eluded linguists for centuries.

As we can see, the creation and analysis of a corpus of Internet Relay Chat is a project which has numerous applications and far reaching implications for the field of linguistics in general and corpus linguistics in particular. These stem from two main aspects of IRC conversations: firstly, the many features of casual spoken conversation and secondly, the easy, low cost manner of data collection from this medium. Both of these qualities have been proven beyond doubt by the present study. Hopefully, the CIRC will be the first step towards a full exploitation of this medium's unique possibilities. This is all the more important considering the medium's burgeoning popularity and likely influence on the process of linguistic change.

WORKS CITED

- Aijmer, K. & Altenberg, B. (Eds.), (1991). *English Corpus Linguistics*. London: Longman.
- Altenberg, B. (1990). Spoken English and the dictionary. Pp. 177-193. In Svartvik, J. ed. (1990). *The London-Lund Corpus of Spoken English: Description and Research*. Lund: Lund University Press.
- Arndt, H. & Janney, R. W. (1987). *InterGrammar*. New York: Mouton de Gruyter.
- Bays, H. (1998). *Framing and Face in Internet Exchanges: a Socio-Cognitive Approach*.
<http://viadrina.euv-frankfurt-o.de/~wjournal/bays.htm>
- Bays, H. (1999). *The Gift Economy in Internet Relay Chat: Giving Immaterial and Material Gifts*. Paper presented at the Cybersociety conference 4-6 July, 1999, Newcastle, England.
- Bechar-Israeli, H. (1995). From <bonehead> to <cLoNehEAd>: Nicknames, Play, and Identity on Internet Relay Chat. *Journal of Computer Mediated Communication*: v1, n2. <http://www.ascusc.org/jcmc/vol1/issue2/bechar.html>
- Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*. V.8, n.4. pp. 243-257.
- Bublitz, W. (1994). In the Eye of the Beholder: 'The Rather Mystical Notion of Coherence'. Pp. 213-231. In: Carlon, K., Davidse, K. & Rudzka-Ostyn, B. eds. (1994). *Perspectives on English*. Peeters: Paris.
- Burgess, K. & Livesay, K. (1998). The Effect of Corpus Size in Predicting Reaction Time in a Basic Word Recognition Task: Moving on From Kucera and Francis. *Behavior, Research Methods, Instruments & Computers*, v. 30, n.2, pp. 272-277.
- Chafe, W. & Danielewicz, J. (1987). Properties of Spoken and Written Language. Pp. 83-116. In Horowitz, R. & Samuels, S. J. eds. (1987). *Comprehending Oral and Written Language*. San Diego, CA: Academic Press Inc.
- Cheepen, C. & Monaghan, J. (1990). *Spoken English: A Practical Guide*. New York: Pinter Publishers.

- Cherny, L. (1994). *Gender Differences in Text-Based Virtual reality*.
[Http://www.oise.on.ca/~jnolan/muds/about_muds/gt](http://www.oise.on.ca/~jnolan/muds/about_muds/gt)
- Collot, M. & Belmore, N. (1996). Electronic Language: A New Variety of English. Pp. 13-28. In Herring, S. C. ed. (1996). *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Davis, L.M. (1990). *Statistics in Dialectology*. Tuscaloosa: University of Alabama Press.
- Eggins, S. & Slade, D. (1997). *Analyzing Casual Conversation*. London: Cassell.
- Ferrara, K., Brunner, H. & Whittemore, G. (1991). Interactive written discourse as an emergent register. *Written Communication*, Vol. 8, No. 1, pp. 8-34.
- Francis, W. N. & Kucera, H. (1967). *Computational Analysis of Present-Day American English*. Providence, Rhode Island: Brown University Press.
- Francis, W. N. & Kucera, H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin Company.
- Garner, R. & Gillingham, M. G. (1996). *Internet Communication in Six Classrooms: Conversations Across Time, Space, and Culture*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.
- Gravetter, F. J. & Wallnau, L. B. (1991). *Essentials of Statistics for the Behavioral Sciences*. New York: West Publishing Company.
- Halliday, M. A. K. (1985). *Spoken and written language*. Oxford: Oxford University Press.
- Halliday, M. A. K. (1987). Spoken and Written Modes of Meaning. Pp. 55-82. In Horowitz, R. & Samuels, S. J. eds. (1987). *Comprehending Oral and Written Language*. San Diego, CA: Academic Press Inc.
- Herring, S.C. (1999). Interactional Coherence in CMC. *Journal of Computer Mediated Communication*. V.4, I.4. <http://ascusc.org/jcmc/vol4/issue4/herring.html>
- Herring, S. C. (1996). *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

- Horowitz, R. & Samuels, S. J. eds. (1987). *Comprehending Oral and Written Language*. San Diego, CA: Academic Press Inc.
- Hymes, D. H. (1971). On Communicative Competence. Pp. 269-293. In Pride, J. B. & Holmes, J. eds. (1972). *Sociolinguistics: Selected Readings*. Baltimore: Penguin Books.
- Inkster, G. (1997). First Catch Your Corpus: Building a French Undergraduate Corpus From Readily Available Textual Resources. Pp.267-276. In Wichmann, A. et al. (Eds.) (1997). *Teaching and Language Corpora*. London: Longman.
- Kelm, O. (1992). The use of synchronous computer networks in second language instruction: A preliminary report. *Foreign Language Annals*, 25 (5), 441-454.
- Kern, R. (1995). Restructuring classroom interaction with networked computers: Effects on quantity and quality of language production. *Modern Language Journal*, 79 (4),
- Kreidler, C.W. (1997). *Describing Spoken English: An Introduction*. New York: Routledge.
- Leech, G. (1991). The State of the Art in Corpus Linguistics. Pp. 8-29. In Aijmer, K. & Altenberg, B. (Eds.), (1991). *English Corpus Linguistics*. London: Longman.
- Leech, G. (1997). Teaching and Language Corpora: a Convergence. Pp. 1-24. In Wichmann, A. et al. (Eds.) (1997). *Teaching and Language Corpora*. London: Longman.
- Leitner, G. (1992). *New Directions in English Language Corpora: Methodology, Results, Software Developments*. New York: Mouton de Gruyter.
- Lundstrom, P. (1995). *Will Internet Talkers Improve the Communicative Competence of ESL/EFL Students?* [Http://elicos.qut.edu.au/truna/lan618/phoenix.txt](http://elicos.qut.edu.au/truna/lan618/phoenix.txt)
- Lunsford, W. (1996). *Turn-taking Organization in Internet Relay Chat*. Unpublished manuscript.
- Macaulay, M.I. (1990). *Processing Varieties in English: An Examination of Oral and Written Speech across Genres*. Vancouver: University of British Columbia.
- McArthur, T. (1992). *The Oxford Companion to the English Language*. London: BCA
- Miller, J. & Weinert, R. (1998). *Spontaneous Spoken Language: Syntax and Discourse*.

Oxford: Clarendon Press.

Murray, D. E. (1985). Composition as conversation: The computer terminal as medium of communication. in Odell, L. & Goswami, D. eds. (1985). *Writing in nonacademic settings*. New York: The Guilford Press. pp. 203-227.

Murray, D.E. (1991a). The composing process for computer conversation. *Written Communication*, v.8, pp. 35-86.

Murray, D. E. (1991b). *Conversation for action; the computer terminal as medium of communication*. Amsterdam: Benjamins.

Murray, D. E. (1995). *Knowledge machines: Language and information in a technological society*. London: Longman.

Paolillo, J. (1999). The Virtual Speech Community: Social Network and Language Variation on IRC. *Journal of Computer Mediated Communication*, v.4, n.4., <http://www.ascusc.org/jcmc/vol4/issue4/paolillo.html>

Pride, J. B. & Holmes, J. eds. (1972). *Sociolinguistics: Selected Readings*. Baltimore: Penguin Books.

Reid, E. (1991). *Electropolis: Communication and Community on Internet Relay Chat*. Unpublished Master's thesis. University of Melbourne. <http://people.we.mediaone.net/elizrs/electropolis.html>

Schaap, F. (2000). *The Words That Took Us There: Not an Ethnography*. Unpublished Master's thesis. <http://people.a2000.nl/fschaap/thesis/index.html>

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Someya, Y. (1999). *A Corpus-based Study of Lexical and Grammatical Features of Written Business English*. Unpublished Master's thesis. University of Tokyo. <http://www2.gol.com/users/ysomeya/>

Stenstroem, A-B. (1991). Expletives in the London Lund Corpus. Pp. 239-253. In Aijmer, K. & Altenberg, B. (Eds.), (1991). *English Corpus Linguistics*. London: Longman.

Stenstroem, A-B. (1990). Lexical items peculiar to spoken discourse. Pp. 137-176. In Svartvik, J. ed. (1990). *The London-Lund Corpus of Spoken English: Description and Research*. Lund: Lund University Press.

- Svartvik, J. (1990). The TESS Project. Pp. 63-86. In Svartvik, J. ed. (1990). *The London-Lund Corpus of Spoken English: Description and Research*. Lund: Lund University Press.
- Svartvik, J. ed. (1990). *The London-Lund Corpus of Spoken English: Description and Research*. Lund: Lund University Press.
- Svartvik, J. et al. (1982). *Survey of Spoken English: Report on Research 1975-81*. Lund: CWK Gleerup.
- Tannen, D. ed. (1984). *Coherence in Spoken and Written Discourse*. Norwood, NJ: Ablex Publishing Corporation.
- Tannen, D. (1982). Spoken and Written Narrative in English and Greek. Pp. 21-44. In Tannen, D. ed. (1982). *Spoken and written language: Exploring orality and literacy*. Norwood NJ: Ablex Publishing Corporation.
- Tannen, D. ed. (1982). *Spoken and written language: Exploring orality and literacy*. Norwood NJ: Ablex Publishing Corporation.
- Tannen, D. (1982). The oral/literate continuum in discourse. in Tannen, D. ed. (1982). *Spoken and written language: Exploring orality and literacy*. Norwood NJ: Ablex Publishing Corporation. pp. 1-16.
- Warschauer, M. & Healey, D. (1998). State of the Art Article: Computers and Language Learning: an Overview. *Language Teaching* v.31 n.2 pp.57-71
- Wichmann, A. et al. (Eds.) (1997). *Teaching and Language Corpora*. London: Longman.
- Werry, C. C. (1996). Linguistic and Interactional Features of Internet Relay Chat. Pp. 47-64. In Herring, S. C. (1996). *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Yates, S. J. (1996). Oral and Written Linguistic Aspects of Computer Conferencing. Pp. 29-46. In Herring, S. C. (1996). *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

CHAPTER 9

APPENDIX

Legend to the appendix:

- 1 = acmel = #melbourne
- 2 = acper = #perth
- 3 = dacan = #canada
- 4 = dache = #cheers
- 5 = escyb = #cyberchat
- 6 = eshel = #help
- 7 = itaus = #australia
- 8 = itphi = #philosophy
- 9 = utbeg = #beginners
- 10 = utcha = #chat-world

Appendix A:

The basic measures for the 200 CIRC files.

Name	Words	Types	↔	***	T/T	Total M	W/M	Hapax	T-Hapax	Hapax/Mille
acmel051 5003	1425	585	147	94	0.285	826	6.057	853	572	170.498
acmel052 5006	1312	687	48	86	0.262	821	6.097	739	573	147.623
acmel061 5003	1317	670	50	127	0.263	847	5.907	776	541	155.107
acmel062 5004	1396	586	73	65	0.279	724	6.912	863	533	172.462
acmel071 5003	1428	728	51	74	0.285	853	5.865	858	570	171.497
acmel072 5001	1410	657	38	74	0.282	769	6.503	836	574	167.167
acmel081 5008	1458	734	48	117	0.291	899	5.571	883	575	176.318
acmel101 5003	1414	657	97	97	0.283	851	5.879	803	611	160.504
acmel111 5004	1303	649	97	93	0.260	839	5.964	727	576	145.284
acmel121 5002	1284	580	51	141	0.257	772	6.479	729	555	145.742
acmel131 5007	1286	673	48	89	0.257	810	6.181	702	584	140.204
acmel132 5006	1296	662	63	113	0.259	838	5.974	682	614	136.237
acmel141 5002	1318	712	46	97	0.263	855	5.850	757	561	151.339
acmel142 5000	1353	592	89	92	0.271	773	6.468	777	576	155.400
acmel143 5001	1196	806	36	77	0.239	919	5.442	697	499	139.372
acmel151 5000	1334	621	42	133	0.267	796	6.281	737	597	147.400

acmel152 5002	1313	497	64	137	0.263	698	7.166	719	594	143.743
acmel153 5001	1253	579	71	142	0.251	792	6.314	661	592	132.174
acmel154 5002	1265	573	57	129	0.253	759	6.590	694	571	138.745
acmel155 5013	1234	580	77	146	0.246	803	6.243	690	544	137.642
acper051 5002	1474	882	37	67	0.295	986	5.073	912	562	182.327
acper052 5007	1456	954	49	38	0.291	1041	4.810	899	557	179.549
acper061 5001	1526	853	49	45	0.305	947	5.281	956	570	191.162
acper062 5006	1579	844	42	60	0.315	946	5.292	946	633	188.973
acper071 5001	1363	882	72	54	0.273	1008	4.961	804	559	160.768
acper081 5007	1423	774	31	90	0.284	895	5.594	822	601	164.170
acper101 4994	1531	579	57	58	0.307	694	7.196	895	636	179.215
acper111 5006	1552	730	116	45	0.310	891	5.618	991	561	197.962
acper112 5002	1473	711	69	67	0.294	847	5.906	889	584	177.729
acper121 5003	1509	525	123	49	0.302	697	7.178	938	571	187.488
acper122 5011	1398	739	54	74	0.279	867	5.780	777	621	155.059
acper123 5006	1496	745	50	48	0.299	843	5.938	915	581	182.781
acper124 4997	1452	866	61	43	0.291	970	5.152	884	568	176.906
acper131 5008	1402	773	51	17	0.280	841	5.955	873	529	174.321
acper132 5004	1160	514	64	93	0.232	671	7.458	661	499	132.094
acper133 5012	1424	745	31	96	0.284	872	5.748	859	565	171.389
acper141 5006	1539	615	46	113	0.307	774	6.468	942	597	188.174
acper142 4997	1415	675	34	107	0.283	816	6.124	847	568	169.502
acper151 5015	1459	623	104	57	0.291	784	6.397	866	593	172.682
acper152 5001	1334	618	57	116	0.267	791	6.322	758	576	151.570
dacan011 5003	1341	760	52	28	0.268	840	5.956	771	570	154.108
dacan012 5010	1330	801	93	37	0.265	931	5.381	777	553	155.090
dacan013 5005	1381	660	16	37	0.276	713	7.020	814	567	162.637
dacan041 5007	1443	790	45	34	0.288	869	5.762	834	609	166.567
dacan042 5010	1456	778	35	34	0.291	847	5.915	861	595	171.856
dacan051 5001	1487	719	59	51	0.297	829	6.033	845	642	168.966
dacan052 5003	1340	693	25	65	0.268	783	6.390	786	554	157.106
dacan053 5013	1385	763	31	32	0.276	826	6.069	815	570	162.577
dacan061 5002	1397	746	38	45	0.279	829	6.034	794	603	158.737
dacan062 5005	1302	799	26	23	0.260	848	5.902	789	513	157.642
dacan063 5002	1479	725	24	60	0.296	809	6.183	896	583	179.128
dacan071 5009	1528	632	26	11	0.305	669	7.487	922	606	184.069
dacan072 5008	1517	652	29	13	0.303	694	7.216	967	550	193.091
dacan073 5000	1572	760	63	13	0.314	836	5.981	1003	569	200.600
dacan081 5003	1479	729	24	30	0.296	783	6.390	868	611	173.496
dacan091 5006	1303	596	68	128	0.260	792	6.321	770	533	153.815
dacan111 5000	1243	524	49	166	0.249	739	6.766	748	495	149.600
dacan121 5009	1311	503	13	190	0.262	706	7.095	799	512	159.513
dacan131 5009	1389	616	24	135	0.277	775	6.463	791	598	157.916
dacan132 5000	1303	643	67	105	0.261	815	6.135	733	570	146.600
dache011 5006	1361	653	78	27	0.272	758	6.604	806	555	161.007
dache012 5006	1330	682	43	20	0.266	745	6.719	784	546	156.612
dache041 5002	1364	713	53	42	0.273	808	6.191	811	553	162.135
dache042 5002	1393	675	55	57	0.278	787	6.356	814	579	162.735
dache051 5001	1403	652	63	32	0.281	747	6.695	811	592	162.168
dache061 5004	1645	653	78	19	0.329	750	6.672	1036	609	207.034
dache081 5002	1361	786	45	31	0.272	862	5.803	823	538	164.534
dache091 5004	1249	631	25	60	0.250	716	6.989	661	588	132.094
dache092 5010	1419	600	62	50	0.283	712	7.037	831	588	165.868

dache1115004	1273	599	53	81	0.254	733	6.827	768	505	153.477
dache1215004	1311	579	41	43	0.262	663	7.548	783	528	156.475
dache1315007	1356	587	26	71	0.271	684	7.320	783	573	156.381
dache1324998	1403	636	77	30	0.281	743	6.727	798	605	159.664
dache1335002	1318	647	53	18	0.263	718	6.967	734	584	146.741
dache1345012	1343	540	70	24	0.268	634	7.905	758	585	151.237
dache1415002	1432	520	62	30	0.286	612	8.173	847	585	169.332
dache1424999	1335	645	88	24	0.267	757	6.604	724	611	144.829
dache1515005	1211	706	48	25	0.242	779	6.425	687	524	137.263
dache1525022	1294	610	59	60	0.258	729	6.889	728	566	144.962
dache1535005	1401	682	50	32	0.280	764	6.551	819	582	163.636
escyb011 5004	1490	929	52	43	0.298	1024	4.887	900	590	179.856
escyb012 5005	1351	809	71	26	0.270	906	5.524	821	530	164.036
escyb013 5005	1331	897	60	18	0.266	975	5.133	810	521	161.838
escyb021 5006	1280	876	120	23	0.256	1019	4.913	747	533	149.221
escyb031 5004	1432	860	69	50	0.286	979	5.111	861	571	172.062
escyb032 5000	1276	870	63	25	0.255	958	5.219	732	544	146.400
escyb033 5012	1351	856	68	32	0.270	956	5.243	811	540	161.812
escyb034 5002	1345	853	107	23	0.269	983	5.089	797	548	159.336
escyb035 5007	1314	877	64	21	0.262	962	5.205	761	553	151.987
escyb041 5000	1310	656	65	33	0.262	754	6.631	707	603	141.400
escyb051 5006	1301	884	54	38	0.260	976	5.129	743	558	148.422
escyb052 5004	1297	999	65	25	0.259	1089	4.595	729	568	145.683
escyb071 5011	1447	727	64	18	0.289	809	6.194	880	567	175.614
escyb072 5003	1493	806	29	21	0.298	856	5.845	923	570	184.489
escyb081 5003	1434	837	55	25	0.287	917	5.456	889	545	177.693
escyb111 4986	1382	639	121	24	0.277	784	6.360	793	589	159.045
escyb112 5008	1373	547	117	29	0.274	693	7.227	781	592	155.950
escyb121 5002	1368	552	65	38	0.273	655	7.637	729	639	145.742
escyb131 5005	1431	549	119	22	0.286	690	7.254	802	629	160.240
escyb132 4994	1306	727	35	36	0.262	798	6.258	752	554	150.581
eshel011 5007	1394	675	10	22	0.278	707	7.082	808	586	161.374
eshel012 5004	1433	662	17	10	0.286	689	7.263	851	582	170.064
eshel031 5001	1372	661	9	14	0.274	684	7.311	800	572	159.968
eshel032 5004	1428	628	14	6	0.285	648	7.722	855	573	170.863
eshel033 5007	1406	662	6	8	0.281	676	7.407	846	560	168.963
eshel051 5002	1403	615	14	23	0.280	652	7.672	805	598	160.936
eshel052 5010	1347	657	10	12	0.269	679	7.379	779	568	155.489
eshel053 5006	1539	638	54	16	0.307	708	7.071	959	580	191.570
eshel071 5003	1398	660	11	14	0.279	685	7.304	809	589	161.703
eshel072 5003	1476	610	31	11	0.295	652	7.673	900	576	179.892
eshel081 5001	1359	618	22	20	0.272	660	7.577	796	563	159.168
eshel091 5007	1381	623	14	29	0.276	666	7.518	817	564	163.172
eshel141 5001	1419	584	3	16	0.284	603	8.294	834	585	166.767
eshel151 5001	1384	572	11	17	0.277	600	8.335	810	574	161.968
eshel161 5005	1458	685	12	27	0.291	724	6.913	848	610	169.431
eshel171 5016	1335	642	8	23	0.266	673	7.453	716	619	142.743
eshel172 5000	1080	580	13	13	0.216	606	8.251	599	481	119.800
eshel191 5003	1437	573	5	29	0.287	607	8.242	851	586	170.098
eshel192 5002	1370	602	4	21	0.274	627	7.978	785	585	156.937
eshel193 5000	1484	740	9	16	0.297	765	6.536	912	572	182.400
itaus031 5009	1499	886	31	16	0.299	933	5.369	904	595	180.475
itaus032 5005	1490	754	43	15	0.298	812	6.164	919	571	183.616

itaus041	5002	1475	739	30	14	0.295	783	6.388	954	521	190.724
itaus061	5002	1328	935	27	35	0.265	997	5.017	809	519	161.735
itaus151	5008	1373	924	41	43	0.274	1008	4.968	798	575	159.345
itaus152	5003	1447	860	24	40	0.289	924	5.415	918	529	183.490
itaus153	5004	1556	850	28	12	0.311	890	5.622	1012	544	202.238
itaus161	5000	1448	821	35	29	0.290	885	5.650	879	569	175.800
itaus162	5003	1413	868	47	31	0.282	946	5.289	854	559	170.698
itaus163	5006	1433	774	32	24	0.286	830	6.031	877	556	175.190
itaus171	5006	1600	792	33	41	0.320	866	5.781	1053	547	210.348
itaus172	4999	1477	856	40	35	0.295	931	5.370	891	586	178.236
itaus173	5006	1398	964	78	33	0.279	1075	4.657	863	535	172.393
itaus181	5000	1359	1023	62	33	0.272	1118	4.472	830	529	166.000
itaus182	5006	1343	821	45	15	0.268	881	5.682	848	495	169.397
itaus183	5005	1380	849	38	18	0.276	905	5.530	823	557	164.436
itaus184	4998	1434	885	55	39	0.287	979	5.105	883	551	176.671
itaus185	5006	1525	796	41	24	0.305	861	5.814	942	583	188.174
itaus191	5004	1199	945	55	35	0.240	1035	4.835	673	526	134.492
itaus192	5003	1490	878	26	42	0.298	946	5.289	940	550	187.887
itphi011	5006	1598	599	8	10	0.319	617	8.113	1049	549	209.549
itphi031	5002	1548	843	31	16	0.309	890	5.620	1017	531	203.319
itphi041	5001	1497	781	12	23	0.299	816	6.129	957	540	191.362
itphi121	5000	1587	720	12	30	0.317	762	6.562	1018	569	203.600
itphi151	5003	1682	537	11	21	0.336	569	8.793	1103	579	220.468
itphi161	5013	1261	612	12	8	0.252	632	7.932	762	499	152.005
itphi162	5014	1442	682	17	7	0.288	706	7.102	880	562	175.509
itphi171	5004	1277	462	16	9	0.255	487	10.275	799	478	159.672
itphi172	5010	1296	515	52	6	0.259	573	8.743	786	510	156.886
itphi181	5001	1401	450	31	4	0.280	485	10.311	905	496	180.964
itphi182	5002	1632	632	46	15	0.326	693	7.218	1064	568	212.715
itphi191	4999	1557	607	26	19	0.311	652	7.667	996	561	199.240
itphi201	5001	1506	605	7	14	0.301	626	7.989	944	562	188.762
itphi202	5010	1538	623	23	16	0.307	662	7.568	1008	530	201.198
itphi203	5010	1533	592	16	12	0.306	620	8.081	993	540	198.204
itphi211	5007	1858	650	12	18	0.371	680	7.363	1253	605	250.250
itphi212	5000	1571	726	19	12	0.314	757	6.605	1048	523	209.600
itphi213	5003	1560	745	13	12	0.312	770	6.497	999	561	199.680
itphi221	4994	1379	501	11	2	0.276	514	9.716	821	558	164.397
itphi222	5003	1502	441	4	1	0.300	446	11.217	1005	497	200.879
utbeg011	5003	1523	571	83	26	0.304	680	7.357	917	606	183.290
utbeg012	5012	1529	467	98	32	0.305	597	8.395	922	607	183.959
utbeg013	5011	1528	549	62	43	0.305	654	7.662	882	646	176.013
utbeg031	5004	1429	489	93	21	0.286	603	8.299	824	605	164.668
utbeg032	5015	1404	571	66	36	0.280	673	7.452	862	542	171.884
utbeg041	5004	1580	557	50	28	0.316	635	7.880	963	617	192.446
utbeg042	4999	1658	516	77	21	0.332	614	8.142	1058	600	211.642
utbeg051	5014	1437	554	57	62	0.287	673	7.450	825	612	164.539
utbeg052	5006	1409	509	118	27	0.281	654	7.654	807	602	161.207
utbeg053	5010	1401	610	82	27	0.280	719	6.968	837	564	167.066
utbeg054	5002	1367	537	88	30	0.273	655	7.637	776	591	155.138
utbeg071	5002	1469	546	74	40	0.294	660	7.579	866	603	173.131
utbeg072	5003	1505	541	93	31	0.301	665	7.523	834	671	166.700
utbeg073	5006	1481	573	84	36	0.296	693	7.224	881	600	175.989
utbeg074	5003	1429	583	81	30	0.286	694	7.209	847	582	169.298

utbeg075 5003	1334	496	78	43	0.267	617	8.109	703	631	140.516
utbeg081 5007	1374	548	47	29	0.274	624	8.024	804	570	160.575
utbeg082 5000	1461	524	74	24	0.292	622	8.039	840	621	168.000
utbeg083 4996	1442	580	63	30	0.289	673	7.423	807	635	161.529
utbeg084 5002	1443	575	85	19	0.288	679	7.367	800	643	159.936
utcha011 5001	1457	716	48	23	0.291	787	6.355	883	574	176.565
utcha012 5003	1330	722	49	27	0.266	798	6.269	757	573	151.309
utcha013 5000	1403	782	54	24	0.281	860	5.814	830	573	166.000
utcha031 5003	1401	671	128	37	0.280	836	5.984	805	596	160.903
utcha041 5006	1358	773	69	16	0.271	858	5.835	825	533	164.802
utcha061 5002	1515	715	70	39	0.303	824	6.070	937	578	187.325
utcha062 5003	1228	783	80	29	0.245	892	5.609	680	548	135.918
utcha071 5016	1337	646	139	8	0.267	793	6.325	765	572	152.512
utcha081 5011	1353	688	100	22	0.270	810	6.186	781	572	155.857
utcha082 5002	1435	734	90	26	0.287	850	5.885	846	589	169.132
utcha121 5001	1424	789	83	63	0.285	935	5.349	852	572	170.366
utcha122 5010	1524	684	71	67	0.304	822	6.095	910	614	181.637
utcha131 5008	1341	885	57	74	0.268	1016	4.929	769	572	153.554
utcha141 5000	1365	763	28	63	0.273	854	5.855	787	578	157.400
utcha151 5001	1408	781	56	66	0.282	903	5.538	817	591	163.367
utcha152 5001	1522	779	28	80	0.304	887	5.638	894	628	178.764
utcha161 5001	1435	822	40	48	0.287	910	5.496	818	617	163.567
utcha162 5003	1396	807	73	47	0.279	927	5.397	820	576	163.902
utcha163 5001	1474	750	76	68	0.295	894	5.594	888	586	177.564
utcha164 5005	1384	701	123	55	0.277	879	5.694	801	583	160.040

Appendix B:

The hundred most frequent words and their frequencies in the CIRC, the LLC, and the Brown corpus.

	CIRC	LLC	Brown
i	1.845	1.241	0.517
is	1.551	0.475	1.009
to	1.484	1.085	2.615
the	1.427	1.774	6.997
a	1.382	0.9	2.323
you	1.125	0.801	0.329
*	1.036	n/a	n/a
***	0.873	n/a	n/a
and	0.836	1.267	2.885
it	0.765	0.699	0.876
now	0.727	0.141	0.131
me	0.691	0.114	0.121
in	0.662	0.717	2.134
as	0.660	0.227	0.725
that	0.622	0.719	1.059
hi	0.619	n/l	n/l
lol	0.593	n/a	n/a
was	0.543	0.412	0.982

known	0.540	n/l	n/l
of	0.531	1.027	3.641
:)	0.530	n/a	n/a
u	0.500	n/a	n/a
for	0.441	0.275	0.949
are	0.425	0.174	0.439
on	0.422	0.323	0.674
what	0.407	0.197	0.191
have	0.405	0.281	0.394
hey	0.394	n/l	n/l
my	0.393	0.124	0.132
all	0.384	0.21	0.3
not	0.378	0.203	0.461
by	0.368	0.118	0.53
with	0.361	0.219	0.729
do	0.351	0.179	0.136
)	0.347	n/a	n/a
here	0.323	n/l	n/l
how	0.318	n/l	n/l
no	0.318	0.219	0.22
from	0.306	0.139	0.437
kicked	0.294	n/l	n/l
so	0.289	0.227	0.198
be	0.284	0.274	0.638
hello	0.283	n/l	n/l
chat	0.280	n/l	n/l
but	0.275	0.339	0.438
just	0.259	0.141	n/l
any	0.257	n/l	0.135
like	0.251	0.109	0.129
there	0.250	0.195	0.272
know	0.250	0.272	n/l
your	0.244	0.082	0.092
can	0.244	0.11	0.177
get	0.238	0.106	n/l
if	0.237	0.176	0.22
anyone	0.230	n/l	n/l
or	0.226	0.166	0.421
one	0.220	0.201	0.323
good	0.216	n/l	n/l
up	0.210	0.122	0.189
wanna	0.210	n/l	n/l
at	0.206	0.237	0.538
this	0.202	0.354	0.515
ya	0.198	n/l	n/l
ok	0.196	n/l	n/l
i'm	0.188	0.124	n/l
go	0.187	0.085	n/l
am	0.169	n/l	n/l
we	0.169	0.268	0.265
msg	0.167	n/a	n/a

well	0.167	0.31	0.089
too	0.164	n/l	n/l
about	0.163	0.179	0.182
they	0.162	0.266	0.362
:	0.162	n/a	n/a
im	0.159	n/l	n/l
out	0.155	0.111	0.207
he	0.155	0.3	0.954
hehe	0.154	n/l	n/l
back	0.153	n/l	0.097
want	0.152	n/l	n/l
oh	0.152	0.152	n/l
<kinoko>	0.150	n/a	n/a
see	0.149	0.141	n/l
an	0.147	0.094	0.375
thanks	0.140	n/l	n/l
<chmod>	0.137	n/a	n/a
think	0.137	0.228	n/l
then	0.136	0.121	0.138
some	0.135	0.083	0.162
got	0.134	0.149	n/l
bye	0.133	n/l	n/l
who	0.132	0.089	0.225
its	0.132	n/l	0.186
where	0.132	n/l	n/l
yeah	0.132	0.154	n/l
dont	0.129	n/l	n/l
will	0.124	n/l	0.224
has	0.124	0.079	0.244
why	0.124	n/l	n/l
2	0.121	n/a	n/a

Appendix C:

The hundred most frequent words (frequencies and ranks) corrected for system messages

words	frequencies			ranks			
	CIRC	LLC	Brown	CIRC	LLC	LOB	Brown
i	1.845	1.241	0.517	1	3	17	20
to	1.484	1.085	2.615	2	4	4	4
the	1.427	1.774	6.997	3	1	1	1
a	1.382	0.9	2.323	4	6	5	5
you	1.125	0.801	0.329	5	7	32	33
*	1.036	n/a	n/a	6	n/a	n/a	n/a
is	1.016	0.475	1.009	7	11	8	8
***	0.873	n/a	n/a	8	n/a	n/a	n/a
and	0.836	1.267	2.885	9	2	3	3

it	0.765	0.699	0.876	10	10	10	12
me	0.691	0.114	0.121	11	66	66	82
in	0.662	0.717	2.134	12	9	6	6
that	0.622	0.719	1.059	13	8	7	7
hi	0.619	n/a	n/a	14	n/a	n/a	n/a
lol	0.593	n/a	n/a	15	n/a	n/a	n/a
of	0.531	1.027	3.641	16	5	2	2
:)	0.530	n/a	n/a	17	n/a	n/a	n/a
u	0.500	n/a	n/a	18	n/a	n/a	n/a
for	0.441	0.275	0.949	19	20	11	11
are	0.425	0.174	0.439	20	42	27	24
on	0.422	0.323	0.674	21	16	16	16
what	0.407	0.197	0.191	22	37	55	54
have	0.405	0.281	0.394	23	19	26	28
hey	0.394	n/a	n/a	24	n/a	n/a	n/a
my	0.393	0.124	0.132	25	58	58	75
all	0.384	0.21	0.3	26	33	39	36
not	0.378	0.203	0.461	27	35	23	23
with	0.361	0.219	0.729	28	32	14	13
do	0.351	0.179	0.136	29	40	73	72
)	0.347	n/a	n/a	30	n/a	n/a	n/a
here	0.323	n/a	n/a	31	n/a	n/a	n/a
how	0.318	n/a	n/a	32	n/a	n/a	n/a
no	0.318	0.219	0.22	33	31	47	49
from	0.306	0.139	0.437	34	53	25	26
so	0.289	0.227	0.198	35	30	46	52
be	0.284	0.274	0.638	36	21	15	17
hello	0.283	n/a	n/a	37	n/a	n/a	n/a
chat	0.280	n/a	n/a	38	n/a	n/a	n/a
but	0.275	0.339	0.438	39	15	24	25
was	0.263	0.412	0.982	40	13	9	9
just	0.259	0.141	n/a	41	52	n/a	n/a
any	0.257	n/a	0.135	42	n/a	75	74
like	0.251	0.109	0.129	43	71	83	78
there	0.250	0.195	0.272	44	38	36	38
know	0.250	0.272	n/a	45	22	n/a	n/a
your	0.244	0.082	0.092	46	93	n/a	98
can	0.244	0.11	0.177	47	70	61	61
get	0.238	0.106	n/a	48	74	n/a	n/a
if	0.237	0.176	0.22	49	41	45	50
anyone	0.230	n/a	n/a	50	n/a	n/a	n/a
or	0.226	0.166	0.421	51	44	31	27
one	0.220	0.201	0.323	52	36	38	32
good	0.216	n/a	n/a	53	n/a	n/a	n/a
up	0.210	0.122	0.189	54	61	56	55
wanna	0.210	n/a	n/a	55	n/a	n/a	n/a
at	0.206	0.237	0.538	56	26	19	18
this	0.202	0.354	0.515	57	14	22	21
ya	0.198	n/a	n/a	58	n/a	n/a	n/a
ok	0.196	n/a	n/a	59	n/a	n/a	n/a
now	0.192	0.141	0.131	60	51	72	76

i'm	0.188	0.124	n/a	61	57	n/a	n/a
go	0.187	0.085	n/a	62	86	n/a	n/a
am	0.169	n/a	n/a	63	n/a	n/a	n/a
we	0.169	0.268	0.265	64	23	40	41
msg	0.167	n/a	n/a	65	n/a	n/a	n/a
well	0.167	0.31	0.089	66	17	95	100
too	0.164	n/a	n/a	67	n/a	n/a	n/a
about	0.163	0.179	0.182	68	39	54	57
they	0.162	0.266	0.362	69	24	33	30
:	0.162	n/a	n/a	70	n/a	n/a	n/a
im	0.159	n/a	n/a	71	n/a	n/a	n/a
out	0.155	0.111	0.207	72	69	53	51
he	0.155	0.3	0.954	73	18	12	10
hehe	0.154	n/a	n/a	74	n/a	n/a	n/a
back	0.153	n/a	0.097	75	n/a	n/a	94
want	0.152	n/a	n/a	76	n/a	n/a	n/a
oh	0.152	0.152	n/a	77	47	n/a	n/a
<kinoko>	0.150	n/a	n/a	78	n/a	n/a	n/a
see	0.149	0.141	n/a	79	50	n/a	n/a
an	0.147	0.094	0.375	80	81	34	29
thanks	0.140	n/a	n/a	81	n/a	n/a	n/a
<chmod>	0.137	n/a	n/a	82	n/a	n/a	n/a
think	0.137	0.228	n/a	83	28	n/a	n/a
then	0.136	0.121	0.138	84	63	68	71
some	0.135	0.083	0.162	85	90	57	65
got	0.134	0.149	n/a	86	49	n/a	n/a
bye	0.133	n/a	n/a	87	n/a	n/a	n/a
who	0.132	0.089	0.225	88	83	50	46
its	0.132	n/a	0.186	89	n/a	n/a	56
where	0.132	n/a	n/a	90	n/a	n/a	96
yeah	0.132	0.154	n/a	91	46	n/a	n/a
dont	0.129	n/a	n/a	92	n/a	n/a	n/a
as	0.125	0.227	0.725	93	29	13	14
will	0.124	n/a	0.224	94	n/a	n/a	47
has	0.124	0.079	0.244	95	94	n/a	44
why	0.124	n/a	n/a	96	n/a	n/a	n/a
2	0.121	n/a	n/a	97	n/a	n/a	n/a
please	0.120	n/a	n/a	98	n/a	n/a	n/a
4	0.120	n/a	n/a	99	n/a	n/a	n/a
did	0.119	0.075	n/a	100	98	n/a	n/a

Appendix D:

Correlations not corrected for system messages

Correlations between frequencies:

CIRC	Pearson Correlation	CIRC	LLC	BROWN
	Sig. (2-tailed)	1.000	.793	.557
	N	100	62	56
LLC	Pearson Correlation	.793	1.000	.845
	Sig. (2-tailed)	.000	.	.000
	N	62	62	52
BROWN	Pearson Correlation	.557	.845	1.000
	Sig. (2-tailed)	.000	.000	.
	N	56	52	56

** Correlation is significant at the 0.01 level (2-tailed).

Correlations between ranks:

Spearman's rho	CIRC	Correlation Coefficient	CIRC	LLC	LOB	BROWN
		Sig. (2-tailed)	1.000	.596	.545	.543
		N	100	62	51	57
	LLC	Correlation Coefficient	.596	1.000	.762	.740
		Sig. (2-tailed)	.000	.	.000	.000
		N	62	62	50	52
	LOB	Correlation Coefficient	.545	.762	1.000	.993
		Sig. (2-tailed)	.000	.000	.	.000
		N	51	50	51	51
	BROWN	Correlation Coefficient	.543	.740	.993	1.000
		Sig. (2-tailed)	.000	.000	.000	.
		N	57	52	51	57

** Correlation is significant at the .01 level (2-tailed).

Appendix E:

Correlations corrected for system messages

Correlations between frequencies:

CIRCFR	Pearson Correlation	CIRCFR	LLCFR	BROWNFR
	Sig. (2-tailed)	1.000	.818	.523
	N	110	70	62
LLCFR	Pearson Correlation	.818	1.000	.814
	Sig. (2-tailed)	.000	.	.000
	N	70	70	57
BROWNFR	Pearson Correlation	.523	.814	1.000
	Sig. (2-tailed)	.000	.000	.
	N	62	57	62

** Correlation is significant at the 0.01 level (2-tailed).

Correlations between ranks:

Spearman's rho	CIRCRA	Correlation Coefficient	CIRCRA	LLCRA	LOBRA	BROWNRA
		Sig. (2-tailed)	1.000	.647	.367	.449
		N	105	67	54	61

LLCRA	Correlation Coefficient	.647	1.000	.675	.715
	Sig. (2-tailed)	.000	.	.000	.000
	N	67	70	54	57
LOBRA	Correlation Coefficient	.367	.675	1.000	.953
	Sig. (2-tailed)	.006	.000	.	.000
	N	54	54	55	55
BROWNRA	Correlation Coefficient	.449	.715	.953	1.000
	Sig. (2-tailed)	.000	.000	.000	.
	N	61	57	55	62

** Correlation is significant at the .01 level (2-tailed).

Appendix F:

Sample passages from the CIRC.

From file acper052:

```

<participantm> *pfft*
<participantm> Ive changed it twice this mornin, big woop
<participantm> swollen glands in right hand side of neck
<{participants}> I don't like Korn i prefer Peas
<participantt> CYA ALLL!!!! hmmm....nauti me
<participantm> possible gland infection
<participantr> shit hey
<participantr> not good
<participantm> mmm sounds like glandular fever
<participantm> *chuckles*
<participants> Penis
<participantr> poiniz
<participantm> no know remedy but rest
<participantr> brb
<participantm> but i aint resting thats bullshit
<participants> soothers
<participantm> *bah*
<participantm> Soothers do shit for a gland infection
<participantt> hehe....you cloning participantr??
<participantt> hehehehehehehe
<participants> no
<^participantv^> bwhahahhahahahaha
<participantm> i'm allergic to most soothers
<^participantv^> rama and sonic clones hahahahaha yer tony :P
<^participantv^> huuhhuhuhu
<^participantv^> they may look alike :P
<participants> mmmmm, cherry soothers
<participantm> *blerk*
*** participantm is now known as participantb
* participantr bootz shauna
<participantr> ehyehehe
<participantb> Oi
<participantb> Stoopid anal monkey
<^participantv^> told gibbo
<participantb> heh
<participants> Penis
<participantk> participantr!!!!!!!!!!!!!!SEXAE!!!!!!!!!!!!!!

```

```

<participantw> hi all
<participantg> hey everyone thanx 4 the chat.....cya'z l8er :):):) !!!!!
<participantr> hi participantk hunnie
<^participantv^> ahhh well participantk has an orgasm when she sees u
<participantb> *gargles at the thought*
<participantr> all chicks do boris
<^participantv^> im sure herc wount be too happy about that
<participantr> :)
<^participantv^> hehehe
<participantr> heheheheh
<participants> sure
<participantb> Pfft
<participantb> Sandy @$@#
<^participantv^> gibbo: then u wake up\
<participantr> open up shauna, im cumming in!
* participantk looks at boris...at least i orgasm!
<participantr> hahahahaha
<participants> oh sweeeeeeeeet

```

From file eshell171:

```

Session Start: Sun Oct 03 19:34:27 1999
*** Now talking in #help
*** Topic is '-=UNDERNET HELP=- ?'s about the *@*.tm.net.my ban or Undernet
    Proxy Check, goto http://help.undernet.org, Need a IRCop /quote who 0 o,
    Email: help@undernet.org '
*** Set by X on Sun Oct 03 14:31:25
<participantm> ta ta
<participantf> i need "Winsock 2.2" to play Mech3 over tcp/ip what do i do to
    get it? please quick
<participantc> any body can help me
<participantc> elm plz help
<participantb> hey
<participantr> heya
<participantb> can I turn a wav file into an mp3 file?
<participanta> hi, I am trying to join the channel #bast, but it says that is
    no channnel
<participants> participantb sure
<participantb> how?
<participantf> i need "Winsock 2.2" to play Mech3 over tcp/ip what do i do to
    get it? please quick
<participants> use mp3 encoder
<participantc> ey i'v scanned my pc and it say that i'm infected my netbus so
    how i can remove it
<participanti> i got a avp
<participants> faster way i know using a software call music match
<participantb> where?
<participantb> ok
<participanti> how do i work avp
<participantb> ho
<participantb> hi
<participantb> what's my IP address?
<participantb> the numeric one?
<participantb> how can i find out
<participantc> yup
<participantw> participantb, do /dns yournickname

```

```

<participanth> thanks
<participantw> easiest way...
<participantw> //say $ip
<participanth> $ip
<participanth> ahh!
<participanth> 194.176.209.212
<participanth> thanks
<participanth> :)
<participantw> no problemo
* participantw will brb
<participantq> Type /server falcon.phix.com and join #falconnet or #freestyle-
warez
<participantq> Please! Perty Perty Please!
<participantq> Type /server falcon.phix.com and join #falconnet or #freestyle-
warez
<participantq> Please! Perty Perty Please!
<participantq> Type /server falcon.phix.com and join #falconnet or #freestyle-
warez
<participantq> Please! Perty Perty Please!
<participantq> Type /server falcon.phix.com and join #falconnet or #freestyle-
warez
*** participantq was kicked by SafeBot (flood )
<participanth> lol
<participantg> hola
<participantd> anyone know how to use BO2K?
<participantn> whats up guyd
<participantn> guys
<participantn> peace
<participantv> can anyone get me .dll files?
<participantj> can u guys help me with games
<^participant4^> maybe
<participantv> i need dsw32.dll
<participantv> please.
<participantj> have u played streetwars?
<^participant4^> i don't think so
<^participant4^> Virus: i don't have dsw32.dll
<participantv> do you no someone who does?
<^participant4^> what program uses it?
<^participant4^> or os
<participantv> i'm wantin it for zmud
<^participant4^> i don't know, keep askin around
<participantv> k
<participantr> someone has seen somebody?
<^participant4^> what
<^participant4^> i haven't
<participantr> ok
<participantz> QUIEN ES DE PMA
<participantu> I need help!
<participantz> yes
<participantu> What is a good place to get JavaScript from?
<participantz> i
<participantu> Anybody here know JavaScript?
<participantz> wath your name
<participantu> Anybody here know JavaScript?
<participantd--> what do i type to gove someone access?
<participantd--> i forgot
<participantd--> i forgot
<participantd--> what do i type to gove someone access?

```

```

<participanty> participantd-- ADDUSER syntax - /msg X@channels.undernet.org
adduser #channel <nick> [user@host] <level> <password> OR /msg
W@channels2.undernet.org adduser #channel <nick> [user@host] <level>
<password> If the person is online user@host can be left out. You can
only add people with access lower than yours. Requires level 400+
<participanty> Oh look! Its Jady!
<participantd--> how do i turn someones auto op on??
<participanty> participantd-- /msg XorW modinfo <#channel> AUTOOP <user>
<OFF/ON> :Changes autoop from ON to OFF. Autoop means that X/W will op
the user when they enter the channel. Requires level 400+
<participanti> what is the adress (url) of the undernet regestering page ?
<participanti> for W or X...
<participanty> participanti: http://www.cservice.undernet.org/regist
<participanty> no
<participanty> participanti: http://cservice.undernet.org/regist
<participanti> thanx

```

Appendix G:

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
TT	Between Groups	2.076E-02	9	2.307E-03	6.987	.000
	Within Groups	6.273E-02	190	3.302E-04		
	Total	8.349E-02	199			
HAPAXMIL	Between Groups	25253.629	9	2805.959	11.459	.000
	Within Groups	46525.917	190	244.873		
	Total	71779.546	199			
WM	Between Groups	148.969	9	16.552	32.140	.000
	Within Groups	97.849	190	.515		
	Total	246.817	199			

Appendix H:

Post Hoc tests

Multiple Comparisons

Dependent Variable	(I)	(J)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
TT Scheffe	C	C					
	H	H					
	A	A					
	N	N					
	E	E					
TT Scheffe	L	L					
	1	2	-2.36401E-02	5.7460E-03	.057	-4.75845E-02	3.0427E-04

	3	-1.38043E-02	5.7460E-03	.761	-3.77486E-02	1.0140E-02
	4	-5.99778E-03	5.7460E-03	.999	-2.99421E-02	1.7947E-02
	5	-7.17395E-03	5.7460E-03	.997	-3.11183E-02	1.6770E-02
	6	-1.30362E-02	5.7460E-03	.819	-3.69806E-02	1.0908E-02
	7	-2.06924E-02	5.7460E-03	.173	-4.46367E-02	3.2520E-03
	8	-3.62423E-02	5.7460E-03	.000	-6.01866E-02	1.22979E-02
	9	-2.59729E-02	5.7460E-03	.019	-4.99173E-02	2.02854E-03
2	10	-1.49224E-02	5.7460E-03	.663	-3.88668E-02	9.0220E-03
	1	2.3640E-02	5.7460E-03	.057	-3.04269E-04	4.7584E-02
	3	9.8358E-03	5.7460E-03	.966	-1.41085E-02	3.3780E-02
	4	1.7642E-02	5.7460E-03	.404	-6.30204E-03	4.1587E-02
	5	1.6466E-02	5.7460E-03	.515	-7.47822E-03	4.0411E-02
	6	1.0604E-02	5.7460E-03	.944	-1.33405E-02	3.4548E-02
	7	2.9477E-03	5.7460E-03	1.000	-2.09966E-02	2.6892E-02
	8	-1.26022E-02	5.7460E-03	.848	-3.65465E-02	1.1342E-02
	9	-2.33281E-03	5.7460E-03	1.000	-2.62772E-02	2.1612E-02
	10	8.7177E-03	5.7460E-03	.985	-1.52267E-02	3.2662E-02
3	1	1.3804E-02	5.7460E-03	.761	-1.01401E-02	3.7749E-02
	2	-9.83582E-03	5.7460E-03	.966	-3.37802E-02	1.4109E-02
	4	7.8065E-03	5.7460E-03	.993	-1.61379E-02	3.1751E-02
	5	6.6303E-03	5.7460E-03	.998	-1.73140E-02	3.0575E-02
	6	7.6807E-04	5.7460E-03	1.000	-2.31763E-02	2.4712E-02
	7	-6.88810E-03	5.7460E-03	.997	-3.08325E-02	1.7056E-02
	8	-2.24380E-02	5.7460E-03	.093	-4.63824E-02	1.5064E-03
	9	-1.21686E-02	5.7460E-03	.874	-3.61130E-02	1.1776E-02
	10	-1.11810E-03	5.7460E-03	1.000	-2.50625E-02	2.2826E-02
4	1	5.9978E-03	5.7460E-03	.999	-1.79466E-02	2.9942E-02
	2	-1.76423E-02	5.7460E-03	.404	-4.15867E-02	6.3020E-03
	3	-7.80650E-03	5.7460E-03	.993	-3.17509E-02	1.6138E-02
	5	-1.17617E-03	5.7460E-03	1.000	-2.51205E-02	2.2768E-02
	6	-7.03843E-03	5.7460E-03	.997	-3.09828E-02	1.6906E-02
	7	-1.46948E-02	5.7460E-03	.684	-3.86390E-02	9.2498E-03
	8	-3.02445E-02	5.7460E-03	.002	-5.41889E-02	6.30012E-03
	9	-1.99751E-02	5.7460E-03	.217	-4.39195E-02	3.9692E-03
	10	-8.92461E-03	5.7460E-03	.982	-3.28690E-02	1.5020E-02
5	1	7.1739E-03	5.7460E-03	.997	-1.67704E-02	3.1118E-02
	2	-1.64662E-02	5.7460E-03	.515	-4.04105E-02	7.4782E-03
	3	-6.63033E-03	5.7460E-03	.998	-3.05747E-02	1.7314E-02
	4	1.1762E-03	5.7460E-03	1.000	-2.27682E-02	2.5121E-02
	6	-5.86226E-03	5.7460E-03	.999	-2.98066E-02	1.8082E-02
	7	-1.35184E-02	5.7460E-03	.783	-3.74628E-02	1.0426E-02
	8	-2.90683E-02	5.7460E-03	.004	-5.30127E-02	5.12394E-03
	9	-1.87990E-02	5.7460E-03	.304	-4.27433E-02	5.1454E-03
	10	-7.74843E-03	5.7460E-03	.994	-3.16928E-02	1.6196E-02
6	1	1.3036E-02	5.7460E-03	.819	-1.09082E-02	3.6981E-02
	2	-1.06039E-02	5.7460E-03	.944	-3.45483E-02	1.3340E-02
	3	-7.68069E-04	5.7460E-03	1.000	-2.47124E-02	2.3176E-02
	4	7.0384E-03	5.7460E-03	.997	-1.69059E-02	3.0983E-02
	5	5.8623E-03	5.7460E-03	.999	-1.80821E-02	2.9807E-02
	7	-7.65616E-03	5.7460E-03	.994	-3.16005E-02	1.6288E-02
	8	-2.32061E-02	5.7460E-03	.068	-4.71504E-02	7.3832E-04
	9	-1.29367E-02	5.7460E-03	.826	-3.68811E-02	1.1008E-02
	10	-1.88617E-03	5.7460E-03	1.000	-2.58305E-02	2.2058E-02

7	1	2.0692E-02	5.7460E-03	.173	-3.25200E-03	4.4637E-02
	2	-2.94773E-03	5.7460E-03	.000	-2.68921E-02	2.0997E-02
	3	6.8881E-03	5.7460E-03	.997	-1.70563E-02	3.0832E-02
	4	1.4695E-02	5.7460E-03	.684	-9.24977E-03	3.8639E-02
	5	1.3518E-02	5.7460E-03	.783	-1.04259E-02	3.7463E-02
	6	7.6562E-03	5.7460E-03	.994	-1.62882E-02	3.1601E-02
	8	-1.55499E-02	5.7460E-03	.604	-3.94943E-02	8.3945E-03
	9	-5.28054E-03	5.7460E-03	.000	-2.92249E-02	1.8664E-02
	10	5.7700E-03	5.7460E-03	.999	-1.81744E-02	2.9714E-02
8	1	3.6242E-02	5.7460E-03	.000	1.2298E-02	6.0187E-02
	2	1.2602E-02	5.7460E-03	.848	-1.13422E-02	3.6547E-02
	3	2.2438E-02	5.7460E-03	.093	-1.50639E-03	4.6382E-02
	4	3.0244E-02	5.7460E-03	.002	6.3001E-03	5.4189E-02
	5	2.9068E-02	5.7460E-03	.004	5.1239E-03	5.3013E-02
	6	2.3206E-02	5.7460E-03	.068	-7.38317E-04	4.7150E-02
	7	1.5550E-02	5.7460E-03	.604	-8.39448E-03	3.9494E-02
	9	1.0269E-02	5.7460E-03	.955	-1.36750E-02	3.4214E-02
	10	2.1320E-02	5.7460E-03	.140	-2.62449E-03	4.5264E-02
9	1	2.5973E-02	5.7460E-03	.019	2.0285E-03	4.9917E-02
	2	2.3328E-03	5.7460E-03	.000	-2.16116E-02	2.6277E-02
	3	1.2169E-02	5.7460E-03	.874	-1.17757E-02	3.6113E-02
	4	1.9975E-02	5.7460E-03	.217	-3.96924E-03	4.3920E-02
	5	1.8799E-02	5.7460E-03	.304	-5.14541E-03	4.2743E-02
	6	1.2937E-02	5.7460E-03	.826	-1.10077E-02	3.6881E-02
	7	5.2805E-03	5.7460E-03	.000	-1.86638E-02	2.9225E-02
	8	-1.02694E-02	5.7460E-03	.955	-3.42137E-02	1.3675E-02
	10	1.1051E-02	5.7460E-03	.928	-1.28938E-02	3.4995E-02
10	1	1.4922E-02	5.7460E-03	.663	-9.02199E-03	3.8867E-02
	2	-8.71772E-03	5.7460E-03	.985	-3.26621E-02	1.5227E-02
	3	1.1181E-03	5.7460E-03	.000	-2.28263E-02	2.5062E-02
	4	8.9246E-03	5.7460E-03	.982	-1.50198E-02	3.2869E-02
	5	7.7484E-03	5.7460E-03	.994	-1.61959E-02	3.1693E-02
	6	1.8862E-03	5.7460E-03	.000	-2.20582E-02	2.5831E-02
	7	-5.76999E-03	5.7460E-03	.999	-2.97144E-02	1.8174E-02
	8	-2.13199E-02	5.7460E-03	.140	-4.52643E-02	2.6245E-03
	9	-1.10505E-02	5.7460E-03	.928	-3.49949E-02	1.2894E-02
	2	-2.36401E-02	5.7460E-03	.003	-4.26674E-02	-4.61280E-03
Bonferroni	3	-1.38043E-02	5.7460E-03	.776	-3.28316E-02	5.2230E-03
	4	-5.99778E-03	5.7460E-03	.000	-2.50251E-02	1.3030E-02
	5	-7.17395E-03	5.7460E-03	.000	-2.62012E-02	1.1853E-02
	6	-1.30362E-02	5.7460E-03	.000	-3.20635E-02	5.9911E-03
	7	-2.06924E-02	5.7460E-03	.018	-3.97197E-02	-1.66507E-03
	8	-3.62423E-02	5.7460E-03	.000	-5.52696E-02	-1.72150E-02
	9	-2.59729E-02	5.7460E-03	.000	-4.50002E-02	-6.94561E-03
	10	-1.49224E-02	5.7460E-03	.456	-3.39497E-02	4.1049E-03
	2	2.3640E-02	5.7460E-03	.003	4.6128E-03	4.2667E-02
	3	9.8358E-03	5.7460E-03	.000	-9.19148E-03	2.8863E-02
	4	1.7642E-02	5.7460E-03	.110	-1.38497E-03	3.6670E-02
	5	1.6466E-02	5.7460E-03	.208	-2.56115E-03	3.5493E-02
	6	1.0604E-02	5.7460E-03	.000	-8.42341E-03	2.9631E-02

3	7	2.9477E-03	5.7460E-031.000	-1.60796E-02	2.1975E-02
	8	-1.26022E-02	5.7460E-031.000	-3.16295E-02	6.4251E-03
	9	-2.33281E-03	5.7460E-031.000	-2.13601E-02	1.6694E-02
	10	8.7177E-03	5.7460E-031.000	-1.03096E-02	2.7745E-02
	1	1.3804E-02	5.7460E-03 .776	-5.22302E-03	3.2832E-02
	2	-9.83582E-03	5.7460E-031.000	-2.88631E-02	9.1915E-03
	4	7.8065E-03	5.7460E-031.000	-1.12208E-02	2.6834E-02
	5	6.6303E-03	5.7460E-031.000	-1.23970E-02	2.5658E-02
	6	7.6807E-04	5.7460E-031.000	-1.82592E-02	1.9795E-02
	7	-6.88810E-03	5.7460E-031.000	-2.59154E-02	1.2139E-02
4	8	-2.24380E-02	5.7460E-03 .006	-4.14653E-02	-3.41069E-03
	9	-1.21686E-02	5.7460E-031.000	-3.11959E-02	6.8587E-03
	10	-1.11810E-03	5.7460E-031.000	-2.01454E-02	1.7909E-02
	1	5.9978E-03	5.7460E-031.000	-1.30295E-02	2.5025E-02
	2	-1.76423E-02	5.7460E-03 .110	-3.66696E-02	1.3850E-03
	3	-7.80650E-03	5.7460E-031.000	-2.68338E-02	1.1221E-02
	5	-1.17617E-03	5.7460E-031.000	-2.02035E-02	1.7851E-02
	6	-7.03843E-03	5.7460E-031.000	-2.60657E-02	1.1989E-02
	7	-1.46946E-02	5.7460E-03 .510	-3.37219E-02	4.3327E-03
	8	-3.02445E-02	5.7460E-03 .000	-4.92718E-02	-1.12172E-02
5	9	-1.99751E-02	5.7460E-03 .028	-3.90024E-02	-9.47835E-04
	10	-8.92461E-03	5.7460E-031.000	-2.79519E-02	1.0103E-02
	1	7.1739E-03	5.7460E-031.000	-1.18534E-02	2.6201E-02
	2	-1.64662E-02	5.7460E-03 .208	-3.54935E-02	2.5611E-03
	3	-6.63033E-03	5.7460E-031.000	-2.56576E-02	1.2397E-02
	4	1.1762E-03	5.7460E-031.000	-1.78511E-02	2.0203E-02
	6	-5.86226E-03	5.7460E-031.000	-2.48896E-02	1.3165E-02
	7	-1.35184E-02	5.7460E-03 .885	-3.25457E-02	5.5089E-03
	8	-2.90683E-02	5.7460E-03 .000	-4.80956E-02	-1.00410E-02
	9	-1.87990E-02	5.7460E-03 .057	-3.78263E-02	2.2834E-04
6	10	-7.74843E-03	5.7460E-031.000	-2.67757E-02	1.1279E-02
	1	1.3036E-02	5.7460E-031.000	-5.99109E-03	3.2064E-02
	2	-1.06039E-02	5.7460E-031.000	-2.96312E-02	8.4234E-03
	3	-7.68069E-04	5.7460E-031.000	-1.97954E-02	1.8259E-02
	4	7.0384E-03	5.7460E-031.000	-1.19889E-02	2.6066E-02
	5	5.8623E-03	5.7460E-031.000	-1.31650E-02	2.4890E-02
	7	-7.65616E-03	5.7460E-031.000	-2.66835E-02	1.1371E-02
	8	-2.32061E-02	5.7460E-03 .004	-4.22334E-02	-4.17875E-03
	9	-1.29367E-02	5.7460E-031.000	-3.19640E-02	6.0906E-03
	10	-1.88617E-03	5.7460E-031.000	-2.09135E-02	1.7141E-02
7	1	2.0692E-02	5.7460E-03 .018	1.6651E-03	3.9720E-02
	2	-2.94773E-03	5.7460E-031.000	-2.19750E-02	1.6080E-02
	3	6.8881E-03	5.7460E-031.000	-1.21392E-02	2.5915E-02
	4	1.4695E-02	5.7460E-03 .510	-4.33270E-03	3.3722E-02
	5	1.3518E-02	5.7460E-03 .885	-5.50888E-03	3.2546E-02
	6	7.6562E-03	5.7460E-031.000	-1.13711E-02	2.6683E-02
	8	-1.55499E-02	5.7460E-03 .334	-3.45772E-02	3.4774E-03
	9	-5.28054E-03	5.7460E-031.000	-2.43078E-02	1.3747E-02
	10	5.7700E-03	5.7460E-031.000	-1.32573E-02	2.4797E-02
8	1	3.6242E-02	5.7460E-03 .000	1.7215E-02	5.5270E-02
	2	1.2602E-02	5.7460E-031.000	-6.42514E-03	3.1629E-02
	3	2.2438E-02	5.7460E-03 .006	3.4107E-03	4.1465E-02
	4	3.0244E-02	5.7460E-03 .000	1.1217E-02	4.9272E-02

HAPAXMIL Scheffe	9	5	2.9068E-02	5.7460E-03	.000	1.0041E-02	4.8096E-02
		6	2.3206E-02	5.7460E-03	.004	4.1788E-03	4.2233E-02
		7	1.5550E-02	5.7460E-03	.334	-3.4774E-03	3.4577E-02
		9	1.0269E-02	5.7460E-03	.000	-8.7579E-03	2.9297E-02
		10	2.1320E-02	5.7460E-03	.012	2.2926E-03	4.0347E-02
		1	2.5973E-02	5.7460E-03	.000	6.9456E-03	4.5000E-02
		2	2.3328E-03	5.7460E-03	.000	-1.6694E-02	2.1360E-02
		3	1.2169E-02	5.7460E-03	.000	-6.8586E-03	3.1196E-02
		4	1.9975E-02	5.7460E-03	.028	9.4783E-04	3.9002E-02
		5	1.8799E-02	5.7460E-03	.057	-2.2833E-04	3.7826E-02
	10	6	1.2937E-02	5.7460E-03	.000	-6.0906E-03	3.1964E-02
		7	5.2805E-03	5.7460E-03	.000	-1.3746E-02	2.4308E-02
		8	-1.0269E-02	5.7460E-03	.000	-2.9296E-02	8.7579E-03
		10	1.1051E-02	5.7460E-03	.000	-7.9767E-03	3.0078E-02
		1	1.4922E-02	5.7460E-03	.456	-4.1049E-03	3.3950E-02
		2	-8.7177E-03	5.7460E-03	.000	-2.7745E-02	1.0310E-02
		3	1.1181E-03	5.7460E-03	.000	-1.7909E-02	2.0145E-02
		4	8.9246E-03	5.7460E-03	.000	-1.0102E-02	2.7952E-02
		5	7.7484E-03	5.7460E-03	.000	-1.1278E-02	2.6776E-02
		6	1.8862E-03	5.7460E-03	.000	-1.7141E-02	2.0913E-02
	1	7	-5.7699E-03	5.7460E-03	.000	-2.4797E-02	1.3257E-02
		8	-2.1319E-02	5.7460E-03	.012	-4.0347E-02	-2.2925E-03
		9	-1.1050E-02	5.7460E-03	.000	-3.0077E-02	7.9768E-03
		2	-22.4683	4.9485	.018	-43.0891	-1.8474
		3	-13.9330	4.9485	.543	-34.5538	6.6879
		4	-6.1865	4.9485	.996	-26.8073	14.4343
		5	-7.8477	4.9485	.980	-28.4685	12.7731
		6	-11.9425	4.9485	.756	-32.5634	8.6783
		7	-24.8445	4.9485	.004	-45.4653	-4.2237
		8	-42.1901	4.9485	.000	-62.8109	-21.5693
	2	9	-18.6536	4.9485	.124	-39.2744	1.9673
		10	-12.8016	4.9485	.669	-33.4224	7.8192
		1	22.4683	4.9485	.018	1.8474	43.0891
		3	8.5353	4.9485	.964	-12.0855	29.1561
		4	16.2817	4.9485	.295	-4.3391	36.9025
		5	14.6206	4.9485	.466	-6.0002	35.2414
		6	10.5257	4.9485	.871	-10.0951	31.1465
		7	-2.3762	4.9485	1.000	-22.9970	18.2446
		8	-19.7219	4.9485	.077	-40.3427	.8990
		9	3.8147	4.9485	1.000	-16.8061	24.4355
	3	10	9.6667	4.9485	.921	-10.9541	30.2875
		1	13.9330	4.9485	.543	-6.6879	34.5538
		2	-8.5353	4.9485	.964	-29.1561	12.0855
		4	7.7464	4.9485	.981	-12.8744	28.3672
		5	6.0853	4.9485	.997	-14.5355	26.7061
		6	1.9904	4.9485	1.000	-18.6304	22.6112
		7	-10.9115	4.9485	.844	-31.5323	9.7093
		8	-28.2572	4.9485	.000	-48.8780	-7.6363
		9	-4.7206	4.9485	1.000	-25.3414	15.9002
		10	1.1314	4.9485	1.000	-19.4894	21.7522
	4	1	6.1865	4.9485	.996	-14.4343	26.8073
		2	-16.2817	4.9485	.295	-36.9025	4.3391
		3	-7.7464	4.9485	.981	-28.3672	12.8744

	5	-1.6611	4.94851.000	-22.2820	18.9597
	6	-5.7560	4.9485 .998	-26.3768	14.8648
	7	-18.6579	4.9485 .124	-39.2788	1.9629
	8	-36.0036	4.9485 .000	-56.6244	-15.3828
	9	-12.4670	4.9485 .704	-33.0878	8.1538
	10	-6.6151	4.9485 .994	-27.2359	14.0058
5	1	7.8477	4.9485 .980	-12.7731	28.4685
	2	-14.6206	4.9485 .466	-35.2414	6.0002
	3	-6.0853	4.9485 .997	-26.7061	14.5355
	4	1.6611	4.94851.000	-18.9597	22.2820
	6	-4.0949	4.94851.000	-24.7157	16.5260
	7	-16.9968	4.9485 .233	-37.6176	3.6240
	8	-34.3424	4.9485 .000	-54.9633	-13.7216
	9	-10.8059	4.9485 .852	-31.4267	9.8149
	10	-4.9539	4.9485 .999	-25.5747	15.6669
6	1	11.9425	4.9485 .756	-8.6783	32.5634
	2	-10.5257	4.9485 .871	-31.1465	10.0951
	3	-1.9904	4.94851.000	-22.6112	18.6304
	4	5.7560	4.9485 .998	-14.8648	26.3768
	5	4.0949	4.94851.000	-16.5260	24.7157
	7	-12.9019	4.9485 .658	-33.5228	7.7189
	8	-30.2476	4.9485 .000	-50.8684	-9.6268
	9	-6.7110	4.9485 .993	-27.3318	13.9098
	10	-8.590	4.94851.000	-21.4799	19.7618
7	1	24.8445	4.9485 .004	4.2237	45.4653
	2	2.3762	4.94851.000	-18.2446	22.9970
	3	10.9115	4.9485 .844	-9.7093	31.5323
	4	18.6579	4.9485 .124	-1.9629	39.2788
	5	16.9968	4.9485 .233	-3.6240	37.6176
	6	12.9019	4.9485 .658	-7.7189	33.5228
	8	-17.3456	4.9485 .206	-37.9665	3.2752
	9	6.1909	4.9485 .996	-14.4299	26.8117
	10	12.0429	4.9485 .746	-8.5779	32.6637
8	1	42.1901	4.9485 .000	21.5693	62.8109
	2	19.7219	4.9485 .077	-8.990	40.3427
	3	28.2572	4.9485 .000	7.6363	48.8780
	4	36.0036	4.9485 .000	15.3828	56.6244
	5	34.3424	4.9485 .000	13.7216	54.9633
	6	30.2476	4.9485 .000	9.6268	50.8684
	7	17.3456	4.9485 .206	-3.2752	37.9665
	9	23.5366	4.9485 .010	2.9157	44.1574
	10	29.3885	4.9485 .000	8.7677	50.0093
9	1	18.6536	4.9485 .124	-1.9673	39.2744
	2	-3.8147	4.94851.000	-24.4355	16.8061
	3	4.7206	4.94851.000	-15.9002	25.3414
	4	12.4670	4.9485 .704	-8.1538	33.0878
	5	10.8059	4.9485 .852	-9.8149	31.4267
	6	6.7110	4.9485 .993	-13.9098	27.3318
	7	-6.1909	4.9485 .996	-26.8117	14.4299
	8	-23.5366	4.9485 .010	-44.1574	-2.9157
	10	5.8520	4.9485 .998	-14.7688	26.4728
10	1	12.8016	4.9485 .669	-7.8192	33.4224
	2	-9.6667	4.9485 .921	-30.2875	10.9541

Bonferroni	1	3	-1.1314	4.94851.000	-21.7522	19.4894
		4	6.6151	4.9485 .994	-14.0058	27.2359
		5	4.9539	4.9485 .999	-15.6669	25.5747
		6	.8590	4.94851.000	-19.7618	21.4799
		7	-12.0429	4.9485 .746	-32.6637	8.5779
		8	-29.3885	4.9485 .000	-50.0093	-8.7677
		9	-5.8520	4.9485 .998	-26.4728	14.7688
		2	-22.4683	4.9485 .000	-38.8545	-6.0820
		3	-13.9330	4.9485 .242	-30.3192	2.4533
		4	-6.1865	4.94851.000	-22.5728	10.1997
	2	5	-7.8477	4.94851.000	-24.2339	8.5386
		6	-11.9425	4.9485 .754	-28.3288	4.4437
		7	-24.8445	4.9485 .000	-41.2307	-8.4582
		8	-42.1901	4.9485 .000	-58.5764	-25.8039
		9	-18.6536	4.9485 .010	-35.0398	-2.2673
		10	-12.8016	4.9485 .469	-29.1878	3.5847
		1	22.4683	4.9485 .000	6.0820	38.8545
		3	8.5353	4.94851.000	-7.8510	24.9215
		4	16.2817	4.9485 .054	-.1045	32.6680
		5	14.6206	4.9485 .159	-1.7657	31.0068
	3	6	10.5257	4.94851.000	-5.8605	26.9120
		7	-2.3762	4.94851.000	-18.7625	14.0100
		8	-19.7219	4.9485 .004	-36.1081	-3.3356
		9	3.8147	4.94851.000	-12.5716	20.2009
		10	9.6667	4.94851.000	-6.7196	26.0529
		1	13.9330	4.9485 .242	-2.4533	30.3192
		2	-8.5353	4.94851.000	-24.9215	7.8510
		4	7.7464	4.94851.000	-8.6398	24.1327
		5	6.0853	4.94851.000	-10.3010	22.4715
		6	1.9904	4.94851.000	-14.3958	18.3767
	4	7	-10.9115	4.94851.000	-27.2978	5.4747
		8	-28.2572	4.9485 .000	-44.6434	-11.8709
		9	-4.7206	4.94851.000	-21.1069	11.6656
		10	1.1314	4.94851.000	-15.2549	17.5176
		1	6.1865	4.94851.000	-10.1997	22.5728
		2	-16.2817	4.9485 .054	-32.6680	.1045
		3	-7.7464	4.94851.000	-24.1327	8.6398
		5	-1.6611	4.94851.000	-18.0474	14.7251
		6	-5.7560	4.94851.000	-22.1423	10.6302
		7	-18.6579	4.9485 .010	-35.0442	-2.2717
	5	8	-36.0036	4.9485 .000	-52.3898	-19.6173
		9	-12.4670	4.9485 .566	-28.8533	3.9192
		10	-6.6151	4.94851.000	-23.0013	9.7712
		1	7.8477	4.94851.000	-8.5386	24.2339
		2	-14.6206	4.9485 .159	-31.0068	1.7657
		3	-6.0853	4.94851.000	-22.4715	10.3010
		4	1.6611	4.94851.000	-14.7251	18.0474
		6	-4.0949	4.94851.000	-20.4811	12.2914
		7	-16.9968	4.9485 .033	-33.3830	-.6105
		8	-34.3424	4.9485 .000	-50.7287	-17.9562

		9	-10.8059	4.94851.000	-27.1921	5.5804
		10	-4.9539	4.94851.000	-21.3402	11.4323
6		1	11.9425	4.9485 .754	-4.4437	28.3288
		2	-10.5257	4.94851.000	-26.9120	5.8605
		3	-1.9904	4.94851.000	-18.3767	14.3958
		4	5.7560	4.94851.000	-10.6302	22.1423
		5	4.0949	4.94851.000	-12.2914	20.4811
		7	-12.9019	4.9485 .443	-29.2882	3.4843
		8	-30.2476	4.9485 .000	-46.6338	-13.8613
		9	-6.7110	4.94851.000	-23.0973	9.6752
		10	-.8590	4.94851.000	-17.2453	15.5272
7		1	24.8445	4.9485 .000	8.4582	41.2307
		2	2.3762	4.94851.000	-14.0100	18.7625
		3	10.9115	4.94851.000	-5.4747	27.2978
		4	18.6579	4.9485 .010	2.2717	35.0442
		5	16.9968	4.9485 .033	.6105	33.3830
		6	12.9019	4.9485 .443	-3.4843	29.2882
		8	-17.3456	4.9485 .026	-33.7319	-.9594
		9	6.1909	4.94851.000	-10.1953	22.5772
		10	12.0429	4.9485 .714	-4.3434	28.4291
8		1	42.1901	4.9485 .000	25.8039	58.5764
		2	19.7219	4.9485 .004	3.3356	36.1081
		3	28.2572	4.9485 .000	11.8709	44.6434
		4	36.0036	4.9485 .000	19.6173	52.3898
		5	34.3424	4.9485 .000	17.9562	50.7287
		6	30.2476	4.9485 .000	13.8613	46.6338
		7	17.3456	4.9485 .026	.9594	33.7319
		9	23.5366	4.9485 .000	7.1503	39.9228
		10	29.3885	4.9485 .000	13.0023	45.7748
9		1	18.6536	4.9485 .010	2.2673	35.0398
		2	-3.8147	4.94851.000	-20.2009	12.5716
		3	4.7206	4.94851.000	-11.6656	21.1069
		4	12.4670	4.9485 .566	-3.9192	28.8533
		5	10.8059	4.94851.000	-5.5804	27.1921
		6	6.7110	4.94851.000	-9.6752	23.0973
		7	-6.1909	4.94851.000	-22.5772	10.1953
		8	-23.5366	4.9485 .000	-39.9228	-7.1503
		10	5.8520	4.94851.000	-10.5343	22.2382
10		1	12.8016	4.9485 .469	-3.5847	29.1878
		2	-9.6667	4.94851.000	-26.0529	6.7196
		3	-1.1314	4.94851.000	-17.5176	15.2549
		4	6.6151	4.94851.000	-9.7712	23.0013
		5	4.9539	4.94851.000	-11.4323	21.3402
		6	.8590	4.94851.000	-15.5272	17.2453
		7	-12.0429	4.9485 .714	-28.4291	4.3434
		8	-29.3885	4.9485 .000	-45.7748	-13.0023
		9	-5.8520	4.94851.000	-22.2382	10.5343
WM Scheffe	1	2	.27479	.22693 .997	-.67087	1.22045
		3	-.13761	.226931.000	-1.08327	.80805
		4	-.66275	.22693 .485	-1.60841	.28291
		5	.44184	.22693 .923	-.50382	1.38750
		6	-1.36174	.22693 .000	-2.30740	-.41608
		7	.76489	.22693 .260	-.18077	1.71055

2	8	-1.78786	.22693 .000	-2.73352	-.84220
	9	-1.48237	.22693 .000	-2.42803	-.53671
	10	.39141	.22693 .964	-.55425	1.33707
	1	-.27479	.22693 .997	-1.22045	.67087
	3	-.41240	.22693 .950	-1.35806	.53326
	4	-.93753	.22693 .055	-1.88320	8.1262E-03
	5	.16705	.226931.000	-.77861	1.11271
	6	-1.63652	.22693 .000	-2.58219	-.69086
	7	.49010	.22693 .860	-.45556	1.43577
	8	-2.06265	.22693 .000	-3.00831	-1.11699
3	9	-1.75716	.22693 .000	-2.70282	-.81150
	10	.11662	.226931.000	-.82904	1.06228
	1	.13761	.226931.000	-.80805	1.08327
	2	.41240	.22693 .950	-.53326	1.35806
	4	-.52513	.22693 .800	-1.47080	.42053
	5	.57945	.22693 .686	-.36621	1.52511
	6	-1.22412	.22693 .001	-2.16979	-.27846
	7	.90250	.22693 .079	-4.31570E-02	1.84817
	8	-1.65025	.22693 .000	-2.59591	-.70459
	9	-1.34476	.22693 .000	-2.29042	-.39910
4	10	.52902	.22693 .793	-.41664	1.47468
	1	.66275	.22693 .485	-.28291	1.60841
	2	.93753	.22693 .055	-8.12618E-03	1.88320
	3	.52513	.22693 .800	-.42053	1.47080
	5	1.10459	.22693 .007	.15893	2.05025
	6	-.69899	.22693 .399	-1.64465	.24667
	7	1.42764	.22693 .000	.48198	2.37330
	8	-1.12512	.22693 .005	-2.07078	-.17945
	9	-.81963	.22693 .170	-1.76529	.12603
	10	1.05415	.22693 .014	.10849	1.99982
5	1	-.44184	.22693 .923	-1.38750	.50382
	2	-.16705	.226931.000	-1.11271	.77861
	3	-.57945	.22693 .686	-1.52511	.36621
	4	-1.10459	.22693 .007	-2.05025	-.15893
	6	-1.80358	.22693 .000	-2.74924	-.85792
	7	.32305	.22693 .991	-.62261	1.26871
	8	-2.22970	.22693 .000	-3.17536	-1.28404
	9	-1.92421	.22693 .000	-2.86987	-.97855
	10	-5.04316E-02	.226931.000	-.99609	.89523
6	1	1.36174	.22693 .000	.41608	2.30740
	2	1.63652	.22693 .000	.69086	2.58219
	3	1.22412	.22693 .001	.27846	2.16979
	4	.69899	.22693 .399	-.24667	1.64465
	5	1.80358	.22693 .000	.85792	2.74924
	7	2.12663	.22693 .000	1.18097	3.07229
	8	-.42613	.22693 .938	-1.37179	.51954
	9	-.12064	.226931.000	-1.06630	.82502
	10	1.75314	.22693 .000	.80748	2.69881
7	1	-.76489	.22693 .260	-1.71055	.18077
	2	-.49010	.22693 .860	-1.43577	.45556
	3	-.90250	.22693 .079	-1.84817	4.3157E-02
	4	-1.42764	.22693 .000	-2.37330	-.48198
	5	-.32305	.22693 .991	-1.26871	.62261

Bonferroni	8	6	-2.12663	.22693	.000	-3.07229	-1.18097
		8	-2.55275	.22693	.000	-3.49842	-1.60709
		9	-2.24727	.22693	.000	-3.19293	-1.30160
		10	-.37348	.22693	.974	-1.31915	.57218
		1	1.78786	.22693	.000	.84220	2.73352
		2	2.06265	.22693	.000	1.11699	3.00831
		3	1.65025	.22693	.000	.70459	2.59591
		4	1.12512	.22693	.005	.17945	2.07078
		5	2.22970	.22693	.000	1.28404	3.17536
		6	.42613	.22693	.938	-.51954	1.37179
	9	7	2.55275	.22693	.000	1.60709	3.49842
		9	.30549	.22693	.994	-.64017	1.25115
		10	2.17927	.22693	.000	1.23361	3.12493
		1	1.48237	.22693	.000	.53671	2.42803
		2	1.75716	.22693	.000	.81150	2.70282
		3	1.34476	.22693	.000	.39910	2.29042
		4	.81963	.22693	.170	-.12603	1.76529
		5	1.92421	.22693	.000	.97855	2.86987
		6	.12064	.22693	1.000	-.82502	1.06630
		7	2.24727	.22693	.000	1.30160	3.19293
	10	8	-.30549	.22693	.994	-1.25115	.64017
		10	1.87378	.22693	.000	.92812	2.81944
		1	-.39141	.22693	.964	-1.33707	.55425
		2	-.11662	.22693	1.000	-1.06228	.82904
		3	-.52902	.22693	.793	-1.47468	.41664
		4	-1.05415	.22693	.014	-1.99982	-.10849
		5	5.0432E-02	.22693	1.000	-.89523	.99609
		6	-1.75314	.22693	.000	-2.69881	-.80748
		7	.37348	.22693	.974	-.57218	1.31915
		8	-2.17927	.22693	.000	-3.12493	-1.23361
	1	9	-1.87378	.22693	.000	-2.81944	-.92812
		2	.27479	.22693	1.000	-.47668	1.02625
		3	-.13761	.22693	1.000	-.88908	.61385
		4	-.66275	.22693	.176	-1.41421	8.8719E-02
		5	.44184	.22693	1.000	-.30963	1.19331
		6	-1.36174	.22693	.000	-2.11320	-.61027
		7	.76489	.22693	.041	1.3426E-02	1.51636
		8	-1.78786	.22693	.000	-2.53933	-1.03640
		9	-1.48237	.22693	.000	-2.23384	-.73091
		10	.39141	.22693	1.000	-.36006	1.14287
	2	1	-.27479	.22693	1.000	-1.02625	.47668
		3	-.41240	.22693	1.000	-1.16387	.33907
		4	-.93753	.22693	.002	-1.68900	-.18607
		5	.16705	.22693	1.000	-.58441	.91852
		6	-1.63652	.22693	.000	-2.38799	-.88506
		7	.49010	.22693	1.000	-.26136	1.24157
		8	-2.06265	.22693	.000	-2.81412	-1.31118
		9	-1.75716	.22693	.000	-2.50863	-1.00570
		10	.11662	.22693	1.000	-.63485	.86809
	3	1	.13761	.22693	1.000	-.61385	.88908

4	2	.41240	.226931.000	-.33907	1.16387
	4	-.52513	.22693 .978	-1.27660	.22633
	5	.57945	.22693 .515	-.17201	1.33092
	6	-1.22412	.22693 .000	-1.97559	-.47266
	7	.90250	.22693 .004	.15104	1.65397
	8	-1.65025	.22693 .000	-2.40172	-.89878
	9	-1.34476	.22693 .000	-2.09623	-.59330
	10	.52902	.22693 .936	-.22245	1.28049
	1	.66275	.22693 .176	-8.87189E-02	1.41421
	2	.93753	.22693 .002	.18607	1.68900
5	3	.52513	.22693 .978	-.22633	1.27660
	5	1.10459	.22693 .000	.35312	1.85605
	6	-.69899	.22693 .107	-1.45046	5.2476E-02
	7	1.42764	.22693 .000	.67617	2.17910
	8	-1.12512	.22693 .000	-1.87658	-.37365
	9	-.81963	.22693 .018	-1.57109	-6.81604E-02
	10	1.05415	.22693 .000	.30269	1.80562
	1	-.44184	.226931.000	-1.19331	.30963
	2	-.16705	.226931.000	-.91852	.58441
	3	-.57945	.22693 .515	-1.33092	.17201
6	4	-1.10459	.22693 .000	-1.85605	-.35312
	6	-1.80358	.22693 .000	-2.55504	-1.05211
	7	.32305	.226931.000	-.42841	1.07452
	8	-2.22970	.22693 .000	-2.98117	-1.47824
	9	-1.92421	.22693 .000	-2.67568	-1.17275
	10	-5.04316E-02	.226931.000	-.80190	.70103
	1	1.36174	.22693 .000	.61027	2.11320
	2	1.63652	.22693 .000	.88506	2.38799
	3	1.22412	.22693 .000	.47266	1.97559
	4	.69899	.22693 .107	-5.24762E-02	1.45046
7	5	1.80358	.22693 .000	1.05211	2.55504
	7	2.12663	.22693 .000	1.37516	2.87809
	8	-.42613	.226931.000	-1.17759	.32534
	9	-.12064	.226931.000	-.87210	.63083
	10	1.75314	.22693 .000	1.00168	2.50461
	1	-.76489	.22693 .041	-1.51636	-1.34262E-02
	2	-.49010	.226931.000	-1.24157	.26136
	3	-.90250	.22693 .004	-1.65397	-.15104
	4	-1.42764	.22693 .000	-2.17910	-.67617
	5	-.32305	.226931.000	-1.07452	.42841
8	6	-2.12663	.22693 .000	-2.87809	-1.37516
	8	-2.55275	.22693 .000	-3.30422	-1.80129
	9	-2.24727	.22693 .000	-2.99873	-1.49580
	10	-.37348	.226931.000	-1.12495	.37798
	1	1.78786	.22693 .000	1.03640	2.53933
	2	2.06265	.22693 .000	1.31118	2.81412
	3	1.65025	.22693 .000	.89878	2.40172
	4	1.12512	.22693 .000	.37365	1.87658
	5	2.22970	.22693 .000	1.47824	2.98117
	6	.42613	.226931.000	-.32534	1.17759
	7	2.55275	.22693 .000	1.80129	3.30422
	9	.30549	.226931.000	-.44598	1.05695
	10	2.17927	.22693 .000	1.42780	2.93074

9	1	1.48237	.22693 .000	.73091	2.23384
	2	1.75716	.22693 .000	1.00570	2.50863
	3	1.34476	.22693 .000	.59330	2.09623
	4	.81963	.22693 .018	6.8160E-02	1.57109
	5	1.92421	.22693 .000	1.17275	2.67568
	6	.12064	.226931.000	-.63083	.87210
	7	2.24727	.22693 .000	1.49580	2.99873
	8	-.30549	.226931.000	-1.05695	.44598
	10	1.87378	.22693 .000	1.12232	2.62525
	10	1	-.39141	.226931.000	-1.14287
	2	-.11662	.226931.000	-.86809	.63485
	3	-.52902	.22693 .936	-1.28049	.22245
	4	-1.05415	.22693 .000	-1.80562	-.30269
	5	5.0432E-02	.226931.000	-.70103	.80190
	6	-1.75314	.22693 .000	-2.50461	-1.00168
	7	.37348	.226931.000	-.37798	1.12495
	8	-2.17927	.22693 .000	-2.93074	-1.42780
	9	-1.87378	.22693 .000	-2.62525	-1.12232

* The mean difference is significant at the .05 level.

Appendix I:

Leave-taking item 'bye'
bye

acmel	70
acper	68
dacan	113
dache	201
escyb	294
eshel	22
itaus	65
itphi	66
utbeg	174
utcha	259
total	1332

Appendix J:

A list of some emoticons encountered in the CIRC

:)	;}	:p~~~~~
:p	;/	:p~~~~~
]:	;-/	:-p~~~~~
]::	;\	:p~~~~~
]:::)	:]	:p~~~~~
]o_	;(	:p~~~~~
:p	;-(-	~~~~~
;-p	;(p	:pp
:p~	;))	:ppp
:o	;-)	:pppp
;-o	;))	:ppppp
:o)	;-))	:-ppppp
:o0	:)))	:pppppp
:od	;-)))	:ppppppp~~~~~
:op	:))))	:pppppppppppppp
:op	:)))))	:r
:p	:)))))	:s
;-p	:)))))	:p
:p;	:)))))	:-p
:p~	:)))))	:p~
:p~	:)))))	
:p~~~~~	:))>	
:p~~~~~	:x	
:p~~~~~`	:-x	
:p~~~~~	:op	
;-pp	:op~	
;-ppp	:op~`	
:pppp	:op~~~~~	
;-ppppp	:op~~~~~	
;-pppppp	:op~~~~~`	
;-ppppppp	:p	
;-pppppppppp	:-p	
;+)	:p)	
;=)	:p,	
:0	:p~	
:0)	:p~	
:0=	:p~	
;-{ }	:p~~~~~	