International Journal of Social Research Methodology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tsrm20

Making sense of big text: a visual-first approach for analysing text data using Leximancer and Discursis

Daniel Angus [a], Sean Rintel [b] & Janet Wiles [c]

[a] School of Information Technology and Electrical Engineering, School of Journalism and Communication, The University of Queensland, Brisbane, Australia

[b] School of Journalism and Communication, The University of Queensland, Brisbane, Australia

[c] School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia

Version of record first published: 03 Apr 2013.

To cite this article: Daniel Angus , Sean Rintel & Janet Wiles (2013): Making sense of big text: a visual-first approach for analysing text data using Leximancer and Discursis, International Journal of Social Research Methodology, 16:3, 261-267

To link to this article: http://dx.doi.org/10.1080/13645579.2013.774186

International Journal of Social Research Methodology, 2013
Vol. 16, No. 3, 261–267, http://dx.doi.org/10.1080/13645579.2013.774186

Routledge
Taylor & Francis Group

Making sense of big text: a visual-first approach for analysing text data using Leximancer and Discursis

Daniel Angus[a], Sean Rintel[b]* and Janet Wiles[c]

[a]School of Information Technology and Electrical Engineering, School of Journalism and Communication, The University of Queensland, Brisbane, Australia; [b]School of Journalism and Communication, The University of Queensland, Brisbane, Australia; [c]School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia

(*Received 30 November 2012; final version received 5 February 2013*)

This article reports on Leximancer and Discursis, two visual text analytic software tools developed at the University of Queensland. Both analyse spatial and temporal relationships in text data, but in complementary ways: Leximancer focuses on thematic analysis, while Discursis focuses on sequential analysis. Our report explains how they work, how to work with them and how visual concepts are relevant to all stages of their use in analytic decision-making.

Keywords: text; analytics; visualization; CAQDAS; big data; concept map; Leximancer; Discursis

1. Introduction

Scientific findings rely on structured processes that support, examine or test theories in the light of different types of evidence. Computational techniques in quantitative content analysis have been used since the 1950s (Krippendorff, 2012). Qualitative computer-aided discourse analysis arrived later, but since the 1990s computer-aided qualitative discourse analysis software (CAQDAS) applications have provided the means for semi-automated analysis of conversation, interview, mass media and new media data (Fielding & Warnes, 2009; Krippendorff, 2012; Ronen Feldman, 2006; Schönfelder, 2011; Seale, 2010; Smith, 2000). In an almost parallel timeline, the field of information visualisation has developed ways to make visual sense of relationships within data-sets (Tufte, 2001). We are now seeing the merging of text analysis and visualisation in visual text analytics (Risch, Kao, Poteet, & Wu, 2008), techniques that visually model text data for interpretation by a researcher. In this introduction to Leximancer and Discursis, two new-generation visual text analytic applications developed at the University of Queensland, we demonstrate their capacity to dramatically widen the scope of analyses. As with many new tools, they introduce new requirements into the workflow, and we also describe the concomitant attention to visuality in the research workflow that underlies their effective use.

Both Leximancer and Discursis use word frequency statistics to generate their respective visualisations, however, they facilitate distinct but complementary ana-

*Corresponding author. Email: s.rintel@uq.edu.au lytic tasks. Common tasks that a researcher can perform using Leximancer include: determining the main topics within a text; highlighting how topics relate to each other; and indicating which source files (or individual authors/speakers) contain particular topics. In contrast, Discursis facilitates tasks that include: determining how topics are used over time (global structure); finding critical time points where topic changes occur; and dividing a conversation/text into regions of interest.

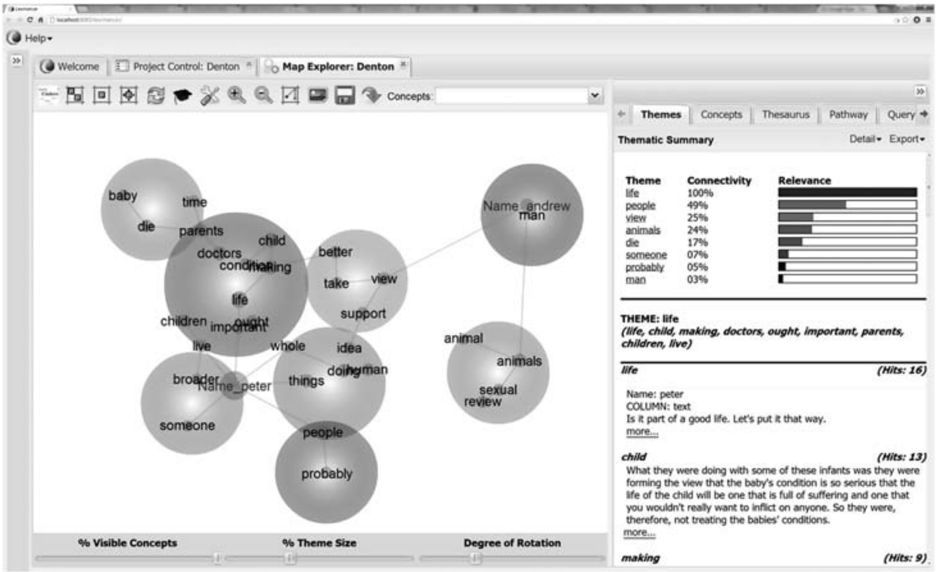
Well-established CAQDAS applications such as NVivo, Atlas.ti and MAXQDA share with Leximancer and Discursis the ability to produce visual representations of connections (see overviews such as Fielding & Warnes, 2009; Seale, 2010). However, Leximancer and Discursis highlight a visual approach to analysis which changes the way text is used as evidence and forms the basis for decision-making. CAQDAS applications are an invaluable tool for semi-automating, flexibly organising and presenting the results of open, axial and selective coding of grounded theory (Glaser & Strauss, 1967; Moghaddam, 2006) and allied qualitative social research methods. However, their workflow is effectively still textual. Automation helps researchers find materials to code and then iteratively split, combine and refine, but researchers do the work of metaphorically handling the text to create themes for discussion. Visuality tends to be relevant mainly for display of the researcher's themes.

Leximancer and Discursis both automatically generate textual relationships that are brought into focus for discussion through visual representations. The task of the researchers is to make sense of the relationships so displayed, rather than defining the relationships. Researchers can look for ways to split, combine and refine the themes that are made apparent. Critically, this is a process of adjusting the software's automation rather than a process of manual arrangement using software. The automated representation does not prevent the researcher from delving into that text to explore and explain the research problem. Rather, it draws attention to value of the visual for researchers as more than simply end-stage output for readers.

2. Leximancer

Instead of requiring analysts to iteratively design lists of concepts and codes, Leximancer generates its own lists and relationships based on the input text. An advantage of generating the concept list automatically is that the list is statistically reliable and reproducible, being generated from the input text itself, whereas manual lists require checks for coding reliability and validity. Additionally, subtle or unusual relationships may be more likely to emerge using automated concept list creation. Leximancer uses word occurrence and co-occurrence counts to extract major thematic and conceptual content directly from an input text. This automated process generates a tailored taxonomy which can be displayed graphically via an interactive concept map, or as tables indicating key concepts and conceptual relationships.

Leximancer supports a visualisation process that enables an analyst to examine concepts in the original text linked to a global perspective of the entire data-set provided by the automatically generated concept map: a typical workflow could first generate a concept map, initially viewing only the largest theme. Then gradually increase the map resolution, revealing additional concepts and noting how they a linked through the spanning tree (also automatically generated). Finally, examine the relative ranking of concepts, and drill down to show how pairs of concepts are used in the original text through the linking buttons.



Theme	Connectivity	Relevance
life	100%	
people	49%	
view	25%	
animals	24%	
die	17%	
someone	07%	
probably	05%	
man	03%	

THEME: life
(life, child, making, doctors, ought, important, parents, children, live)

life ***(Hits: 16)***

Name: peter
COLUMN: text
Is it part of a good life. Let's put it that way.
more...

child ***(Hits: 13)***

What they were doing with some of these infants was they were forming the view that the baby's condition is so serious that the life of the child will be one that is full of suffering and one that you wouldn't really want to inflict on anyone. So they were, therefore, not treating the babies' conditions.
more...

making ***(Hits: 9)***

Figure 1. Example Leximancer plot.

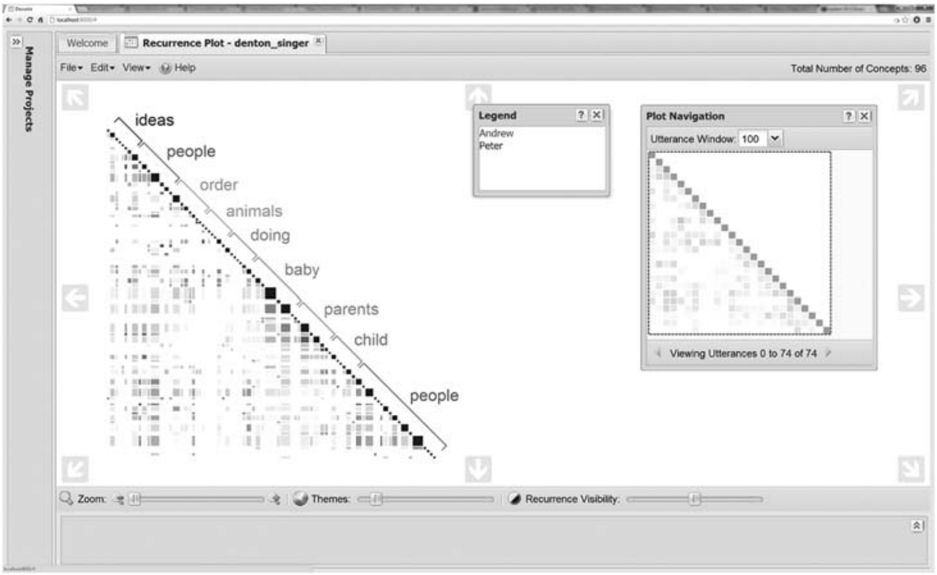
Leximancer has been used previously to analyse opinion polling and political commentary (McKenna, 2007); to evaluate incident reporting in a maritime environment (Grech, Horberry, & Smith, 2002); and to explore communication strategies employed by care providers of persons with schizophrenia (Cretchley, Gallois, Chenery, & Smith, 2010). Many more examples of the application of Leximancer are available from the corporate website[1].

The Leximancer plot (Figure 1) was generated from a transcript of the ABC *Enough Rope* Programme that aired on Australian television on 4 October 2004. The conversation is between the host Andrew Denton and interviewee Prof. Peter Singer. The Leximancer map highlights the prominent concepts discussed during the episode, and how each conversational agent relates to those concepts (how much they use them). Nodes represent individual concepts with the size of the node reflecting a measure of the prominence of the concept in the input text. Nodes are grouped according to similarity with other concepts, and connecting lines added to those concepts that share the strongest conceptual similarity. Large coloured circles group concepts into themes.

3. Discursis

Discursis is an information visualisation technique that produces informative visualisation of input text that has an inherent temporal structure, for example, conversation transcripts. Discursis automatically builds an internal language model from an input text (using a statistically based concept engine similar to Leximancer), tags each temporal unit (a single turn in the context of a conversation) based on the conceptual content and generates an interactive visual representation of the input text by linking similar temporal units. The Discursis visual representation enables an analyst to quickly overview an entire text and see at a glance the turn dynamics (who speaks when and for how long), the thematic content of the text over time, and regions of thematic coherence over short-time (turn-by-turn), medium-time (10 temporal units) and long-time (whole conversation) scales. Discursis is useful for locating periods of conversation where participants engage in similar topics or repeat their own content, in addition to periods that lack topical coherence. It then facilitates detailed examination of regions of interest.

A significant advantage of Discursis over alternative visualisation techniques is the ability to visualise topic usage patterns across a range of time scales



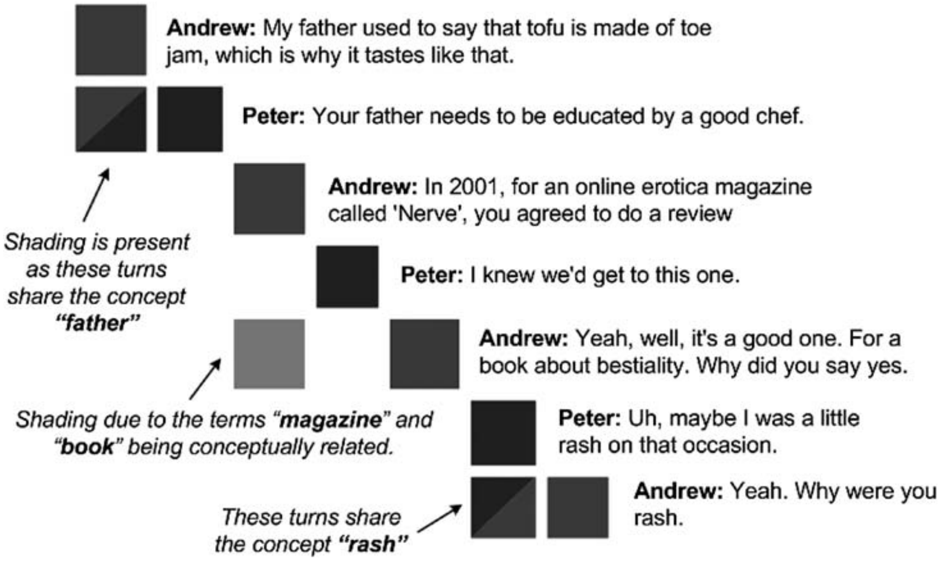
Total Number of Concepts: 96

Andrew
Peter

Utterance Window: 100

Viewing Utterances 0 to 74 of 74

Figure 2. Example Discursis plot – full overview.



Andrew: My father used to say that tofu is made of toe jam, which is why it tastes like that.

Peter: Your father needs to be educated by a good chef.

Andrew: In 2001, for an online erotica magazine called 'Nerve', you agreed to do a review

Peter: I knew we'd get to this one.

Andrew: Yeah, well, it's a good one. For a book about bestiality. Why did you say yes.

Peter: Uh, maybe I was a little rash on that occasion.

Andrew: Yeah. Why were you rash.

*Shading is present as these turns share the concept **"father"***

*Shading due to the terms **"magazine"** and **"book"** being conceptually related.*

*These turns share the concept **"rash"***

Figure 3. Example Discursis plot – zoomed.

simultaneously. Recent studies (Angus, Smith, & Wiles, 2012a, 2012b; Angus, Watson, Smith, Gallois, & Wiles, 2012) have analysed conversations from Australian television talk shows, medical consultations, cockpit recordings and telephone conversations, to understand topic convergence (on a turn-by-turn time scale) between the participants, and whole conversation call-back behaviours (where early mentioned topics were revisited much later in the conversation). The studies indicated that a discourse analyst could use the visuals to confirm theoretically motivated hypotheses about the type and magnitude of interaction (in terms of topic reuse) between conversation participants, and as a forensic tool to discover patterns of interaction and interesting time periods where conversation participants demonstrated topic convergence characteristics.

The Discursis plots (Figures 2 and 3) were generated using the same data-set as Figure 1. The light grey (normally red) blocks correspond to utterances by host Andrew Denton, the dark grey (normally blue) blocks to utterances by interviewee Peter Singer. Blocks are sized according to the amount of text in each utterance and off-diagonal blocks correspond to conceptual consistency between pairs of utterances on the diagonal. In the zoomed section of the Discursis plot (Figure 3), annotations highlight the salient visual features together with the original text.

4. Conclusion and future directions

Social research methodology that uses visual-first analytic methods is still in its infancy, but tools such as Leximancer and Discursis are powerful techniques for developing evidence-based global analyses that would otherwise tax the cognitive abilities of an analyst. This report has drawn attention to the value of visual analytics in existing studies, and shown that the field has already begun to shift from simply using the visual as an end-stage output for readers, to one which can be tied into critical reasoning and decision-making about study data.

The future direction for visual text analytics is to expand the number of theoretical frameworks that can be enhanced through the use of evidence-based visuality. This is primarily a kind of new operationalisation of evidentiary procedures. One of the more interesting likely requirements and outputs of such operationalisation will be a library or taxonomy of visual motifs that can be drawn upon to help researchers make visual sense of their data. Such motifs would allow researchers from very disparate fields to find patterns of relevance, both expected and unexpected. Over time, the motifs could either allow for rapid re-evaluation of new data and to improve the cumulative relevance of qualitative findings.

Lastly, and most importantly, we reiterate that techniques such as Leximancer and Discursis are not designed to replace the role of the human analyst in the social sciences; rather they are tools to help analysts perform and draw greater insight from their data. The workflows required for such advances are an ongoing research agenda.

Note

[1] http://www.leximancer.com

Notes on contributors

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, Melbourne, Australia, in 2004 and 2008, respectively. He is currently a lecturer in the School of Information Technology and Electrical Engineering and School of Journalism and Communication, The University of Queensland, Brisbane, Australia. His research focuses on the development of text analytic techniques, with a specific focus on techniques for analysing human discourse.

Sean Rintel received his BA (Hons 1) and MA in English from The University of Queensland in 1995 and 2000, respectively, and PhD in Sociology with a specialisation in Communication from the University at Albany, SUNY, in 2010. He is currently a lecturer in the School of Journalism and Communication at The University of Queensland, Brisbane, Australia. His research focuses on how the affordances and constraints of communication technologies interact with language, social action and culture.

Janet Wiles received the BSc (Hons I) and the PhD degrees in computer science from The University of Sydney, Sydney, Australia, 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, Brisbane, Australia, and Director of the Thinking Systems Project. Her research programme involves using computational modelling to understand complex systems with particular applications in biology, neuroscience and cognition.

References

Angus, D., Smith, A., & Wiles, J. (2012a). Human communication as coupled time series: Quantifying multi-participant recurrence. *IEEE Transactions on Audio, Speech, and Language Processing, 20*, 1795–1807. doi: 0.1109/TASL.2012.2189566

Angus, D., Smith, A. E., & Wiles, J. (2012b). Conceptual recurrence plots: Revealing patterns in human discourse. *IEEE Transactions on Visualization and Computer Graphics, 18*, 988–997. doi: 10.1109/TVCG.2011.100

Angus, D., Watson, B., Smith, A., Gallois, C., & Wiles, J. (2012). Visualising conversation structure across time: Insights into effective doctor–patient consultations. *PLoS ONE, 7*, e38014. doi: 10.1371/journal.pone.0038014

Cretchley, J., Gallois, C., Chenery, H., & Smith, A. (2010). Conversations between carers and people with schizophrenia: A qualitative analysis using Leximancer. *Qualitative Health Research*. doi: 10.1177/1049732310378297

Fielding, N., & Warnes, R. (2009). Computer-based qualitative methods in case study research. In D. Byrne & C. C. Ragin (Eds.), *The Sage handbook of case-based methods* (pp. 270–288). Los Angeles, CA: Sage.

Glaser, B. G., & Strauss, A. (1967). *Discovery of grounded theory: Strategies for qualitative research*. Mill Valley, CA: Sociology Press.

Grech, M. R., Horberry, T., & Smith, A. (2002). Human error in maritime operations: Analyses of accident reports using the Leximancer tool. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 46*, 1718–1721. doi: 10.1177/154193120204601906

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology* (3rd ed.). Thousand Oaks, CA: Sage.

McKenna, B. (2007). Mediated political oratory following terrorist events: International political responses to the 2005 London bombing. *Journal of Language and Politics, 6*, 377–399.

Moghaddam, A. (2006). Coding issues in grounded theory. *Issues in Educational Research, 16*, 52–66.

Risch, J., Kao, A., Poteet, S., & Wu, Y. (2008). Text visualization for visual text analytics. In S. Simoff, M. Böhlen & A. Mazeika (Eds.), *Visual data mining* (Vol. 4404, pp. 154–171). Berlin: Springer.

Ronen Feldman, J. S. (2006). *The text mining handbook: Advanced approaches in analyzing unstructured data*. New York, NY: Cambridge University Press.

Schönfelder, W. (2011). CAQDAS and qualitative syllogism logic – NVivo 8 and MAX-QDA 10 compared. *Qualitative Social Research, 12*. Retrieved from: http://www.qualitative-research.net/index.php/fqs/article/viewArticle/1514

Seale, C. (2010). Using computers to analyse qualitative data. In D. Silverman (Ed.), *Doing qualitative research: A practical handbook* (pp. 251–267). London: Sage.

Smith, A. E. (2000, December). *Machine mapping of document collections: The Leximancer system*. Paper presented at the proceedings of the Fifth Australasian Document Computing Symposium, Sunshine Coast, Australia.

Tufte, E. R. (2001). *The visual display of quantitative information* (Vol. 2). Cheshire: Graphics Press.