Conceptual Recurrence Plots: Revealing Patterns in Human Discourse

Daniel Angus, Andrew Smith, and Janet Wiles, *Member*, *IEEE*

Abstract—Human discourse contains a rich mixture of conceptual information. Visualization of the global and local patterns within this data stream is a complex and challenging problem. Recurrence plots are an information visualization technique that can reveal trends and features in complex time series data. The recurrence plot technique works by measuring the similarity of points in a time series to all other points in the same time series and plotting the results in two dimensions. Previous studies have applied recurrence plotting techniques to textual data; however, these approaches plot recurrence using term-based similarity rather than conceptual similarity of the text. We introduce conceptual recurrence plots, which use a model of language to measure similarity between pairs of text utterances, and the similarity of all utterances is measured and displayed. In this paper, we explore how the descriptive power of the recurrence plotting technique can be used to discover patterns of interaction across a series of conversation transcripts. The results suggest that the conceptual recurrence plotting technique is a useful tool for exploring the structure of human discourse.

Index Terms—Concept map, recurrence, concept, plotting, conversation analysis, text analysis.

1 Introduction

Human discourse creates a rich and often complex narrative for the sharing of concepts between communicating participants. Similar concepts can be evoked using many different terms and it is the concepts that recur throughout the discourse even when different terms are used.

Visualization of a narrative can help reveal the patterns of communication within a conversation, such as the degree of engagement between conversation participants or the location of key concepts within the conversation. The identification and tagging of concepts within language utterances can be performed in several ways, but the resulting high-dimensional information space is often very complex even without considering the dynamics of change. Techniques such as Arc Diagrams [16], Recursive Pattern [10], and ThemeRiver [7] are very good at highlighting repeated sequences in structured data but are limited to displaying Boolean matches, representing information streams in a single dimension or have difficulty displaying unstructured data. An alternative solution is to consider the similarity of the concept feature vectors between any two utterances and to show the full $O(n^2)$ set of pairwise matches in a grid. By comparing in a pairwise fashion it can be shown when a set of concepts recur in the conversation, and the relationship can be visualized in two dimensions.

- *D. Angus and J. Wiles are with the School of Information Technology and Electrical Engineering, University of Queensland, Brisbane St Lucia, Queensland 4072, Australia. E-mail: d.angus@uq.edu.au, janetw@itee.uq.edu.au.*
- *A. Smith is with the School of Information Technology and Electrical Engineering and also with the Institute for Social Science Research, University of Queensland, Brisbane St Lucia, Queensland 4072, Australia. E-mail: andrew@leximancer.com.*

Manuscript received 24 June 2010; revised 27 Jan. 2011; accepted 5 June 2011; published online 13 June 2011.
Recommended for acceptance by S. Carpendale.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org, and reference IEEECS Log Number TVCG-2010-06-0127.
Digital Object Identifier no. 10.1109/TVCG.2011.100.

This two-dimensional pair-wise comparison is known as a recurrence plot, and it is a useful information visualization technique for identifying trends and features in complex time series data [5], [17]. Recurrence plot strategies can be applied to a variety of complex dynamic systems as a way of identifying trends that are otherwise difficult to interpret given their multidimensional nature. Previous application of recurrence plots to text data uses Boolean matches of specific terms to create term-based recurrence plots [3], [4], [8], [12]. In this paper, conceptual information in utterances is explored by creating and interpreting conceptual recurrence plots.

The paper begins by outlining existing text-based recurrence plot methods in Section 2. A conceptual similarity metric and the conceptual recurrence plotting technique are introduced in Section 3. A series of conversation transcripts are analyzed in Section 4 to demonstrate how these plots can be interpreted from a user's point of view, and to demonstrate the effects of different conceptual models on the composition of conceptual recurrence plots. Concluding remarks and future work are provided in Section 5.

2 Background

The recurrence plotting technique was introduced by Eckmann et al. [5] to display and identify trends within time series data from complex dynamical systems. Human discourse can be considered a complex dynamical system, and thus, it follows that these communications may contain a variety of recurrence structures. The analysis of such structures may then provide insight into the nature of the human discourse.

Previous approaches that use recurrence plots on text [3], [4], [8], [12] focus on displaying recurrence at a purely symbolic or term-based level, measuring recurrence of single words or symbols. At the term-based level, recurrence comparisons are Boolean, examples being $\mathrm{a} = \mathrm{a}$, $\mathrm{a} \neq \mathrm{b}$, $\mathrm{cat} = \mathrm{cat}$, and $\mathrm{cat} \neq \mathrm{dog}$, rather than at a conceptual

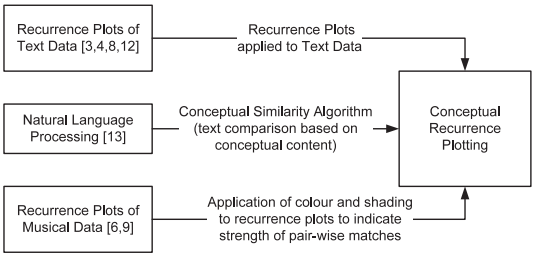

Fig. 1. Inspiration for the design of the conceptual recurrence plotting system.

level where *dog* may be similar to but not quite equal to *cat* in the context of both of these terms referring to domesticated animals. As an example, consider the following different yet conceptually similar sentences:

- We're going out to get some food. We'll be back soon.
- Me and the guys are heading into town to grab some dinner. See you later.

Using existing text-based recurrence approaches, these sentences would not show recurrence given that they don't share sequences of words. However, like the cat-dog example above, the statements are similar in a conceptual sense, and thus, at a higher level of abstraction, some recurrence is present. This distinction is important for human discourse analysis given that there is no guarantee that multiple participants will share a common vocabulary and thus may use different terms to mean the same semantic construct. If recurrence visualization is used to measure the degree of engagement between participants, term-based similarity as opposed to conceptual similarity may miss important engagements and critical points within a spoken dialogue.

Conceptual similarity measurements present a challenge to the creation of a recurrence plot, given that matches are not Boolean. Pairwise comparisons return a real value that can vary between an upper bound (absolute exact conceptual similarity) and a lower bound (no conceptual similarity). Many of the existing techniques for graphing recurrence information in a recurrence plot in both textual and numeric domains rely on the use of a dot-plot style recurrence plot. Such a dot-plot is comprised of individual pixels that, for text, indicate 1 if a word comparison is a match or 0 if it is a nonmatch. Dot plots are useful for identifying chains of repeated terms, however have limited utility when displaying conceptual similarity given that conceptual measurements vary on a continuous number line. Techniques for displaying partial matches in a recurrence plot exist in the field of music pattern identification [6], [9]. These music visualization techniques use recurrence plots for visualization of patterns in musical data and use shading and color to better differentiate between channels of information and to highlight partial matches. Music and natural language share many characteristics and as such these techniques are an excellent inspiration for the development of a conceptual recurrence plotting technique.

The conceptual recurrence plotting technique draws inspiration from a variety of visualization techniques, indicated in Fig. 1. Text-based recurrence techniques have successfully applied recurrence plots to text [3], [4], [8], [12], and we extend these approaches through the incorporation of a conceptual similarity calculation [13], in addition to incorporating color and shading [6], [9]. While the conceptual recurrence system reuses many components from existing visualization systems, this novel combination enables analyses of human discourse that are not possible using any of these individual components.

3 Methodology

3.1 Conceptual Similarity

The input text data are processed to build a semantic model that can be used to compare any two sections of text (utterances) for conceptual similarity. In theory, any conceptual similarity algorithm could be used to generate the semantic model, including but not limited to Leximancer [15], Latent Semantic Analysis [11], and Latent Dirichlet Allocation [2]. For this study, the technique chosen is taken from the work of Salton [13] which uses lexical statistics including term co-occurrence and term occurrence to build a semantic model. The algorithm and its use in determining the similarity of two utterances are described below.

A document (D) is processed to remove stop words and punctuation with sentence boundaries preserved. All unique terms in the processed document (D') are inserted into a term list (T) which has length T_m. D' is decomposed into a series of sentence windows that can vary from a single sentence to multiple sentences (for this study, we used a sentence window size of 3). The occurrence count of each individual term (t_i) is recorded in an occurrence vector (O) with duplicate terms within a sentence window being ignored. The co-occurrence matrix $\mathbf{C}$, records co-occurrence values ($c_{t_i t_j}$) that reflect how many times any pair of terms (t_i, t_j) co-occur in a single sentence window. For term-term co-occurrence calculation, repeated terms within a sentence window are treated as only appearing once within that particular set of sentences. The similarity $S(t_i, t_j)$ of term t_i to t_j can be calculated according to the semantic similarity model proposed by Salton [13, pg. 275], which uses a probabilistic model of term co-occurrence

$$
\begin{aligned}
S(t_i, t_j) &= \frac{P(t_i, t_j) \times P(\overline{t}_i, \overline{t}_j)}{P(\overline{t}_i, t_j) \times P(t_i, \overline{t}_j)} \\
P(t_i, t_j) &= C_{t_i t_j}/N \\
P(\overline{t}_i, \overline{t}_j) &= \begin{cases} 1, & \text{if } O_{t_i} + O_{t_j} = C_{t_i t_j} + N \\ (N - O_{t_i} - O_{t_j} & \\ \quad + C_{t_i t_j})/N, & \text{otherwise} \end{cases} \\
P(t_i, \overline{t_j}) &= \begin{cases} 1, & \text{if } O_{t_i} = C_{t_i t_j} \\ (O_{t_i} - C_{t_i t_j})/N, & \text{otherwise} \end{cases} \\
P(\overline{t_i}, t_j) &= \begin{cases} 1, & \text{if } O_{t_j} = C_{t_i t_j} \\ (O_{t_j} - C_{t_i t_j})/N, & \text{otherwise,} \end{cases}
\end{aligned}
\tag{1}
$$

where N is the number of sentence windows, O_{t_i} is the occurrence of t_i, and $C_{t_i t_j}$ is the co-occurrence of terms t_i and t_j.

Sentence windows are used to calculate semantic similarity but after this processing step these sentences are

$$\mathbf{V} = \mathbf{S} \times \mathbf{B}$$

$$\begin{array}{c} \\ \text{Key Terms} \end{array} \begin{array}{c} \text{Utterances} \\ \begin{bmatrix} v_{k_1u_1} & \cdots & v_{k_1u_p} \\ \vdots & \ddots & \vdots \\ v_{k_nu_1} & \cdots & v_{k_nu_p} \end{bmatrix} \end{array} = \begin{array}{c} \\ \text{Key Terms} \end{array} \begin{array}{c} \text{Terms} \\ \begin{bmatrix} s_{k_1t_1} & \cdots & s_{k_1t_m} \\ \vdots & \ddots & \vdots \\ s_{k_nt_1} & \cdots & s_{k_nt_m} \end{bmatrix} \end{array} \times \begin{array}{c} \\ \text{Terms} \end{array} \begin{array}{c} \text{Utterances} \\ \begin{bmatrix} b_{t_1u_1} & \cdots & b_{t_1u_p} \\ \vdots & \ddots & \vdots \\ b_{t_mu_1} & \cdots & b_{t_mu_p} \end{bmatrix} \end{array}$$

Fig. 2. Calculating concept vectors from term vectors. A concept vector for each utterance u_i is calculated using the semantic similarity of each key term K_n to the list of all terms T_m, and the occurrence of all terms in the utterance.

reorganized into groups of different sentences called utterances (U), where an utterance can range in size from a single sentence to many sentences, and sentences have membership to only one utterance. The total number of utterances is U_p.

The organization of sentences into utterances is domain specific, and for conversation text, sentences within contiguous blocks of speech by a single user are assigned together into single utterances. If the text was a single author text, paragraph boundaries could be used to define finite utterances rather than speaker turns. Conversation transcripts have an implicit ordering but often no other specific timing information; given this, each utterance is numbered in the order that it is spoken. Timing could be used as extra metadata in each utterance for later rendering in the recurrence plot, and could allow for the introduction of blanks for periods of silence, and block sizes proportional to the actual time spent talking. Without access to timing information, block size can be set proportional to the number of terms in the utterance (excluding stop words as they are removed) or it can be left uniform for all utterances.

The most frequent terms list (K) is constructed by taking a subset of K_n terms with the highest frequency of occurrence from the term list T. After constructing K, a matrix of similarity ($\mathbf{S}$) is constructed such that each element contains the similarity of a key term to a term in the term list. Thus, $\mathbf{S}$ has dimension $K_n \times T_m$. The document D' is also reduced to a Boolean matrix ($\mathbf{B}$) which indicates the presence of individual terms in each utterance (1 if a term is present, 0 if the term is not present), such that it's dimensionality is $T_m \times U_p$. The feature matrix, $\mathbf{V}$, is calculated by

$$\mathbf{V} = \mathbf{S} \times \mathbf{B}, \tag{2}$$

where $\mathbf{S}$ and $\mathbf{B}$ are as defined previously, and each column of this matrix (V_{*j}) reflects the similarity of utterance j to the most frequent terms (K) within document D'. Thus, the dimensionality of $\mathbf{V}$ is $K_n \times U_p$. The calculation of $\mathbf{V}$ is illustrated in Fig. 2.

The similarity of any two utterances is calculated by taking the dot product of any two columns of $\mathbf{V}$, for example: $V_{*1} \cdot V_{*2}$ measures the similarity of utterance 1 and 2. It is also possible to use a subset of the matrix elements to obtain a reduced dot product, for example: $V_{11} \cdot V_{12}$ measures the similarity of utterance 1 and 2, using only key term 1.

3.2 Conceptual Recurrence Plot

Given a pairwise conceptual similarity measurement, the construction of the concept recurrence plot is a straightforward procedure. The conceptual similarity of every utterance to every other utterance is measured and the results stored in an $U_p \times U_p$ matrix of pairwise conceptual similarity. The comparison is limited to one side of the diagonal given the symmetry of the comparison. If the resulting matrix is displayed by shading each element in the matrix according to its conceptual similarity score ($1 = \text{black}$, $0 = \text{white}$, with shades of gray between), a conceptual recurrence plot is obtained. These values can be scaled using a nonlinear transform to emphasize regions of high similarity and de-emphasize regions of low similarity.

Recurrence plots can be enhanced by using different colors for specific speakers, groups, types, or other categorizations within a conversation. For example, if a conversation contains two speakers, one can be tagged red and the other tagged blue. Then, each element of the recurrence plot can be colored in the speaker-specific color for any element that corresponds to a conceptual comparison between two utterances by this same speaker. If an element corresponds to a conceptual comparison between two different speakers, a third color or a gradient of color between the two speaker specific colors can be used (see Fig. 3).

3.3 Term-Based Recurrence Plot

Similar to measuring conceptual recurrence, the term-based recurrence of any two utterances can be calculated by a number of ways. Term-based recurrence is different to conceptual recurrence as it indicates the degree to which two utterances use the exact same terms, rather than conceptually similar terms [3], [4], [8], [12], and is included here as it is a useful baseline to compare with conceptual recurrence. One method to calculate term-based recurrence is to take the dot product of any two columns of $\mathbf{B}$, for example, $b_{*i} \cdot b_{*j}$ measures the term-wise similarity of utterance i and j. A simplification of this idea is to only use the rows of $\mathbf{B}$ that correspond to the terms in the Key Term Vector, K, in the calculation. Using this term-based pairwise similarity measurement means that the resulting recurrence plot provides an indication of the number of terms that any pair of utterances share. Similarly to the conceptual recurrence plot, utterances can be colored according to the speakers.

4 Analysis

Visualization is effective to the extent that it supports an analyst in their typical activities and extends their capabilities. Visualization can support decision making and learning by explaining confidence in perceived data relationships and enabling representational conclusions for decision support [1]. A series of case studies designed

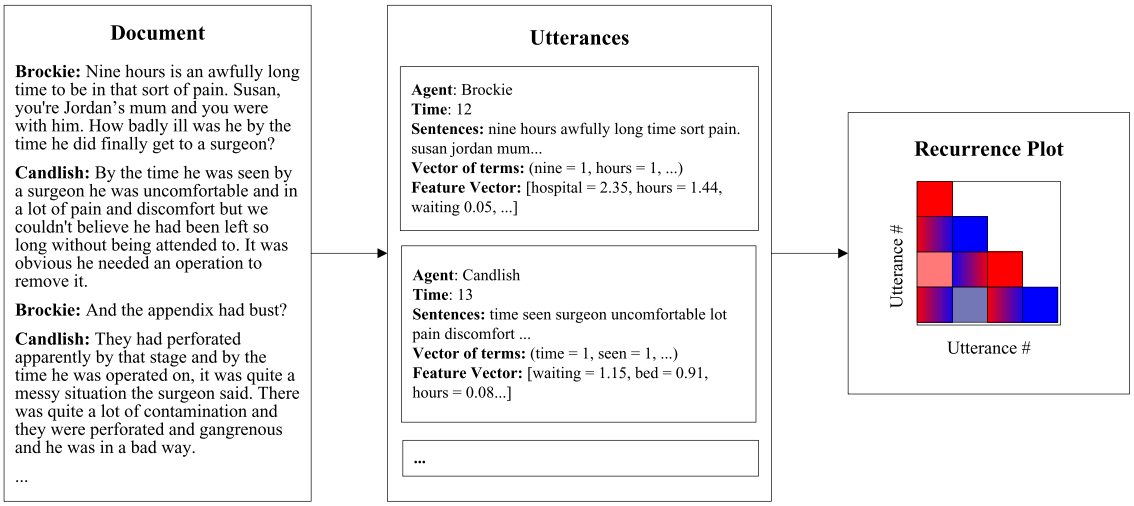

Fig. 3. Conceptual Recurrence Plot System Diagram. Text, in the form of a single document, is input into the system and broken into a series of finite utterances. These utterances are coded with conceptual information which can be used to compare any two utterances for conceptual similarity and thus construct the conceptual recurrence plot. Brockie's utterances are colored red; Candlish's are blue. Below the diagonal, a gradient between blue and red indicates similarity between Brockie and Candlish, and solid red or blue indicate self-similarity.

to explore how the conceptual recurrence system supports these goals are presented here.

The first study presents an overview of the visualization system and the basic functions available to users. The second study demonstrates the differences between using conceptual similarity and term-based similarity. The third and fourth studies highlight how major qualitative features can be identified using the conceptual recurrence system, how these features can be linked to speaker behavior, and how recurrence plots can be explored through the isolation of individual key concepts.

4.1 Data

The data sets used in studies 1, 3, and 4 are from the long running Australian television series *Enough Rope*, produced by Zapruder's Other Films Pty. Ltd. and broadcast by the publicly funded Australian Broadcasting Corporation (ABC). The program is hosted by Andrew Denton and consists of the host interviewing prominent celebrities including influential musicians, politicians, authors, actors, and members of the public who may have an interesting life story to tell. The interviews are interesting as texts for analysis of conversation as there are key points of most interviews when Denton gets interviewees to discuss topics that they would not normally share openly. It is also interesting as many interviewees have topics that they don't want to talk about and try to avoid questions, sometimes becoming openly aggressive or defensive.

The interviews are with ethicist and animal rights campaigner Professor Peter Singer[1] (used in studies 1, 3, and 4), and former politician and then spokesperson for depression awareness organization Beyond Blue, Jeff Kennett[2] (used in studies 3 and 4). The interview with Singer is characterized into several stanzas centered on different ethical issues including: food choices, abortion, personal choice, and sexuality. The Kennett interview was chosen due to Denton's commenting postinterview about how difficult the interview had been. In this interview, Denton indicated how he had wanted to discuss elements of Kennett's political career, whereas Kennett was only comfortable discussing his position as spokesperson for Beyond Blue. Denton tries to switch the conversation to politics at various points of the interview and in response Kennett becomes openly defensive, insisting that he only wishes to talk about Beyond Blue and depression.

The data used in study 2 are from the *Insight* program, a panel discussion program broadcast by an Australian independent national broadcaster, Special Broadcast Services (SBS). Insight consists of an audience of approximately 40 people discussing a controversial topic for 1 hour. The audience often includes parliamentarians, industry and academic experts, and other members of the public who are in some way affected by the issue being discussed. Brockie's interview style is to ask a question to a particular audience member, then use the response to frame a follow-up question to that same person or another member of the audience. The program included here is titled *Emergency*, and is on the topic of hospital emergency departments within Australia.[3]

4.2 Case Study 1: Overview

4.2.1 Aims

A fundamental task of a discourse analyst is to understand the pattern of interaction between participants [18]. The task is memory intensive for the analyst as there are $K_n U_p(U_p - 1)/2$ potential engagements, where K_n is the size of the conceptual vocabulary and U_p the total number of utterances. The challenge for a visualization is to represent the engagement patterns in a form that reduces memory load and supports exploration. The aim of the first study is to present the basic design and functionality of the visualization system and show how it supports investigation of the patterns of engagement.

1. Transcript available at: http://www.abc.net.au/tv/enoughrope/transcripts/s1213309.htm.

2. Transcript available at: http://www.abc.net.au/tv/enoughrope/transcripts/s1152967.htm.

3. Transcript available at: http://news.sbs.com.au/insight/episode/index/id/112#transcript.

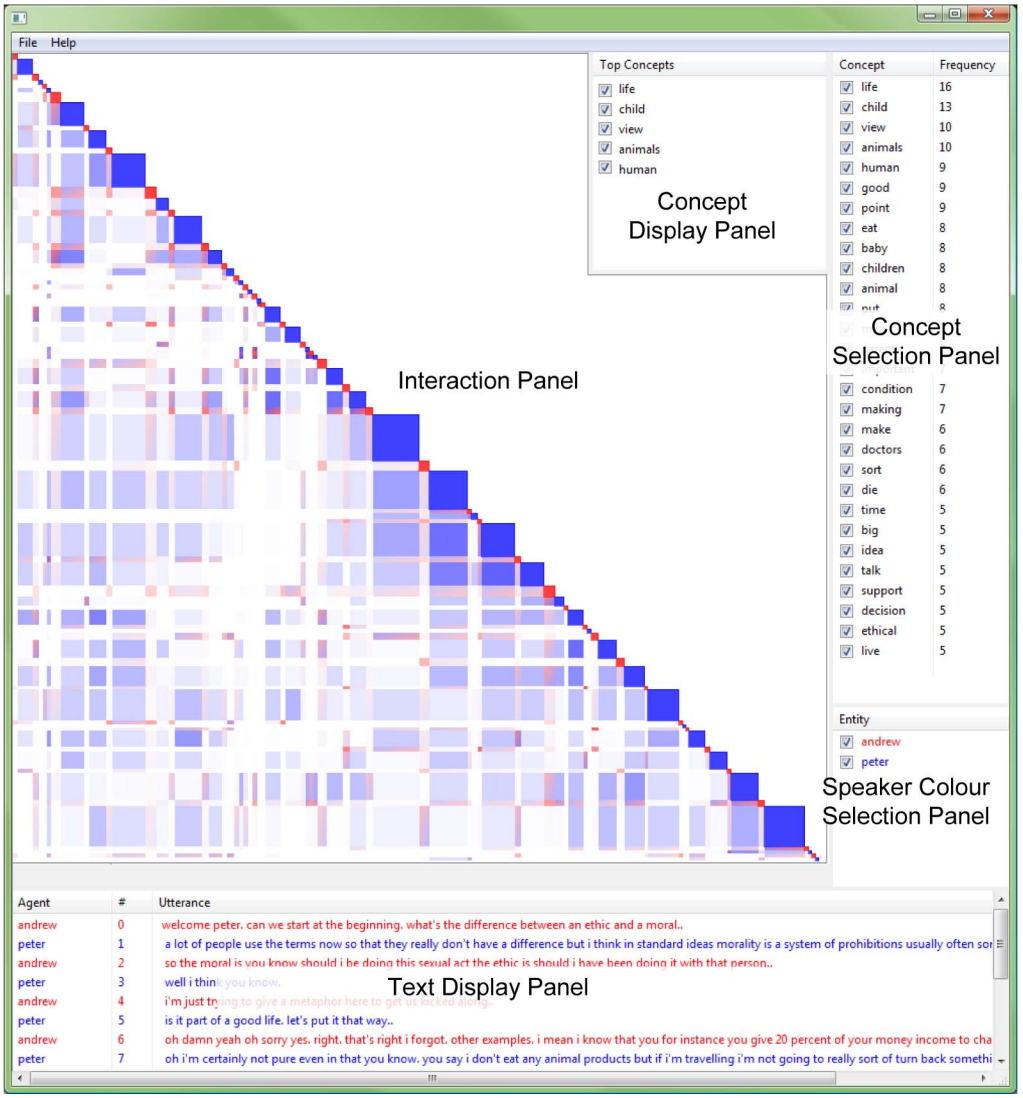

Fig. 4. The conceptual recurrence visualization system (*Denton/Singer*). In this default view, the user can see the entire pattern of recurrence at a glance in the Interaction Panel, with details displayed in the text panel below. The most highly recurring concepts are listed in the concept display panel. The rightmost concept selection panel can be used to select major concepts to see how these concepts are distributed throughout the entire text (by default all are selected).

4.2.2 Method and Results

The conceptual recurrence visualization system can support a range of data visualization and interrogation tasks. For this study, the entire text transcript of the Denton/Singer interview was used to create a semantic model and individual utterance feature vectors. The interview was an hour long with 74 turns (utterances) by each speaker.

The analysis system is designed to first present the user with an overview of the data, then to provide a series of interactions that enable the user to zoom and filter, enabling exploration of details in their own time (a design philosophy consistent with many information visualization systems [14]). A screenshot of the system in use on the Denton/Singer data set is included in Fig. 4. In Fig. 5, an example of selective filtering is demonstrated where

1. the user identifies an area of interest and selects this area of the recurrence plot via a drag box;
2. the system identifies the relevant utterances and displays those utterances' text below the plot;
3. shading is applied to the recurrence plot to better highlight the selected region's connecting utterances; and,
4. the top concepts from those utterances are isolated and displayed.

The rightmost panel is responsible for the selection of major concepts (Key Terms) that are used in measuring the conceptual similarity of any two utterances, a feature which is explored in detail in study 4. The default behavior uses proportionally sized utterances where the size of each block is proportional to the utterance length; however, a user can select uniform sizing in which case the original text is made visible next to each utterance as is done in Fig. 6. By presenting users with proportional sizing in the first instance, they are able to get a quick view of the turn taking dynamics of the conversation. The system supports zooming and the addition and removal of utterances. Coloring can be specified by the user in addition to the opacity of shading using nonlinear transforms.

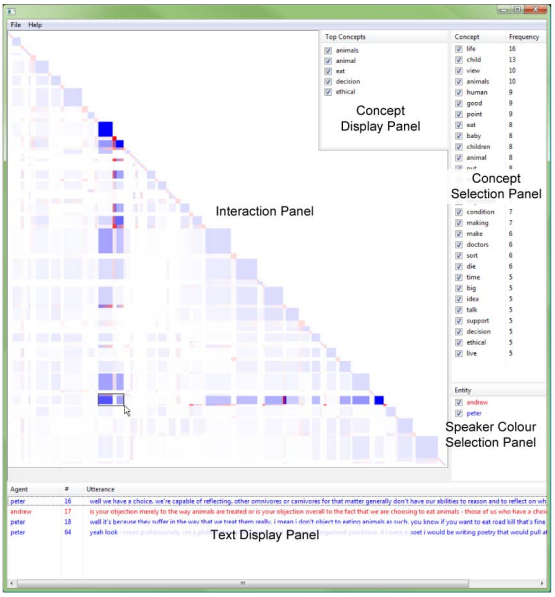

Fig. 5. Interacting with the visualization system (*Denton/Singer*). In this example, a user has selected a section of recurrence in the interaction panel and the utterances responsible for this recurrence are highlighted on the plot, their text is displayed in the text display panel, and the concept display panel has been updated to indicate the most highly recurring concepts in the selected utterances.

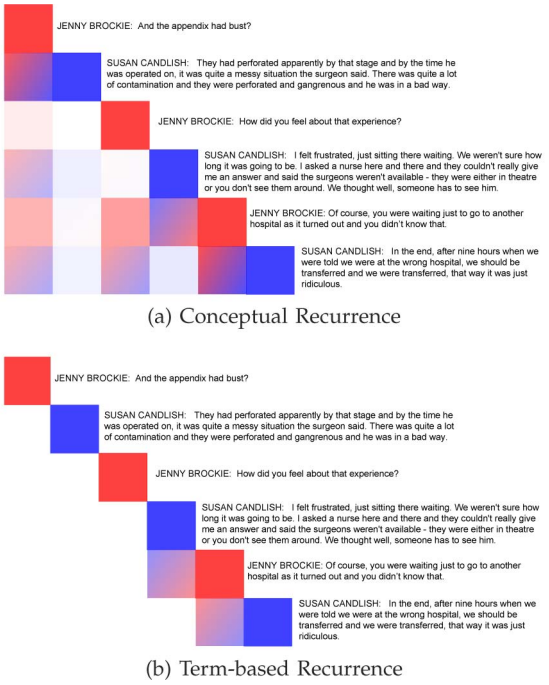

Fig. 6. Conceptual and term-based recurrence plots (*SBS Insight*). Jenny Brockie is indicated in red and an audience member is indicated in blue. Recurrence between the two participants is shown as a gradient between their respective colors and self-recurrence is their own color.

The conceptual recurrence visualization system supports discourse analysts in determining the accuracy of any existing hypotheses of input data by displaying recurrence patterns alongside the text and concepts which generated these patterns. As an example, if a discourse analyst believed that participants engaged strongly around a few key concepts in a particular interview, they can readily select these particular concepts and visualize where these concepts appear, and to whom the recurrence can be attributed. The visualization system also enables exploratory data analysis by displaying patterns that an analyst may not be aware of. All of the basic interactions listed in this section are designed to concretize relationships that exist within the input data and provide users with faithful representations of input data.

4.3 Case Study 2: Term versus Conceptual Similarity

4.3.1 Aims

A possible activity for a discourse analyst is to search a text for utterances by one or more participants that refer to similar topics [18]. The challenge for term-based similarity matching is that concepts may recur but different words may be used. We have addressed this challenge by using conceptual similarity rather than term-based similarity in the recurrence plot. The aim of the second study is to compare conceptual and term-based measures of similarity when used in the creation of recurrence plots.

4.3.2 Method and Results

The entire text transcript of the Insight Emergency program was used to create a semantic model and individual utterance feature vectors. A subset of six utterances were used to create recurrence plots to highlight the difference between the conceptual and term-based recurrence plots. Plots of conceptual and term-based recurrence are created, using uniform block sizing for each utterance.

When comparing conceptual and term-based recurrence plots (Fig. 6), the extra conceptual recurrence stemming from Brockie's question "And the appendix had bust?" is most apparent. The entire excerpt displayed is related to an audience member Candlish's son Jordan who had a burst appendix. In many of the utterances displayed, the specific term *appendix* doesn't appear, and as such on a purely term-based level Brockie appears to not recur with Candlish when asking the simple question "And the appendix had bust?" When conceptual information is used, this utterance appears to recur strongly with neighboring utterances due to the conceptual model linking terms such as: *surgeon*, *perforated*, and *contamination* to *appendix* and *bust*. The extra recurrence creates a near contiguous block of recurrence to highlight this interaction.

4.4 Case Study 3: Feature Identification

4.4.1 Aims

A possible activity of a discourse analyst is to determine patterns of interaction in a conversation that can be related to conversation participant behaviors [18]. Inspired by the work of Webber and Zbilut [17] on characteristic features of recurrence plots in physiology, the aim of study 3 is to identify characteristic features of a conceptual recurrence plot, and to determine what meaning they convey about the conversations being analyzed. The presence of such features has potential to enable the formation of hypotheses about conversation behaviors.

4.4.2 Method and Results

The entire text transcripts of each Denton interview were used to create semantic models and conceptual recurrence

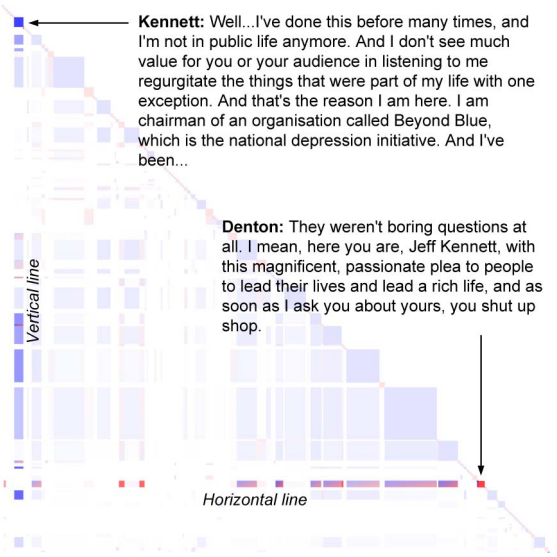

Fig. 7. Recurrence lines (*Denton/Kennett*).

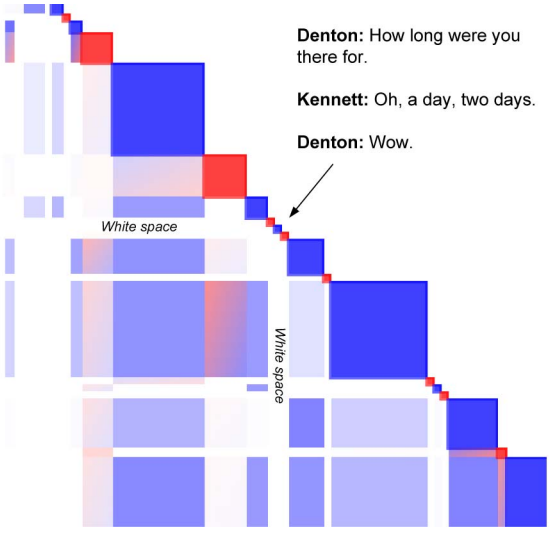

Fig. 8. White space (*Denton/Kennett*).

plots. The resultant plots were inspected manually to identify interesting features and these features were annotated by hand onto the recurrence plots. Block sizing proportional to the utterance length is used to highlight the magnitude of the contribution by each conversation participant throughout the interview. Recurrence is colored according to the interviewer and interviewee and shared recurrence is a gradient of color between the speakers. In the examples included for analysis only, the recurrence plot itself has been reproduced, but where relevant the top concept (or concepts) for particular utterance groupings has been reproduced.

1. Horizontal/vertical lines. A horizontal line appears when a single utterance has conceptual similarity to many other utterances backward in time; a vertical line is the same except that it recurs with utterances forward in time. These patterns imply that a single utterance contains subject matter that has conceptual similarity to previously or later occurring utterances. It is often the case that an utterance will be connected to both vertical and horizontal lines; such a statement would be seen as contributing to the main body of the discussion, whereas an utterance that is weighted toward more horizontal (backward) recurrence may be seen as being a closing statement, and one weighted more to vertical (forward) recurrence an opening statement. The positioning of these recurrence lines conveys meaning as to the nature of an utterance. Depending on the conversation genre, an early occurring utterance with a strong vertical line may indicate an influential statement that is responsible for framing much of the later conversation; a horizontal line late in the conversation may indicate a summarization statement, recapping previously discussed material.

In Fig. 7, a positioning statement by Kennett and an utterance by Denton with strong horizontal recurrence are highlighted. Kennett's utterance is interesting given the absence of recurrence for the first third of the interview. In this part of the interview, Denton began a line of questioning about Kennett's personal life, rather than discussing depression. It isn't until later in the interview that the conversation turns back to the issue of depression and at that point we can observe strong recurrence by Kennett with his initial statement. In this example, the vertical line can be used as a way of highlighting exactly when the interview began to return to the issues introduced in Kennett's early utterance. Denton's highlighted utterance has strong recurrence with earlier utterances by Kennett; in this statement, Denton is trying to convince Kennett to divulge details about his personal life which Kennett had resisted for most of the interview. Kennett had just given a long period of engaging and passionate speech about depression and its affect on people's lives and Denton is trying to bootstrap off this long period of engagement to encourage Kennett to talk about his own life.

2. Bands of white space. A band of white space appears when a single utterance has limited or no conceptual similarity to all other utterances forward and backward in time. An utterance that contains concepts that are not present anywhere else or has too little text to strongly connect it to any of the major concepts used for comparison of utterance pairs can generate white space in a horizontal and vertical direction.

In Fig. 8, the utterance identified contains a quick back and forth exchange between Denton and Kennett. The text contained in the utterances has no relationship to the major concepts contained throughout the interview. These white spaces can be a useful way to identify a break in the flow of a conversation and we can observe far more white space in the Kennett interview (Fig. 7) than in the Singer interview (Fig. 4). In the Kennett interview, white space punctuates topic changes or appears when Denton tries to move the interview toward Kennett's personal life.

3. Downward diagonals. Downward diagonals are an indication of repeated patterns of interaction distributed in time [17]. In the case of the conceptual recurrence plots analyzed here, we do not observe isolated diagonals, instead we observe a different pattern that we call an *Engagement Block*. An engagement block is a section of connected recurrence that is strongly adjacent to the diagonal. These blocks indicate strong engagement by one or both conversation participants around a particular set of concepts. Engagement blocks may be distinct or could overlap with other engagement blocks. In Fig. 9, Denton begins an engagement block by asking Kennett a question about his first days in power as the Victorian Premier, very

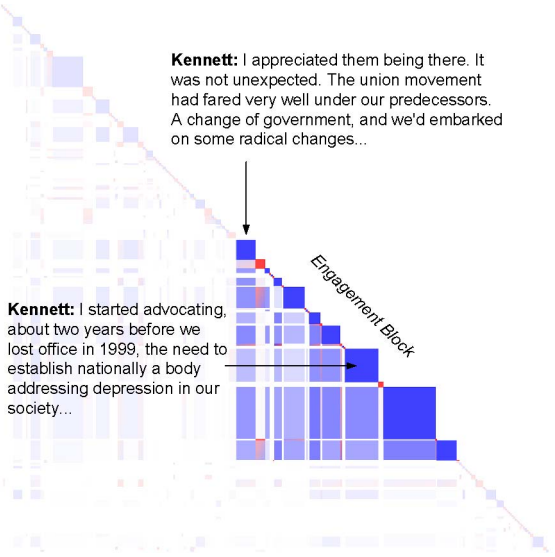

Fig. 9. Engagement block (*Denton/Kennett*).

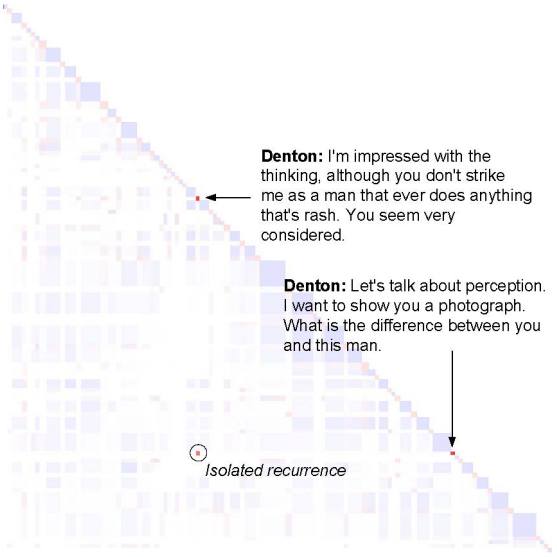

Fig. 10. Random scattering (*Denton/Singer*).

quickly the conversation turns to Kennett talking about depression and the recurrence plot shows a large engagement around these conceptually connected utterances.

4. Random scattering. Random scattering (isolated recurrence) are points of recurrence that are not connected to other recurrence patterns. Random scattering can be due to chance conceptual overlap between conversation participants rather than direct engagement between them. Put simply, if two sections of otherwise unconnected speech were spliced together, we would expect to see some amount of polysemous conceptual recurrence due to the presence of high-frequency terms. Random scattering is hard to isolate, but can be observed by looking for single points of recurrence that are surrounded by white space. In the Denton/Singer data set, a single point of recurrence was identified that linked Denton's statements at $\text{time} = 30$ & 62 (Fig. 10). For this example, the concepts *perception*, *considered*, and *thinking* link the two utterances; however, these are infrequent concepts and thus no larger recurrence patterns are observed elsewhere for these particular utterances. Random scattering as a characteristic feature is included for completeness with Webber and Zbilut [17], but in the texts we present here it has not played a significant role.

5. Nonuniform texture outward from diagonal (dark to light). Nonuniform texturing is a horizontal or vertical line that diminishes in intensity away from the diagonal. Nonuniform texturing is an indication of an utterance's importance increasing (in the horizontal) and diminishing (in the vertical) through time. In a normal conversation, temporally connected utterances are likely to recur, whereas utterances that are a long time before or after the current utterance are less likely to recur if the conversation progresses through multiple topics. If this drift is gradual, the recurrence pattern will appear as an utterance having bold recurrence within a short time period forward and/or backward, then slowly diminishing further away from the utterance.

In Fig. 11, a large engagement block is highlighted that includes discussion around the central concept of ethics. In this section of the interview, Denton quizzes Singer about a review he wrote for a book on bestiality, then segues from

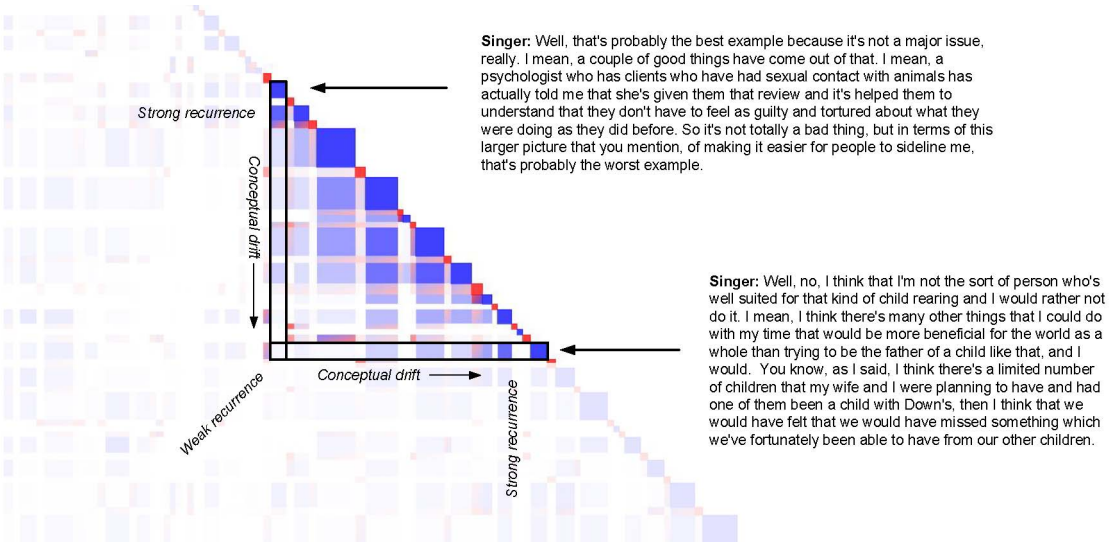

Fig. 11. Concept drift (*Denton/Singer*).

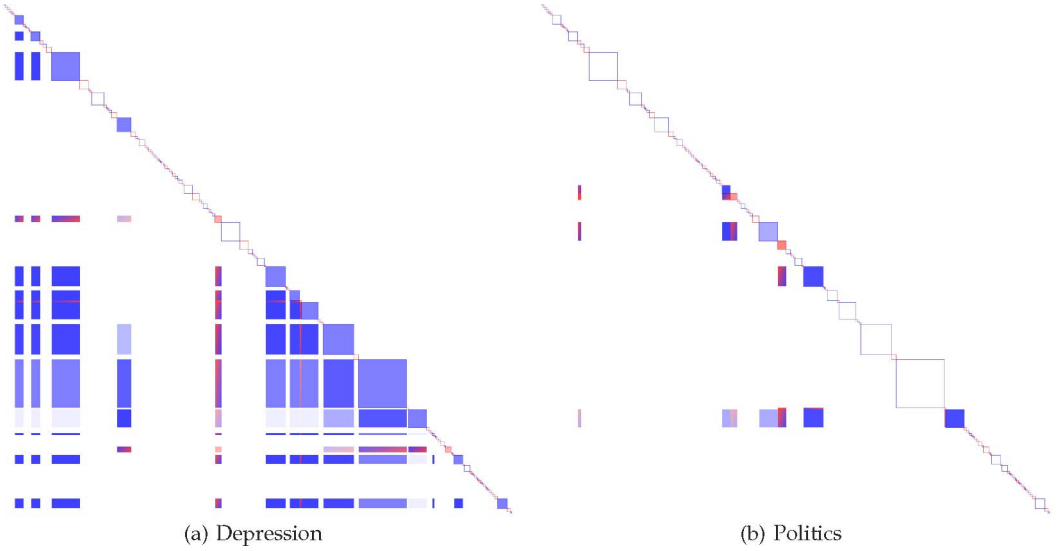

Fig. 12. Features of a difficult interview. Single concept recurrence plots (*Denton/Kennett*). Kennett is indicated in blue and Denton is indicated in red. (a) Conceptual recurrence plot where the concept *depression* is the only concept used to classify utterances. (b) Conceptual recurrence plot where the concept *politics* is the only concept used to classify utterances.

this topic, using questions about personal criticism that Singer has received for his views, to get Singer to comment on his controversial views on abortion. The early utterance identified shows progression in the vertical direction indicating that the conversation is progressing away from the concepts contained in this utterance as it continues forward; the later utterance shows concept progression in the horizontal direction indicating that the conversation is progressing toward the concepts contained within it.

4.5 Case Study 4: Concept Selection

4.5.1 Aims

Discourse has a high dimensionality due to the number of concepts that are raised in a conversation. In some activities, the discourse analyst needs to isolate interaction between participants around one or a small set of concepts [18]. The aim of the fourth and final study is to explore how the conceptual recurrence system supports multivariate exploration through the use of single concept recurrence plots.

4.5.2 Method and Results

The entire text transcript of the Denton/Kennett interview was used to create a semantic model. The semantic model was created using 50 key concepts, and the single concepts *depression* and *politics* were isolated for analysis. The conceptual recurrence plots use a block size proportional to utterance length. The shading of the diagonal blocks indicates the degree to which the individual utterances match each of the selected concepts.

Denton commented postinterview that he had wanted Kennett to talk about his political past, but that Kennett was most interested in discussing depression. By constraining the utterance feature vectors (V) to the individual concepts *depression* (Fig. 12a) and *politics* (Fig. 12b), we can observe these concepts throughout the conversation and how they recur with surrounding utterances.

In the single concept recurrence plots, Kennett is observed to talk at length around the topic of depression, engaging in a long discussion around this key concept; however, on the topic of politics, the amount of recurrence is severely limited. By looking along the diagonal, it is observed that *depression* was an early mentioned concept (Fig. 12a) in this interview; however, rather than using the *depression* concept to frame the remainder of the interview, Denton changed topic which can be seen by the horizontal recurrence from these early utterances being punctuated by a large band of white space.

5 Discussion and Future Work

Conceptual recurrence plots are a useful visualization technique to aid the analysis of text data. The studies presented in this paper illustrated how the conceptual recurrence visualization system could assist in qualitative assessment of conversation transcripts. The analyses demonstrated several characteristic recurrence patterns and offered interpretations to relate these patterns to conversational participants behaviors. The case studies highlighted the advantages of using conceptual similarity in addition to term-based similarity when identifying recurrence patterns within natural language utterances.

The conceptual recurrence visualization system is a decision-support tool, and a discourse analyst can use the system to confirm preheld hypotheses about the type and magnitude of interaction between conversation participants or as a forensic tool to discover patterns of interaction. In this study, the conceptual recurrence technique was used to: identify the major concepts used within a conversation and how participants interact over the course of a conversation (Study 1); identify areas of recurrence not detected by simple term-based matching techniques (Study 2); link key features including engagement blocks, horizontal lines, white space, and conceptual progression to interesting conversation participant behavior (Study 3); and, limit utterance comparisons to individual concepts to identify the distribution and recurrence of these concepts throughout a text (Study 4).

The conceptual similarity algorithm used in this study is a simple, yet effective algorithm for detecting conceptual overlap between conceptually related texts. The design of the visualization system allows this algorithm to be readily

exchanged with different conceptual similarity algorithms if so required. Such a modification may be useful for the analysis of data sets where a background corpus is available or for short texts where co-occurrence data are limited. Future implementations of the system could also incorporate classification technologies to detect and categorize recurrence patterns such as engagement blocks, horizontal lines, white space, and conceptual progression. Automated classification could also be used as the basis for quantitative metrics that measure the degree of recurrence between participants at various time scales. While the example data sets analyzed in this paper are of human discourse, it is straightforward to apply the technique to single author texts by using fragments of the text such as paragraphs for each utterance.

The data sets used in this paper are of length between 75 and 175 utterances, and illustration of the technique was limited to this number for the visualizations in a printed publication. It is possible to handle larger numbers of utterances through the use of zooming or feature detection to selectively highlight areas of interest in a larger data set and future works will explore these issues. The number of agents that can be represented is limited by the number of easily distinguishable colors and in future implementations this issue could be addressed by grouping agents using a single color or limiting the number of agents that are displayed at any one time. The system could also be augmented with back channel information (pauses, murmurs, etc.) to reflect nonverbal interaction.

The possible uses of the tool are far reaching and we have been working with discourse analysts in areas including medical consultation analysis, cockpit communications, online forum discussions, and professional interviews. The intended scope of the work we present here included how the technique works, how it can be used, and what value it adds for an analyst. Future works will explore each domain in more detail, as each domain has genre specific questions and requires targeted exploration to link recurrence features with behavioral qualities in each of these particular contexts. Access to domain specialists will allow future works to explore the usefulness of the conceptual recurrence plot when compared to other visualization and analysis strategies.

Acknowledgments

The authors thank Cindy Gallois, Faculty of Social and Behavioural Sciences, University of Queensland. This work was funded by the Australian Research Council Thinking Systems Grant TS0669699.

References

[1] R. Amar and J. Stasko, "A Knowledge Task-Based Framework for Design and Evaluation of Information Visualizations," *Proc. IEEE Symp. Information Visualization (InfoVis '04),* 2004.

[2] D.M. Blei et al., "Latent Dirichlet Allocation," *J. Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

[3] K.W. Church and J.I. Helfman, "Dotplot: A Program for Exploring Self-Similarity in Millions of Lines of Text and Code," *J. Computational and Graphical Statistic,* vol. 2, no. 2, pp. 153-174, 1993.

[4] R. Dale and M.J. Spivey, "Unraveling the Dyad: Using Recurrence Analysis to Explore Patterns of Syntactic Coordination between Children and Caregivers in Conversation," *Language Learning,* vol. 56, no. 3, pp. 391-430, 2006.

[5] J.-P. Eckmann et al., "Recurrence Plots of Dynamical Systems," *Europhysics Letters,* vol. 5, pp. 973-977, 1987.

[6] J. Foote, "Visualizing Music and Audio Using Self-Similarity," *Proc. Seventh ACM Int'l Conf. Multimedia,* pp. 77-80, 1999.

[7] S. Havre et al., "Themeriver: Visualizing Theme Changes over Time," *Proc. IEEE Symp. Information Visualization (InfoVis '00),* pp. 115-123, 2000.

[8] J.I. Helfman, "Similarity Patterns in Language," *Proc. IEEE Symp. Visual Languages,* pp. 173-175, Oct. 1994.

[9] V.K. Jacek Wolkowicz and S. Brooks, "Midivis: Visualizing Music Pieces Structure via Similarity Matrices," *Proc. Int'l Computer Music Conf. (ICMC '09),* 2009.

[10] D. Keim et al., "Recursive Pattern: A Technique for Visualizing Very Large Amounts of Data," *Proc. IEEE Conf. Visualization (Visualization '95),* pp. 279-286, 1995.

[11] T. Landauer et al., "Introduction to Latent Semantic Analysis," *Discourse Processes,* vol. 25, pp. 259-284, 1998.

[12] F. Orsucci et al., "Orthographic Structuring of Human Speech and Texts: Linguistic Application of Recurrence Quantification Analysis," *Int'l J. Chaos Theory and Applications,* vol. 4, nos. 2/3, pp. 21-28, 1999.

[13] G. Salton, *Automatic Text Processing : The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley, 1989.

[14] B. Shneiderman, "Inventing Discovery Tools: Combining Information Visualization with Data Mining," *Information Visualization,* vol. 1, no. 1, pp. 5-12, 2002.

[15] A.E. Smith and M.S. Humphreys, "Evaluation of Unsupervised Semantic Mapping of Natural Language with Leximancer Concept Mapping," *Behavior Research Methods,* vol. 38, no. 2, pp. 262-279, 2006.

[16] M. Wattenberg, "Arc Diagrams: Visualizing Structure in Strings," *Proc. IEEE Symp. Information Visualization (InfoVis '02),* pp. 110-116, 2002.

[17] C.L. Webber and J.P. Zbilut, "Dynamical Assessment of Physiological Systems and States Using Recurrence Plot Strategies," *J. Applied Physiology,* vol. 76, no. 2, pp. 965-973, 1994.

[18] S. Yates et al., *Discourse as Data: A Guide for Analysis.* Sage Publications, 2001.

Daniel Angus received the BS/BE double degree in research and development, and electronics and computer systems, and the PhD degree in computer science from Swinburne University of Technology, in 2004 and 2008, respectively. He is currently a postdoctoral researcher with the Thinking Systems Project at the University of Queensland. His research focuses on the development of visualization and analysis methods for conceptual data, with a specific focus on human discourse.

Andrew Smith received the BSc(Hons) and PhD degrees in physics in 1987 and 1993, respectively, and the MPhil degree in information science in 2002, from the University of Queensland. He is the creator of Leximancer, a software application for text analysis. He is currently Leximancer chief scientist, adjunct researcher within the School of Information Technology and Electrical Engineering at UQ, and a senior research officer within the Institute for Social Science Research (ISSR) in the UQ Faculty of Social and Behavioural Sciences. His research focuses on developing new ways to visualize and quantify the temporal dynamics of communication.

Janet Wiles received the BSc Hons I and the PhD degrees in computer science from The University of Sydney in 1983 and 1989, respectively. She is a professor of Complex and Intelligent Systems in the School of Information Technology and Electrical Engineering at The University of Queensland, and director of the Thinking Systems Project. Her research program involves using computational modeling to understand complex systems with particular applications in biology, neuroscience, and cognition. She is a member of the IEEE.