

Visualizing Usenet: A Factor-Analytic Approach

John C. Paolillo
 University of Texas at Arlington
 john@ling.uta.edu

Abstract

Usenet is a widely popular system for organizing and distributing asynchronous discussions, but the newsgroup hierarchies and messages "threads" of Usenet do little to assist users in navigating through the hundreds of thousands of messages that appear on a single day. Messages may be cross-posted to newsgroups in different, even unrelated hierarchies, and threads wander off-topic as people follow up to prior messages. The social practices afforded by Usenet are not readily apparent in the interface of the typical newsreader.

In this paper, I utilize corpus linguistic methods to suggest strategies for visualizing the underlying social practices of Usenet discussions. Using common factor analysis of word frequency lists extracted from a corpus of soc.culture Usenet messages, I present displays that are organized by language, topic of discussion, and frequency of interaction. Thus, by using information that is already present in Usenet messages themselves, new, potentially more meaningful displays of topic and social interaction can be obtained.

1. Usenet News

Usenet News is a widely used asynchronous computer-mediated discussion protocol in which messages are organized into hierarchically-structured newsgroups. Messages are stored and copied among a network of servers, and users connect using a "newsreader" client which downloads messages from an appropriate server, when requested by the user. Usenet news networks may be any size, but the most prominent network, simply known as "the Usenet" today consists of tens of thousands of newsgroups on a broad range of topics — social, cultural, political, academic, etc. Usenet News was designed to be open-ended, and as the Usenet has grown to the point where it now reaches millions of users, the number of newsgroups has also grown.

Of course, such growth is never problem-free. For example, the soc.culture hierarchy doubled in size, with more than 130 newsgroups by 1996. Some large groups, such as soc.culture.indian, splintered and spun off several new newsgroups, all the while carrying over 100 messages per day in traffic. At the same time, the spin-

off newsgroups, such as soc.culture.punjab, are not always clearly distinct from the parent and sibling newsgroups, since users persist in cross-posting a large proportion of their messages. Another manifestation of growth is especially prevalent in the alt hierarchy, where many of the thousands of newsgroups carry no message traffic. Often these are "joke" newsgroups, sometimes created as meta-commentary on the Usenet's expansion, e.g. alt.is.dead, etc. [10].

In addition to the problems of growth, there are also usability problems with Usenet News itself, especially topic drift and conversational incoherence [4]. Usenet clients generally list the messages in a newsgroup by a construct known as a "thread", identified by the subject line.¹ When a user follows up on a message, s/he can also edit the message headers, to indicate a change of topic. However, users seldom employ this option when following up on tangential points in others' messages. The result is that long threads frequently contain many off-topic messages. Or sometimes a message may become "disconnected" from the other messages on a closely related topic, thus causing the discussion to splinter and lose coherence.

Thus, in the turbulent social environment of the Usenet, users must contend with poorly-defined and confusing discussion spaces. Newsgroups and threads are, in effect, convenient fictions that are only respected by the Usenet News client and server software. They fail to serve as accurate indexes to the content of the discussions taking place, because they require too much maintenance on the part of the user. What is needed is a means of indexing messages according to their actual content, and a transparent means of displaying the relationships among messages. Ideally, this system should be more generally applicable than the heavily hand-tailored newsreader Loom [2], and it should provide more transparent means of identifying meaningful groupings of messages.

2. Usenet, noise and information

Identifying coherent, meaningful conversations on the Usenet is analogous to isolating signals in a very noisy environment. To isolate a signal, it is necessary to

¹ Although RFC 1036 [5] endorses the use of the "references" header line to group related messages, most news clients do not display this information transparently.

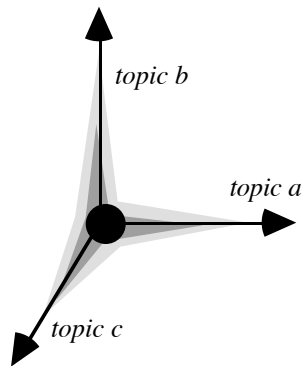


Figure 1. Idealization of Usenet message space.

identify characteristics of it that are different from that of the noise, and to use those characteristics to filter the noise from the signal. What counts as “noise” or signal depends in large part on the user’s purposes of the moment. One might have different purposes for accessing Usenet at different times, and so want to attend to different kinds of messages.

From the user’s point of view, only a very small part of what is available on the Usenet, or even on a given newsgroup, may be of interest. The vast majority of the messages constitute noise, and are only very weakly related to any particular topic. Messages tend to be more about one topic than about any other; the more a message concerns a particular topic, the less it tends to be associated with any other topics, however closely related they may be. This leads to a conception of the message space in which different topics are represented by different, orthogonal dimensions. The idealized distribution of messages resembles Figure 1, where the density of the pattern in a particular region of the graph corresponds to the density of messages found there. If a user could simply nominate a number of topics to be used in a display like Figure 1, s/he could then readily identify messages of interest by navigating through the message space graphically, starting from the tails of the graph and moving along each, toward the center.

While only three dimensions are shown in Figure 1, there should be one ray for every coherent conversation in the message space, so the entire Usenet would require a very large number of dimensions to fully characterize it. If the user were able to nominate the number and kind of topics s/he were interested in, then the display could be limited to a selection of those that could be easily represented at once.

The conception of the message space in Figure 1 has other advantages as well. One such advantage is that the statistical technique of Common Factor Analysis (CFA) is well suited to identifying systems of orthogonal dimensions like those in Figure 1 [7, 8, 13]. If messages in the message space could be represented using data in the appropriate form, it should prove relatively easy to construct graphs like Figure 1 for navigation. The

Table 1. Data and samples for different displays of Usenet message space.

Data	Samples		
	message	poster	newsgroup
<i>Newsgroups</i>	crosspost patterns (1)	how people “hang out”	crossposting patterns
<i>Paths</i>	where msgs come from	where pstrs post from	news traffic patterns
<i>Grammar</i>	message by language (2)	posters by language (3)	newsgroups by lg.
<i>Content</i>	msg by topic	posters by topic (4)	newsgroups by topic

association of the dimensions with topics suggests that the underlying dimensions might be identified through word frequencies. Each topic might be represented by a relatively small number of words. Messages in which those words appear sufficiently often can then be construed to be about that topic. Thus, it should be possible to construct a view of the message space like that in Figure 1 merely by conducting CFA on word frequencies that are extracted from a suitable corpus of Usenet messages.²

A further advantage of this approach is that it lends itself to other, similar views of the message space, representing different kinds of information. For example, if high-frequency grammar words are chosen (e.g. “the”, “of”, “by”, etc. for English, “et”, “je” “le”, etc. for French, “en”, “es”, “lo”, etc. for Spanish), then the message space would be displayed in dimensions corresponding to languages, rather than topics. Moreover, Usenet message headers can be processed in the same way, to conduct analyses based on the senders, the subject lines, the newsgroups to which the message was posted, the path of servers through which each message arrives, and other information. These kinds of information could be recombined and used to generate a range of different displays of the message space. Some of these possibilities are indicated in Table 1.

Each of these different possible displays offers a view of Usenet that users might want to use in selecting messages for perusal, especially those views that reflect the social practices of Usenet: topic coherence, cross-posting and hanging out patterns. The procedures for constructing each of these views of the message space differ primarily in the nature of the data extracted and the kinds of samples that the messages are organized into.

3. Methods

In order to explore the use of displays like Figure 1 in navigating through Usenet data, I conducted an analysis

² A similar approach has been used in literary stylistics [1].

of a 2.5 million word corpus of Usenet data collected in October and November 1996. This corpus represents the bulk of messages posted to 130 newsgroups under the soc.culture hierarchy for a two-week period. The messages were downloaded from the Usenet server at the University of Texas at Arlington, which filtered its Usenet newsfeed fairly extensively at the time in an effort to prevent pornography trafficking and other abuses, and to limit the message capacity that the server needed to handle. As a consequence, some soc.culture newsgroups in existence at that time were not carried by the UTA server, and so their data are only represented by the messages that were cross-posted to the other groups. In addition, the UTA server was fairly far downstream from most of the news traffic, meaning that some message traffic would only less reliably reach the UTA server.

The messages were downloaded using NewsWatcher, a freeware Macintosh news browser. The entire message traffic for each newsgroup was downloaded into a single file. Thus, 130 files, ranging in size from 15 KB to 2 MB were created. These were then processed with specialized computer programs to allow the messages to be imported into a Filemaker relational database. To simplify processing, all upper-case alphabetic characters were rolled into lower case. Since there is a large amount of cross-posting on the soc.culture hierarchy, it was also necessary to cull out a large number of duplicate messages, using a script that sorted and matched the messages by their ID numbers. This resulted in a final corpus of 10,000 messages. Once imported, the posters' email addresses (often faked, but nonetheless unique and serviceable identifiers of the poster), the newsgroups and the message ID of each message were extracted, as well as the message text itself. This information was then exported and further processed, using other specialized programs, for input to Conc 1.8, a freeware Macintosh concordancing program, which was used to generate the necessary word-frequency lists for the different samples. These frequency lists were then re-imported into relational databases to facilitate their display as tables of the appropriate form, which were then submitted to the CFA procedure of Minitab 11 Xtra.

For the purposes of this study, only four of the analyses described in Table 1 were conducted; these are indicated by the numbered, boxed cells. The nature of the samples and data used for each is described further below.

(1) Messages by newsgroup: For this display, the entire set of 10,000 messages was too large, so a randomly selected sub-sample of 1,000 messages was used. The 100 most frequently mentioned soc.culture newsgroups were listed, and each message was identified as being posted to that group (by entering "1" in the database) or not (by entering "0"). To facilitate the factor analysis, the set of 100 newsgroups was broken into four sets of 25 newsgroups each; CFA was conducted separately on each set to identify those newsgroups most likely to result in stable, orthogonal factors. Once

identified, the remaining newsgroups (49) were combined into a single set for further factor analysis. Because of the large number of dimensions hypothesized to be in the message space, and because of the likelihood that newsgroups are self-contained discussions, there is a tendency for each newsgroup to be assigned to a separate factor, a separate dimension of its own, as it were. All such single-newsgroup factors were culled from the analysis, so that the presentation would focus on the cross-posting patterns among newsgroups.

(2) Messages by language: for this display, I used the same random sub-sample of 1,000 messages as in the previous analysis, and a selection of 100 grammar words from different languages. This list was compiled by extracting the 25 most frequent words from each newsgroup and combining the resulting lists. Topic words that happened to occur in this list were excluded, and the list was narrowed to the 100 most frequent remaining grammar words. The data were submitted to CFA again in an iterative fashion as above.

(3) Languages by poster: For this analysis, the same 100 grammar words of the second analysis were used, but the data were aggregated this time according to poster, so that the language patterns of individuals could be tracked. There are 4172 unique posters in the corpus, so again it was necessary to take a sub-sample for the analysis to be tractable. For this reason, only posters who had contributed at least three messages in the corpus were considered, resulting in a reduced list of 625 posters. Again, an iterative process of CFA was conducted.

(4) Topics by poster: For this analysis, I used the same the aggregation of messages according to the 625 posters as in the previous analysis. However, instead of 100 grammar words, I selected 100 topic words from a master word frequency list; each word selected had a frequency greater than 120 tokens per million words, so that it would be likely to occur in at least a few messages. Again CFA was applied iteratively.

4. Results

Overall, the results of the factor analyses came out as expected, and more or less, each one fits the idealized factor graph in Figure 1. The analysis that shows this most dramatically is the cross-posting analysis of messages by newsgroup. The other analyses exhibit some degree of non-orthogonality among some of the dimensions, i.e., some of the space in the back and bottom planes of the idealized Figure 1 is occupied with points. In the case of the language analyses, the reason for this is that different languages often share the same grammar word forms, although they are used in different functions. For example, the English indefinite article "a" has the same word form as the case-marking preposition "a" in Spanish (generally written *á*, but for various reasons represented merely as "a" in Usenet messages). In the case of the topic analysis, different conversations may

use the same topic words, either through cross-posting, or through coincidence. Both of these conditions lead to regions of space other than the chief axes being represented in the data.

4.1. Cross-posting

In the cross-posting analysis, eight factors consisting of two or more newsgroups each were identified. These factors grouped a total of 41 out of the 100 groups considered into fairly recognizable regional clusters. In Figure 2, the newsgroups are plotted according to their loadings on the first three factors: Latin-America and Iberia, East Asia, and South Asia and the Middle East. The fit with the idealized model of orthogonal dimensions in Figure 1 is very good. It should be noted however that the eight factors identified only account for about 40% of the observed variance. Therefore, there is a great deal of noise in the message space that is not accounted for by this analysis. However, since the purpose of the analysis is to provide a means for identifying relatively coherent

behaviors in a rather noisy environment, the analysis seems acceptable.

Figure 3 shows the distribution of the messages from which the eight newsgroup factors were derived, according to their first three factor scores. The distribution of messages is even more clearly orthogonal than that of the newsgroups themselves. This suggests a far stronger coherence among the messages exchanged in a given discussion, than of the newsgroups that they are posted to. This interpretation makes sense, given that many different conversations may take place on a newsgroup at any one time. It also suggests that a graph of the message space such as that of Figure 3 would be a useful way to navigate to or through the messages of a discussion. By making the points in the graph active elements of a user-interface, one could access the messages themselves directly, and have a good sense as to what other messages are likely to be relevant.

The knot near the origin in Figures 2 and 3 represents whatever is "noise" from the perspective of the three

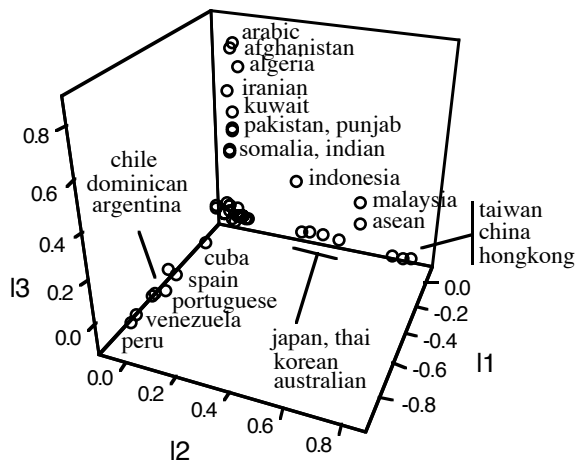


Figure 2: Newsgroups on first three factors.

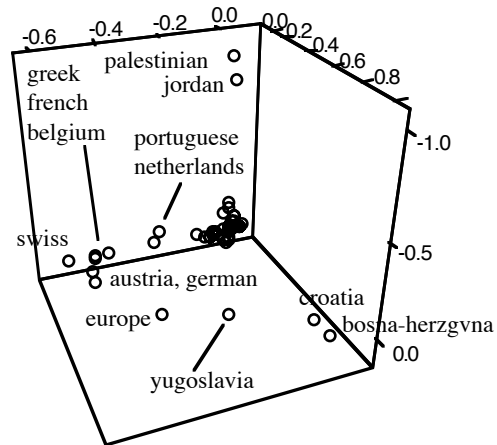


Figure 4: Loadings on the Balkan, Western European and Middle Eastern factors.

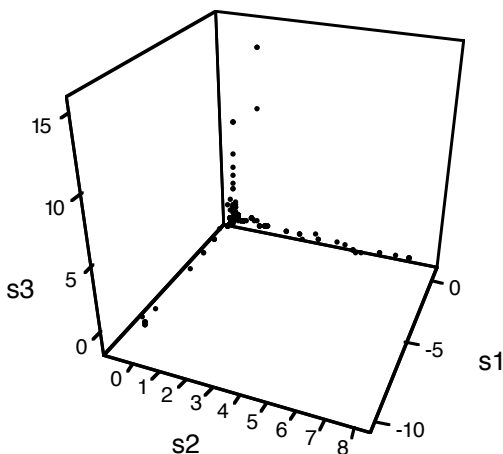


Figure 3: Message density in the first three factors

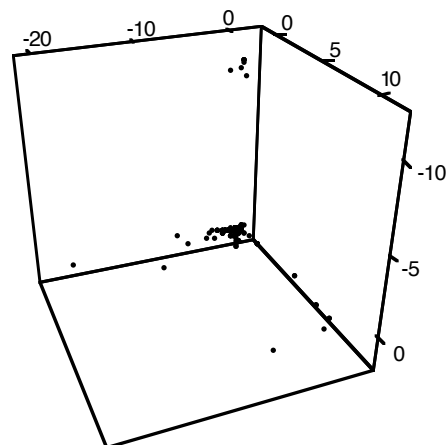


Figure 5: Message density in the Balkan, Western European and Middle Eastern factors.

selected newsgroup factors, including, in this case, the other five factors of newsgroups. In practice, it is difficult to display more than three dimensions at a time, so alternative views need to be provided to illustrate the relationships resulting from the other factors. Such a view is provided for three more newsgroup factors, namely the Balkan, Western European and Middle Eastern newsgroup factors, in Figures 4 and 5. As before, Figure 4 shows the factor loadings of the newsgroups, and Figure 5 shows the factor scores of the messages. There is a small degree of non-orthogonality among the Balkan and Western European factors, possibly on account of cross-posting of some messages to newsgroups on both factors, as indicated by the dots on the lower plane of Figure 5. A display such as this would allow users to identify such cross-talk between the different newsgroups, and either investigate it filter it out, as desired.

4.2. Languages

At present, displays according to language factors are probably more interesting for research purposes than for the orientation of users to the message space. At the same time, it is an important direction to look toward as communication networks such as the Usenet become larger and more widespread. Even today, one cannot say in principle what language(s) a Usenet poster may wish to communicate in, other than that it is most likely that s/he will use English or another major European language. It is important, in any case, to put the linguistic analysis to test, to see if the factor-analytic approach will serve the purpose of identifying discussions in languages that a user might want to orient toward.

In the analysis of the random sample of individual messages by grammar words, three main factors of

grammar words were identified, corresponding to Spanish, English and French. The factor loadings and factor scores of this analysis show somewhat less orthogonality than the last analysis, and this particularly between the Spanish and French factors, on account of the fact that some grammar word forms are shared in the two languages (e.g. “la”, “de”, “un”, etc.). In addition, we notice a tendency for the words to cluster away from the origin point, unlike the analysis of the newsgroups above, and the idealized message space in Figure 1. This is probably on account of a tendency for messages to be in one language or another, and to avoid code-mixing [11, 12]. In any case, the model of orthogonal dimensions still appears to fit fairly well. The fit might possibly be improved by using two-word (or larger) collocations (i.e., n-grams), rather than single words, since collocations are more likely to be unique from language to language (e.g., English, French and Spanish all have a wordform “a”, but “of a” is a distinctly English collocation, “a la” a distinctly French one, etc.).

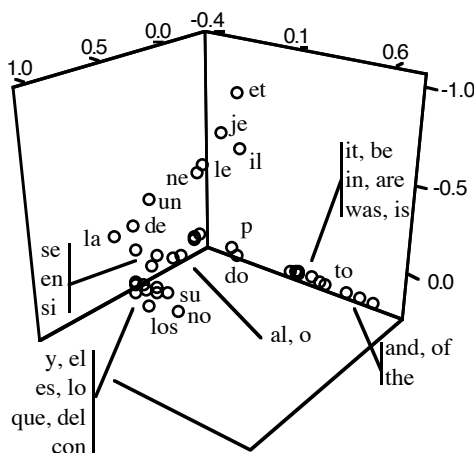


Figure 6: Factor loadings of grammar words in random sample of messages.

Table 2. Selection of 100 grammar words.

word	freq.	word	freq.	word	freq.	word	freq.
a	45989	he	6528	au	1099	por	2400
i	28576	we	6432	cu	1008	its	2357
e	5478	en	6193	the	97323	una	2186
y	5418	me	5649	and	42191	new	2072
o	4199	el	5622	you	21359	her	2062
u	2394	an	5604	for	15827	she	1740
p	1564	so	5368	are	13565	las	1622
m	1489	my	5131	not	12751	les	1387
v	1173	un	4299	was	9223	god	1073
to	50220	se	4117	que	9161	war	1040
of	49248	si	3015	but	7822	that	23092
in	38647	up	3008	all	6774	have	11768
is	27375	le	2860	who	6149	this	11190
de	17202	es	2366	one	5451	with	10398
it	16226	da	2353	has	5283	from	8476
on	13000	il	2144	can	4825	they	7723
la	11717	al	2055	his	4294	your	6679
as	10992	ca	2040	any	3606	what	5995
be	10970	lo	1985	los	3065	will	5517
no	8882	na	1866	had	3001	were	4145
by	8843	ne	1686	bit	2933	more	4128
or	8214	su	1648	our	2763	some	3791
if	7096	je	1349	che	2647	been	3308
at	6874	te	1201	del	2628	know	3009
do	6674	et	1165	con	2577	said	2695

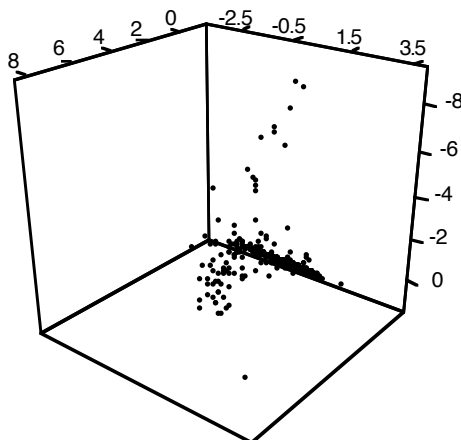
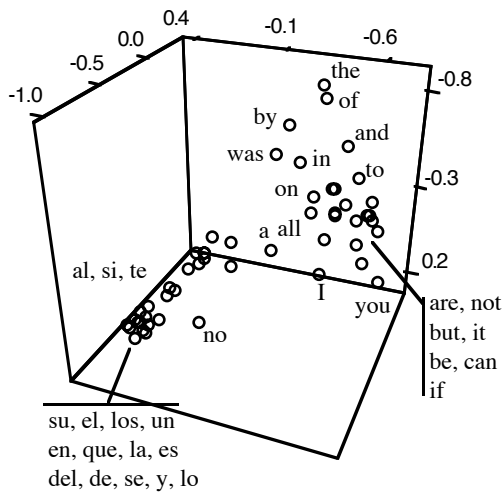


Figure 7: Distribution of random message sample by factors of grammar words.

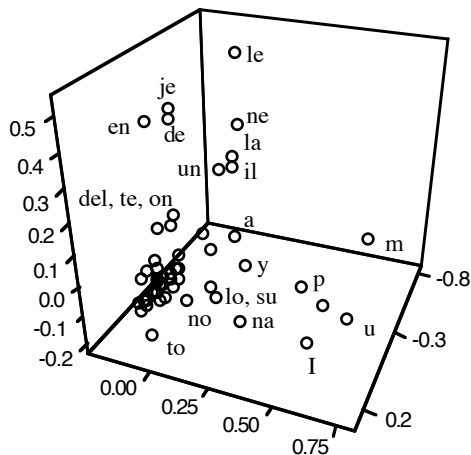
In Figure 7, where the message sample is plotted according to its factor scores, the non-orthogonality of the language factors is seen even more clearly. In particular, the messages on the vertical French dimension are “bent” toward the English axis. This could potentially be from some use of English in French messages, since English and French do not share very many grammar word forms.

The second analysis of grammar words aggregated messages by the 625 most frequent posters, in order to see if a more representative distribution of languages could be observed. In this analysis, 51 grammar words were retained in six factors, accounting for 26% of the observed variance. The factors appear to differentiate more than languages, since there are two distinct factors each for both English and French grammar words. Possibly these factors serve as indicators of authorial style [1].

Figure 8 shows two views of the loadings of grammar words on the factors in Table 5. Each view again shows three different factor loadings. View (a) shows the



(a) Spanish, English 1 and English 2

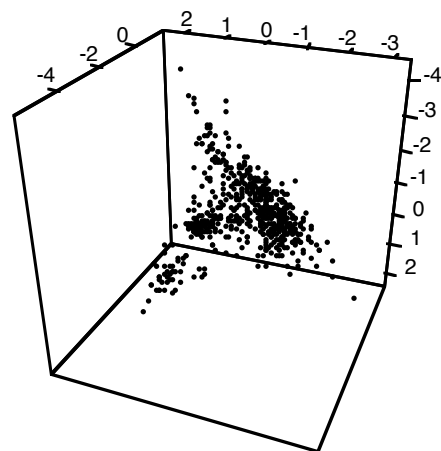


(b) French 1, French 2 and unidentified

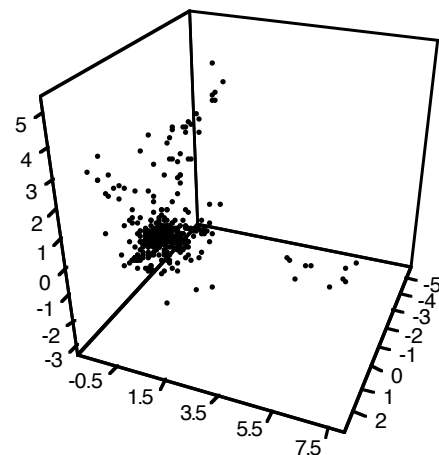
Figure 8: Loadings of grammar words by poster.

Spanish and two English factors, and view (3) shows the two French factors and the remaining unidentified factor. In view (a), the greatest non-orthogonality occurs between the two English factors, suggesting a continuum of different grammatical styles of messages. In view (b), the unidentified dimension appears to be largely orthogonal to the two French dimensions. The six factors of this analysis thus largely confirm those of the previous analysis of grammar words, possibly elaborating it further by adding stylistic dimensions to the English and French factors.

Figure 9 shows two parallel views of the message space, indicating that the density of posters is distributed in much the same way as the factor loadings of the grammar words. At the same time, it becomes clearer that the two English factors forming the back plane of view (a) exhibit a continuous, linear cluster of posters oriented at a 45° angle to the two English factor axes. In other words, most of the English samples lie on an axis which is skew to the axis of the Spanish factor. In view (b), there is a less prominent, though perceptible diagonal



(a) Spanish, English 1 and English 2

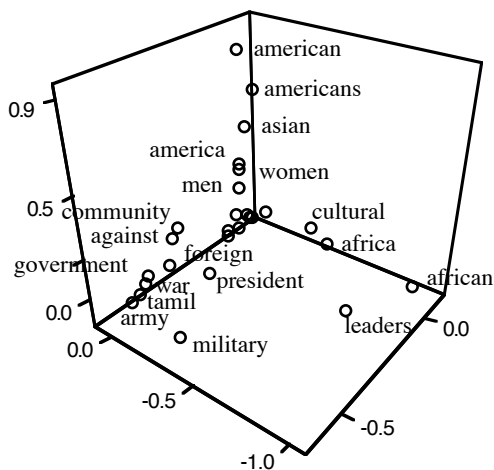


(b) French 1, French 2 and unidentified

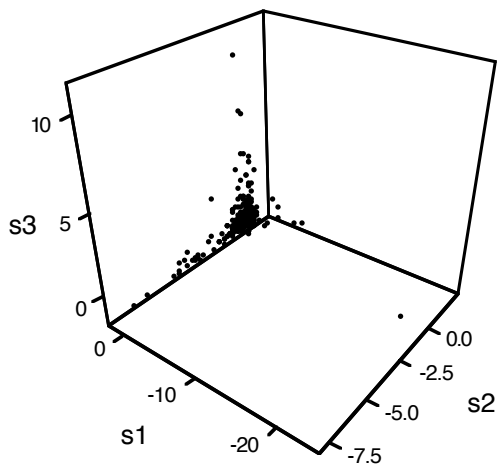
Figure 9: Posters by grammar words.

(perhaps even a “v”) on the two dimensions of French grammar words. Again, an interpretation in terms of stylistic patterns might be possible.

The set of views in Figures 8 and 9 illustrate what may be an important limit to the technique of using word frequencies to identify dimensions of the message space. Words must be carefully selected so that clearly interpretable factors are identified. While there are factors corresponding to the major languages in the sample there is differentiation among some of them in what are possibly stylistic dimensions, yet such an interpretation would need to be supported by closer examination of the data. Furthermore, unless users desire to identify stylistic dimensions of variation in addition to languages, the propensity of grammar words to identify stylistic or other dimensions may be too much of an inconvenience. Further research is therefore necessary to optimize the means for identifying languages in the message space.



(a) Factor loadings of topic words



(b) Factor scores of posters by topic

Figure 10: Factor loadings and factor scores for topics by poster.

4.3. Topic

The final analysis addresses the identification of topics of discussion, a potential alternative to using subject lines to identify threads. Seven topic factors were identified as having two or more topic words from the initial list of 100 topic words. The first three factors are displayed in Figure 10, where view (a) represents the factor loadings of the topic words and view (b) represents the factor scores of each poster on those words. The axes of the factors are occupied somewhat more sparsely than in the previous analyses, in particular the first factor, pertaining to Africa, is occupied by only a single poster who posted a total of four messages, all of which are duplicates of a single message posted separately to four different African soc.culture newsgroups. Unless it turns out that other posters (i.e. from those posting fewer than two messages in the two weeks when these messages were collected) are engaged in conversation with this poster, this is probably a rather uninteresting dimension. However, the absence of other posters along this factor axis is a clear indication to the user of a dearth of coherent discussion along that dimension. The second dimension, a discussion about the ongoing civil war in Sri Lanka, shows a substantially richer discussion, with more posters involved in the interaction. The third dimension, also involving a number of participants, consists of messages to the newsgroup soc.culture.asian.american about interracial relationships in the Asian American community.

As with the grammar words analyses, the selection of words for analysis will make a great difference in the outcome. For this particular analysis, the means of selecting the words was not particularly systematic. Fortunately, there are some simple improvements one could make in the selection procedure, such as using a thesaurus program to assist in nominating word forms for counting and analysis.

5. Discussion

On the whole, this study reveals that representations of the message space similar to the idealized Figure 1 are relatively easy to construct merely by applying factor analysis to catalogued word frequencies of each message in an appropriate sample. The cross-posting patterns, a useful indication of the larger social practices of a collection of newsgroups in the message space, conforms quite closely to the idealized conception in Figure 1, and usefully identifies discussions that are both coherent and potentially of interest to the user. Depending on the words that are selected as a basis for the factor analysis, correspondingly different views of the message space can be obtained. The factor analysis itself seems to be an appropriate procedure for filtering coherent, discursive signals for the large amount of noise on the Usenet. Two types of issues remain to be considered, however: design

issues, and the effects of wide-scale use these techniques in the message space.

5.1. Design Issues

There are three general sorts of design issues that would need to be addressed more fully in any implementation of these techniques for actual use in newsreaders: (i) the selection of appropriate wordforms for analysis, (ii) whether the indexing and analysis is to be accomplished on the server or client side of the system, and (iii), presentation of the results to the user.

Regarding the first issue, while the present study has demonstrated that coherent discussions *can* be identified on the basis of the frequencies of a selected set of words, it has not demonstrated a more than haphazard means for specifying those words. More research is needed to determine what techniques will best serve to reliably identify useful dimensions of coherent conversation in the message space. In part, the techniques which have been used here are responsible for this lack of understanding, since all of the statistical and database computations needed to be tractably performed using off-the-shelf software. This requirement imposed practical limits on the size of the data files that could be used, and meant that certain, potentially useful aggregations in Table 1 could not be attempted (e.g. aggregations of messages within newsgroups). Consequently, it was only possible to view a fraction of the corpus for most analyses (though a larger fraction when aggregated by poster — about half of the 10,000 messages in the corpus were included in these analyses). Applying this technique to a larger selection of newsgroups would require solutions to a number of technical problems.

Design issues of the second sort further amplify the need to address the technical resource-management problems. A certain amount of convenience might be afforded if all of the work needed to present the views of the message space could be accomplished on the client side. After all, the information needed to identify useful dimensions of the message space is already contained in the messages available to the client, and in the case of the first analysis, merely in the headers of those messages. However there are serious problems with this, even beyond downloading large numbers of messages. The analysis of 130 newsgroups here began with a data set of approximately 50 MB of messages. There was substantial redundancy in this data set, which was reduced to approximately 17 MB after headers and redundant messages were removed. The database listing each word according to the message it appears in, which is the most important form of data to all of these analyses, is 270 MB, while the supporting databases that count the words by the appropriate aggregations are between 2 and 3 MB each, all this for only a very small fraction of the total Usenet newsfeed. While storage and speed requirements can no doubt be optimized to a large extent, these figures

should indicate that the burden of performing factor analyses on large samples of Usenet feed is probably too great to be acceptable on a client system. In addition, servers such as Dejanews (now Deja.com) already satisfactorily perform much of the needed indexing of messages; only small modifications would be needed to accommodate the needs of a system that automatically performs the analyses described here. Thus, a more desirable approach would have the data management and statistical computations performed on the server side, leaving only the display functions to the client.

Displaying the statistical analyses for perusal by the user presents its own set of problems, as readers of this paper will no doubt be aware. Traditional graphical modes of presentation in two dimensions (plus or minus one) impose limits on the information that the user can have simultaneous access to. While one can conceive of interfaces that would give users control over which dimensions were displayed at once, displays which are enhanced through animation, etc., it is not immediately clear what techniques would work best for representing the message space. While the possibility of an active, graphic display is appealing, research and testing would need to be conducted to determine what sorts of displays work best and usefully provide users with appropriate information for navigating the message space.

5.2. Usage Issues

A different set of issues concerns the effects of using a system that makes information about social, topical and linguistic coherence available directly to users. Again, three general issues may be identified: (i) ethical aspects of making such information available to large numbers of users, including privacy concerns of users, (ii) possible misuse of the technology by seeding messages, similar to the way that “spider food” is used to attract hits to web pages, and (iii) feedback influences of users exploiting these forms of information on the structures of coherence in the message space.

Opinions differ on the extent to which participants in Usenet discussions should enjoy privacy rights (or even copyright) to the information they post on the Usenet.³ In terms of legalities, all of the information on the Usenet is considered public, and there is no information that is required for the analyses presented in this study that is not already somehow present in the Usenet messages, headers and newsgroups. Thus, there should be no legal obstacle to application of these techniques, since they merely involve extracting information that is already present, and, as it were, part of the public record. On the other hand, people regularly conduct discussions on Usenet newsgroups with a sense of privacy or protection, in the belief that the people listening and responding —

³ See, e.g. [3] and [9], for two quite different views of cyber-ethics in research.

the people actually *using* the messages — will respect the wishes of the people posting regarding how those messages are to be used. Does the automatic indexing of messages, to identify social, linguistic and discursive patterns of behavior, infringe upon this perception of privacy? And does it expose users to potential or actual abuses that they would otherwise not be exposed to? It is difficult to answer these two questions. On the one hand, automatic indexing of messages already takes place, on the aforementioned Dejanews and other websites. Anyone with a web browser can go to Dejanews and look up messages from the same “private” discussions, by the same topic words or newsgroups that would be used for a system based on the present study. On the other hand, the proposed system makes certain kinds correlational information more readily available than it otherwise might be. One cannot rule out the possibility that this would make abuses of peoples’ privacy more possible. However, it should be acknowledged that encrypted messages, or merely messages encoded from a non-ASCII format are not easily handled by the techniques described here; people who wish to protect their conversations from prying mechanical eyes can easily do so through these means (or by using a more private form of messaging). The problems raised are not new problems in the design and use of message systems, though they deserve careful consideration in the design of new Usenet systems, particularly on the server side, where information is potentially broadcast to large numbers of people.

The proposed method of identifying conversations has one noteworthy and not-so-obvious advantage: it automatically filters out spam and advertising that tends to clutter unmoderated newsgroups.⁴ However, this advantage raises the possibility that all an advertiser needs to do to get his/her message read by members of a discussion is to “seed” the offending message with enough topic words of the appropriate sort to attract the participants of a given discussion. One could invade different discussions by including other seed-words in a copy of the same message sent separately. The problem is analogous to the use of “spider food” in commercial websites to attract users searching for current “hot” topics. To a certain extent, these problems are preventable if one knows the content of the advertising or spam messages; these messages themselves can show up as a factor in their own right, as the single-poster “Africa” factor, of Figure 10 illustrates. Of course, we can imagine an escalating arms race in which the advertisers and spammers use more clever means to hide the offending messages, and systems use more elaborate means to allow users to weed and filter them out. These problems are also not unique to the Usenet, nor are they an affliction specific to the kind of system proposed here. Solutions to these problems, if they come at all, are

⁴ Unless one were to select topic words like “get rich”, “envelope” and perhaps “scam” (as in “this is not a scam”).

likely to come from the regulation of advertising on the Internet more generally.

A final and more interesting question concerns the potential effects of providing information about discursive coherence on the structure of the message space itself. It is impossible to do more than speculate about these effects at this point. However, two possible paths of development are imaginable. One possibility is users would reinforce existing coherence patterns, so that a given discussion would become more and more separate, effectively shrinking its connection with the region of discursive noise at the graph origin. In the extreme, this could lead to the fragmentation of the message space into entirely separate, non-overlapping discussions. The other possibility goes in the opposite direction: whatever coherence emerges from any given discussion could be reduced by people coming from outside discussions and discovering it, or discovering an interest in it; they might then address more posts to both discussions, and thereby contribute to non-orthogonality in the discussion factors, bending them towards each other and eventually collapsing them back into the incoherent noise at the origin. What makes both possibilities all the more speculative is that such developments might even occur at present, without the enhanced information about social and discursive coherence. More research is needed to reveal the temporal processes of discussions unfolding in a large message space like the Usenet.⁵ Such research itself could fruitfully employ analyses such as the factor analysis of word frequencies that has been employed here, taking samples that are temporally organized and comparing them.⁶

6. Conclusion

This study has shown that factor analysis of word-frequency lists can provide a rough-and-ready analysis of coherence among messages in a message space on many different levels, including cross-posting patterns, topics, and languages employed. These initial results are in many ways merely suggestive, and need further refinements. In particular, techniques need to be developed that will improve the selections of wordforms for analysis. Other refinements were suggested above, such as using collocations of words, rather than individual wordforms, and using an online thesaurus for nominating topic words for analysis. Other applications of the techniques can be imagined as well, such as organizing the results reported by web search engines.

Much further research is needed, however, to identify the most useful ways to discover and present information about the structure of the message space. In addition, we

⁵ While [4] addresses the development discursive coherence to some extent, it does so only on a micro-analytic level, and not on the scale of larger temporal and social units.

⁶ This is the technique adopted by Burrows in an analysis of the historical developments in fiction on the stylistic level [1].

need an improved understanding of the development of coherence in discursive interactions over time, especially as such coherent dimensions emerge out of large, noisy message spaces such as the Usenet. Only when such things are known can we assess the ultimate impact of publicly available information about the social, discursive, and linguistic dimensions of the Usenet message space.

7. References

- [1] Burrows, John F. 1992. Computers and the study of literature. In C. S. Butler, ed. *Computers and Written Texts*, 167-204. Oxford: Blackwell.
- [2] Donath, Judith; Karrie Karahilos; and Fernanda Viegas. 1999. Visualizing conversation. Proceedings of the 32nd Hawaii International Conference on Systems Sciences. Los Alamitos, CA: Institute of Electrical and Electronics Engineers (IEEE) Computer Society.
- [3] Herring, Susan C. 1996. Linguistic and critical analysis of computer-mediated communication: Some ethical and scholarly considerations. *The Information Society*, 12.2:153-168.
- [4] Herring, Susan C. 1999. Interactional coherence in CMC. Proceedings of the 32nd Hawaii International Conference on Systems Sciences. Los Alamitos, CA: Institute of Electrical and Electronics Engineers (IEEE) Computer Society.
- [5] Horton, Mark, and R. Adams. 1987. Standard for interchange of USENET messages. Network Working Group RFC 1036 (<http://www.ietf.org/rfc/rfc1036.txt>).
- [6] Kantor, Brian, and Phil Lapsley. 1986. Network News Transfer Protocol. Network Working Group RFC 977 (<http://www.ietf.org/rfc/rfc0977.txt>).
- [7] Kim, Jae-On; and Charles W. Mueller. 1978a. Introduction to factor analysis: What it is and how to do it. Sage University Paper series on Quantitative Applications in the Social Sciences., 07-013. Newbury Park, CA: Sage.
- [8] Kim, Jae-On; and Charles W. Mueller. 1978b. Factor analysis: Statistical methods and practical issues. Sage University Paper series on Quantitative Applications in the Social Sciences., 07-014. Newbury Park, CA: Sage.
- [9] King, Storm A. 1996. Researching Internet communities: Proposed ethical guidelines for the reporting of results. *The Information Society*, 12.2:119-127.
- [10] Paolillo, John C. 1996. Language choice on soc.culture.punjab. *Electronic Journal of Communication/Revue Electronique de Communication*, 6(3). (<http://www.cios.org/>)
- [11] Paolillo, John C. 1997. Meta-joke newsgroups on Usenet. Paper presented at the Annual Meeting of the International Society for Humor Studies, Oklahoma City, OK, July 1997.
- [12] Paolillo, John C. Forthcoming. 'Conversational' codeswitching on Usenet and Internet Relay Chat. To appear in S. Herring, ed., *Computer-Mediated Conversation*. Oxford: Oxford University Press.
- [13] Reitveld, Toni; and Roeland van Hout. 1993. *Statistical Techniques for the Study of Language and Language Behavior*. Berlin: Mouton De Gruyter.