How few is too few? Determining the minimum acceptable number of LSA dimensions to visualise text cohesion with Lex

Caroline McKinnon
School of Communication and Arts
The University of Queensland
Brisbane, Australia
c.mckinnon@uq.edu.au

Ibtehal Baazeem
School of Information Technology and Electrical Engineering
The University of Queensland
Brisbane, Australia
ibtehal.baazeem@uq.net.au

Daniel Angus
School of Communication and Arts
The University of Queensland
Brisbane, Australia
d.angus@uq.edu.au

National Center for Computer Technology and Applied Mathematics
King Abdulaziz City for Science and Technology
Riyadh, Saudi Arabia
ibaazeem@kacst.edu.sa

Abstract

Building comprehensive language models using latent semantic analysis (LSA) requires substantial processing power. At the ideal parameters suggested in the literature (for an overview, see Bradford, 2008) it can take up to several hours, or even days, to complete. For linguistic researchers, this extensive processing time is inconvenient but tolerated—but when LSA is deployed in commercial software targeted at non-specialists, these processing times become untenable. One way to reduce processing time is to reduce the number of dimensions used to build the model. While the existing research has found that the model's reliability starts to degrade as dimensions are reduced, the point at which reliability becomes unacceptably poor varies greatly depending on the application. Therefore, in this paper, we set out to determine the lowest number of LSA dimensions that can still produce an acceptably reliable language model for our particular application: Lex, a visual cohesion analysis tool. We found that, across all three texts that we analysed, the cohesion-relevant visual motifs created by Lex start to become apparent and consistent at 50 retained dimensions.

1 Introduction

Latent Semantic Analysis (LSA) is a well-established method for describing the semantic content in textual data as a set of vectors in a high dimensional semantic space (Wade-Stein & Kintsch, 2004). It is used for a range of applications across a range of fields, including linguistics, cognitive science, education, information science and text analysis (Evangelopoulos, Zhang, & Prybutok, 2012), and it has been verified as an effective method in the majority of these fields both practically and theoretically (Evangelopoulos et al., 2012; Wade-Stein & Kintsch, 2004).

2 Lex: an overview

The application of LSA we are focusing on, pioneered by Foltz, Kintsch, & Landauer (1998), is its use in predicting the coherence of a piece of text by identifying and measuring its lexical cohesive ties. Building on this work, we have designed an LSA-based tool, which we have called Lex, to allow writers and editors to visually analyse the cohesion—and, by extension, coherence—of their own text. Users upload their text, and the tool derives the relatedness of meaning that occurs in each sentence throughout the text using a LSA language model by investigating word usage patterns in a large text corpus

Caroline Mckinnon, Ibtehal Baazeem and Daniel Angus. 2015. How few is too few? Determining the minimum acceptable number of LSA dimensions to visualise text cohesion with Lex . In *Proceedings of Australasian Language Technology Association Workshop*, pages 75−83.

(McCarthy, Briner, Rus, & McNamara, 2007), then maps out the strength of the conceptual match between every pair of sentences to a recurrence plot visualisation. The intensity of shading in each block increases with the strength of the match: shading represents more shared concepts and a higher level of cohesion between the two sentences, and paler shading or whitespace represents fewer shared concepts and less cohesion. Users can then use the visualisation to assess the overall cohesion level of their document, quickly locate areas of low cohesion that may need improving, or discover any other cohesion-relevant patterns that would otherwise have been difficult to detect.

Though it has yet to be subjected to thorough empirical testing at this early stage, we theorise that this visualisation-based method should provide a more efficient method of cohesion analysis than the traditional manual approach, because it takes advantage of the high-bandwidth, pre-attentive processing that visual perception enables (Ware, 2013). Especially in larger documents of more than a few pages, an editor's ability to detect cohesion problems is limited by their working memory capacity—by the time they get to the end of the document, they have forgotten what was at the beginning (Kintsch, Patel, & Ericsson, 1999).

In practice, we see Lex as particularly useful for a large organization such as, for example, a Queensland Government department. we most likely see Lex being used by communication staff as part of their editing process. It could help them diagnose potential problems and identify areas requiring editing or restructuring in documents intended for the public, thereby helping to ensure that the documents are cohesive enough to be clearly understood. Government organisations in particular stand to gain from clear communication: studies have shown links with improved public understanding of and increased compliance with regulations, reduced time and resources devoted to answering questions, and even greater support for government and its initiatives (Watson & Lynch, 1998). Especially in the case of guidelines or policies, unclear communication can have ethical and legal consequences, raising the question of whether citizens can be expected to comply with guidelines or laws that they are not able to fully understand (Austen, Gilbert, & Mitchell, 2001).

3 The problem

Using a pre-generated language model to analyse the user's text is not ideal for Lex's purposes. To be most useful, it needs to be able to provide reliable results when analysing any text genre or style (within reason), but the literature clearly establishes that the reliability of a result is affected significantly by the semantic similarity of the corpus text that a language model is generated from. The more similar the corpus is to the text being analysed, the more reliable the results (Biber, 1993; Koester, 2010). One way to get around this problem is to supply a range of readymade language models based on broad genres (fiction, academic, journalistic) but also offer the user the option to supply a corpus of their own that is more similar to the text they wish to analyse, and have the tool build a language model from that in run time. However, building a language model at the specifications that most literature recommends is a resource-intensive, time-consuming computational process, beyond the capability of the average desktop PC (not to mention the average user's patience) (Bradford, 2008).

One impediment is the need to use a very large corpus: the literature often recommends, on the whole, using very large corpora in the vicinity of 10 million words (Landauer, Foltz, & Laham, 1998), which can be extremely resource intensive to process. However, quality, rather than quality, is more important when it comes to corpus size: in other words, the size of the corpus could be reduced significantly without sacrificing too much by way of performance if it is highly semantically similar to the text to be analysed (Biber, 1993; Koester, 2010). The other restriction is the number of dimensions retained in the semantic space—the higher the number of dimensions retained, the more resource-intensive the process (Bradford, 2008). The bulk of studies conducted broadly appear to recommend 300 dimensions as the ideal number for LSA, but individual studies have settled on anywhere between six (Lerman, 1999) and close to 2000 (Efron, 2005). The experiments conducted to arrive at these specifications vary broadly in purpose, and use vastly different corpora types and sizes, though, which explains the large variation in findings. Reducing the number of dimensions required to produce acceptably reliable results for Lex could make this 'custom corpus' option viable, by reducing the processing time to within

reasonable limits. We suspected that a highly semantically similar, small corpus would require fewer retained dimensions to perform at acceptable accuracy levels than a large, generalised one—potentially far fewer than the industry standard of 300. What we needed to determine, though, was just how few dimensions we could retain in our semantic space before the analysis results became unreliable.

4 Method

In order to find out where the acceptability threshold lies, we generated eight LSA recurrence plots each for three different samples of text, setting the number of dimensions retained to a different threshold each time (10, 20, 50, 70, 100, 300, 500, and 700)—in total, 24 recurrence plots. We then conducted qualitative visual analyses to identify several lexical-cohesion-relevant patterns—which we will call 'motifs'—that were readily apparent in the 300-dimension versions of the plot. Three hundred was the threshold we chose as the 'gold standard' because, as discussed, it is most often recommended in the literature, and what it showed aligned most closely to our own expert assessment of the cohesion patterns in the text. We then searched for the motifs in the plots generated at successively lower dimensions, aiming to determine the lowest dimension interval at which they were still easily recognisable.

The texts we used for analysis are small sub-sections (of between 700 and 1400 words) of three different Queensland government-affiliated reports: *Delivering continuity of midwifery care to Queensland women: A guide to implementation* (Queensland Government, 2012) (herein known as the "Midwives report"); *A shared challenge: Improving literacy, numeracy, and science skills in Queensland Primary Schools* (Masters, 2009), (the "Education report"); and *Not Now, Not Ever: Putting an end to domestic violence in Queensland* (Special Taskforce on Domestic and Family Violence in Queensland, 2015), (the "Domestic violence report"). The reports in full were all around 45,000 words in length each (before pre-processing), and for each text sample we analysed, we used the full text of the report from which it came as a corpus to generate the language model. Based on our experience, 40 – 50,000 words is likely to be as large a corpus as most non-specialists could conveniently locate, so these reports imitated the conditions under which Lex would likely be used—and, as previously discussed, these corpora may be small compared to what is often recommended for building LSA language models, but what they lack in size, they make up for in specificity.

We chose government reports because government writers and editors are potential target users for technology such as Lex: they regularly produce long, complex documents for audiences with limited domain knowledge, a scenario in which cohesion is known to significantly affect readers' comprehension (McNamara, Kintsch, Songer, & Kintsch, 1996).

To appropriately test this tool for its intended purpose, we deliberately selected naturalistic data—documents that are, on the whole, fairly cohesive to begin with (as opposed to, for example, putting together random groups of sentences to artificially create or exaggerate the presence of motifs). They all certainly meet the minimum threshold to be coherent, so we knew that any detectable motifs were likely to be subtle.

The Lex plots were compared using a mixed-methods approach. Qualitative interpretation was used to determine the presence or absence of macro and meso-scale features (motifs), and a quantitative distance measure was used to summarise the magnitude of difference between the plots. For the quantitative measure all possible pairings of plots from the same test document were calculated. The measure designed for this study was the absolute difference between the plots, expressed as a percentage. The magnitude of the difference between all paired cells was calculated and averaged as:

$$\%\text{dif} = \frac{\sum_{i=0}^{n}\sum_{j=0}^{n}\left|\boldsymbol{M}_{ij} - \boldsymbol{N}_{ij}\right|}{n(n-1)/2} \times 100$$

Where: $\boldsymbol{M}$ and $\boldsymbol{N}$ are Lex matrices being compared, and n is the total number of plot elements.

5 Results

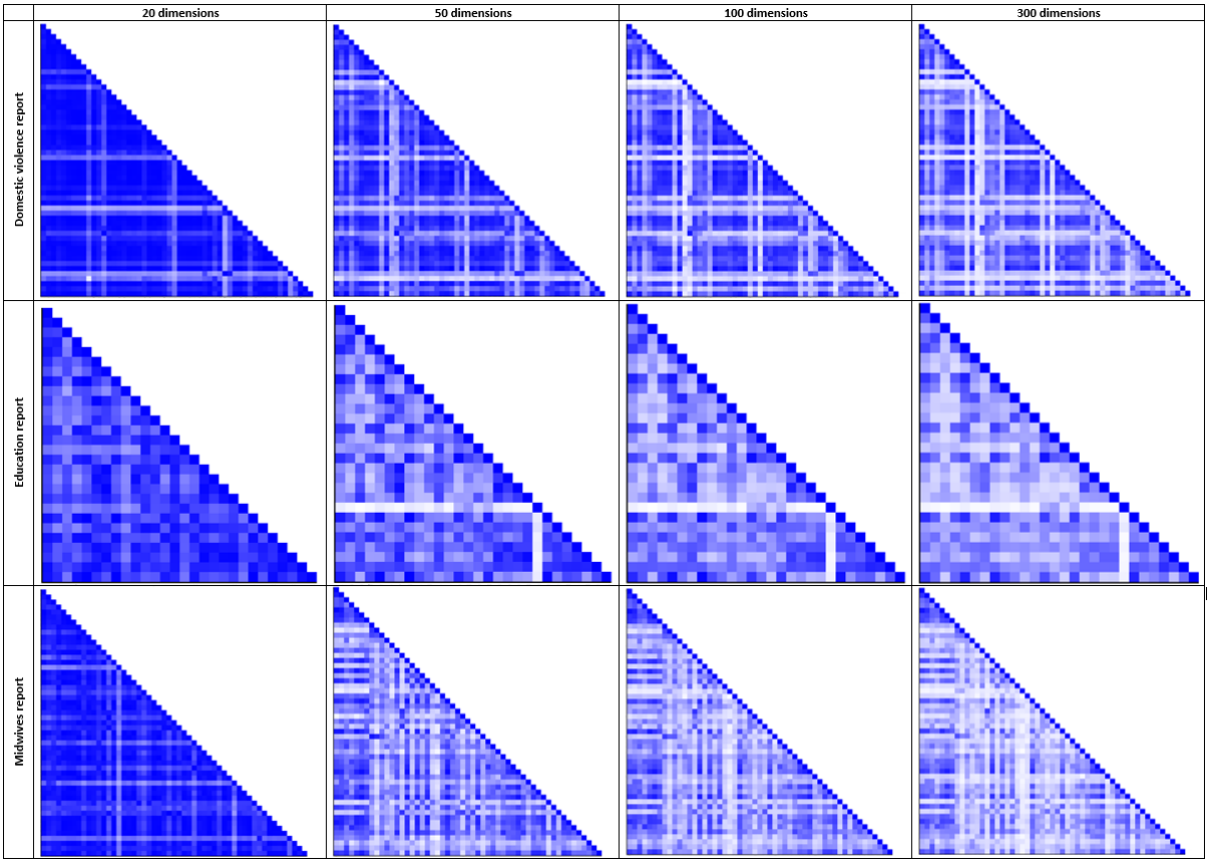

Figure 1. Lex plots for all texts at 20, 50, 100, and 300 dimensions

Domestic Violence:

	20	50	100
50	23%		
100	34%	11%	
300	43%	20%	9%

Education:

	20	50	100
50	20%		
100	28%	9%	
300	36%	16%	8%

Midwives:

	20	50	100
50	36%		
100	45%	9%	
300	53%	17%	8%

Table 1: Absolute difference between the plots

5.1 Reading the Lex plot

Each block along the right diagonal edge of the plot represents a sentence in the document. The document is laid out as a time series, progressing from the first sentence in the top left moving down toward the last sentence in the bottom right. Each variously shaded block in the plot represents the presence (or absence) and strength of the tie between the pair of sentences at whose intersection it sits. The more saturated the shade, the more shared concepts between that pair of sentences. Fainter shading, fewer shared concepts. If no link at all is present, it shows up as white space. In this way, the plot shows the degree of relatedness between every pair of sentences in the document.

At a broad level, a more densely shaded plot can be seen to represent a more globally cohesive document, and a sparse, pale plot represents a less globally cohesive one. But it's the plot's ability to show mesoscale *patterns* of cohesion that are otherwise difficult to detect that separates it from existing methods, such as, for example, the set of cohesion metrics provided by CohMetrix (Graesser, McNamara, Louwerse, & Cai, 2004). The analyses below demonstrate several examples of cohesion-relevant motifs, but only those that happen to be present in the texts we are analysing here: these are by no means an exhaustive set.

5.2 Domestic violence report

At 300 dimensions, the most obvious motifs in the Domestic violence report are the grid-like series of pale stripes criss-crossing the plot at sentences 2, 4-5, 10, 12-17, 25, 27, 35, 37-38, 42-46, 50-51, and 53:

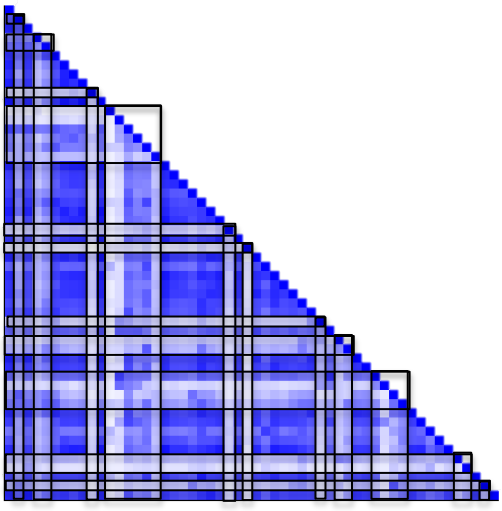

Figure 2: Motifs in 300-dimension Lex plot of Domestic violence report

Although these may present at first glance as problematic low cohesion, on closer inspection, they are actually false alarms—or at least, examples of when lexical cohesion alone cannot always tell the whole cohesion story. Almost all are quite short sentences: for example, sentence 27 reads 'It must not be accepted or excused'. Shorter sentences obviously provide fewer opportunities for content words to occur, which in turn provides fewer opportunities for lexical repetition—though other forms of cohesion may be present, such as the co-reference occurring in sentence 27 with the word 'it'. This highlights a limitation of the method, which we may need to address in future iterations of Lex by normalising for sentence length. Nevertheless, these short sentences are justifiably detected by the algorithm as having little to no semantic similarity to other sentences in the text, and are represented prominently in the visualisation at 300 dimensions, so we have included them in our definition of a motif for the purposes of this exercise.

At ten dimensions, the plot was more or less solid dark blue, with no visible motifs at all. (This was the case for all three texts, so we did not include any of the ten-dimension plots in the results pictured in figure 1.) By 20 dimensions, as pictured in figure 1, the criss-cross pattern had appeared in much the same shape, but lacking a significant amount of detail. It is not until we get to 50 dimensions that it starts to more or less accurately resemble the patterns shown at 100 and 300 dimensions. Of the pale stripes that were obvious in the plot at 300 dimensions, only sentences 14, 16, and 53 become dark enough at 50 dimensions to appear cohesive—the overall pattern remains intact. The 70-dimension plot was, almost identical to the 50-dimension plot, and this was the case for all three texts, so we did not include any of the 70-dimension plots in the results pictured in figure 1. At 100 dimensions, the patterns are slightly more defined than at 50, but overall, it is clear that an analyst would reach the same conclusions about the text, whether they were guided by the plot at 50, 100, or 300 dimensions.

5.3 Education report

The most prominent motifs in the 300-dimension education report recurrence plot are the three examples of local (intra-paragraph) cohesion, which present as darker triangles along the outside edge of the plot:

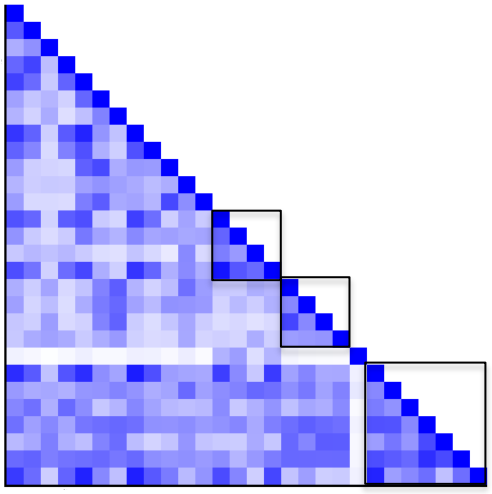

Figure 3. Local cohesion motifs in 300-dimension Lex plot of the Education report

By its original definition, a paragraph is the sustained development of a single idea (Rodgers, 1965), so it stands to reason that the sentences within a paragraph should share more concepts with each other than with sentences in other paragraphs. In this instance, however, the first two motifs represent just one paragraph, as well as the first sentence of the following paragraph. Examining this excerpt of the text offers some insight into why the plot may have divided this paragraph into two distinct motifs:

13. Deep Knowledge Highly effective teachers have a deep understanding of the subjects they teach.
14. These teachers have studied the content they teach in considerably greater depth than the level at which they currently teach and they have high levels of confidence in the subjects they teach.
15. Their deep content knowledge allows them to focus on teaching underlying methods, concepts, principles and big ideas in a subject, rather than on factual and procedural knowledge alone.
16. Highly effective teachers not only have deep knowledge of the subjects they teach, they also have deep understandings of how students learn those subjects (that is, pedagogical content knowledge).
17. They understand how learning typically progresses in a subject: for example, the skills and understandings that are pre-requisites for progress, and common paths of student learning.
18. They are familiar with the kinds of learning difficulties that some students experience and with appropriate interventions and available professional support for those difficulties.
19. And they are aware of common student misunderstandings and errors and know how to diagnose and address obstacles to further learning.
20. Targeted Teaching The most important single factor influencing learning is what the learner already knows.

In sentences 13 to 16, the subject 'teachers' (or variations thereof) is repeated throughout. In sentences 17 to 19, however, 'teachers' is replaced by the pronoun 'they', and the focus shifts to 'students' or 'learners'. Sentence 20 continues the theme, using both 'learner'/'learning' and 'teaching'.

The third local cohesion motif is formed mostly by the last two paragraphs, which together form a sub-section of the report entitled 'Targeted teaching'—though the section begins two sentences before the motif. When the low cohesion stripe discussed below (see figure 4) is accounted for, however, this motif aligns very well with the deliberate sectioning of the text.

The other noticeable motif in the Education report is the pale stripe in sentence 21 (figure 4), which, as in the Domestic violence report, is seemingly evidence of a short sentence containing few content words ("Ascertain this and teach him accordingly"), rather than a true example of low cohesion.

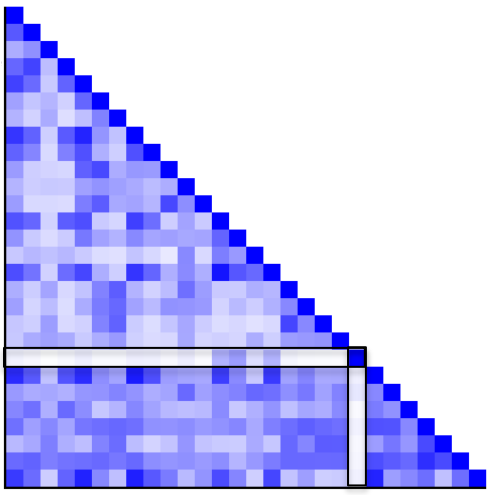

Figure 4. Low cohesion stripe motif in 300-dimension Lex plot of the Education report

Figure 1 demonstrates that the motifs for this text begin to disappear at 20 dimensions—whereas at 50 dimensions, the motifs in the 100- and 300-dimension plots are darker, but still clearly visible. Again, the threshold appears to be 50 dimensions.

5.4 Midwives report

The midwives report plot, at 300 dimensions, shows a dense introductory stripe, which is formed by the first two paragraphs of the text:

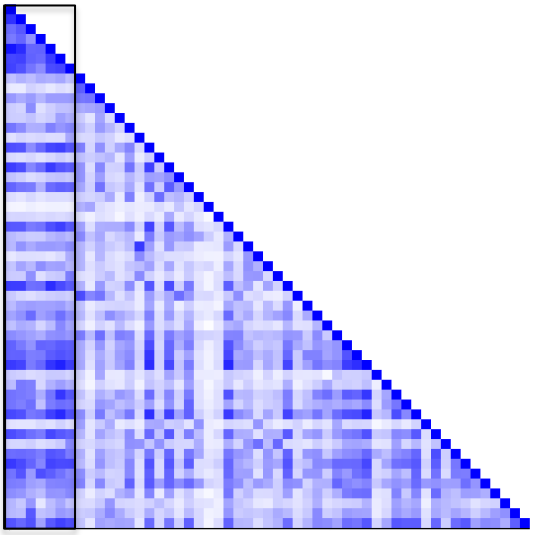

Figure 5. Introductory stripe motif in the 300-dimension Lex plot of the Midwives report

Although in the original document this text was split into two segments, together they can broadly be seen to represent the introductory section of the text, in that they set out the document's purpose and introduce and define the key terms heavily used throughout the rest of the document ('continuity', 'midwifery', 'care', 'birth', 'women', 'models', 'work', and variations thereof). The real business of the text is conducted after these two sections. Therefore, it is not surprising to see that it shows a greater level of cohesion both locally—within itself—and globally, with the entire rest of the document.

The second motif of interest is the two distinct pale stripes at sentences 21-23 and 39-40, signalling a group of sentences that do not share many concepts with either those preceding or following them. This pattern flags the possibility of low cohesion.

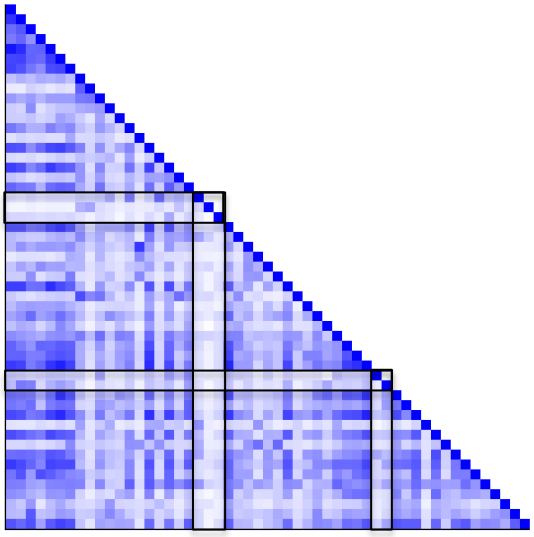

Figure 6. Low cohesion stripe motifs in 300-dimension plot of Midwives report

The full text of these two sentences are as follows:

Stripe 1

21. Communication within and beyond the service builds collaboration and understanding.
22. Engagement of stakeholders helps align expectations and manage divergent motivations.
23. 5. A guide to implementation

Stripe 2

39. This requires a different philosophy and skill set.
40. Relationships with women are close, continuous (sometimes for more than one baby), responsive to women's needs and very effective in supporting women's ability to birth and mother.

The palest stripes are again red herrings, caused by short sentences with few content words (sentences 23 and 39). The remaining sentences, especially 21 and 22, use a high proportion of abstract terms such as 'communication, 'collaboration', 'understanding', 'expectations', 'motivations', rather than the specific terms that more routinely occur throughout the text (variations of 'midwives', 'continuity models', 'birth', and 'women').

As with the plots for the other two texts, the motifs that are readily apparent at 300 dimensions hold steady until 20 dimensions, at which point they disappear completely. At 50 dimensions, it is likely that an analyst would reach the same conclusions as they would at 100 or 300 dimensions, but this would not be possible at 20.

6 Discussion and Conclusion

It is evident that, across all three texts, the visual motifs created by Lex start to become apparent and consistent at 50 dimensions. They are arguably a little clearer at 100 dimensions, and may even begin to fade out again at 300 dimensions. This finding is also supported in the quantitative data in Table 1, which shows that, for all three text-sample-and-corpora pairs, the absolute difference between 20 and 50 dimensions is much greater than between 50 and 100, or 100 and 300 dimensions.

This finding has implications for the original stated problem of whether allowing users to upload a custom corpus to a visual language analysis tool is a viable option. Using a MacBook Air running OSX Yosemite version 10.10.3 with a 1.7 GHz Intel Core i7 processor and 8GB of memory, the average processing time to build the semantic space from the corpus with 50 dimensions retained, analyse the input text, and render the visualisation for each of our three samples was 10.48 seconds, which we consider a reasonable time for commercial deployment. This suggests that allowing users to upload a custom corpus *is*, in fact, viable. Increasing the number of dimensions retained to 100 possibly brings with it a very small gain in performance, but a significant increase in processing time, given that the LSA algorithm utilises Singular Value Decomposition, which has an order O(n^3) complexity, where n is the number of dimensions.

Obviously, the findings outlined here are limited by a reliance on our own perception of the presence or absence of visual motifs. The next step will be to repeat this exercise on multiple texts, under controlled conditions involving external participants. We also plan to conduct further research exploring the effect of the size and specificity of the corpus.

7 References

Austen, L., Gilbert, B., & Mitchell, R. (2001). Plain English—an Ethical Issue? *Legal Ethics*, *4*(1), 5–7.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243–257.

Bradford, R. B. (2008). An empirical study of required dimensionality for large-scale latent semantic indexing applications. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 153–162). ACM.

Efron, M. (2005). Eigenvalue-based model selection during latent semantic indexing. *Journal of the American Society for Information Science and Technology*, *56*(9), 969–988.

Evangelopoulos, N., Zhang, X., & Prybutok, V. R. (2012). Latent semantic analysis: five methodological recommendations. *European Journal of Information Systems*, *21*(1), 70–86.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, *25*(2-3), 285–307. http://doi.org/10.1080/01638539809545029

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202.

Kintsch, W., Patel, V. L., & Ericsson, K. A. (1999). The role of long-term working memory in text comprehension. *Psychologia*, *42*(4), 186–198.

Koester, A. (2010). Building small specialised corpora. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 66–79). Abingdon, UK: Taylor & Francis.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259–284.

Lerman, K. (1999). Document clustering in reduced dimension vector space. *Unpublished Manuscript*. Retrieved from http://patwa.googlecode.com/svn-history/r15/trunk/docs/ref/Lerman99.pdf

Masters, C. (2009). *A shared challenge: Improving literacy, numeracy, and science skills in Queensland Primary Schools*. Australian Council for Educational Research. Retrieved from http://education.qld.gov.au/mastersreview/pdfs/final-report-masters.pdf

McCarthy, P. M., Briner, S. W., Rus, V., & McNamara, D. S. (2007). Textual signatures: Identifying text-types using latent semantic analysis to measure the cohesion of text structures. In *Natural language processing and text mining* (pp. 107–122). Springer.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*(1), 1–43.

Queensland Government. (2012). *Delivering continuity of midwifery care to Queensland Women: A guide to implementation*. Brisbane,

Queensland. Retrieved from http://www.qcmb.org.au/media/pdf/Midwives%20Imp%20guide_web.pdf

Rodgers, P. C. (1965). Alexander Bain and the rise of the organic paragraph. *Quarterly Journal of Speech*, *51*(4), 399–408. http://doi.org/10.1080/00335636509382739

Special Taskforce on Domestic and Family Violence in Queensland. (2015). *Not Now, Not Ever: Putting an end to domestic violence in Queensland*. Retrieved from https://www.qld.gov.au/community/documents/getting-support-health-social-issue/dfv-report-vol-one.pdf

Wade-Stein, D., & Kintsch, E. (2004). Summary Street: Interactive Computer Support for Writing. *Cognition and Instruction*, *22*(3), 333–362. http://doi.org/10.1207/s1532690xci2203_3

Ware, C. (2013). Foundations for an Applied Science of Data Visualization. In *Information Visualization* (pp. 1–30). Elsevier. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/B9780123814647000016

Watson, R. P., & Lynch, T. D. (1998). Plain English and Public Administration. *Public Policy and Administration*, *13*(1), 107–114. http://doi.org/10.1177/095207679801300108