A Noun Phrase Analysis Tool for Mining Online Community Conversations

Caroline Haythornthwaite and Anatoliy Gruzd

University of Illinois at Urbana-Champaign, USA

1. Introduction

Online communities are creating a growing legacy of texts in online bulletin board postings, chat, blogs, etc. These texts record conversation, knowledge exchange, and variation in focus as groups grow, mature, and decline; they represent a rich history of group interaction and an opportunity to explore the purpose and development of online communities. However, the quantity of data created by these communities is vast, and to address their processes in a timely manner requires automated processes. This raises questions about how to conduct automated analyses, and what can we gain from them: Can we gain an idea of community interests, priorities, and operation from automated examinations of texts of postings and patterns of posting behavior? Can we mine stored texts to discover patterns of language and interaction that characterize a community?

This paper presents a prototype tool for on-the-fly analysis of online conversations using text mining techniques, specifically noun and noun phrase analysis, to discover meaningful techniques for summarizing online conversational content. To pursue this work, we use as a test case data from a corpus of bulletin board postings from eight iterations of a graduate class for students earning an online graduate degree in library and information science (LIS). Although we aim to analyze this environment using the techniques outlined here, work to date has concentrated on building the application environment and applying the methodology of noun phrase

analysis to community conversations. Thus, this paper addresses methodology as much as our preliminary findings.

The underlying assumption in this kind of analysis is that language can reveal characteristics of community. This follows on investigations of online language and community by a number of researchers, including Cherny's examination of the role of chat in supporting an online community (1999), Herring's work on gender and CMC language (1996, 2000, 2003), and Crystal's on language and the Internet (2001). Language plays an important role in creating community out of design (Stuckey & Barab, forthcoming), bootstrapping and reinforcing norms of behavior and of community (McLaughlin et al, 1995; DeSanctis & Poole, 1994; Hearne & Nielsen, 2004). Analyzing conversations provides a way into these communities, discovering what structures and supports the community and how conversational genres are constructed. Recent work on analyzing persistent conversations includes analyses of blogs by Herring, Scheidt, Kouper & Wright (in press), chat by Herring (2003), and bulletin boards from online classes by Fahy and colleagues (Fahy, 2003; Fahy, Crawfor & Ally, 2001).

Traditionally, researchers use a descriptive type of analysis to study online communities. However, due to its ad hoc nature, such analysis is primary done manually. This makes assessment very time consuming and often cost prohibitive. Furthermore, as Hmelo-Silver (2006) has pointed out, descriptive analysis by itself does not accurately reflect the overall picture of the entire community and its processes. To overcome these limitations, researchers are starting to utilize fully- or semi-automated approaches to analyze online discourse (e.g. Dönmez et al, 2005). However, researchers in the field have yet to reach a consensus as to what methods of automatic content analysis to use, and more importantly, how these methods could or should be used to analyze online collaborative communication. Most of the current approaches apply a human-expert heuristic to build production rules. These production rules are then used to find certain patterns (speech acts) in the raw text which are believed to correspond to certain processes or categories of behavior (e.g. motivation, learning, conflict resolution). For example, a production rule 'to find all instances where a community member is seeking information' can be constructed by selecting sentences ending with question marks. Although such an *automatic categorical coding* may allow a significant reduction in work time, it has its limitations. Production rules are normally constructed and validated on a single textual corpus, making transfer inappropriate and ineffective for another corpus of data. This is especially true for corpora of different genres.

In our research, we are exploring low-cost time-efficient techniques for on-the-fly, genre-independent, content analysis of online communities. Specifically, we are using noun phrases extracted from the text to explore and visualize communal processes found in community corpora.

2. Noun-Phrase Extraction Method

The approach to automatic analysis presented here is content analysis (Krippendorf, 2004; Weber, 1985), with text analyzed at the noun and noun phrase level. Since the early 80s, the noun phrase extraction method has been successfully used in various applications including back-of-the-book indexing (e.g. Salton, 1988), document indexing for information retrieval purposes (e.g. Fagan, 1989; Zhai, 1997), and, most recently, text visualization and text summarization (e.g. Boguraev et al, 1998, 1999). It is different from other light-weight types of content analysis that often use only single words as the unit of analysis. Unlike single words, noun phrases allow researchers to disambiguate the meaning of component words. For example, a common word like *information* can be used in many different contexts (e.g. 'travel information', 'information center'); thus, it is not useful as a stand-alone single term. However, when found and extracted as a phrase, e.g., 'information science' the technique provides a better understanding of meaningful terminology found within the corpus, and used within the community.

There are several significant benefits of using noun phrases while studying online communities. First, nouns and noun phrases tend to be the most informative elements of any sentence. As a result, they are the most ideal candidates for providing a very "concise, coherent, and useful representation of the core information content of a text" (Boguraev & Kennedy, 1999, p. 20). Second, using state-of-the-art computational linguistic statistical parsers, we can effectively extract meaningful noun phrases across different genres (e.g. chat, bullet boards) and even different domains (e.g. biology, math). This generally means that once implemented, a noun phrase extractor requires very little or no human interventions to be applied to different datasets. And finally, it is now possible to parse on-the-fly semantic structures in large-size corpora. A demonstration of such a powerful parser is Yahoo! term extractor (http://developer.yahoo.com), a free online service for extracting meaningful terms from documents submitted by users.

3. Corpus

The corpus used in the current analyses consists of bulletin board postings from eight iterations of the same online class, given by the same instructor from 2001 to 2004. The bulletin boards examined are those public to all class members (password protected and so not open to the general public; see below regarding permissions for use of these postings). Classes last 15 weeks. During that time the 31 to 54 class members, the professor, and 3-4 teaching assistants posted 1200 to 2100 public messages per class. Along with the public boards, students posted 2-3000 messages a semester in bulletin boards set aside for smaller student work groups. For research purposes, these are not part of the dataset examined here, but are examples of places where participants might want to make use of the tool presented here. Beyond bulletin boards, the community is also maintained via other online means, including email, and online chat during live weekly class sessions. Future work will include examination of chat logs public to the class, but there are no plans to include email.

This corpus provides an ideal test environment because, as an environment familiar to the authors, we can use our knowledge of the environment to evaluate results. Moreover, although not explored in depth here, a primary reason for examining this environment is a particular interest in the practical issue of understanding learning processes in these online classes, and exploring ways of providing information back to instructors and students to help them make sense of the vast number of postings created per semester.

This is a learning community, and hence the postings include discussion about what to do, what things mean, how to go about work, as well as information on the topics at hand, and social interaction. All students are relative novices when they begin, both in the subject matter and in taking classes online; however, many students have or are spending time in LIS environments as employees or assistants and thus are not all naïve about the topic. This environment is known as supportive pedagogically, technically, and socially. Thus, another question is whether this can be seen in the text of postings, and whether local, supportive phrases can be identified that might prove useful for other sites for indicating support behaviors. Such an indicator might include a list of words and phrases that can be used for a quick look at the overall tone of online interaction.

Class size and posting activity in class-wide bulletin boards

	No. students	No. of instructors / TAs	No. of unique msgs
2001A	33	5	1205
2001B	42	5	1581
2002A	39	4	1469
2002B	46	4	1895
2003A	52	4	1280
2003B	54	4	1242
2004A	31	4	1493
2004B	34	4	2157

3.1 Permissions

At the beginning of each class, students were alerted that postings public to the class would be made available for analysis. Permission was obtained from the researchers' Institutional Review Board before the first alert was posted. The alert was posted in each class, each year from 2001 to 2004. Part of that text read:

> "This message is to alert you that transcripts of the [this course] class chat (main room only, not including whispers or chat in other rooms), and postings to [class] webboards (whole class webboard, not sub-group webboards) will be examined for research on how [program] students learn to communicate online. The only transcripts being examined are those that are already recorded as part of [program] class records... In our research, we are interested in trends in expression via chat and webboards in [this class], the first course most students take at a distance, and how students learn to interact online."

Students were given contact information for the lead researcher. If students did not want their text quoted in any way they were asked to contact the researcher. No requests have been received.

4. Data Processing

For the analyses presented here, data from the bulletin boards were retrieved as one file and pre-processed to create the workable dataset. Future work will address connecting this tool directly to ongoing discussions for actual use in classes as they are in progress. Part of the work conducted here is to discover what kinds of pre-processing will be necessary to create a workable dataset on the fly.

As the first step in pre-processing, individual postings were separated from each other and into different meaningful fields such as *subject, email, date, message*. This was accomplished with a Python script that uses

regular expressions to locate and extract the needed text information. Additionally, this script was used to *remove the inclusion of quotes from reply-messages*. Thus, the text analyzed excludes the repeated text often included in replies to other's posts. In the bulletin board system used such replies are indicated by a starting colon.

The resulting data were then imported into a MySQL database of three tables. One table includes information of each individual posting; a second table includes a list of all academic classes for which were imported into the database; and finally, to uncover potential coding categories to be used in content analysis, a third table includes noun phrases extracted from every posting.

4.1 Noun Phrase Extractor

In order to extract noun phrases from postings, the Natural Language ToolKit (NLTK), developed at the University of Pennsylvania (freely available from http://nltk.sourceforge.net) was used. This toolkit is a set of Python modules for symbolic and statistical Natural Language Processing (NLP).

Conventional NLP consists of six main steps (Liddy, 1998): morphological, lexical, syntactic, semantic, discourse, and pragmatic analysis. To identify noun phrases, only lexical and syntactic analysis are usually required. Lexical analysis is used to assign word classes (e.g. noun, verb, adjective) to words in a sentence; syntactic analysis is used to uncover grammatical structures of sentences (e.g. noun phrases).

For the first step (lexical analysis), we used a *probabilistic tagger* trained on a subset of the Penn Treebank corpus (http://www.cis.upenn.edu/~treebank) that consists of 99 articles from the Wall Street Journal. Each of the words in this corpus has pre-assigned lexical information known as part-of-speech tags. For each unique word in the corpus, the tagger used the manually assigned tags to calculate the probability that a word belonged to a specific part of speech. Then, using the values of these probabilities, the tagger assigned a part-of-speech tag with the highest probability to each word in our raw data. A sample output of this step is presented below.

Part-of-Speech Analyzer

Original Text:

One of the things that keeps hitting me about these recents [misspelled] events is how much virtually instantaneous communication has impacted the level of knowledge by the public.

Tagged Text:

<One/CD>, <of/IN>, <the/DT>, <things/NNS>, <that/WDT>, <keeps/VBZ>, <hitting/VBG>, <me/PRP>, <about/IN>, <these/DT>, <recents/NN>, <events/NNS>, <is/VBZ>, <how/WRB>, <much/JJ>, <virtually/RB>, <instantaneous/JJ>, <communication/NN>, <has/VBZ>, <impacted/VBN>, <the/DT>, <level/NN>, <of/IN>, <knowledge/NN>, <by/IN>, <the/DT>, <public/NN>

For the second step (syntactic analysis), we used our own syntactic rule [<JJ.*>*<NN.*>+] to identify and extract meaningful noun phrases, where 'JJ' stands for 'adjective' and 'NN' for 'noun'. This rule recognizes as a phrase any sequence of words consisting of zero or more adjectives and at least one noun (e.g. 'school', 'reference librarian' or 'mental models').

The quality of extracted noun phrases is directly dependent on the accuracy of the part-of-speech tagger used in step two above. In turn, the tagging accuracy is heavily dependent on a particular implementation of the tagger. In general, probabilistic taggers yield a relatively high (over 95%) level of tagging accuracy (see, for example, Brants, 2000; Schmid, 1994). However, due to the exploratory nature of our study, we did not formally measure the accuracy level of the probabilistic tagger that was used. Despite this fact, our initial examination of the data output suggested that the results were highly reliable. We discovered that the most common errors were not due to the weaknesses in the tagging algorithm, but were instead a limitation of the tagged corpus. For instance, words commonly used in online exchanges such as emoticons and acronyms (e.g. lol = "laughing out loud") were not present in the tagged corpus; as a result, they were not interpreted correctly by our tagger. Taking this fact into consideration, we intentionally decided to mark all words unknown to the tagger as *nouns*. This design decision itself had its own drawback. For instance, when the tagger encountered misspelled words such as *recents* (see the example above) or words with contractions such as *don't*, the tagger automatically marked these unknown words as *nouns*. Nevertheless, despite this weakness, we decided to keep this design decision to ensure that potentially important but unrecognizable nouns were not missed. And by chance, this decision was later proven to be very useful for other reasons (see below in the section 'Community Style').

5. The Application Environment

The Internet Community Text Analyzer (ICTA) text mining tool is operationalized through a web-based environment that facilitates searching the stored versions of the text from the eight classes. ICTA represents a prototype for automatic inquiry of ongoing processes in online classes. The main screen of this tool provides the user with a means to select the class and bulletin board to be analyzed, performs that analysis, and returns the top 100 noun phrases found in the selection with their frequency counts (ordered by the frequency counts). A tag cloud gives an immediate visual representation of the relative importance of particular words (see figure 1).

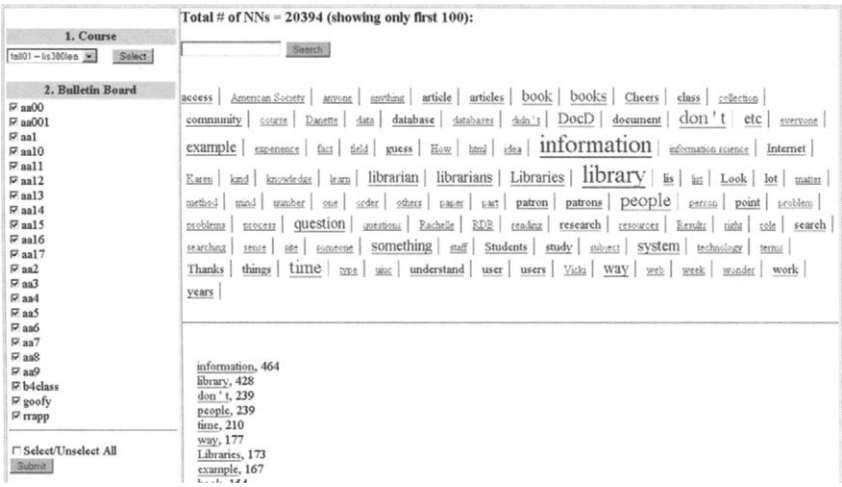

Figure 1: ICTA tool: Main screen. showing selection of course and bulletin boards with returned top 100 nouns/noun phrases as tag cloud and list with frequency of occurrence.

Clicking on any noun phrase from the initial list of 100 returns a list of the words in context, giving the phrase of interest and the 50 preceding and following symbols (see FIGURE 2). Also included for each use of the noun phrase in context is a set of nouns and noun phrases automatically extracted from the corresponding posting, the ID of the poster and the bulletin board the phrase appeared in. Clicking on the ID of the bulletin board returns the full posting, with instances of the selected noun phrase highlighted. Also presented in this display are the number of unique messages in which the phrase appears, and a list of the IDs of posters with the number of unique messages in which they used the phrase (the latter is not shown in the figure below to maintain confidentiality).

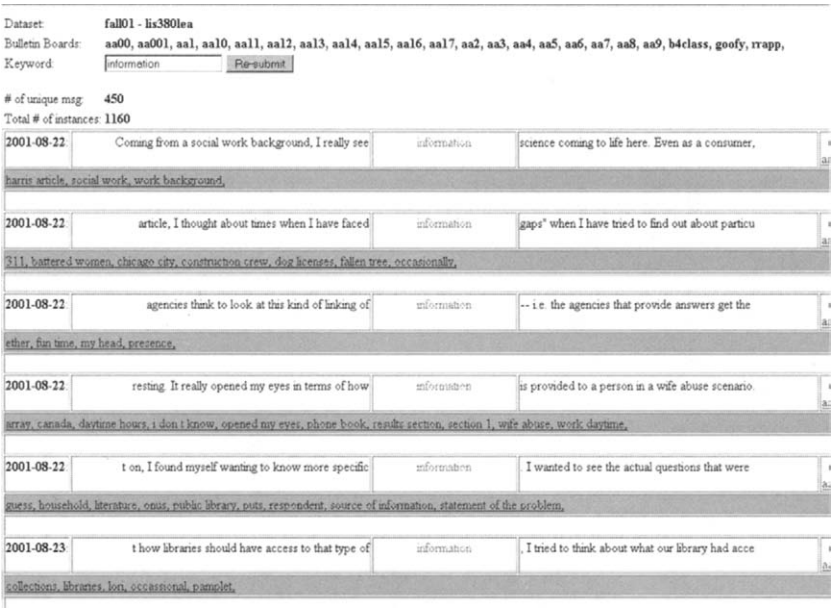

Figure 2: ICTA tool: Noun/Noun Phrase in context: "Information"
Note: Bulletin board and identification of the poster appear at the right, but this is intentionally hidden here to retain anonymity.

As an alternative approach, the user may also search by individual noun phrase, instead of performing the top 100 analysis. In this case, the user types the phrase into a text box, and the system returns details in the same way as clicking on a noun phrase in the top 100 list.

6. Analyzing Word Use and the Online Community

The following presents several exploratory analyses conducted using the ICTA tool and the results it generates. First, the overall structure of communication by bulletin board and by class is presented. This provides insight into interaction patterns over the semester and the degree of activity on the boards. Second, the most frequently occurring nouns and noun phrases are evaluated for what they reveal about the community. Third, words of particular significance to the community are taken as examples and the use of these words examined across the semester.

The natural language processing returned lists of the 100 top words for each class. Data from four classes, one each in 2001 to 2004, were examined in detail. The following refers to these four classes only.

The top 75 words are given in the table below. These are ordered as they appear in the list starting with the top 50 words for 2001, adding in words in the top 50 from subsequent years. (Words that appears multiple times are listed only once.) The first thing to note is that while the NLP extraction process was tasked to look for both nouns and noun phrases, only single nouns appear in the list of the top 50; noun phrases such as "information science" appear but not in the 50 most frequently used words. Also due to our intentionally broad definition of noun phrases (see the previous section), there are a few false-positive results in the top 75 words list (e.g. *don't*). These false-positive results are words that are undefined in the part-of-speech category, marked by a parser as 'None'. [0]Because of the exploratory nature of our study, we decided at the time to include these false-positive words for further analysis. As you will see in the following section, this turned out to be a good decision; these "none"-noun terms allowed us to make couple interesting observations.

The top 75 words by occurrence

library/ies; information; book/s; librarian/s; user/s; patron/s; don't; people; question/s; article/s; time; database/s; way; example; something; system; study; class; thanks; things; students; lot; Internet; understand; search; years; community; work; access; research; lis; guess; document/s; point; cheers; look; course; how; part; sense; fact; problem; technology; web; someone; others; order; subject; paper; experience; idea; type; reading; person; resources; knowledge; museums; site; learn; place; case; need; materials; evaluation; data; results; html; service/s; process; topic; collection; kind; method; journal; list

Analysis also revealed the need to compensate for the following: proper names often present when some participants sign their posts while others do not; journal names; partial words, particularly prevalent due to spelling errors or shorthands; and link addresses. General methods for dealing with names, partial words and link addresses will be generalizable to other environments, although journal names may be specific to an academic discussion.

A primary question is whether these words reveal characteristics of the community. Knowing the environment, we are able to shed light on the significance of these words and their usefulness to someone from the community seeking to assess activity and discussion. These words reveal several aspects of the community and include words associated with the profession, learning in an academic setting, and the supportive character of the community:

Profession. Not surprising for a community with a focus on library and information science, the top words identified are those associated with the

profession: library/libraries, information, book/s, librarian/s; as well as words concerned with those who make use of libraries: user/s, and patron/s, people, as well as community. Also evident are topics of particular importance for the field and for the student: database/s, search, document/s.

Learning. Other top words are associated with the learning and student orientation of the community: question/s, article/s, example/s, way, study, class, course, research, journal, reading, method, results.

Interaction. A few words tell us more about the character of this particular community. For example, the word "don't" appears high in the list for all classes, as does the word "thanks". Both indicate a way of phrasing and approaching interaction with others, e.g., the use of "don't" shows a deference to one's knowledge and a reluctance to declare an opinion. The high use of the word "Thanks" shows a supportive environment, concerned with interacting positively with others. This is examined further below by looking also at the use of "agree" and "disagree" in these classes.

These classes are primarily composed of women. Although there is no comparison group of classes of men, it is possible that this style of interaction is a result of women's ways of communicating. Gender greatly impacts communication style, particularly in avoiding making assertions. The kind of language seems to follow how Herring described the female-gendered style of online communication:

> "The female-gendered style, in contrast, has two aspects which typically co-occur: supportiveness and attenuation. 'Supportiveness' is characterized by expressions of appreciation, thanking, and community-building activities that make other participants feel accepted and welcome. 'Attenuation' includes hedging and expressing doubt, apologizing, asking questions, and contributing ideas in the form of suggestions." (Herring, 1994, no page)

Change over time. Also telling for the community are the way in top words change over time, following changes in emphasis in the program. Knowing the community it is possible to understand how "museums" comes in higher in the order for later years; this is the case as the ideas of museum informatics are being adopted as an area of research and teaching at this particular institution. The term "web" also appears to give way to "Internet", following general trends in word usage. Changes in use of other words are discussed below.

The next sections look at a few specific words/phrases to explore this community's use of words. In what follows, the number of unique messages containing the word(s) is used as an indicator of quantity of use because it provides the least ambiguous interpretation of use given the potential for multiple uses of words.

6.1 Important Topics: Databases, Books

Two aspects of LIS are highly important, and at times in competition: the digital world of databases and online resources, and the world of books. Thus, it is interesting to see how these compare in the classes over time.

The topic of databases is of particular relevance and importance to the field of LIS and to the students in the class. Many are learning about database structures for the first time, as well as learning to use databases, both those generally available and those specifically relevant to an LIS career. Along with the words "database" and "databases" another term that appears in the list of top words (although not the top 50) is "RDB" for relational database. Examining the overall use of these three terms shows that at a rough estimate the terms appear in 14-23% of messages (searches are not case sensitive).

It is surprising to find that the use of these words actually *declines over time* in terms of the percent of messages in the semester containing at least one occurrence of the word. At first this does not seem appropriate given the importance of the term; however, the percentage reduction in use may reflect the increased familiarity with databases that students have in general as these become more part of the overall curriculum. Further research is needed to understand this changing pattern.

Number of unique messages using RDB, database, databases

	RDB	% msgs with RDB	data-base	% msgs with 'data-base'	data-bases	% msgs with 'da-tabases'	Total % *
2001	61	5.06	135	11.20	85	7.05	23.32
2002	56	3.81	141	9.60	92	6.26	19.67
2003	34	2.66	106	8.28	65	5.08	16.02
2004	48	3.22	109	7.30	58	3.88	14.40

*Note: Combining these numbers could overestimate the percentage use where more than one of these terms appears in the same posting.

"Book" or "books" also appear in 14-23% of postings. This varies across years, although not in a linear pattern. Use is noticeably lower in 2002 and 2004. Again, this pattern is hard to interpret. It may be related to the particular topics for each semester. Again further research is needed to interpret the year to year pattern.

Number of unique messages using book, books

	book	% with 'book'	books	% with 'books'	Total %
2001	146	12.12	139	11.54	23.65
2002	108	7.35	98	6.67	14.02
2003	128	10.00	131	10.23	20.23
2004	112	7.50	126	8.44	15.94

What is evident is that both of these words are highly used in these classes. With some variation there appears to be a persistent balance in use, with approximately 23% of messages containing use of each in 2001, and 15% in 2004. Thus, we can see that despite potential change in the relevance and place of databases versus books in LIS, for this community the two have remained relatively in equilibrium.

6.2 Community Style: Don't Think, Don't Know, Don't Have

Although the intention was to capture nouns and noun phrases, as explained above, misspelled and unanticipated words were included in the noun lists. There was an unanticipated benefit to this in that certain words showed up with such frequency that they seemed worth examining in more detail for their significance to the community. This was the case for "don't" which appeared so highly in the list of top words for all years that its use was examined in context. Phrases detected as most common are "don't think", "don't know", and "don't have", appearing in 9 to 15% of messages. Although it is not possible to say with certainty why this word and these phrases occur, they suggest a hesitancy and deference both about opinions and about personal knowledge. Our speculation is that this is a key attribute of this community, one largely composed of women, known to be supportive and non-confrontational, and fully composed of individuals who are unsure of their knowledge base. It remains to future research to see if the word has significance in other communities and to explore more specifically its meaning in this one.

Number of unique messages using "Don't *"

	Total no. msgs	don't think	don't know	don't have	% msgs*
2001	1205	51	31	33	10.48
2002	1469	35	38	23	15.30
2003	1280	49	50	43	9.01
2004	1493	32	27	41	14.93

* Note: % msgs may overestimate the use of "don't" as it is based on the presence of any of the three phrases, and hence may count a message 1-3 times.

6.3 Community Interaction and Support: Agree/Disagree, Thanks

Pursuing further our investigation of the friendliness and supportiveness of the community, we asked: What evidence is there in the words used that might allow identification of that attribute from an outside perspective? To explore this, we used the ICTA tool to compare the occurrence of the words "**agree**" and "**disagree**" to see the balance of these two sentiments. As well, given that "**thanks**" appears in the top words used overall, the occurrence of this word was also examined. Each instance of "agree" and "disagree" was examined in context to ensure that the meaning was correctly interpreted. Some uses of the words refer to opinions on readings, and some to other's postings. The tabulation for agree given below includes 'totally agree', 'completely agree', and 'couldn't agree more'. Disagree includes 'don't agree', 'do not agree', 'don't necessarily agree', 'not agree', 'not * agree', and 'don't * agree'.

As can be seen, the number of postings in which agreement is expressed greatly outweighs the disagreements: 5 to 12% of messages include agreement, compared to less than 1% expressing disagreement. This seems good evidence for the presence of supportive relations in this community. The same can be said about "Thanks", which appears in 7 to 18% of messages.

It is interesting to note that as the percentage of messages containing agreement decrease over time, the use of thanks increases. This may represent ways in which particular forms of support are expressed in these classes. In the 2004 class, with over 17% of messages including the word "Thanks", but only 5% expressing agreement (at least in the specific use of the word "agree"), it seems likely that these indicate a local class pattern about choice of means of expressing support. Different words may appear based on local usage, perhaps one word bootstrapping its further use. This is something in need of further examination.

Number of unique messages agreeing or disagreeing

	Agree	% msgs	Disagree	% msgs
2001	140	11.62	11	0.91
2002	127	8.65	7	0.48
2003	95	7.42	5	0.39
2004	80	5.36	11	0.74

Number of unique messages containing 'T/thanks'

	Thanks	% of messages
2001	106	8.80
2002	109	7.42
2003	146	11.41
2004	261	17.48

6.4 CMC language

One further aspect of community interaction of interest is the use of online language, such as emoticons or acronyms. In the early years of this program, students were primarily new to online communication. This is not so true currently, but many are still picking up consistent use of CMC language for the first time. What is surprising is how little use there is of emoticons and paralanguage. A search for "lol" for example (for 'laughing out loud') revealed no use in the bulletin boards although reports from students had suggested this was in common use. This may represent the difference between chat interaction and bulletin board postings – two different genres of online conversation, the former much more like conversation, and the latter more like memos or homework. (Examining chat logs for these classes is a future project).

There may also be more idiosyncratic use that is not revealed by searching on known emoticons or paralanguage. For example, Barrett, LaPointe & Greysen (2004) discuss how a new local smilie emerged in an online class following discussion of Thanksgiving holiday and pie. The combination **___** was used and came to mean 'eating too much pie.' This is true also for local word combinations. Students in a few of these years came to use the term "brass monkey" to signify an inadvertent leaking of whispers into chat conversation. The term emerged from a whispered discussion of pubs. However, the term faded over time as new students coming into the program did not know its connotations, and thus did not reinforce its use (Anna Nielsen, personal communication).

Continuing the notion of agreement and disagreement, the bulletin boards were examined for the use of **smilies** and **frownies**. There is also the convention of 'nose' / 'no-nose' to examine in these emoticons, i.e., whether or not an intervening hyphen is used between the colon eyes, and the parenthesis mouth. As can be seen in the table below, these classes overwhelmingly adopt the no-nose smilie, and they very rarely use frownies. This use of smilies for positive over negative affect continues the language use as demonstrated above for agree and disagree – again, this

community, perhaps because composed of women, perhaps because composed of future librarians – exhibits few negative sentiments.

Number of unique messages using these smilies and frownies

	:)	% of msgs for :)	:-)	:(	:-(	% msgs
2001	59	4.90	1	1	1	5.15
2002	56	3.81	14	1	1	4.90
2003	62	4.84	11	0	1	5.78
2004	83	5.56	3	0	0	5.76

Overall about 5-6% of messages contain one of these four emoticons. It is difficult to say if this is a lot or a little use. One other estimate of a similar environment found 10% of posts using emoticons (a sample of 5626 postings in 12 online graduate classes; Sixl-Daniell & Williams, 2005). In comparison, use of emoticons in these classes appears low, but again may indicate differences in genre between the bulletin boards examined here, and chat communication. These classes have a set time each week when they can converse synchronously. It may be that the emotional exchange symbolized by the emoticon characters is handled synchronously, thus allowing a different in genres to emerge. However, this is only speculation at present while the chat logs have not been mined.

7. Future work

Our preliminary work has shown that the idea of exploring community through analysis of word usage opens up many possibilities for analysis and discovery, but the work is at a very preliminary stage. We have ideas on improving our own tool and the methods and corpus on which it is based. Here are a few areas of future work we are considering.

Disambiguation. Originally we used *tag clouds* as a way just to visualize the communal processes occurring in corpora. However, they have proven to be very useful also as a technique to 1) build a concise and coherent representation of online community, and 2) provide entry points to explore the community in greater details. For example, one can quickly grasp important issues in a community by just simply skimming terms in its tag cloud. To investigate different contexts in which a concept has been used by the community members, a researcher can just click on a corresponding tag. This will take him or her to a list of corresponding entries in corpora where the concept was used. To further increase the readability of

a *tag cloud* representation, there needs to be a quick way to disambiguate single-word terms that often appear there. For example, the word 'library/ies' found in Fall 2001 dataset has 1543 occurrences. To identify all possible usages of this word, we would have to examine all 1543 occurrences. To avoid such a time consuming procedure, we are proposing to display the most frequent noun phrases that include this word. When a researcher hovers a mouse over a single-word term in a *tag cloud*, a pop up window will show noun phrases and their frequency counts. For example, for 'library/ies', there will be 'public library/ies' (116 occurrences), 'digital library/ies' (85), 'library system/s' (19), library school (14), etc.

Verbs and Verb Phrases. Another area that needs further investigation is the role of verbs and verbs phrases in the content analysis in online learning communities. In our exploratory study, we found some verbs and verb phrases can reveal many important characteristics of interaction between class members as well as the structure of students' arguments (e.g. *agree*, *disagree*). In the future version of our research tool, we are considering adding the capability of extracting and visualizing verb phrases from corpora. This will help us to determine what specific verbs or groups of verbs are the most useful for content analysis.

Clustering Algorithms. To further speed up the analysis of corpora, we are considering various clustering algorithms to group together related nouns and noun phrases. An example of related terms could be 'reference services', 'reference librarian' and 'reference interview'. Once grouped together, these terms create a subject category or topic that can be used as a new unit of analysis. Thus, instead of evaluating all unique noun and noun phrases extracted from corpora, we will only need to focus on a relatively smaller number of topics. There are also other benefits of having terms grouped together. For instance, we will be able to study the emergence and evolution of different topics over time. This type of analysis usually referred to as *temporal text mining* (Mei & Zhai, 2005). A sample question that can be answered through this analysis is whether or not topics appearing in clusters correspond to weekly topics initially assigned by the instructor in the syllabus. Another possible use of clusters is to identify experts among community members in different subject areas. Somewhat related work has been done by researchers studying online bookmarking communities (e.g. Wu et al 2006).

CMC Corpus. Our work shows that another problem to be addressed is the lack of a *tagged corpus of computer-mediated communication.* Because of this, our part-of-speech tagger could not recognize some of the newer, commonly used online words and expressions. To effectively apply NLP techniques to Internet specific discourse, we need to develop more

sophisticated language models that would "understand" the language used by members of online communities. Unfortunately, there is still a lot of work that needs to be done in this area of research (Ooi, 2000). Therefore, we see a need to collect and manually annotate a corpus consisting of documents from a variety of online genres and domains presented in computer-mediated communication as part of future work.

8. Conclusion

This paper has presented an exploratory study using a prototype tool for mining communal conversations, including descriptions of the noun phrase extraction methodology, and a first look at the bulletin board posting behavior for one community. Results show that content analysis at the noun phrase level selects words that can be identified as important for this community, relating both to the profession and the status of participants as learners. We are also able to find results that confirm our understanding of the environment regarding interaction style. The high use of words such as "agree" and "thanks", and of smilies ":)", combined with the low use of words such as "disagree" and frownies ":(" concurs with our understanding of this as a supportive environment. Although preliminary at this point, we take this result as supportive of our efforts to use text mining techniques to identify major interests and character of online communities. Our work continues with more refinement of the text mining algorithms used to evaluate this kind of text, and looks to future work for application of this technique in practice for members of this online community, and to examining other online communities.

9. References

Barrett, K. LaPointe, D. & Greysen, K. (Jan. 2004). *Speak2Me: Using synchronous audio for ESL teaching in Taiwan.* Report R28/0401, Athabasca University, Centre For Distance Education.

Boguraev, B. & Kenned, C. (1999). "Applications of term identification technology: domain description and content characterization", *Natural Language Engineering* 5(1): 17–44.

Boguraev, B., Wong, Y. Y., Kennedy, C., Bellamy, R., Brawer, S., and Swartz, J. (1998). *Dynamic presentation of document content for rapid on-line browsing. AAAI Spring Symposium on Intelligent Text Summarization,* Stanford, CA. 118-128.

Brants, T. (2000). "TnT: A statistical part-of-speech tagger", in *Proceedings of the 6th Conference on Applied Natural Language Processing* (Seattle, WA), pp. 224-231.

Cherny, L. (1999). *Conversation and community: Chat in a virtual world.* Stanford, CA: CSLI Publications.

Crystal, D. (2001). *Language and the Internet.* Cambridge, UK: Cambridge University Press.

DeSanctis, G. & Poole, M. S. (1994). "Capturing the complexity in advanced technology use: Adaptive structuration theory", *Org. Science, 5*(2), 121-47.

Dönmez, P., Rosé, C., Stegmann, K., Weinberger, A. & Fischer, F. (2005). "Supporting CSCL with automatic corpus analysis technology", *CSCL '05: Proceedings of Th 2005 Conference on Computer Support for Collaborative Learning,* Taipei, Taiwan. 125-134.

Erickson, T. Herring, S. & Sack, W. (2002). *Discourse Architectures: Designing and Visualizing Computer-Mediated Communication.* Workshop at the CHI 2002 Conference, Minneapolis, MN.

Fagan, J. L. (1989). "The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval", *Journal of the American Society for Information Science* 40(2): 115-132.

Fahy, P.J. (2003). "Indicators of support in online interaction", *The International Review of Research in Open and Distance Learning*, 4(1). Retrieved June 13, 2006 from: http://www.irrodl.org/index.php/ irrodl/article/view/129/209

Fahy, P.J., Crawford, G. & Ally, M. (2001). "Patterns of interaction in a computer conference transcript", *International Review of Research in Open and Distance Learning*, 2 (1). Retrieved June 13, 2006 from: http://www.irrodl.org/index.php/ irrodl/article/view/36/74

Garrison, D. R. & Anderson, T. (2003). *E-Learning in the 21st Century.* London: RoutledgeFalmer.

Hearne, B. & Nielsen, A. (2004). "Catch a cyber by the tale: Online orality and the lore of a distributed learning community", in Haythornthwaite, C. & Kazmer, M. M. (Eds.)(pp. 59-87). *Learning, Culture and Community in Online Education: Research and Practice.* NY: Peter Lang.

Herring S. C. (1996). "Gender and democracy in computer-mediated communication", in R. Kling (Ed.) *Computerization and Controversy.* 2nd edition. San Diego: Academic Press.

Herring, S.C. (1994). "Gender differences in computer-mediated communication:_Bringing familiar baggage to the new frontier." Presented at American Library Association convention, Miami, FL. Retrieved June 13, 2006 from: http://www.cpsr.org/prevsite/cpsr/gender/herring.txt

Herring, S.C. (2000). "Gender Differences in CMC: Findings and Implications", CPSR Newsletter, 18(1). Retrieved June 13, 2006 from: http://www.cpsr.org/issues/womenintech/herring

Herring, S.C. (2003). "Dynamic topic analysis of synchronous chat", *Symposium on New Research for New Media*, University of Minnesota, Minneapolis. Retrieved June 5, 2006 from: http://ella.slis.indiana.edu/~herring/dta.html

Herring, S.C., Scheidt, L.A., Kouper, I. & Wright, E. (in press). "A longitudinal content analysis of weblogs: 2003-2004", in M. Tremayne (Ed.), *Blogging, Citizenship and the Future of Media.* London: Routledge.

Hmelo-Silver, C. E. (2006). Analyzing collaborative learning: Multiple approaches to understanding processes and outcomes. *ICLS '06: Proceedings of the 7th International Conference on Learning Sciences,* Bloomington, Indiana. 1059-1065.

Krippendorff, K. (2004). *Content Analysis.* Thousand Oaks, CA: Sage.

Liddy, E.D. (1998). "Enhanced text retrieval using natural language processing", *Bulletin of the American Society for Information Science*, 24(4). Available at: http://www.asis.org/Bulletin/Apr-98/liddy.html

McLaughlin, M. L., Osborne, K. K. & Smith, C. B. (1995). "Standards of conduct on usenet", in S. G. Jones (Ed.), *CyberSociety: Computer-Mediated Communication and Community* (pp 90-111). Thousand Oaks, CA: Sage.

Mei, Q. & Zhai, C. (2005). "Discovering evolutionary themes patterns from text – an exploration of temporal text mining", *KDD'05* (Chicago, Illinois). 198-207.

Ooi, V. B.Y. (2000). "Aspects of computer-mediated communication for research in corpus linguistics", *Language and Computers, 36*, 91-104.

Rafaeli, S. & Sudweeks, F. (1997). "Networked interactivity", *Journal of Computer-Mediated Communication, 2(4).* Available online: http://www.ascusc.org/jcmc/vol2/issue4/rafaeli.sudweeks.html

Salton, G. (1988). "Syntactic approaches to automatic book indexing", in *Proceedings of the 26th Annual Meeting on Association for Computational Linguistics,* Buffalo, New York. 204-210.

Schmid, H. (1994). "Probabilistic part-of-speech tagging using decision trees", in *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.

Sixl-Daniell, K. & Williams, J.B. (May 2005). *Paralinguistic Discussion in an Online Educational Setting: A Preliminary Study.* Retrieved June 13, 2006 from: http://www.u21global.edu.sg/portal/ corporate/docs/wp_010-2005.pdf

Stuckey, B. & Barab, S. (forthcoming). "Why good design isn't enough for web-supported communities", in R. Andrews & C. Haythornthwaite (Eds.), *Handbook of Elearning Research*, Sage.

Weber, R.P. (1985). *Basic Content Analysis.* Beverly Hills, CA: Sage.

Wu, H., Zubair, M., & Maly, K. (2006). "Harvesting social knowledge from folksonomies", in *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia* (Odense, Denmark, August 22 - 25, 2006). 111-114.

Zhai, C. (1997). "Fast statistical parsing of noun phrases for document indexing", in *Proceedings of the Fifth Conference on Applied Natural Language Proessing,* Washington, DC. 312-319.