Computer-mediated Communication: Experiments with E-mail Readability

Philip Sallis and Diana Kassabova
Department of Information Science, University of Otago
PO Box 56, Dunedin, New Zealand

The emergence of Computer-mediated Communication

In recent years the global web of computer networks has expanded at an exponential rate, linking education institutions, businesses and individuals. It has become a medium for unprecedented human interaction that takes different forms: one-to-one E-mail messages, computer conferencing, Internet, Intranet, Usenet (also known as newsgroups), Electronic distribution lists and voice mail. Although these forms of communication serve different purposes and therefore, have different characteristics, they all can be described as part of Computer-mediated communication (CMC). In [12] it is argued that *"CMC in its broadest sense covers any kind of human communication involving the transmission of electronic signals between computers"*.

In [2] the author argues that *"Computer-mediated communication (CMC) is a relatively new area of study, but as computers have become an integral part of society, spanning education, industry and government, the field is growing significantly. The lowered costs of and easier access to computer technologies has increased the number of users. This in turn is accompanied by a rapid growth of scholarly study of CMC. Because CMC scholarship spans many fields, and because of its rapid and continuing development, there is a variety of CMC terminology. [...] In general, the term computer-mediated communication refers to both task-related and interpersonal communication conducted by computer. This includes communication both to and through a personal or a mainframe computer, and is generally understood to include asynchronous communication via E-mail or through use of an electronic bulletin board; synchronous communication such as 'chatting' or through the use of group software; and information manipulation, retrieval and storage through*

computers and electronic databases." This lengthy quotation provides a comprehensive definitional insight to the emergence of the CMC process.

Characteristics of electronic mail

In everyday life the notion of Computer-mediated Communication is usually linked to its most common form, namely electronic mail (E-mail). Millions of people around the world use electronic mail for business and personal communication. As a matter of interest, according to some surveys, New Zealand is the second biggest user of E-mail per capita after the US which is fascinating for a country of only 3.5 million people. The reason why electronic mail is so appealing as a means of communication is obvious to all who have used it. The electronic mail is first of all a great means for reaching across distances other organisations and individuals and is much faster and cheaper than the alternative forms of communication such as postal mail or telephone conversations. It is also non-intrusive when compared with face-to-face or telephone conversations. Individuals can choose when to read the received messages and when to answer them and also have the opportunity to think before answering. They can send a single message to many recipients and can easily share data with other people. By simply sending E-mail messages, users can participate in a wide variety of forums on the global computer network, using for instance the so-called newsgroups on the Usenet.

While all the advantages of this comparatively new means of communication have become popular with a wide community of devoted users, the electronic mail can also have some downside effects. The fact itself that it is readily available to so many users means that it could be misused. It can cause information overload as users may receive information that is not relevant to them, particularly when participating in group discussions that waiver from one focus to another. Many users tend to include lengthy signatures in their mail that do not provide valuable information. For instance, some signatures contain lyrics from favourite songs or funny drawings. There is also a tendency to neglect grammar, spelling, and 'good' vocabulary, to write text with incomplete

sentences or to use shorthand script. All these can lead to misunderstanding and ambiguity, or at least to difficulties in comprehending the meaning of a particular message. The convenience of using electronic mail may also lead to decreasing personal contacts and to de-personalising of communication. As electronic mail is increasingly used instead of telephone or face-to-face conversations, misunderstanding can occur due to lack of intonation, facial expression and completeness due to dialogue, although there have been some remedies for this problem, using the so-called smilies or emoticons /ee-moh'ti-kon/. These are a combination of ASCII characters used in E-mail correspondence to indicate an emotional state, for instance :-) represents humour, laughter, friendliness, occasionally sarcasm, :-(is used to express sadness, anger, or upset.

Another negative side of the E-mail phenomenon is the threat to the privacy of individuals engaged in E-mail correspondence. As it is pointed out in [16] *"...one data security expert has noted that E-mail has 'the same security level as a postcard'. Thus, users of E-mail may be exposed to breaches of confidentiality of their communications. In addition, E-mail creates an electronic trail of messages that can be used to monitor individuals. Complex legal and ethical questions have emerged about the right to privacy of E-mail users, particularly in the workplace."*

Brief overview of electronic mail research issues

The fact that an enormous number of people use and will be using electronic mail as their main means for communication makes all these issues well worth investigating as a sociological phenomenon, let alone for the technical insight provided. An increasing number of studies concerned with communication over the global computer network have been carried out by many researchers in different areas: linguists, sociologists, psychologists, specialists in communication, computer and information science. It could be argued that the speed of development of CMC and the sheer volume of it in a way have 'swamped' the efforts of researchers. This is an area that is yet to be explored. Rudy [12] argues that *"despite a great deal of published work though, the field [of E-mail research] still has an unsatisfactory, piecemeal feel to it."* New theories, methodologies and

techniques are yet to be developed and applied to CMC by the wider scientific community in order to achieve a greater understanding of its nature and to be able to utilise it more efficiently.

Although there is plenty of material for research as millions of E-mail messages are sent around the world every day, the issue of privacy prevents researchers from getting access to this wealth of research material. This is one of the reasons for many researchers either to simulate E-mail correspondence [15] or to use data from publicly available forums, such as the Usenet newsgroups for their investigation [9] and [1]. Wilkins reports in [15] that although her study was based on a publicly accessible conferencing network, one of the participants in her study still objected to the study as an invasion of privacy. A discussion between the participants followed and they came to agreement with a position attributed to Usenet that *"anything posted to a publicly readable topic becomes public domain [...]"*

Another difficulty for the researchers in the area of CMC is the large volume of work involved in just pre-processing the available data, let alone studying it. One of the larger studies [1] involved the effort of 100 people who worked on more than 4000 messages [10]. The research described in these papers involved using a variety of methods, including statistical and connectionist approaches. Another new and very interesting research on message classification and retrieval, based on contextual similarities between the individual texts, is described in [9]. The researchers utilised a particular kind of neural networks, self-organising maps, for producing a document map. A demonstration of this work is available on the Web at the address: http://websom.hut.fi/websom/.

Electronic mail readability

A research project on analysis of E-mail traffic has been carried out recently by a small team in the Information Science Department at the University of Otago. The aims of the project are:

- to create a text corpus consisting of E-mail messages;
- to obtain stylometric statistics for the messages;

- to utilise statistics for profiling the authorship characteristics of individual message originators and newsgroups.

The source set of data used for this research contains a set of messages extracted from a large number of Newsgroups on the Usenet. It was first complied in the U.K. for a text retrieval and indexing research project funded by The British Library [11]. This was part of an international project TREC, originated by The National Institute of Standards and Technology (NIST), in the USA [5]. The aims of that work are concerned with information retrieval issues.

The data used for the experiments described in this paper were extracted from a database containing 46621 messages that were posted by 21006 senders to 2240 newsgroups on the Usenet. Given its comparatively large size, and the large number of newsgroups from which the messages have been extracted, this data set could be considered an indicative sample of the population of messages posted to newsgroups on the Usenet.

To begin with, all texts were computationally processed to remove unnecessary lines (for instance, lengthy signatures or lines predominantly containing numbers). Then the texts were placed in a relational database along with other relevant information for each message: sender ID, newsgoroup(s), and subject line, thus producing a text corpus. The pre-processed texts in the database contain 5,681,386 words altogether.

For the purpose of conducting analysis of the texts, a set of stylometric characteristics was obtained for each text. These include: number of running words, number of common words, number of unique word forms, readability scores, passive voice usage, number and length of sentences and paragraphs. The combination of these characterises the individual texts in a unique way. Together they can be used for compiling text profiles.

It is interesting to note that many messages contain some number of words that are not part of completed sentences. This is due to the fact that senders often do not care much about punctuation,

capitalisation, or other formal text attributes. There are also some cases when the automatic pre-processing of messages has removed lines that contain not only a large proportion of numbers, but also some words which may be part of sentences. This peculiarity of the texts under investigation has to be taken in account when obtaining stylometric statistics. For instance, readability scores are only obtained for those parts of texts that consist of completed sentences. This is the only sensible approach, as one cannot (and should not) measure readability for a string of words that do not form a sentence.

It is reasonable to assume that the number of words in messages is a quantitative measure for texts. On the other hand, the readability of a text could be considered a measure for its quality in the sense that the more readable a text is, the more likely it is to be comprehended by its readers.

For the purpose of investigating the behaviour of Usenet newsgroups in terms of quantity and quality was first undertaken using a traditional statistical approach to build one-dimensional profiles of selected groups of messages. The groups of messages whose attributes were considered for profile building are shown in Table 1. The first three groups were selected from the database because there were similar number of messages in these groups. Each of the message originators in the group of 'Frequent senders' had more than 20 messages.

Table 1. Names and sizes of groups of messages

GROUP	Number of messages
Group 1 (alt.politics.equality)	237
Group 2 (comp.os.ms-windows.apps.utilities)	239
Group 3 (alt.journalism.criticis)	212
Frequent senders	5951
All	46621

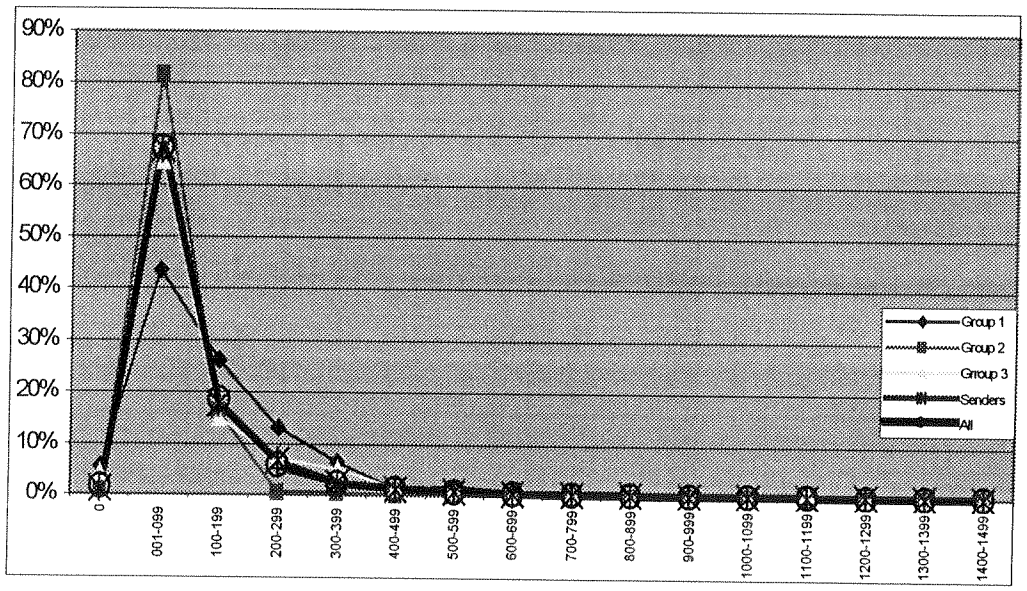

Figure 1. Graphical representation of relative frequency of messages per a hundred-word interval for the individual groups.

By finding the relative frequency of messages per a hundred-word interval, one-dimensional profiles were created for each group (Figure 1). The figure shows that the peak value for number of words in the majority of messages in all investigated groups is around 100. It also reveals that the profiles of groups 'Frequent senders' and 'All' are identical.

Next the relative frequency of messages per readability score interval was found for the same five groups. These intervals were defined according to the Flesch Reading Ease method, one of the best-known readability measures [6]. The formula used by this method produces a difficulty index which relates to comprehension score on a scale of 0 - 100.

Reading ease score=206.835 - (0.846 x SYLLS/100W) - (1.015 x WDS/SEN)

where SYLLS/100W = syllables per 100 words and WDS/SEN = average number of words per sentence.

Note that the higher the score, the more readable is the text.

For the purposes of this research the readability scores are interpreted as it is shown in Table 2. The frequency of messages per readability score interval is depicted in Figure 2. As can be seen in

Figure 2, the graph for 'Frequent senders' has a very close similarity to that for 'All'. On the other hand, from the rest of the groups only the graph for group 3 is vaguely similar to 'All' and 'Senders'. Group 1 demonstrates surprisingly low readability for a comparatively high percentage of messages. Although most messages in group 2 belong to 'standard' and 'fairly easy' intervals, there is a sudden increase of number of messages in the area of 'very difficult'.

Table 2. Readability score interpretation

Score	Reading Difficulty
90 -100	Very easy
80 - 90	Easy
70 - 80	Fairly easy
60 - 70	Standard
50 - 60	Fairly difficult
30 - 50	Difficult
0 - 30	Very Difficult

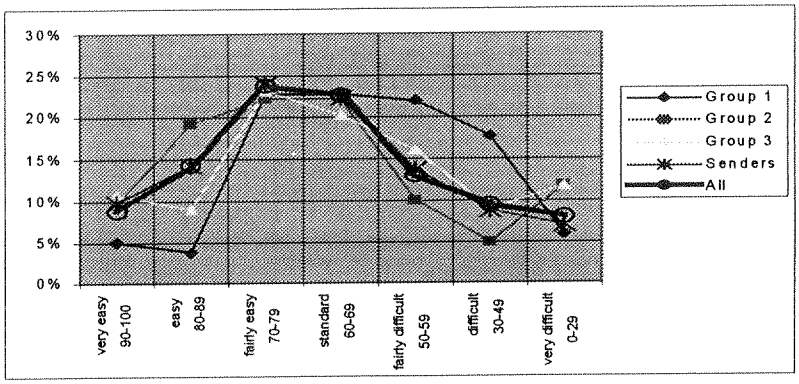

Figure 2. Message frequency per readability score interval for individual groups

Although the graphs in Figure 1 and Figure 2 indicate clearly the similarities and differences between groups of messages under investigation, they can only do this in relation to one attribute at a time – either number of words or readability.

As it is apparent from Figure 1 and Figure 2, the graphs for groups 'Frequent senders' and 'All' are practically identical. It could be argued that the group of 'Frequent senders' pre-determines (and therefore represents) the characteristics of the whole sample both in terms of readability and number of words. Next a two-dimensional profile for each of groups 1, 2 and 3 is built and compared to the profile of group 'Senders' assuming that it closely represents the whole sample (group 'All').

There is a number of newsgroups on the Usenet where the participants discuss educational issues. In the database containing E-mail text corpus there are messages from 14 newsgroups related to the educational issues whose names are shown in column 1 in Table 3. Their topics of discussion and some statistics (where available) are shown in columns 2-6 in Table 3 [14].

Table 3. Statistics for related to education newsgroups in the E-mail database

1	2	3	4	5	6	7	8	9
Newsgroup name	**Topic**	**Num of readers**	**Num of messages received per day**	**Percentage of Internet sites who receive this group**	**Crossposting**	**Num of messages in the database**	**Avg readability**	**Avg Number of words**
alt.education.alternative	School doesn't have to suck	-	-	-	-	38	62.15	527.87
alt.education.disabled	Education for people with physical/mental disabilities	9100	12	53%	6%	30	54.67	141.8
alt.education.distance	Learning from teachers who are far away	9000	15	49%	9%	2	21	680.5
alt.education.home-school.christian	Christian home-schoolers (Moderated)	-	-	-	-	11	64.82	197.54
alt.education.research	Studying about studying	7200	22	40%	62%	19	58.73	147.42
alt.education.university.vision2020	-	-	-	-	-	3	41.5	79.25
comp.ai.edu	Applications of Artificial Intelligence to Education	33000	2	73%	39%	8	29.25	528.125
comp.edu	Computer science education	35000	21	77%	10%	209	57.45	169.89
comp.edu.languages.natural	Computer assisted languages instruction issues	6600	1	57%	18%	8	45.37	207.25
misc.education	Discussion of the educational system	87000	47	72%	43%	2	42	570
misc.education.language.english	Teaching English to speakers of other languages	11000	14	61%	8%	1	62	125
misc.education.medical	Issues related to medical education	7700	33	53%	6%	2	46	322
sci.edu	The science of education	24000	7	75%	40%	1	48	234
uk.education.teachers	For discussion by/about teachers	-	-	-	-	3	44.33	53.33

Columns 7-9 in Table 3 depict the average values of readability and text length for the messages from these groups as they appear in the E-mail database. As can be seen, there are very few messages available in some of the newsgroups in the education area, for example, *misc.education.language.english, sci.edu, misc.education, misc.eduication.medical*, which do not provide sufficient data for statistical analysis. From the data for the three groups with the highest

number of available messages, *alt.education.alternative*, *alt.education.disabled* and *comp.edu*, one can see that the average readability for these groups is in the range of 'standard' - to - 'fairly difficult'. This is an indication that the participants in these groups tend to write in a less readable style than the average contributor to the Usenet whose readability is mostly in the range of 'fairly easy' - to - 'standard' (see the graph for 'All messages' in Figure 2). Very intriguing is the result for the average number of words in messages. In 12 out of 14 newsgroups related to educational issues, the average number of words is higher than 100, in most cases well above the number of words used by most contributors to the Usenet (Figure 2). One might wonder if it is at all possible for readers of these groups to properly follow and take part in such verbose discussions particularly if the readability is not particularly good. This highlights the issue of subject line relevance to the message content and how important subject lines are. Obviously, if the subject line is relevant to the content of the message, it gives the readers a powerful tool to select messages of interest. So often this correlation is not present; sometimes due to the 'multiple reply syndrome' of E-mail users, some times due to the participation in the so-called 'threads', where an initial subject is unaltered during a multiple exchange of messages with content variation; in some other cases it is due to crossposting when a message originator sends the same message to more than one newsgroup.

Alternative methods of analysis

The results described above were obtained by utilising standard statistical methods and they only deal with two of the characteristics of the messages, readability and number of words. For investigating a more complex issue such as the individual writing profile of message originators, a larger set of characteristics has to be used. In [7], a study of senders profiles is described that involved traditional statistical methods along with some connectionist methods, such as fuzzy clustering and self-organising maps. By using self-organising maps, five-dimensional senders profiles were compiled. It could be argued that similar methods could be used for authorship

profiling and comparison for a variety of texts and further used for document retrieval. The documents could be not only E-mail messages but also scientific, literary or religious texts, and epistolary, on both the global computer network and other more traditional information sources. Some work is also being carried out using these methods for computer software authorship determination [3], [4], [8] and [13].

Conclusion

At present, experiments are being conducted using artificial neural networks to help discriminate between user profiles based on CMC authorship characteristics. It is believed that a mix of both traditional statistical methods and some connectionist techniques will provide a better tool-set for result acquisition. The data used in these experiments provides an insight into the issue of E-mail readability. It indicates an informality in E-mail messages that can lead to considerable ambiguity of information transfer, let alone poor grammar, vocabulary and written expression. The results given here reflect this reality and provide a commentary on the dynamics of CMC as an information transfer process.

Finally, although observing the nature of E-mail message formulation is of interest to both originators and recipients, this authorship data also provides a basis for experiments with traditional and non-traditional information processing methods.

Reference:

[1] Berthold, M.R., F. Sudweeks, S. Newton, R. Coyne (1997) It makes sense: Using an autoassociative neural network to explore typicality in computer mediated discussions, in *Network and Netplay: Virtual Groups on the Internet*, (eds. F. Sudweeks et al), AAAI/MIT Press Menlo Park, Ca.

[2] Ferris S.P., What is CMC? An Overview of Scholarly Definitions in *Computer-Mediated Communication Magazine* ISSN 1076-027X / Volume 4, Number 1 / January 1, 1997, http://www.december.com/cmc/mag/1997/jan/

[3] Gray, A.R., P.J. Sallis, and S.G. MacDonell (1997) Software forensics: extending authorship analysis to computer programs. In *Proceedings of Third Biannual Conference of the International Association of Forensic Linguists*, Durham NC.

[4] Gray, A.R., P.J. Sallis, and S.G. MacDonell, IDENTIFIED (Integrated Dictionary-based Extraction of Non-language-dependent Token Information for Forensic Identification,

Examination, and Discrimination): a dictionary-based system for extracting source code metrics for software forensics. Submitted to SE:E&P '98 (Software Education Conference), Dunedin, New Zealand. Forthcoming.

[5] Harman D. (1995) Overview of the Third Text REtrieval Conference (TREC-3) in *The Third Text REtrieval Conference*, April 1995 (ed. D.K. Harman), NIST, Gaithersburg, MD.

[6] Harrison C. (1980) Readability in the Class Room, Cambridge University Press, Cambridge.

[7] Kassabova, D. and P.J. Sallis, (1997) Connectionist Methods for Stylometric Analysis: A Hybrid Approach, in *Neuro-Fuzzy Tools and Techniques for Information Processing*, Springer Verlag , Singapore, in press.

[8] Kilgour, R.I., A.R Gray,., P.J. Sallis, and S.G. MacDonell, A fuzzy logic approach to computer software source code authorship analysis. Accepted by the ICONIP/ANZIIS/ANNES'97 Conference, Dunedin New Zealand. Forthcoming.

[9] Kohonen T. (1997) Exploration of Very Large Databases by Self Organising Maps, in *Proceedings of the 1997 International Conference on Neural Networks (ICNN'97)*, Houston, June 1997, vol.1, IEEE.

[10] Rafaeli, S. and F. Sudweeks (1997) Interactivity on the Nets, in *NetWork and NetPlay: Virtual Groups on the Internet,* (eds Sudweeks et al), AAAI/MIT Press, Menlo Park, Ca.

[11] Robertson S.E. et al. (1995) OKAPI at TREC-3, in *The Third Text REtrieval Conference*, April 1995, (ed. D.K. Harman), NIST, Gaithersburg, MD, USA.

[12] Rudy I.A. (1996) A Critical Review of Research on Electronic Mail, in *European Journal of Information Systems*, 4, 198-213.

[13] Sallis, P.J., S.G MacDonell,. and A Aakjaer (1996) Software Forensics: old methods for a new science. In *Proceedings, Software Education Conference (SE:E&P '96).* Dunedin, New Zealand, January 1996.

[14] The Comprehensivehttp://tile.net/ Internet Reference to Discussion Lists, Newsgroups, FTP Sites, Computer Products Vendors and Internet Service & Web Design Companies at http://tile.net/.

[15] Wilkins H. (1991) Computer Talk: Long-Distance Conversations by Computer in *Written Communication*, 8(1), Sage Publications, Inc., 56-78.

[16] Wright, T., Privacy Protection Principles for Electronic mail systems, Report of the Information and Privacy Commissioner for Ontario, Canada, *The Computer Law and Security Report*, March-April,1995