

Discourse Diagrams: Interface Design for Very Large-Scale Conversations

Warren Sack
 MIT Media Laboratory
 wsack@media.mit.edu

Abstract

On the Internet conversations between hundreds or thousands of people exchanging thousands of messages have become commonplace. The size of these very large-scale conversations (VLSCs) make them difficult for participants and interested observers to understand and critically reflect upon. To facilitate understanding of the social and semantic structure of VLSCs two tools from the social sciences – social networks and semantic networks – have been used and extended for the purposes of interface design. As interface devices, social and semantic networks need to be flexible, layered diagrams that are useful as a means for exploring and cross-indexing the large volumes of messages that constitute the archives of VLSCs. This paper discusses the design criteria necessary for transforming social scientific representations into interface devices. The discussion is illustrated with the description of the Conversation Map system, an implemented system for browsing VLSCs.

1. Introduction: Social and Semantic Networks as Generative Devices

In architecture the diagram is historically understood in two ways: as an explanatory or analytic device and as a generative device. [1]

Even only ten years ago electronic mail was a novelty outside of computer science departments. But now, with tens of millions of people online, email addresses and URLs are posted on billboards, MTV “call-in” shows print email extracts at the bottom of the screen, and non-nerds exchange email as a matter of day-to-day social and professional life. This rapid increase in the number of Internet inhabitants has made possible the unprecedented phenomenon of very large-scale conversations (VLSC) in which hundreds, even thousands, of people, internationally exchange thousands of messages in daily, many-to-many communications. The most obvious manifestations of this phenomenon are Usenet newsgroups hosted on servers throughout the Internet and archived by a handful of large sites (e.g., www.dejanews.com).

From a social scientist's perspective VLSC is a fascinating phenomenon because it poses a fundamental challenge to many of the analytic tools and descriptive methodologies that have been built up over the past century to understand conversations and discourse. These tools of social science have often been developed to analyze interactions involving far fewer people than the number of participants now taking part in various online VLSCs. It may be possible to scale-up these tools and methodologies from linguistics, psychology, sociology, media studies, and anthropology, but it is not *a priori* obvious how this can be done.

From the perspective of a participant, the idea that VLSC might be a phenomenon is hard to appreciate. From the participant's point of view VLSCs -- e.g., busy Usenet newsgroups -- simply look like huge numbers of email messages clogging the inbox of a news reader.

In this paper it is argued that, from a designer's perspective, the challenge before us is to use and improve upon the tools of social science to create new interfaces for VLSCs that provide a means for participants and interested observers (e.g., social scientists) to understand and critically reflect on them.

The more theoretical discussion of this paper is grounded with the presentation of an implemented Usenet newsgroup browser, the *Conversation Map* system. In principle, the system can be used just like a conventional news or mail reader (e.g., Eudora, Netscape Messenger, RN, etc.). But, the Conversation Map system offers several summaries of newsgroup messages that standard browsers do not provide. It performs a series of combined computational linguistic and sociologic analyses on the messages of a newsgroup and presents the results of these analyses in the form of a graphical interface. The graphical interface can be used to browse the messages and explore connections between the messages.

After the system has been run on an archive of a few thousand messages from a Usenet newsgroup, one can use the system to get an idea of who is “talking” with whom, the central players in the newsgroup, the important themes of discussion, and some of the emergent metaphors or definitions that might be peculiar to the language exchanged in the messages of the newsgroup.

To some, compared to a conventional email reader, the Conversation Map system might look like a fancy appliance overloaded with a bunch of spurious, new features. However, these features are derived from standard tools of the social sciences; tools which have been useful for scientists attempting to understand the social and linguistic structures of individuals and groups.

Thus, the obvious first “users’ group” for the Conversation Map system is to be found in those scientists for whom these tools are already familiar. In fact, a handful of projects are now underway in which social scientists are using the system to gain a better understanding of various VLSCs (e.g., [2]). The system is being adapted and extended to meet their needs in a sort of participatory design process.

Simultaneously, several laypeople are experimenting with the system. Why might a layperson want to use tools from social science? Imagine wanting to join in on a VLSC that has been in progress for several months or even years with an archive of thousands of messages. Having a set of machine-generated summaries of the archive provides one with a relatively quick way to decide if the group is focussed on what is expected or desired. The summaries provided also give some indication of how various subjects are being discussed and who is central to the discussion. In addition, there are ways of using the interface to steer around “spam” and focus in on threads of discussion that address specific themes. In short, the social science-derived tools of the interface give one a means of seeing the “forest” of the conversation before diving into the “trees.”

So, what a layperson wants to know about an archive of messages can overlap with the goals of social science. Consequently, it seems possible to start at the same point (i.e., with the same prototype system) and then evolve the system in two – perhaps only slightly – different ways: one way for the scientists, another way for the non-scientists. The difficulty of the interface design challenge is proportional to the size of the divergence between these two groups of people. The design challenge lies in the attempt to stretch the tools of the social sciences both (a) to meet the demands of social scientists interested in studying the phenomenon of VLSC; and, (b) to render the tools in a form that makes it possible for the layperson – the non-specialist participant in these VLSCs – to achieve some of the insights that these tools make possible for the social scientist.

This paper is an attempt to do two things: (1) to explain the analysis procedures and interface of the Conversation Map system; and, (2) to explain the design rationale used to borrow graphical elements and analyses procedures from the social sciences and rework and extend them for use as interface devices.

In particular, in this paper, two tools from the social sciences -- *social networks* (e.g., [3, 4]); and, *semantic*

networks (e.g., [5, 6]) -- will be re-examined as possible interface components for a VLSC browser.

Rendering a semantic network generally entails sketching edges between a set of nodes labeled with words or concepts. The edges are understood to be a representation of the semantic relations between the words or concepts.

Social networks can be drawn in a similar manner although, obviously, for very different representational purposes. The nodes in a social network are often labeled with the names of people. The edges between nodes represent, for instance, social interactions between the people.

In the social sciences, social and semantic networks can be used and discussed as *analytic devices*, as representations of observed or hypothesized phenomenon, i.e., as scientific models. They are, in short, proposed as answers to outstanding scientific questions. For instance, in cognitive psychology, linguistics, and artificial intelligence, semantic networks have been understood to be an answer to this question: “What constitutes a reasonable view of how semantic information is organized within a person’s memory?” [5]. Within sociology, social networks are used to model and summarize empirical studies of interactions and relationships between people, groups, and institutions.

In addition to their use as analytic, scientific representational devices, I am proposing that social and semantic networks be designed as *generative devices*; i.e., as interfaces with which one can embark upon an exploration of VLSCs. This use of diagrams as generative -- as well as analytic -- devices is relatively well understood in other design disciplines as is illustrated by the quote at the top of this paper from Peter Eisenman on the role of diagrams in architecture. Generative devices, as they are to be discussed in this paper, serve as possible conversational resources: they provide a set of possible conversational foci and a means to reflect upon an on-going discussion. Such devices are generative insofar as they are evocative objects meant to engender discussion. One can think of this generative use of social and semantic networks as an interface into the archive of a discussion instead of (or in addition to) analytic summaries of a discussion. Used in this manner, social and semantic networks are *diagrams of discourse*.

Here is an outline of the remainder of this paper: First, a longer description of semantic and social networks will be undertaken to explain why they are useful as diagrams of discourse. Second, several examples of semantic and social networks used as interface elements are shown. These examples are computed and diagrammed by an implemented Usenet newsgroup browser -- the *Conversation Map* system -- that allows one to explore large archives of newsgroup messages. Finally, after describing how social and semantic networks are

automatically generated, diagrammed, and incorporated into the Conversation Map system, several design criteria for discourse diagrams will be sketched out.

2. Social and Semantic Networks as Useful Descriptions of Discourse

To explain why a combination of social and semantic networks might be a useful description of a VLSC, it is helpful to examine the claims of the linguist Michael Halliday. According to Halliday [7], language has at least three meta-functions: (1) *ideational*: language can represent ideas; (2) *interpersonal*: language functions as a medium of exchange between people; and, (3) *textual*: language functions to organize, structure, and hold itself together; this function allows the various devices of cohesion, including citation, ellipsis, anaphoric reference, etc. to be used. Thus, for example, I can write here, in this present sentence, about the first sentence of the present paragraph and the reader can infer that I am referring to the sentence that begins like this: “To explain why...”

I claim that any interesting interface for VLSCs is incomplete if it does not incorporate all three of these

meta-functions (ideational, interpersonal, and textual). In particular, I claim that a suitably improved implementation of social networks can represent, even if only very roughly, the interpersonal and textual aspects of a VLSC; and, that semantic networks can be an approximation of the ideational content of a VLSC.

Another way of explaining the usefulness of social and semantic networks as a description of VLSCs is in the terms of Paul Dourish and Matthew Chalmers, researchers in human-computer interaction and computer-supported cooperative work. Dourish and Chalmers claim there are at least three forms of navigation mechanism which can be combined in information systems: (1) spatial; (2) semantic; and, (3) social navigation mechanisms [8]. Most graphical interfaces make use of spatial layout (and thus facilitate spatial navigation); some use semantic navigation (e.g., hypertexts); and some social navigation (e.g., a variety of work in the sociology of social networks). However, very few interfaces combine all three sorts of navigation. The Conversation Map system provides the means to spatially navigate through social networks, semantic networks and intersections of the two, thus, implementing all three of Dourish and Chalmers’ forms of navigation.

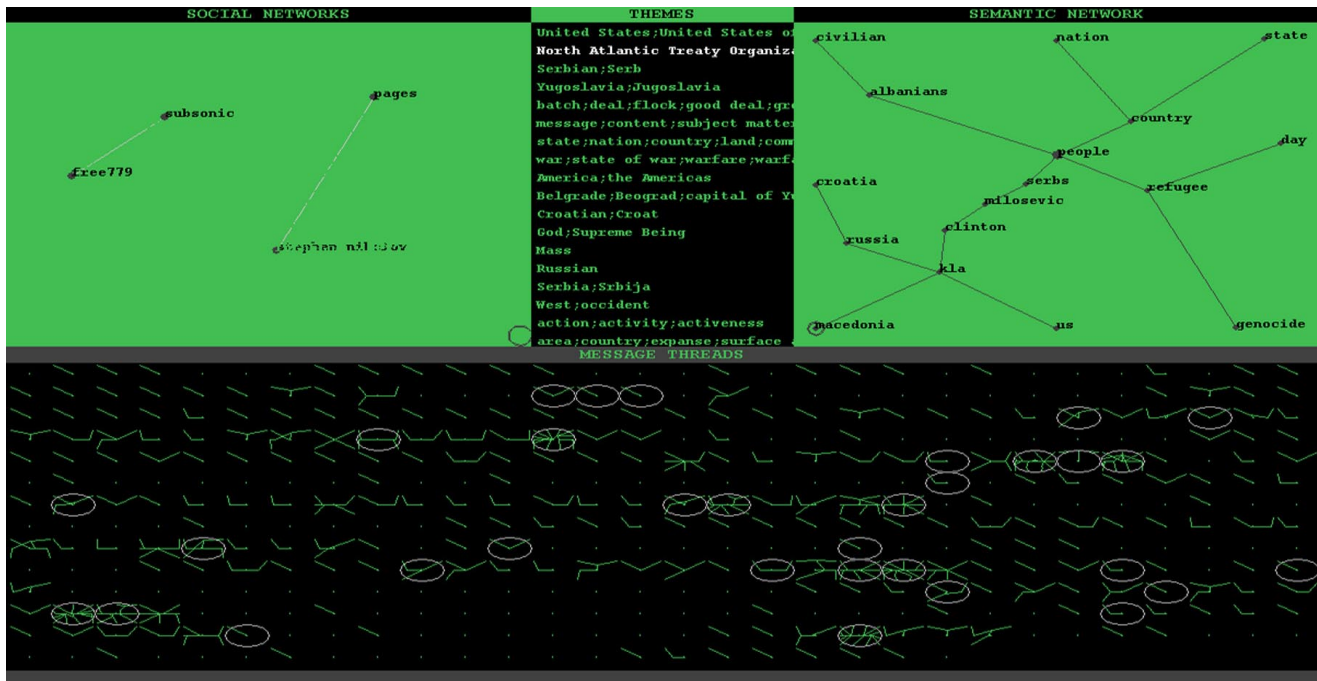


Figure 1: Interface with a discussion theme selected and a subset of the social network visible

2.1. Social Networks

One of the “results” of a VLSC is a social network. After a while, participants in an online discussion get to

know one another and exhibit characteristic patterns of interaction with one another. Some participants post messages that often strike others as interesting, evocative, or worthy of a reply and so these people tend to garner more responses to their messages than others do. Some

participants take pains to reply to the posts of newcomers and regulars alike and so build a reputation as virtual moderators for the discussion (even in groups with no officially designated moderators). Some people post what others consider to be “spam” and so, even though they may tend to post frequently, they are ignored (i.e., their posts do not earn replies from other participants). These collections of posting profiles for the participants of a group, when graphically assembled together, constitute a diagram that would be recognizable to social scientists as a social network.

The simplest networks possible for VLSCs are those which can be computed through an analysis of references between messages; i.e., an analysis of the “threading” of the messages. So, if participant A responds to a message posted by participant B, then a link can be drawn between A and B. Threading is easy to compute given the format of Usenet newsgroup messages and the information computed can be graphically presented. Donath, Karahalios, and Viegas describe a well-designed graphical presentation of Usenet newsgroup threading information in a system called *Loom* [9].

Obviously it would be useful to know something more than simply the number of times any two given participants exchanged messages. It would be nice to know, for instance, in their responses to each other whose messages were cited or quoted by whom. This sort of analysis has been extremely useful in science studies and is generally referred to as *citation indexing* [10]. However, while instances of citation are relatively easy to spot automatically in scientific papers because citations are required to appear in a standard format and must be listed in the references at the end of a paper, this is not the case for Usenet newsgroup messages. Within Usenet newsgroup messages, citations often occur without mention of the original author, citations are often nested inside one another, and citations do not have a standard format, even though it is quite common (but not required) to precede each line of a quotation with some punctuation, like

```

this:
>>> On 31 February 2001 Warren Sack
>>> <wsack@media.mit.edu> wrote:
>>> Hi guys! The future is really great!
    
```

These complicating factors make an automatic citation analysis procedure difficult, but not impossible, to implement for Usenet newsgroup messages. An analysis procedure of this sort has been implemented in the Conversation Map system. By automatically identifying who has either responded to and/or quoted from whom, the Conversation Map system calculates a social network given an archive of Usenet newsgroup messages. The nodes in the network represent people -- i.e., participants in the online discussion -- and the links represent *reciprocating* quotations and/or responses.

Thus, if participant A responds to or quotes a message from participant B and then, later in the discussion, participant B quotes from or responds to a message from participant A, a link is drawn between nodes labeled “A” and “B.” If A and B have reciprocated frequently, the link between them will be shorter than if they have only quoted from or responded to one another once or twice. Figure 2 shows the social network computed by the Conversation Map system after the system was run on two weeks’ (16 April 1999 - 4 May 1999) worth of messages (over 1200 messages from about 260 participants) from a Usenet newsgroup devoted to a discussion of the situation in Kosovo (soc.culture.albanian). Note that there are certain “hubs” in the social network. These hubs represent participants who post many messages but who also receive many responses to their messages. They are virtual moderators of a sort for the newsgroup even though the group depicted has no official moderators. In Figure 2 the social network is displayed with all of the names turned off so that the overall shape of the network can be seen. If one wants to select a part of the network, a node (representing a particular participant) can be clicked on. When this is done, every other participant who has reciprocated replies or quotes with the selected participant is shown and the rest of the social network disappears.

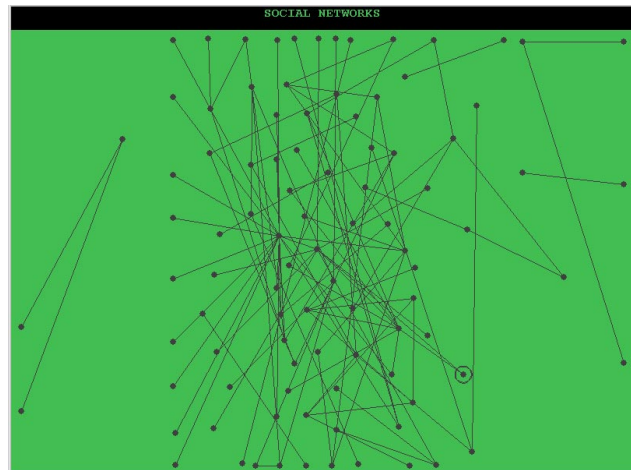


Figure 2: Social networks showing who has reciprocally replied to and/or cited whom

Performing a citation analysis produces social networks, like the one shown above, but it also allows another interface feature to be implemented as well. Once quotations and their sources have been identified, messages in the archive can be hyperlinked so that clicking on a quote in a message opens a window containing the text of the cited message. Within the Conversation Map system one can move between messages by clicking on quotations.

While the metric who-cites-whom is more sensitive

than the simpler metric of who-responds-to-whom, one can imagine an increasingly sophisticated series of metrics based upon more and more sensitive readings of message-to-message linkages. For example, it is possible for me to implicitly cite the Shakespeare play about the two young lovers from feuding families without mentioning the name of the play. These sorts of more or less subtle linkages between texts and within texts are termed *ties of cohesion* in the terminology of *systemic functional linguistics* [11]. While there has been some progress in the area of computational linguistics techniques for inter-textual cohesion analysis (e.g., [12]), it is necessary to merge such an analysis with a social network computation in order to be able to label the links of the computed social network with the sorts of ties that characterize the relationships manifest between participants in a VLSC.

As a first step in this direction a procedure for *social cohesion analysis* has been implemented in the Conversation Map system [13]. This procedure performs an analysis of lexical cohesion [14] between messages and then overlays this information on top of the social network so that a rough approximation of the “theme” of the conversation that exists between posters can be detected. A step-by-step description of this procedure will be given in section 3.1.

In the upper middle panel of Figure 1, the results of the cohesion analysis of the Conversation Map system can be seen: the automatic analysis of social cohesion produces a menu of “discussion themes.” If, for example, participant A mentioned the word “skiing” in a post that also quoted a part of a message from participant B wherein B wrote about skating, and then, later in the conversation participant B wrote about cycling in response to a message by A concerning wrestling, then the link between A and B in the social network would be labeled with the term “sports” since skiing, skating, cycling, and wrestling are all sports. This analysis requires, of course, the use of a thesaurus, specifically WordNet [15].

In the interface, when one clicks on the menu item “sports” the link between A and B is highlighted (along with the links between any other pairs of posters who are connected through a discussion of sports). Figure 1 shows the same social network pictured in Figure 2 along with the menu of “discussion themes” (or, more accurately, the list of lexical ties) that links messages, and thus, people together in conversation about the larger topic of Kosovo and Albanian culture in general. In Figure 1 most of the social network has been hidden so that the two pairs of posters who have exchanged messages concerning NATO are highlighted.

Labeling the links of computed social network with the content of messages, has some precedents in the literature of social network analysis (cf., [16]; see also, for example, [17]). However, no such computations of social cohesion have been implemented in a computer program as has

been done for the Conversation Map system. The precedent research in social network analysis has been performed by hand. Conversely, although procedures for automatic cohesion analysis have been implemented by researchers in computational linguistics (cf., [14, 18]), no such procedures have merged the results of a lexical cohesion analysis with a computed social network as has been done for the Conversation Map system. While the social cohesion analysis and display procedures in the Conversation Map system could be improved upon, even in their present form they effectively compute and interweave an analysis of the *interpersonal* and *textual* dimensions of the language of Usenet newsgroups.

2.2. Semantic Network

A second “result” of a VLSC is a semantic network. Over the course of many exchanges, participants in a VLSC coin new acronyms (e.g., IMHO, “in my humble opinion”), new punctuation (e.g., the (in)famous smiley faces :-), and new or idiosyncratic ways of using existing vocabulary terms (e.g., “to flame” means something rather particular online). Semantic relations between new and existing lexical items (i.e., words and abbreviations) can be represented in a semantic network. In general, I will argue that, with a semantic network, it is possible to diagram some of the *ideational* meta-functions of a conversation.

For reasons that will soon become apparent, it is possible to understand the semantic networks of a VLSC as a diagram of the emergent *metaphors* or *definitions* of participants’ discussion. To clarify this insight it is worthwhile examining an example of how the cognitive scientists George Lakoff and Mark Johnson [19] demonstrate the common usage of specific metaphors. Lakoff and Johnson offer the following sentences to support their claim that, in “our” (presumably U.S.) culture, the metaphor TIME IS MONEY is a common presupposition:

TIME IS MONEY
You're wasting my time.
I don't have the time to give you.
How do you spend your time these days?
I've invested a lot of time in her.
I don't have enough time to spare for that.
You're running out of time.
You need to budget your time.

In these sentences the word “time” could meaningfully be replaced with the word “money.” The hypothesis is that if two words or concepts are thought about in the same way by a group of people, then they will be systematically used in the same way in the discourse of those people. Examples of language use can be used as symptoms for the diagnosis of semantically related words and phrases.

By mapping out which words in a VLSC are used like which other words, a diagram of some of the semantics -- i.e., the meaning -- of the conversation can be displayed. The resulting semantic network for a given VLSC is a way to begin to investigate how the statements made by different participants in the conversation are similar to one another.

The semantic network pictured above -- in the upper right-hand corner of Figure 1 -- was computed automatically by the Conversation Map system given the archive of Usenet newsgroup messages described above concerning Kosovo and the Albanians (from the Usenet newsgroup soc.cult.albanian). The network is a tree. The tree is plotted like a spider web, so that the child nodes of the root (in Figure 1 the root is the node labeled "people") are drawn at a certain radius out from the root, the children of the children drawn a bit further out in a ring around the children, and so forth until the edges of the available area are reached. If two nodes in the semantic network are connected, then they often used in the same way in the archive. Thus, for example, "people" and "Serbs" are connected.

The Conversation Map system parses all of the messages of the archive and for every noun in the archive it builds a "profile." A noun's profile consists of a weighted vector including verbs (with which it appeared as a subject, a direct object, or an indirect object), adjectives (which modify it in the language of the messages of the group), and nouns (which have been used to modify it in noun-noun phrases). All of the nouns' profiles are compared to one another and a nearest neighbor is computed for each noun.

A semantic tree, like the one shown above, is formed when one term -- e.g., "people" -- is the computed nearest neighbor for several other nouns (i.e., in this case, for the nouns Serbs, Albanians, country, and refugee).

The procedure described produces many semantic networks for every archive of messages analyzed. All of the unique nouns encountered in the corpus of messages are included in a semantic network, but not all semantic networks are displayed in the interface. The semantic networks that are picked for display in the Conversation Map interface are those networks that are used most frequently as themes of discussion. This ordering criterion on the semantic networks insures that the networks seen first are those that are formed by group discussion rather than, for example, the verbose postings of a single spammer. So, the ordering criterion knits together the ideational, textual, and interpersonal dimensions of the VLSC interface rendered by the Conversation Map system.

The profiles of the terms that appear in the semantic networks can be examined and compared by selecting one or more terms in the semantic network. For instance, if

the terms "Serbs" and "people" are selected, a list of statements can be constructed like the one Lakoff and Johnson wrote for the TIME IS MONEY metaphor. If the list is restricted to only those verbs for which both "people" and "Serbs" appeared as a subject of the verb, then the resultant list looks like this:

SERBS ARE PEOPLE (terms appear as subject for each of the verbs one or more times)
allow, be, destroy, die, do, drive, exist, flee, get, give, have, keep, know, lay, leave, live, make, need, pay, remember, tell, think, turn

In other words, by looking at the archive of messages one can find many places where, for instance, both "people" and "Serbs" appear as subjects of the same verb. From the intersected lists of verbs one can see that, in the archive of soc.cult.albanian messages "Serbs" and "people" are discussed in similar terms because there exist one or more statements in the archive for both "Serbs" and "people" where they are describe separately as agents which allow, destroy, die, do, drive, exist, etc.

The verb "to need" is one of these shared verbs found in the intersection of the "Serbs" and "people" profiles. Clicking on a verb in the intersected profiles (not shown here, but displayed by the Conversation Map interface when two terms in the semantic network are selected) reveals the following two, example sentences that partially underpin the link between "Serbs" and "people" in the diagram: (1) "You have to realize that Greeks and *Serbs* need a just solution, and not just Serbia has a solution: Serbia." (2) "It is not enough to be alive, *people* need normal life."

Similar verbs lists are computed on-demand by the Conversation Map system for any other pair of terms in the semantic network and, if desired, example sentences of the terms in use can also be viewed.

ALBANIANS ARE PEOPLE (terms appear as subject for each of the verbs one or more times)
cross, displace, do, flee, have, hate, hide, leave, lose, say, suffer, think, walk

SERBS ARE ALBANIANS (terms appear as subject for each of the verbs one or more times)
do, flee, found, have, insist, leave, shoot, think, want

One optimistic way of reading the semantic network computed by the Conversation Map system for the soc.cult.albanian group is this: "people" is a neutral term and both "Serbs" and "Albanians" are people. This is a sort of thin humanism ("after all we are all people") that begins to explain why any sort of exchange can happen in this VLSC concerning Kosovo, even though, admittedly, the exchange is a very heated and argumentative one.

The approach described above for the calculation and display of semantic networks is related to a variety of contemporary, corpus-based techniques of computational linguistics (cf., [20]). These techniques, are, in turn, related to a variety of older work in linguistics including

the lexical fields approach of Trier [21] the work done on collocation by Firth [22], and the distributional approach to discourse analysis introduced by Harris [23] and pursued in a restricted but computational manner by, for example, Pêcheux [24]. See also the discussion by Lyons [25] concerning the longer history of these approaches to semantics in linguistics. More recent work by Callon et al. [26] is in a similar vein and has been carried out as a project in the sociology of science in order to summarize and analyze the discourse of science.

The computational techniques for computing semantic networks in the Conversation Map system largely follow the techniques developed by Grefenstette [27]. Technically, the main advance over Grefenstette's techniques discussed here is the use of computed, social networks and lexical cohesion to automatically select a subset of the calculated semantic networks for display.

3. On the Construction and Display of Social and Semantic Networks

In the following subsections the procedure for calculating social and semantic networks is sketched out and then a short overview of the Conversation Map system interface is presented.

3.1. Construction of Social and Semantic Networks

The analysis procedure of the Conversation Map system performs the following steps on an archive of Usenet newsgroup messages in order to compute the social and semantic networks described above:

- (a) Messages are "threaded."
- (b) Quotations in the messages are identified and their sources (in other messages) are found.
- (c) The "signatures" of posters are identified and distinguished from the rest of the contents of each message.
- (d) An index of posters (i.e., newsgroup participants) to messages is built.
- (e) For every poster, the set of all other posters who replied to the poster -- or quoted from messages authored by the poster -- is recorded. Posters who reply to and/or quote from one another are linked together in the social network. Reciprocity is therefore highlighted in the computed social network.
- (f) The words in the messages are divided into sentences, tagged with part-of-speech information, and their roots are identified. To divide the words into sentences, a tool built at the University of Pennsylvania is used [28]. To accomplish the part-of-speech tagging, a simple trigram based tagger has been constructed (cf., [29]). The morphological analyzer built for the Conversation Map system uses a freely-available morphology and syntax

database [30].

(g) Discourse markers (e.g., connecting words like "if," "therefore," "consequently," etc.) are tagged in the messages. The Conversation Map system employs a list of discourse markers compiled by Marcu [31].

(h) The words of the messages are parsed into sentences using a partial parser. The partial parser is a re-implementation of the parser described in [27].

(i) An analysis of lexical cohesion is performed on every pair of messages where a pair consists of one message of a "thread" and another message that either immediately follows the first message in the thread (i.e., is a reply to the first message) and/or follows the first message in the thread and contains a quotation from the first message. This analysis produces a series of lexical ties between messages that can be understood as a crude approximation to the theme of the conversation in a sequence of messages. The lexical database WordNet [15] is used in the lexical cohesion procedure. See [11] for a definition of lexical cohesion. See [14] for an example implementation of a lexical cohesion routine.

(j) By using the index created in step (d) with the results of step (i) a set of lexical ties are computed for every pair of posters who have replied to and/or quoted from one another over the course of time represented by the Usenet newsgroup archive under analysis. These aggregated lexical ties are layered on top of the social network computed in step (e). The result is that most of the links between pairs of posters are labeled with one or more lexical ties (i.e., one or more "discussion themes"). The combination of social networks and lexical cohesion results is called *social cohesion*. The social cohesion analysis procedure developed for the Conversation Map system is partially described in [13].

(k) The lexicosyntactic context of every noun in the archive is compared to the lexicosyntactic context of every other noun in the archive. Nouns that are used or discussed in the same manner are calculated to be similar and are placed close to one another in the semantic network. An algorithm similar to the one described in [27] is used. Once all of the noun-noun pairs have been compared and a nearest neighbor for each noun computed, a subset of the semantic networks computed are selected for display by ranking the semantic networks. The top-ranked semantic network contains a set of terms (used as "discussion themes") that connect the greatest number of poster pairs linked in step (j). A full description of this procedure will be the subject of a forthcoming paper.

3.2. Display of Social and Semantic Networks

After analyzing an archive of Usenet newsgroup messages, the Conversation Map system generates a display of the archive and the social and semantic

networks that constitute the analysis of the archive. An example of such a display is shown in Figure 1.

The top portion of the screen displays the social and semantic networks. The bottom portion is a graphical representation of the analyzed messages sorted by subject thread. Each thread is allotted a small rectangle of screen space and the threads are laid out chronologically from upper-left to lower-right. The threads are graphically plotted with the same tree-as-spider-web algorithm used to lay out the semantic network. The first message of a thread is plotted in the middle of the allotted rectangle. Responses to the first message are plotted around it and further towards the edges of the rectangle; and so on for the responses to the responses.

A small window containing an enlargement of one of the threads in the archive is created if one double-clicks on one of the threads displayed in the lower-half of the screen. Double-clicking on any of the nodes of the thread (labeled with the posters' names) causes the text of the message to be displayed in a third window.

The social and semantic networks can be used to explore the archive of messages. For instance, clicking on one node of the social network (i.e., clicking on the name of a poster in the network), highlights all of the threads in the archive to which the poster has contributed one or more messages.

In the screen shot shown in Figure 1, the menu item labeled "North Atlantic Treaty Organization" in the menu of discussion themes has been selected. Links between two pairs of posters in the social network have been highlighted because messages concerning NATO have been reciprocally exchanged by those posters. Also, many threads in the archive are highlighted with a white border because NATO has appeared as a discussion theme in them. Note that a discussion theme might appear throughout the archive, but only link one or two pairs of posters. Linking two posters is a more involved requirement than simply linking two messages in a thread. The link joining two posters in the social network is labeled with a discussion theme if and only if A has responded to B concerning the theme and vice versa.

A similar functionality allows one to explore the archive by clicking on nodes in the semantic network. Moreover, as was discussed above, the semantic networks can be "unfolded" to reveal the features (i.e., verbs, adjectives, etc.) shared by two or more terms. When two nodes in the semantic network are highlighted, a window appears containing lists like the one describing the profile intersections for "Serbs" and "Albanians" discussed above.

4. Conclusions

There are several design criteria that have been used to construct and display the social and semantic networks of VLSCs with the Conversation Map system. These criteria negotiate a divergence between a social scientific use of the social and semantic networks of the Conversation Map and a possible, popular, non-scientific use of the same.

Ideally, for the sake of science, the system-generated, social and semantic networks would be constructed as carefully and rigorously as the "hand built" analyses of, for instance, ethnographic or sociolinguistic studies of online discussions (e.g., [32]). However, clearly, the system-generated networks will never be as precise as analyses accomplished by hand. Thus, while, from a scientist's viewpoint, it might at first appear to be a good idea to attempt to automate much of the process of online conversation analysis, a closer look at the pragmatics of such a design goal shows it to not be such a good idea.

However, since the system-generated results are quicker and easier to attain than comparable results compiled by hand, the results of the Conversation Map system could serve different needs for scientists and non-scientists. For the scientist – e.g., for someone who is trained as a discourse or conversation analyst or an ethnographer – the diagrams produced by the Conversation Map system could be understood as a rough sketch of where one might begin to explore an archive of messages. With such a "sketch" in hand, the scientist could begin a set of more rigorous close readings of the archive of messages.

For the non-scientist, the discourse diagrams produced probably represent a much more detailed analysis of a VLSC than anything the non-scientist would ever produce on their own. However, there are pitfalls associated with non-scientific usages of scientific-looking images. Scientific images have always been open for interpretation and put to alternative uses by non-scientists; e.g., journalists, lawyers, politicians, interested laypersons, and non-specialists (e.g., scientists or doctors who are not from the discipline directly responsible for the production of a set of scientific images). For example, the anthropologist of science and technology, Joseph Dumit, has examined how PET scans of the brain are used and understood both within science and "outside" in medicine, law, journalism, and popular culture [33].

The common pitfall associated with many vernacular presentations of scientific images concerns the manner in which the images are often "untethered" from the data used to produce them. Untethered scientific images -- i.e., images that have been unlinked from supporting data -- sometimes become too easy to manipulate because they are no longer manipulated within the rigorous constraints of science; e.g., when a popular magazine recolors an

image received from a biologist to make the image easier to print, or more colorful. At other times, these untethered images become too hard to manipulate because a layperson has no access to the phenomenon pictured; e.g., how would a non-scientist redraw the traces of subatomic particle collisions produced by physicists with an instrumented particle accelerator?

One crucial design question is therefore this: How can the images of science be used for interface devices without untethering the images from the supporting data? Furthermore, how can the social scientific images of semantic and social networks be rendered as *generative diagrams* for use as interface devices that cross-index and provide access to thousands of email messages from the archives of very large-scale conversations? The approach to this design problem taken in the construction of the Conversation Map system has been this: make all of the original data (i.e., the newsgroup messages) accessible through the diagrams (i.e., through the act of double-clicking on the diagrams).

A second tension that exists between possible scientific versus possible popular uses of the Conversation Map system is this: to automatically generate the discourse diagrams – the social and semantic networks – the Conversation Map produces an enormous amount of quantitative data on the messages and the participants of the newsgroup analyzed. Social scientists often want to see the numbers produced. For example, the Conversation Map system generates a set of statistics which could be useful to get at least preliminary answers to the following questions: What was the mean number of responses sent to a message? How many other participants did a given participant respond to? How many times did a given theme show up as a theme of discussion? What proportion of the population of participants contributed towards the discussion on a given topic? Is there a specific subset of participants who started most of the threads concerning a given theme of discussion?

As interfaces for VLSCs come to be more and more representative of the social structure of the conversing group, many in the group may feel that an interface display impinges on their privacy by rendering transparent the history of their interactions with the group. Even today, the poster profiles computed and indexed at sites like www.dejanews.com incite these worries for some posters. Preferably then the construction of discourse diagrams will follow an aesthetics of social translucence [34] and encourage an ethics of social reflection, rather than aiming at an aesthetics of realism and transparency that would make all of us feel we are under a microscope of surveillance. Displaying all of the statistics calculable by the Conversation Map system would probably render the participants' profiles too transparent for comfort.

To negotiate this tension between the need for numbers and the larger scientific and non-scientific need to better understand the linguistic and social structures of VLSCs, the quantitative results calculated by the Conversation Map system are displayed as qualitative, flexible diagrams. These diagrams can be moved and redrawn by the participant or interested observer, but they do not directly yield summary statistics on the group or individual participants of the group.

Finally, while it is often the case that analyses of conversation and discourse have been done by scientists for other scientists, it was a specific design choice to create an interface for the Conversation Map system that can be, at least in principle, accessed by everyone who might be participating in a public, Internet-mediated VLSC. Several Conversation Maps and a user's manual can currently be found on the web: <http://www.media.mit.edu/~wsack/CM/index.html>. To use the Conversation Map system, a newer, Java 1.2 enabled browser is necessary (e.g., Netscape 4.6 on Windows or Linux machines; Internet Explorer 4.5 on Macs).

5. References

- [1] Peter Eisenman "Diagram: An Original Scene of Writing" in *Architecture New York* (23:27, 1998)
- [2] Warren Sack and Joseph Dumit, "Very Large-Scale Conversations and Illness-based Social Movements" presented at *Media in Transition*, MIT, Cambridge, MA, October, 1999.
- [3] Stanley Wasserman and Joseph Galaskiewicz (editors) *Advances in Social Network Analysis: Research in the Social and Behavioral Sciences* (Thousand Oaks, CA: Sage Publications, 1994).
- [4] Barry Wellman "Living in a Wired World" *IEEE Intelligent Systems*, January/February, 15-17, 1999.
- [5] M.R. Quillian "Semantic Memory" In M. Minsky (editor) *Semantic Information Processing* (Cambridge, MA: MIT Press, 1968, p. 80).
- [6] A.M. Collins and E.F. Loftus "A Spreading Activation Theory of Semantic Processing," *Psychological Review*, 82: 407-428. 1975.
- [7] Michael A. K. Halliday. *An Introduction to Functional Grammar, Second Edition* (London: Edward Arnold, 1994, p. 179).
- [8] Paul Dourish and Matthew Chalmers. "Running Out of Space: Models of Information Navigation." Short paper presented at *HCI'94* (Glasgow, UK, 1994).
- [9] Judith Donath, Karrie Karahalios, and Fernanda Viegas "Visualizing Conversations" *Proceedings of HICSS-32*, Maui, HI, January 5-8, 1999.
- [10] E. Garfield. *Citation Indexing: Its Theory and Applications in Science, Technology and Humanities* (New York: John Wiley, 1979).
- [11] Michael A. K. Halliday and Ruqaiya Hasan. *Cohesion in English* (New York: Longman, 1976).

- [12] Amit Bagga and Breck Baldwin "Entity-Based Cross-Document Coreferencing Using the Vector Space Model" In *Proceedings of ACL-COLING'98*, June 1998, Montreal, Canada.
- [13] Warren Sack "Diagrams of Social Cohesion" In *Descriptions of Demonstrated Systems, Association for Computational Linguistics, ACL'99*, University of Maryland, College Park, June 1999.
- [14] Graeme Hirst and David St-Onge. "Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms" in Christiane Fellbaum (editor) *WordNet: An Electronic Lexical Database* (Cambridge, MA: MIT Press, 1998).
- [15] Christiane Fellbaum (editor) *WordNet: An Electronic Lexical Database* (Cambridge, MA: MIT Press, 1998).
- [16] S. Black, J. Levin, H. Mehan, and C. Quinn "Real and non-real time interaction: Unraveling multiple threads of discourse." *Discourse Processes*, 6, 59-75, 1983.
- [17] Sheizaf Rafaeli and Fay Sudweeks "Interactivity on the Nets" In F. Sudweeks, M. McLaughlin, and S. Rafaeli (editors) *Network and Netplay: Virtual Groups on the Internet* (Cambridge, MA: MIT Press/AAAI Press, 1998).
- [18] Mark A. Stairmand. "Textual context analysis for information retrieval" in the *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, August 1997
- [19] George Lakoff and Mark Johnson. *Metaphors We Live By* (University of Chicago Press: Chicago, 1980, pp. 7-8).
- [20] Judith Klavens and Philip Resnik *The Balancing Act: Combining Symbolic and Statistical Approaches to Language* (Cambridge, MA: MIT Press, 1996).
- [21] Jost Trier "Das sprachliche Feld. Eine Auseinandersetzung" *Neue Jahrbucher fur Wissenschaft und Jugendbildung* 10, 428-49, 1934
- [22] J.R. Firth *Papers in Linguistics, 1934-1951* (London: Oxford University Press, 1957).
- [23] Zelig Harris "Discourse Analysis," *Language*, 28: 1-30 and 474-94, 1952.
- [24] Michel Pêcheux *Automatic Discourse Analysis* (Amsterdam: Editions Rodopi, 1995).
- [25] John Lyons, *Semantics, Volume 2*, New York: Cambridge University Press, 1977.
- [26] Michel Callon, John Law, Arie Rip (editors) *Mapping the Dynamics of Science: Sociology in the Real World* (London: Macmillan Press, Ltd., 1986).
- [27] Gregory Grefenstette, *Explorations in Automatic Thesaurus Discovery*. Boston: Kluwer Academic Publishers, 1994.
- [28] Jeffrey C. Reynar and Adwait Ratnaparkhi. "A Maximum Entropy Approach to Identifying Sentence Boundaries" In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, March 31-April 3, 1997. Washington, D.C.
- [29] Eugene Charniak *Statistical Language Learning* (Cambridge, MA: MIT Press, 1993, pp. 39-40).
- [30] Daniel Karp, Yves Schabes, Martin Zaidel, and Dania Egedi. "A Freely Available Wide Coverage Morphological Analyzer for English" In *Proceedings of COLING-92*, 1992.
- [31] Daniel Marcu. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*, Ph.D. Thesis (Toronto: Department of Computer Science, University of Toronto, December 1997)
- [32] Susan Herring, Deborah A. Johnson, Tamra DiBenedetto. "'This discussion is going too far!': Male resistance to female participation on the Internet" In K. Hall and M. Bucholtz (editors) *Gender Articulated: Language and the Socially Constructed Self* (New York: Routledge, 1995).
- [33] Joseph Dumit *Whose Brain Is This? PET Scans and Personhood in Biomedical America* (forthcoming; see also, <http://symptom.mit.edu>).
- [34] Thomas Erickson, David N. Smith, Wendy A. Kellogg, Mark Laff, John T. Richards, Erin Bradner "Socially Translucent Systems: Social Proxies, Persistent Conversation, and the Design of 'Babble'" In *Proceedings of CHI'99*, May 1999, Pittsburg, Pennsylvania.