2003. *Symposium on New Research for New Media,* University of Minnesota, Minneapolis, September 8.

Dynamic Topic Analysis of Synchronous Chat

Susan C. Herring
Indiana University, Bloomington
herring@indiana.edu

Introduction

Dynamic topic analysis is one of a set of computer-mediated discourse analysis techniques developed by the author for analyzing coherence in text-based computer-mediated communication (CMC).[1] Its purpose is to track what participants in synchronous discussion forums are talking about as conversation unfolds dynamically over time, and to quantify and visually represent the resultant patterns. Dynamic topic analysis allows CMC researchers to evaluate the nature and quality of online conversations taking place in Web chat, IRC, MUDs/MOOs, Instant Messaging, and graphical virtual worlds, as well as through specialized chat systems in educational and business environments. Unlike asynchronous discussion forums in which messages containing subject and reply lines allow for automated threading, most chat systems have no formal mechanism for indicating to what a message is responding; rather, cross-message coherence must be inferred from meaning. Dynamic topic analysis identifies the semantic relationships between ideas by reconstructing the message producer's intent to communicate a certain idea in response (or not) to a previously expressed idea, and codes and visually links these relationships. The resulting set of semantic links in temporal sequence reveals the topical flow or "threads" of a chat conversation.

In this paper, I describe how dynamic topic analysis (henceforth, DTA) was employed in a comparative study of recreational and pedagogical uses of Internet Relay Chat (Herring & Nix, 1997). Previous research had advanced optimistic claims about the potential of chat systems to enhance online learning, at the same time as empirical research was bringing to light some of their limitations, including a tendency for chat conversations to be fragmented, chaotic, and topically digressive. At that time, most chat research was based on recreational chat in unstructured, public access Internet contexts. We were interested in finding out if serious, focused, on-topic chat were possible, and if so, what it would look like, as a means to evaluate the extent to which synchronous chat can facilitate learning in educational contexts.

Three extended sequences were selected for analysis from each of two Internet Relay Chat (IRC) contexts, one an on-line course providing advanced instruction in pharmacology, and the other a chat channel devoted exclusively to light social interaction. So that the two samples could be compared, a method was devised to describe and measure topical coherence. It was not enough simply to identify the topic of each sequence, as most topic analysis does (e.g., Chafe, 1994), or to code on-topic as opposed to off-topic messages, on the assumption that each sequence had a single global topic. Chat conversations can shift quickly, for example when new participants enter the chat environment, and multiple conversations often take place simultaneously in the same chat space (Herring, 1999). For multi-participant chat data, therefore, it was necessary to conceptualize topic locally, at the level of the individual exchange, and dynamically, as potentially re-directed by each new message that is posted to the chat channel.

In what follows, I describe the methods we devised to meet this challenge, how they were implemented, and what they reveal in relation to the larger research question. I then discuss their limitations and propose refinements for improving DTA as a tool for CMC research in the future.

The IRC Study

Theoretical Assumptions and Hypotheses

Our analysis of topical coherence in synchronous chat is based on the following theoretical assumptions. First, we assume that chat is "conversation-like", such that methods of conversation analysis can appropriately be applied to it. (The relationship between moves and move sequences involving multiple participants has been previously studied primarily in spoken discourse contexts; e.g., McLaughlin, 1984). Second, with regard to "on-topicness", we assume that an ideal conversation is *not* comprised of a sequence of turns all of which talk about the same thing (that would be uninteresting), but rather that there is an optimal balance of narrowly on-topic moves and moves that add to or otherwise shift the topic in new directions (cf. Hobbs, 1990; Jefferson, 1988; Sacks, 1992). (The precise nature of this balance is a matter for empirical investigation.) Third, in as much as discourse patterns vary in general according to speakers and communicative purpose (cf. Schiffrin, 1994), we assume this will also be the case for topicality, and thus there may be different manifestations (and ideals) of topical coherence depending on the context. Fourth and last, we assume that these manifestations are potentially systematic, and that characteristic topic trajectories can be identified for different contexts of chat communication.

These assumptions, when applied to the two types of IRC data we wish to compare, lead to the following hypothesis:

H1: The purpose of chat communication makes a difference to topical coherence, such that pedagogical chat and recreational chat will display different topical trajectories. Specifically, pedagogical chat will have longer continuous threads, fewer threads, and stay more on-topic than recreational chat, which will have shorter, overlapping threads and more topical digression.

Coding the Text Data

Three sessions from each of two IRC channels were analyzed to test the above hypothesis. The total corpus consisted of 1878 messages, of which 1540 were contributed to the pharmacy group, and 338 were contributed to the social chat group. These were convenience samples that were found on each group's Web site as examples of the group's activity; the researchers did not record or directly observe the original chats. They were selected to optimize the contrast in purpose: We looked for the most serious and the most lighthearted chat channels we could find. The size of each session reflects natural boundaries: The pharmacy sessions were complete class meetings, and the social chat sessions were each more-or-less focused around a single theme. In this sense, the data are biased in favor of coherence, thus any digression or fragmentation found probably under-represents the incidence of such phenomena in IRC as a whole. Each session involved a handful of participants (the pharmacy sessions were taught by one of two male teachers, with four to five students in each session; the social chat averaged seven participants in each session), and there was little joining or leaving of the channel during the portion of the session that was logged. The nature of the data as self-selected, naturally-bounded units of (relatively) coherent chat conversation with relatively limited, stable participation should be kept in mind in interpreting the results of the analysis. Such a configuration is not uncommon in educational uses of chat, but previous research suggests that recreational IRC is typically less stable and coherent.

The messages in the corpus were coded for topical coherence by assessing the relation between each independent proposition (typically a single message, expressed as a single sentence or sentence fragment) and the previous proposition or message in the conversation to which it appears to have been most directly intended to relate. Each proposition was coded for two kinds of information: 1) the *type* of the relation, and 2) the semantic *distance* between the two propositions. *Types* of topic relations were adapted from those identified by Hobbs (1990). A proposition can be narrowly *on-topic*, shift the topic through *parallelism*, *explanation*, or *metatalk*, or *break* from the previous topic altogether. The topical *distance* between propositions was coded on a four-point scale from 0 to 4, with '0' representing a maximally topically-related proposition and '4' a maximally unrelated proposition.

There is a logical interaction between the topic relation types and semantic distance. Narrowly on-topic messages, which include simple agreements, reactions, rephrasings and clarification questions, by convention are assigned a semantic distance of '0'. Breaks, which include propositions which can not be directly related through plausible inference to any previous utterance, are conventionally assigned the maximum value of '4'. Parallelisms and explanations by definition introduce new information into the conversation, and therefore shift the topic; the shift can be slight, or it can be a stretch that challenges the imagination to supply linking inferences. These two types are assigned a value, depending on the context, ranging from '1' to '3'. Metatalk, which is talk about the conversation itself or its organization, is similar, but includes the possibility of a '0' value, in case it comments directly on the previous topic without adding any new information beyond a higher level of abstraction.

The identification of degrees of distance in parallel shifts, explanations and metatalk is somewhat subjective and is one of the most difficult aspects of the coding to implement consistently across samples. To illustrate different degrees of distance of parallel shift, consider the following extract from one of the social chat sessions. The theme of this session is playful banter about blow-up dolls. Messages preceded by nicknames in parentheses are normal 'utterances'; messages preceded by an asterisk are 'action descriptions' produced by the participant whose nick name appears in the subject position in the message (cf. Cherny, 1995). Intervening messages that were not part of this conversation are omitted to simplify the presentation.

Example 1:

14 (poosh) do those things have hair? or do you supply a wig?
15 (sigh) *snicker*
16 (blot) hair optional!
19 (poosh) blot: cool!
20 (poosh) Sinead O'Connor blow up doll!
21 (blot) lol-poosh
22 (happy1) poosh: That one would gather dust on the shelf!
23 (sigh) 9 ball, side pocket!
27 * blot is behind 8 ball!
28 * poosh is under the table

We coded this sequence as show in Table 1:[2]

Proposition	Responds to	Relation Type	Distance
14	n/a	--	–
15	14	On-topic	0
16	14	Parallel	1
19	16	On-topic	0
20	16	Parallel	2
21	20	On-topic	0
22	20	Parallel	1
23	20	Parallel	3
27	23	Parallel	1
28	27	Parallel	2

Table 1. Coding of a sample sequence

The on-topic messages in this extract are straightforward reactions of amusement and appreciation; their coding poses no interpretive difficulty. The parallel shifts, in contrast, require the analyst to supply inferential links based on his or her real-world and cultural knowledge; to the extent that such knowledge varies among coders, assessments of degrees of relatedness may also vary. The two researchers who coded the data for this study, native English-speaking American females in their 30's and 40's, agreed after some discussion that proposition 16 was a direct answer to the question in 14, and that 22 was a direct evaluation of 20 (i.e., that a Sinead O'Connor blow up doll would not be popular/attractive); each of these was assigned a value of '1'. The relation of 27 to 23 is less straightforward, but we ultimately agreed that blot's reference to being 'behind the 8 ball' was a cooperative continuation of poosh's playful invocation of a game of eight ball (pocket pool), and assigned it a distance of '1'. Relating proposition 20 ("Sinead O'Connor blow up doll!") to the previous context required further inferencing: it was necessary to know that Sinead O'Connor was a popular female singer at the time who shaved her head. This adds two new notions to the conversation: that O'Connor is bald, and that blow-up dolls can be modeled after real people; it was thus assigned a distance of '2'. Similarly, poosh's comment in 28 that she is 'under the table' adds not only a new location (the last location was 'behind the 8 ball', hence on the pool table), but plays off of the idiomatic sense of both locative expressions, thereby introducing the additional idea of being disadvantaged or incapacitated; this was assigned a '2' as well. Finally, although it is tempting to treat sigh's remark in proposition 23 ('9 ball, side pocket!') as a complete non-sequitur, upon reflection we were able to reconstruct the inference that bald-headed people resemble pool balls, and decided that the proposition was intended to continue the ongoing topical sequence (this interpretation is supported by the fact that the subsequent conversation continued on the theme of blow-up dolls modeled after real people). A minimum of three inferential steps is required to connect this utterance with proposition 20: 1) Sinead O'Connor is bald, 2) bald people resemble pool balls, and 3) pool balls are used to play eight ball. Thus we assigned this proposition a value of '3'.

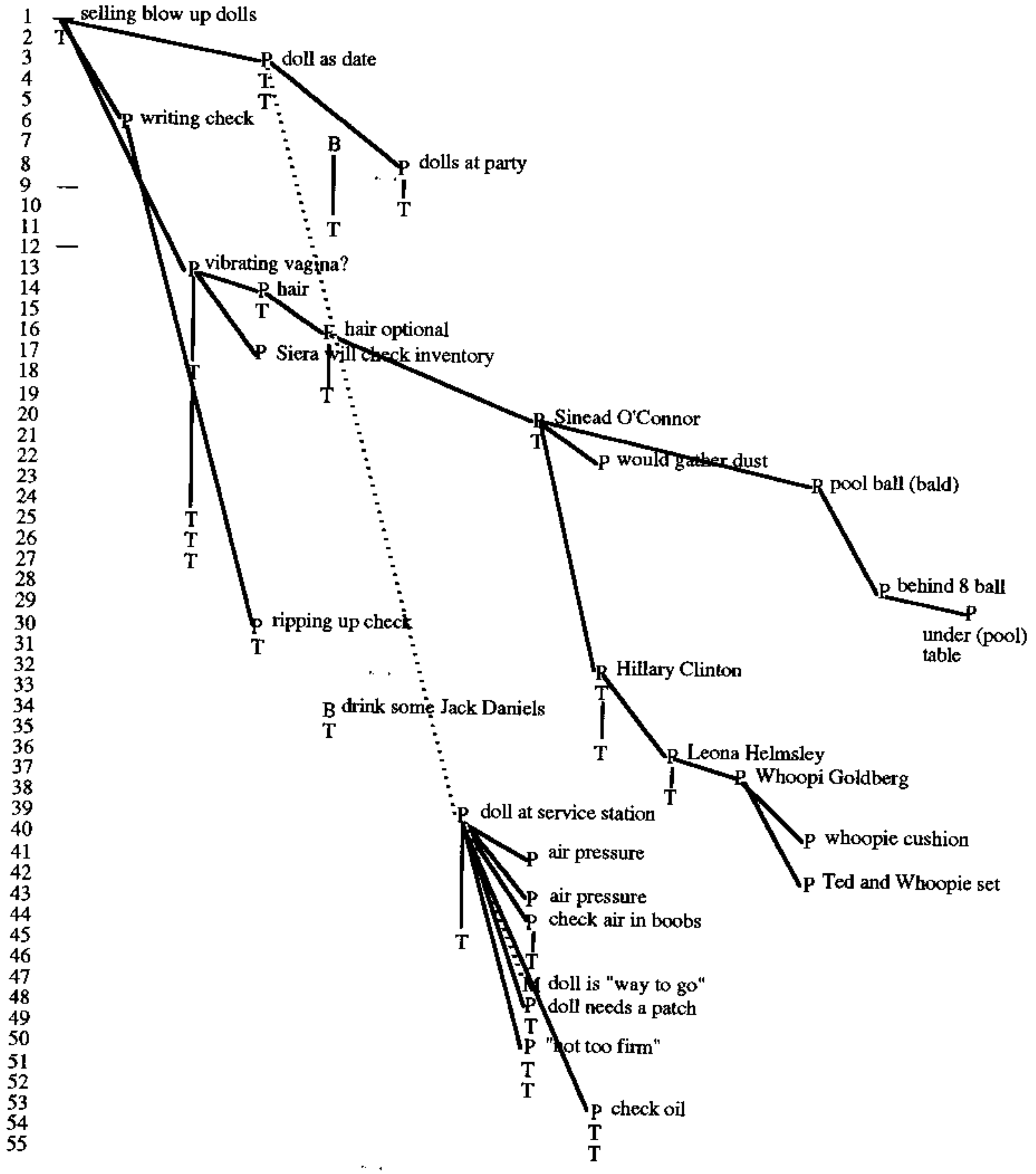
All of the propositions in the corpus were coded in the above manner.

Creating the Visual Representation

The coded information as presented in Table 1 can be transformed into a graphical representation of topical structure over time. We did this by representing time (message/proposition number) along the y-axis, and semantic distance along the x-axis of a two-dimensional grid. Distance was represented cumulatively from left to right; that is, the distance for each new proposition was added to the cumulative distance of the most directly related previous proposition. Relation type, finally, was indicated in the diagram itself, by means of a letter (T for 'on-topic', P for 'parallel shift', E for 'explanation', M for 'metatalk', and B for 'break'), and by placement either directly below (for T) or to the right (E, M, or B) of the proposition to which it is responding. P, E, and M moves are connected to that proposition by a diagonal line. B moves (breaks) are not connected visually to any previous proposition.

As an illustration of this method of visual representation, consider Figure 1 below, which represents the first 55 propositions of the 'blow up doll' session. Propositions 16-28 show how example 1 above looks when diagrammed according to this method. In the diagram, all non-T ("off-topic") moves are labeled as a convenience to remind the viewer of their content. A dotted line indicates a tenuous connection between propositions.

Figure 1. Visual representation of topical structure of 'blow up doll' conversation



In this method of diagramming, off-topic or digressive sequences appear to extend horizontally, while on-topic sequences extend vertically. The blow up doll session has a rather strong horizontal orientation, an indication that topics in this chat conversation tended to digress.

Results

Having described the methodology, we are now in a position to consider what it shows in relation to the research question. As it turns out, example 1 above is especially topically digressive. Most messages in both groups were found to be narrowly on-topic in relation to the previous discourse, although the percentages differ considerably for the two groups. Whereas over three-quarters of messages are on-topic in the pharmacy chat, this is true for only half of the social chat messages. The "off-topic" messages are mostly parallel shifts, but the social chat also has a high percentage of breaks compared to the pharmacy chat. When the distance of each message from the previous message is calculated and averaged, we find that social chat messages are more remotely related to their antecedents than pharmacy messages by a ratio of 4 to 1, resulting in lesser topical coherence and more rapid topic decay.

The teacher in the pharmacy class, as the person responsible for structuring the discourse, plays a key role in maintaining topic coherence. Through his questions and follow-up comments, he repeatedly returns the class discussion to the topics determined in advance by his lesson plan, as illustrated in the following sequence (the students' responses are consolidated to highlight the teacher's contributions):

Example 2:

58 <Teacher> Okay, so the first question is, are there any definitions that anyone did not understand?

60-66 (Students respond)

68 <Teacher> Any other problems with abbreviations?

70-84 (Students respond)

85 <Teacher> Any other problems with abbreviations?

86 <StudentB> Oh, yeah lots.

87 <Teacher> If not, then let's come up with a list of this patient's medical problems.

This pattern is reflected in the overall vertical orientation of the pharmacy chat, as represented visually for the beginning of the same course session in Figure 2. Not only does the selection of chat diagramed in Figure 2 show a strong vertical orientation, but there is a single coherent thread (the teacher's organization of the main points of the lesson outline) that runs through it and that serves as its central focus.

In comparison, all three recreational chat sessions are highly digressive. This was seen in Figure 1 above; a similar pattern is evident in Figure 3, a session in which the overall theme is 'people who live in trailers'. Note the tendency of the topical threads to branch in a horizontal direction, and the fragmented nature of the conversation, as shown by the presence of breaks.

These results support the research hypothesis. In the larger project, we consider the implications of these findings for pedagogical uses of synchronous chat, the extent to which seriousness of purpose can compensate for the tendency for chat to be digressive, and the effects of moderation on topical coherence. For our present purposes, the results have methodological implications in that they constitute a proof-of-concept of the dynamic topic analysis method, which in this study clearly and usefully distinguished between the two chat types.

Figure 2. Topical structure of a pharmacy course session

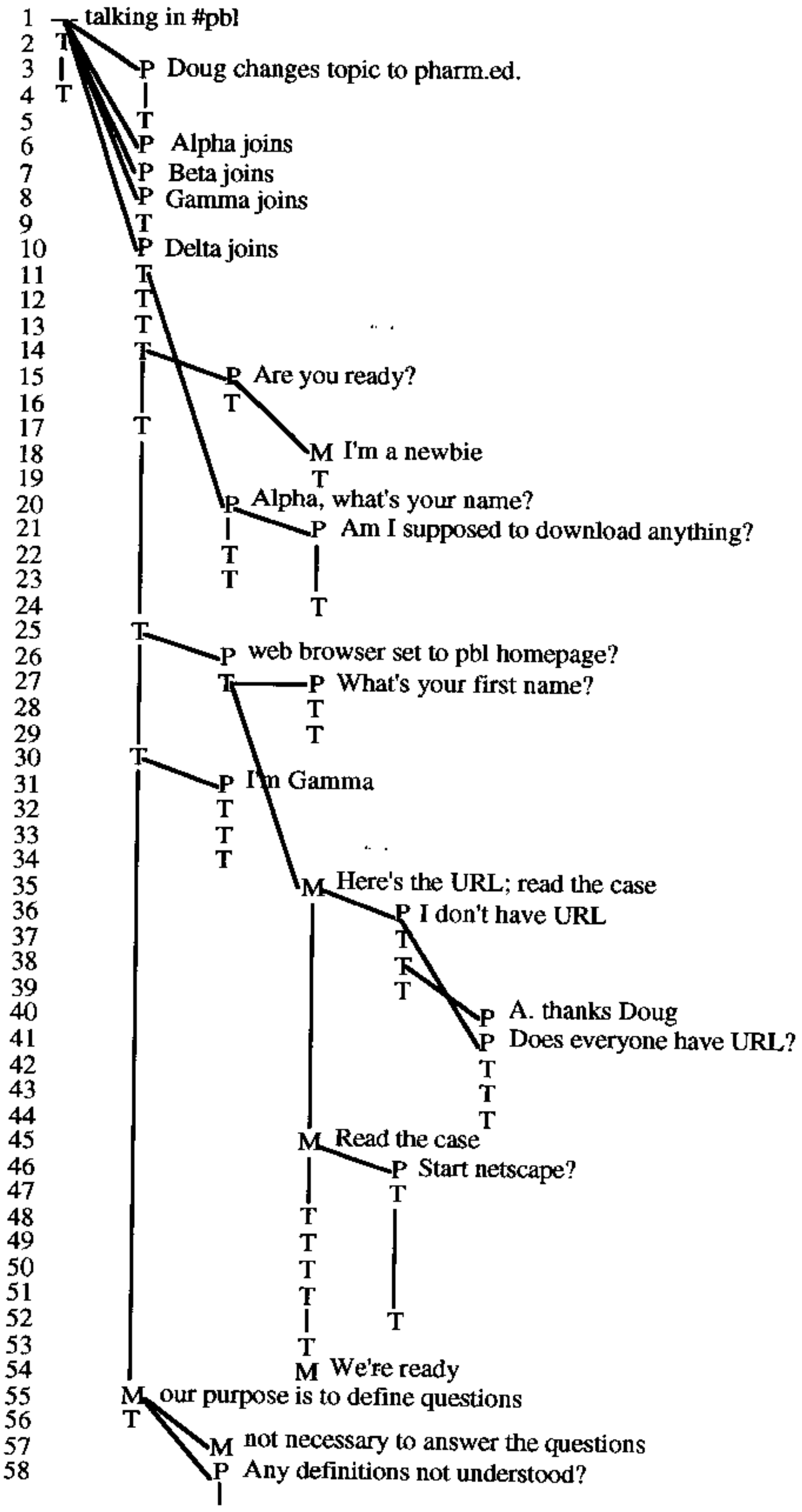
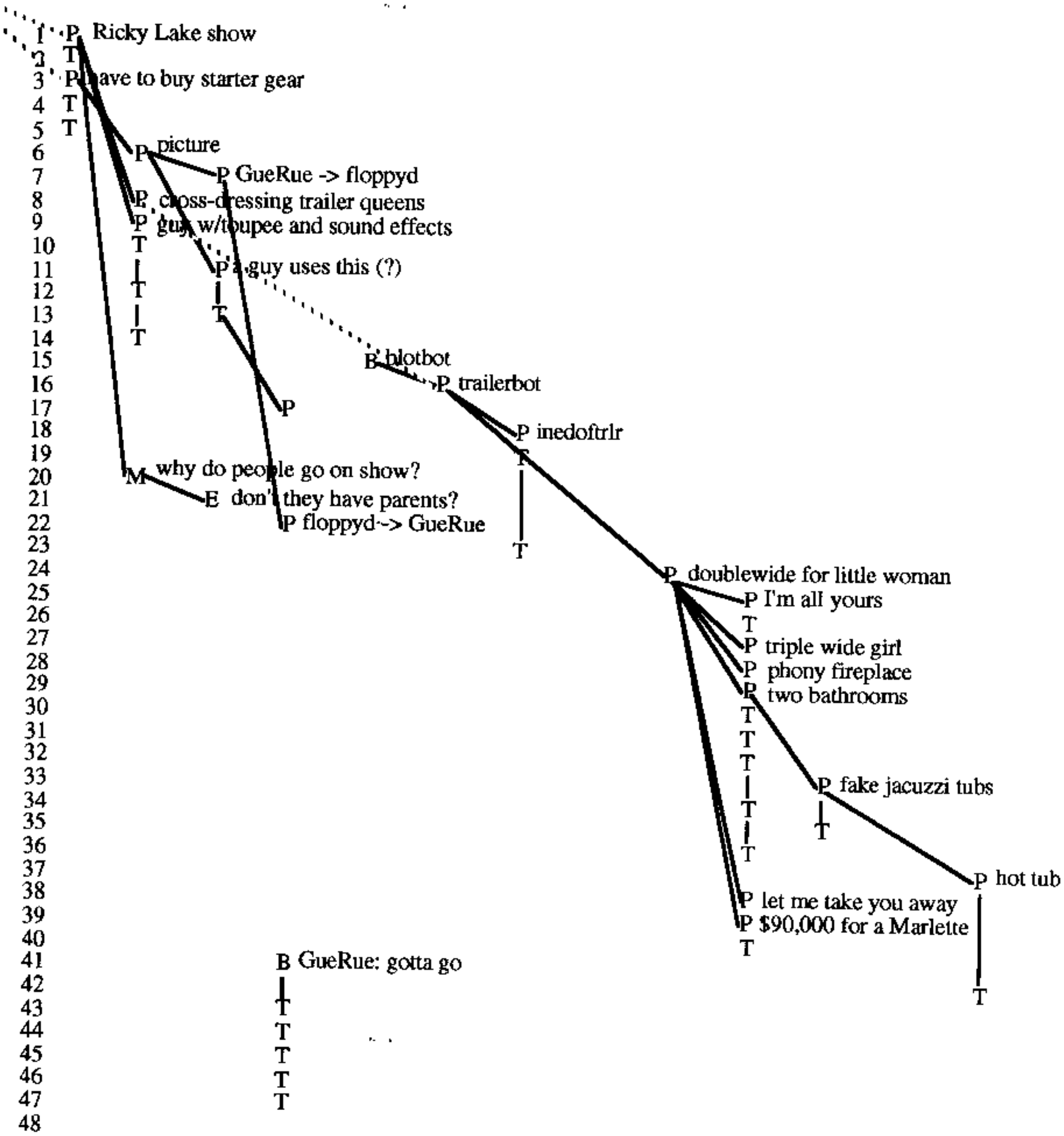

1 talking in #pbl
3 Doug changes topic to pharm.ed.
6 Alpha joins
7 Beta joins
8 Gamma joins
10 Delta joins
15 Are you ready?
18 I'm a newbie
20 Alpha, what's your name?
21 Am I supposed to download anything?
26 web browser set to pbl homepage?
27 What's your first name?
31 I'm Gamma
35 Here's the URL; read the case
36 I don't have URL
40 A. thanks Doug
41 Does everyone have URL?
45 Read the case
46 Start netscape?
54 We're ready
55 our purpose is to define questions
57 not necessary to answer the questions
58 Any definitions not understood?
T
P
M

Figure 3. Topical structure of a recreational chat session



Limitations of DTA

Although it is undoubtedly useful in contexts such as the one described above, DTA has a number of non-trivial limitations. First, the coding of semantic relations and human intentions is inherently subjective. Multiple coders are required to establish reliability. Second, the tree visual representations work best for small samples; lengthy sessions do not fit easily on a standard page or computer screen, and lengthy digressive conversations are especially potentially awkward in that they extend horizontally, which is a dispreferred direction for extended visual scanning. Thus consideration should be given to how the tree diagrams can be made to represent longer conversations in a way that is easy to view. Last, the coding of topical relations and distances between propositions works better for synchronous CMC, in which each message typically contains only one proposition, than for asynchronous CMC, in which a single message may contain many propositions. At the very least, there is a need to distinguish between intra- and inter-message relations in asynchronous CMC that DTA would need to address in order to be useful in analyzing emails, newsgroups, and discussion lists. In short, at the present time, DTA appears to be well-suited to the close, qualitative analysis of synchronous chat conversations, but there are challenges that must be overcome before the method can be as usefully extended to larger and asynchronous CMC samples.

Future Directions

Currently, efforts are being made to extend and improve DTA. In addition to overcoming the inherent limitations of the method identified above, I seek to expand the potential of dynamic topic analysis of online conversations in three domains: quantification, automation, and enhanced visual representation.

The coding in Table 1 lends itself naturally to *quantification*. For example, average distance measures could be calculated and used to compare samples, and eventually to measure the coherence of samples in isolation, once benchmark data are available for a variety of online conversation contexts (cf. Wiley, 2002). Similarly, the topical behavior of participants as individuals and in roles (such as teacher and student) could be quantified and compared to lend empirical support to informal observations that some participants shape the thread of conversation more than others and in characteristic ways.

Such measures could be *automated*, or partially automated, on the basis of coded information entered into a database or spreadsheet. Moreover, while semantic relations—and thus the coding of topic—are notoriously difficult to automate, it may be possible to partially automate DTA by automatically extracting structural information such as participant identity (indicated in chat through nicknames found at the beginning of each message) and using addressivity (nicknames of intended next "speakers", Panyametheekul & Herring, 2003; Werry, 1996) to predict the next related proposition, although the latter would need to be checked by a human coder as it could generate errors, e.g., when the intended addressee does not reply, or someone else replies instead. One solution to the automation problem is to design chat systems in which users themselves select the message to which they are responding (Smith, Cadiz & Burkhalter, 2000).

Finally, the visual representations produced by DTA could be enhanced in several respects. Computer *animation* would allow information in the tree diagrams to be highlighted selectively (for example if one wanted to view only those propositions posted by a single individual), hidden, or viewed from different angles. *Three-dimensional representation* would obviate the problem of crossed and overlapping lines, and would allow additional layers of information to be incorporated into the representations. *Color* could also be used to differentiate participants, topical threads, or indicate the degree of activity in conversations, to mention but a few possibilities.

Quantification would add rigor; automation would add speed and the ability to analyze more data; and enhanced visualizations would add information to the analytical representations. These improvements are within our grasp. At the same time, it is important to keep the nature of the data and the research questions in mind. DTA is currently a focused method of analysis designed to answer questions about discourse coherence in synchronous CMC. Its usefulness for this purpose should not be sacrificed to the desire for speed, information, and large sample sizes, unless those attributes can better enhance our understanding of topical coherence in online contexts.

Notes

[1] Others are tum adjacency analysis (Herring, 1999; Panyametheekul & Herring, 2003) and response-relevance analysis (Herring, 2000). For a meta-methodological discussion of computer-mediated discourse analysis, see Herring (2003).

[2] Approximately 50 propositions from each chat environment were coded by two researchers, with an initial inter-coder agreement rate of 65%. The coding categories were then refined, and an additional 50 messages double coded, until an 80-85% rate of agreement was reached.

References

Chafe, W. L. (1994). *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: University of Chicago Press.

Cherny, L. (1995). The modal complexity of speech events in a social MUD. *Electronic Journal of Communication*, 5 (4). Retrieved August 4, 2003 from: http://www.cios.org/www/ejc/v5n495.htm

Herring, S. C. (1999). Interactional coherence in CMC. *Journal of Computer Mediated Communication, 4* (4).

Herring, S. C. (2000). Norms of computer-mediated conversation: Whither relevance? *7th International Pragmatics Conference*, Budapest, Hungary, July 9, 2000.

Herring, S. C. (2003). Computer-mediated discourse analysis: An approach to researching online behavior. In S. A Barab, R. Kling, & J. H. Gray (Eds.), *Designing for Virtual Communities in the Service of Learning*. New York: Cambridge University Press.

Herring, S. C., & Nix, C. G. (1997). Is 'serious chat' an oxymoron? Pedagogical vs. social uses of Internet Relay Chat. *American Association of Applied Linguistics (AAAL) Conference*, Orlando, FL, March 11, 1997.

Hobbs, J. 1990. Topic drift. In B. Dorval (ed.), *Conversational Organization and its Development,* 3-22. Norwood, NJ: Ablex.

Jefferson, G. (1988). On stepwise transition from talk about a trouble to inappropriately positioned matters. In Atkinson J. M. & Heritage J. (eds.) *Structures of social action: Studies in conversation analysis*. Cambridge and New York: Cambridge University Press. Pp 191-222.

McLaughlin, M. (1984). *Conversation: How Talk is Organized*. Sage.

Panyametheekul, S., & Herring, S. C. (2003). Gender and turn allocation in a Thai chat room. *Journal of Computer Mediated Communication, 9* (1).

Sacks, H. (1992). *Lectures on Conversation*, Volumes I & II. (G. Jefferson, ed.). Oxford, UK, Cambridge, USA: Blackwell.

Schiffrin, D. (1994). *Approaches to Discourse*. Oxford University Press.

Smith, M. A, Cadiz, J. J., & Burkhalter, B. (2000). Conversation trees and threaded chats. *CSCW* 2000.

Werry, C. (1996). Linguistic and interactional features of Internet Relay Chat. In S. C. Herring (ed.), *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*, 47-64. Amsterdam: John Benjamins.

Wiley, D. A. (2002). A proposed measure of discussion activity in threaded discussion spaces. Retrieved August 6, 2003 from: http://wiley.ed.usu.edu/docs/discussion09.pdf