American Communication Journal
2016 Spring (Volume 18, Issue 1)

Text Network Analysis & the Computational Landscape Model: A study of concurrent validity

Edward T. Vieira, Jr.
Simmons College

Susan Grantham
University of Hartford

ABSTRACT

This study tested the concurrent validity between the Landscape Model of Reading and Comprehension (LMRC) computational model and a text network analysis (TNA) computational model both conceptualized by the Landscape Model of Reading Comprehension theory. The TNA computational model was operationalized based on network analysis principles. A comparison revealed that both models identified similar influential nodes and latent themes. Differences rested in the assignment of nodes to themes. LMRC model nodal connectedness is based on the amount of weighted co-occurrence; whereas, TNA is based on betweenness centrality or density of the relationships among nodes. The LMRC model requires sufficiency validity intercoder agreement (unitization), a subjective and time consuming activity. TNA required very little time and no unitization. TNA lends itself to big data and the discovery of latent messages in the text.

Key Words: Content analysis, text network analysis, Landscape Model of Reading Comprehension, concurrent validity

***Contact information:** Please address all communication to the corresponding author. Edward T. Vieira, Jr., Simmons College, School of Management, M-435, 300 The Fenway, Boston, M.A. 02115, Edward.vieira@simmons.edu or 617.521.2833.

Introduction and Literature Review

Krippendorff defines content analysis (CA) as "... a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use" (Krippendorff, 2013, p. 24). More specifically, quantitative CA can be relatively simple from a tabulation of word frequencies to more complex examinations involving the relationships between and among words and the strength of such relations resulting in latent themes (Krippendorff, 2013).

As traditional CA moves from nodal counts to nodal clusters of meaning incorporating manifest and/or latent analysis, there are increased challenges (Neuendorf, 2002). Starting at the unitization stage to final coding operations, there is more reliance on inter-subjective agreement assessments, employing agreement or covariance measures depending on the coding structure. These tasks can be time consuming and labor-intensive.

Various attempts have been made to minimize subjectivity and reduce coding processing time through unsupervised methods with varying degrees of success (Kriffendorff, 2013). CA computational models attempt to minimize subjectivity and maximize efficiencies by reducing coding times and the related coder training and coding effort by using mathematical formulae in place of traditional human coding (Gildea & Jurafsky, 2002; Landauer & Dumais, 1997). For example, the computational Syntagmatic Paradigmatic model (Dennis, 2004) based on String Edit Theory uses an algorithm to extract propositional information from text into manageable sectors. Durbin, Earwood, and Golden (2000), developed a Hidden Markov Model algorithm to compare with human coding decision-making on text data. These models are able to replicate the human coding of words into complex themes with various degrees of success. However, many require extensive pre-editing and unitization of the original text (Krippendorff, 2013).

In the case of high dimensional word context methods, Landauer and colleagues (Landauer, Foltz, & Laham, 1998) developed the Latent Semantic Analysis procedure. High dimensional word context methods fail to account for the integrated properties of the entire text. That is, they do not analyze the relationships between and among words in creating latent themes, the connections among themes, and these themes in relation to the entire text. For a review of unsupervised models see Goldman, Golden, & Van den Broek (2007) and McNamara (2011).

This study focuses on the comparison of two such computational models both informed by Landscape Model of Reading Comprehension (LMRC) theory by testing their concurrent validity. They are the Landscape Model of Reading Comprehension computational model (LMRC) (Rapp & van den Broek, 2005; van den Broek, Rapp, & Kendeou, 2005; van den Broek, Risden, Fletcher, & Thurlow, 1996) and text network analysis (TNA) approach (Paranyushkin, 2011, 2012, 2013).

Both methods focus on the reader's comprehension of the text (Tzeng, van den Broek, Kendeou, & Lee, 2005; van den Broek, 2010). In other words, the LMRC model and TNA are different ways to measure reading comprehension informed by the LMRC theory. LMRC theory concentrates on memory and concept construction during the reading process (van den Broek, Risden, Fletcher, & Thurlow, 1996). The reader identifies and activates semantic connections between and among each word, phrase, and/or concept in the text while taking advantage of existing knowledge structures.

From a LMRC theory perspective, Yeari and van den Broek (2011) describe four types of mechanisms that activate the reader's focus of attention. The first mechanism is the act of reading. Second, text elements from immediately preceding reading cycles can retain some of their activation. Third, elements from prior text can be (re)activated through an automatic associative process. Finally, readers may engage in intentional, constructive processes, by which they strategically (re)activate concepts from prior reading cycles. Therefore, the relationships among units of observation and the strengths of the relationships are critical to text comprehension. The LMRC model operationalizes the theory via computation once the text is unitized. That is, once validity sufficiency is established, the text can be analyzed using real data through mathematical calculations. At that point, it objectively identifies influential and co-occurring nodes based on subjective unitization including co-occurring weighting. There is no mechanism to discover higher order meaning within the text for thematic groupings of nodes.

Although conceptualized along an LMRC framework, the TNA computational model is operationally guided by network analysis principles. It provides operational potential not offered by the LMRC model. TNA borrows from contingency analysis (Osgood, 1959) and semantic network analysis (Hays, 1960, 1969; Kleinnijenhuis, De Ridder, & Rietberg, 1997) methodologies. There is no need for unitization. Inductively based, the initial validity coding and reliability coding are automated and thus data driven leaving little room for subjectivity. Influential nodes and thematic communities are based on betweenness centrality (*BC*) or the relationships and density among nodes rather than simple co-occurrences as is the case with the LMRC computational model. TNA provides high order themes of meaning, which may include latent messages. In other words, it is not simply co-occurrences of words or nodes, but the inter-relationships of nodes and their role in constructing meaning and the interconnected of these themes or latent meaning to the text as a whole.

In short, LMRC theory is sound and has been validated in many studies (van den Broek, Rapp, & Kendeou, 2005). Unfortunately, its operationalization via the LMRC computational model is labor intensive and requires unitization coding thus introducing potential bias although it automates the actual final text coding. For large text corpus, it is not practical. On the other hand, TNA, informed by the same LMRC theory, rectifies these concerns and leaves subjectivity to the final interpretation of quantitative results. With the above in mind, the following hypotheses and research question are submitted.

> Hypothesis 1: Nodal frequencies, LMRC model-based activations, TNA *BC*, and TNA degrees (*D*) or "co-occurrences" will be highly correlated.
>
> Hypothesis 2: The LMRC computational model "Connections" results and a principal components analysis of the text analysis ascribed nodal communities results will not be significantly different.
>
> Research Question 1: What benefits do the methodologies have over each other?

Methods

LMRC computational model and TNA approaches followed by a concurrent validity comparative analysis will be applied to a short text. The initial unit of analysis are the

nodes or words followed by the relationships among node clusters of contextual meaning. Therefore, the most influential words and how these words connect to each other and other words to create higher order meaning (including the overall textual theme) will be explored. In sum, this approach will be applied using the LMRC computational model and TNA followed by a number of concurrent validity analyses including correlations, principal components analysis, and cross-tabulation chi-square testing. So that qualitative comparisons can be easily conducted, the three analyzes will be conducted on an 85 word text document, which is located in Table 1. Since "A Knight's Tale" LMRC analysis and measures were available from the van den Broek, Risden, Fletcher, and Thurlow (1996) study, it served as the text that could also be analyzed using TNA so that a comparison could be made.

Table 1
Text Corpus

A Knight's Tale
A young knight rode through the forest.
The knight was unfamiliar with the country.
A dragon appeared suddenly.
The dragon was kidnapping a beautiful princess.
The knight wanted to free the princess.
The knight wanted to marry the princess.
The knight hurried after the dragon.
The knight and dragon fought for life and death.
Soon, the knight's armor was completely scorched.
At last, the knight killed the dragon.
The knight freed the princess.
The princess was very thankful to the knight.
The princess married the knight.

Note: The text consists of 86 words. For the LMRC computational model, each sentence constitutes a reading cycle for 13 cycles. For TNA, the paragraph is analyzed as a single reading cycle.

The LMRC application requires three types of data files: original text, unitization, and activation. First, the text corpus is required as a reference. Second, the nodes to be analyzed are entered. This is an a priori approach involving the "elaborative pre-editing of raw text" (Krippendorff, 2013, p. 238). Third, the activation area, which measures the number of times a node is represented explicitly or implicitly, involves ascribing weighted values to nodes (Tzeng, 2007; van den Broek, Risden, Fletcher, Thurlow, 1996). The default settings and values are literal mention of the node = 5; referential inferences = 4; causal inferences = 4; enabling conditions = 3; subsequent no mention after one of the aforementioned activation = ½ of the immediately previous activation value (accounts for residual); and no immediately previous mention = 0 per sentence cycle. In the case of "A Knight's Tale," there are 13 cycles (sentences).

The Knight's Tale analysis and measures were taken from the van den Broek, Risden, Fletcher, and Thurlow (1996) study. Both the second and third procedures involve the initial coding and provide for validity sufficiency (Graneheim & Lundman,

2004; Kyngas & Vanhanen, 1999; Potter & Levine-Donnerstein, 1999; Weber, 1990) so that the nodes actually denote what they are posited to represent. This thus becomes a task of inter-subjective agreement and can be exceedingly cumbersome when analyzing a large text corpus.

In *A Knight's Tale*, Cohen kappa reliability for unitization was .94 and for activation strength, it was .88. The LMRC computational model approach generated two outputs: nodal activation and bi-nodal connections. The activation calculation was described above (see Table 2). The connection matrix represents the cross-product of activations' values for co-occurring nodes per reading cycle. In the case of the diagonal values, they are the cross-products of the node times itself for that reading cycle.

Table 2
Complete LMRC Computational Model and TNA Nodal Characteristics for A Knight's Tale

Nodal Theme	Frequencies	Activation	Betweenness Centrality	Degree
Knight	11	65.5	292.70	37
Princess	6	29.0	40.86	16
Dragon	5	34.0	93.77	18
Rode	1	9.5	.50	5
Forest	1	11.0	1.50	6
Country	1	9.5	4.95	5
Unfamiliar	1	9.5	8.28	6
Kidnapped	1	9.5	4.31	5
Appeared	**1**	**9.5**		
Want to Free	1	19.5	.48	6
Want to Marry	1	12.5	.33	6
Hurried After	1	9.5	2.28	6
Fought	1	12.5	2.03	6
Life and Death	1	12.0	3.28	6
Armor	1	7.5	1.17	6
Scorched	1	7.5	3.28	6
Killed	1	9.5	2.57	6
Freed	1	9.5	8.85	5
Thankful	1	8.0	5.58	4
Married	1	5.0	.00	2
Beautiful	**1**		**7.13**	**6**
Suddenly	**1**		**3.97**	**6**
Fought	**1**		**2.03**	**6**
Completely	**1**		**1.17**	**6**

Note: Frequencies were tabulated using NVivo 10.0; activation was calculated using the LMRC computational model; and Textexture (http://textexture.com) and Gephi (http://www.gephi.org) were deployed to generate *BC* and D. Boldface Nodes were dropped for the comparative analysis.

The bi-nodal covariance connection matrix was examined for thematic communities using principal components analysis with Varimax rotation (Boik, 2012; Cheung & Brandes, 2011). A three component solution was discovered which explained 82% of the variance among the nodal components. The rotated component structure

revealed factor loadings from .59 to .97 demonstrating desirable convergent validity and inter-component correlations ranged from .20 to .73 indicating acceptable discriminant validity. Cronbach alpha reliabilities ranged from .89 to .98. Based on the van den Broek, Risden, Fletcher, and Thurlow (1996) study analysis, the unit coders did not include *young*, *suddenly*, *beautiful*, and *completely*. One may assume that they did so because these nodes are attributive adjectives and adverbs and perhaps do not make a substantive contribution to the analysis. Heretofore, these four nodes were removed from further analysis. As depicted in Table 3, three clear themes emerged. First, the knight enters the unfamiliar forest country. Next, the knight then battles the dragon. Last, the freed, thankful princess marries the knight. The context implies that they live happily ever after.

Table 3
LMRC Principal Components Analysis of Thematic Communities

Nodal Theme	The Knight's Entry into the Forest	The Battle	The Aftermath
The Knight's Entry into the Forest	**.98**	-.05	-.01
Unfamiliar	**.98**	-.05	-.01
Country	**.88**	-.02	-.01
Forest	**.86**	-.01	.00
Rode	**.71**	.11	.20
Appeared	-	-	-
The Battle			
Fought	-.06	**.98**	.06
Life and Death	-.05	**.97**	.12
Armor	-.07	**.92**	-.10
Scorched	-.07	**.92**	-.10
Hurried After	.08	**.79**	.31
Dragon	.34	**.77**	**.42**
Killed	-.02	**.72**	**.45**
The Aftermath			
Princess	.13	.21	**.94**
Thankful	-.08	-.07	**.89**
Freed	-.09	.17	**.88**
Want to Marry	.09	.36	**.80**
Married	-.09	-.15	**.79**
Want to Free	.12	**.60**	**.73**
Knight	.37	**.57**	**.68**
Kidnapping	.34	.13	**.59**
Cronbach Alpha	.97	.91	.89

Note: Factor loadings > .40 are in boldface.

The following steps were deployed in order to prepare the text for the TNA application run (Paranyushkin, 2011, 2012, 2013). In the first step, prepositions and other words that bind the text corpus together but did not specifically relate to the content were removed. Second, the remaining words were replaced with their relevant morphemes or stem words using the Krovets Stemmer algorithm (Krovetz, 1993). Then, all words were changed to lowercase in order to avoid counting the same word as two different terms. Superfluous spaces, symbols, and punctuation, and numbers were removed. For this study, some words were combined into single nodes representing single constructs in

order to correspond with the LMRC unitization. They were “lifeanddeath,” “hurriedafter,” “wantedtomarry,” and “wantedfree.”

The normalized data were encoded in the XML format, which would allow the content to be network and graphical analyzed. Encoding involved using an automated two-pass approach (Paranyushkin, 2011, 2012). More specifically, the text data were scanned using two consecutive node scans. Each node that first appears in the scan was recorded as a new node and a value of one was assigned to the connection (or edge). If the pair existed already, the weight of the corresponding edge was incrementally increased by one. In the case of *A Knight’s Tale*, the single paragraph story represented a reading cycle. The second pass used a 5 consecutive node scan which followed a similar procedure as the initial two-word scan. The 5-word pass started at the first word of the paragraph and terminated when it reached the last word of the tale. The 2-word pass approach permits us to discover the general text network structure as well as general themes within the entire text. Paranyushkin (2011, 2012) found that the 5-word scan further differentiates latent themes into more fine-tuned meaningful groups by weighing the closest contiguous connections and then extending the connections to five words. The XML file was run in Gephi, an open source application (Bastian et al., 2009). Gephi produces graphical visualizations and offers a number of analytic procedures.

BC, *D*, and modularity network measures are conductive to receiving and interpreting messages from the reader perspective. The *BC* measure represents how often a node appears on the shortest paths between two nodes in the network (Freeman, 1979; Izquierdo & Hanneman, 2006). In this study, it is calculated on two and five consecutive word passes. The higher the *BC* score, the more influential the node because it links other nodes and disseminates meaning associated with those nodes. *D*s are the number of edges a node possesses; these are number of direct connections to other nodes. Modularity refers to the presence of prominent thematic communities within the entire text.

Table 4
“A Knight’s Tale” Structural Properties

Property Measures	
Nodal Count	23
Edge Count	90
Average Betweenness Centrality	25.09
Average Degree	3.91
Cluster Count	3
Modularity	.28
Strongly Connected Components	2

Note: After the text corpus was processed, the word count was 24.

Table 4 depicts the structural properties of *A Knight’s Tale*. Some of these measures are relative. There are 90 edges. Again, edges are the number of nodal connections based on the 2 and 5 consecutive word scans. The average *BC* was 25.09. The average *D* was 3.91, which is in the low edges per node range (Paranyushkin, 2011, 2012). Three thematic communities were detected. Modularity was .28, which is indicative of less dense communities. Any modularity measure > 0.40 suggests the

presence of prominent (intra-connectedness) thematic communities (Blondel et al, 2008; Freeman, 2000).

Again, *BC* was the basis for determining nodal influence. The key nodes were "knight" (*BC* = 292.70, *D* = 37), "dragon" (*BC* = 93.77, *D* = 18), and "princess" (*BC* = 40.86, *D* = 16). These three nodes serve to connect the three communities and circulate meaning throughout the tale.

Table 2 depicts each node's *BC* and *D*. The nodal *BCs* ranged from 0 to 292.70 (*M* = 21.30, *SD* = 62.53). *D*s ranged from 2 to 37 (unique *M* = 7.83, *SD* = 7.24). There were no significant differences in average *BC*s or *D*s among the three thematic communities, which are described below ($F(2,16) = .360$, $p = .703$ and $F(2,16) = .267$, $p = .769$, respectively).

Thematic Community 1 ($n = 6$): The knights entry into the country ($M_{BC} = 20.40$, $SD_{BC} = 35.98$; $M_D = 7.67$, $SD_D = 5.09$). *Dragon* (*BC* = 93.77, *D* = 18), *Country* (*BC* = 4.95, *D* = 5), and *Kidnap* (*BC* = 4.32, *D* =5) are the nodes. This community comprises 26.09% of *BC*. The Dragon node served as a hub and provided a junction for circulating meaning throughout the tale. In short, the suddenly appearing dragon kidnapped a beautiful princess in unfamiliar country. The princess is not mentioned explicitly.

Thematic Community 2 ($n = 6$): The battle ($M_{BC} = 2.25$, $SD_{BC} = .96$; $M_D = 6.00$, $SD_D = 0$). The nodes comprising this theme are life and death (*BC* = 3.28, *D* = 6, scorch (*BC* = 3.28, *D* = 6), kill (*BC* = 2.57, *D* = 6), fought (*BC* = 2.30, *D* = 6), completely (*BC* = 1.17, *D* = 6) and armor (*BC* = 1.17, *D* = 6). This community accounts for 26.09% of *BC*. This theme describes the life and death struggle between the knight and dragon without including a direct reference to them. By the end of the conflict, the knight's armor is completely scorched and the dragon is killed.

Thematic Community 3 ($n = 11$): The aftermath ($M_{BC} = 32.18$, $SD_{BC} = 87.21$; $M_D = 8.91$, $SD_D = 9.95$). The nodes are knight (*BC* = 292.70, *D* = 37), princess (*BC* = 40.86, *D* = 16), free (*BC* = 8.85, *D* = 5), thankful (*BC* = 5.58, *D* = 4), hurried after (*BC* = 2.28, *D* = 6), forest (*BC* = 1.50, *D* = 6), ride (*BC* = .50, *D* = 5), wanted to free (*BC* = .48, *D* = 6), wanted to marry (*BC* = .33, *D* = 6), young (*BC* = 1.00, *D* = 5), and married (*BC* = 0, *D* = 2). The knight pursued the dragon so that he could liberate a thankful princess and marry her. This thematic community accounts for 47.83% of *BC*.

Figure 1 provides a graphical depiction of the influential nodes and their relationships to other nodes in the story.

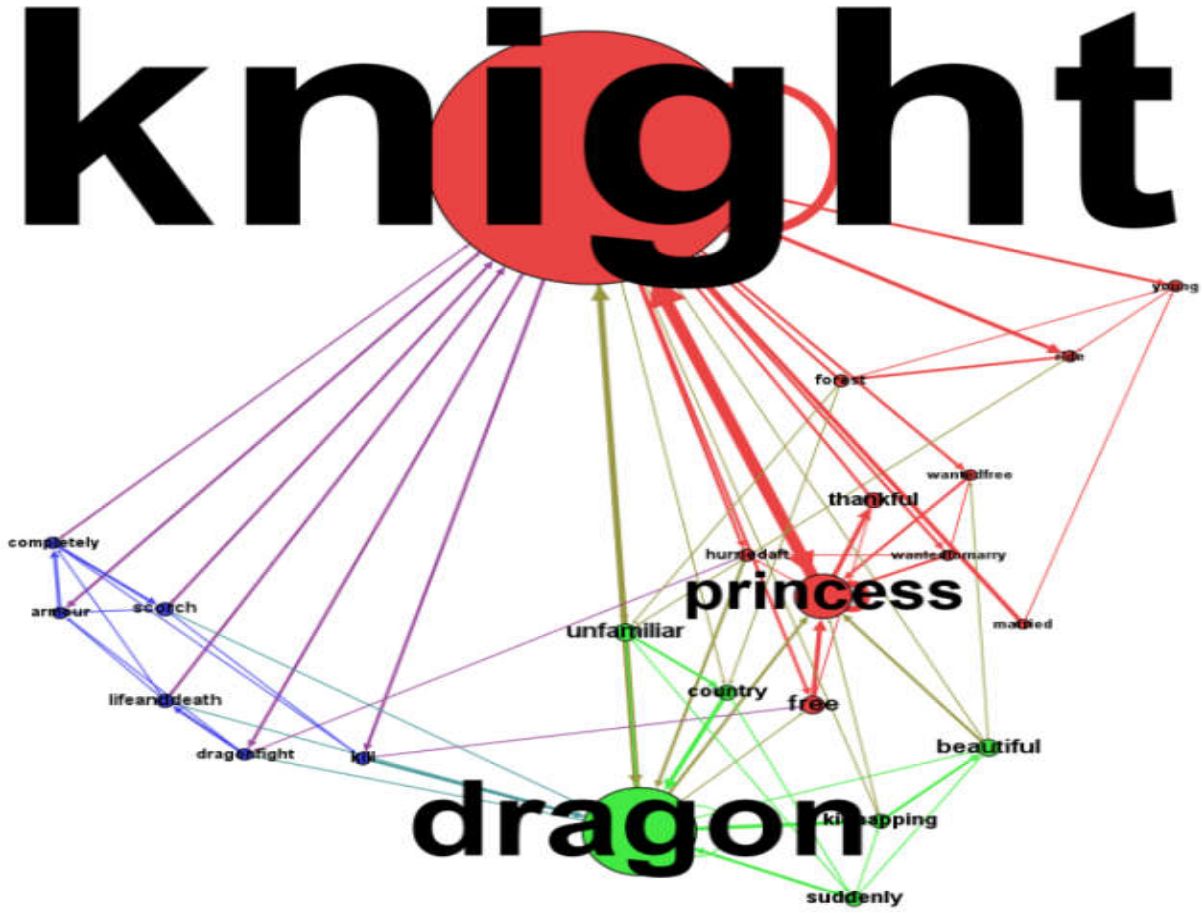

Figure 1
TNA nodes and themes

A frequency table was generated in Nvivo 10.0 based on the nodes produced by the LMRC and TNA analyses. Table 2 contains nodal frequencies along with other measures.

Correlational analyses, principal components analysis, and Cohen kappa agreement assessment were conducted to evaluate concurrent validity. The LMRC unit coders did not include "young," "suddenly," "beautiful," and "completely" because these nodes are adjectives and adverbs, and concluded that these nodes would not contribute to the text's meaning. The TNA did not account for the node "appeared." Therefore, these five nodes are not included in the comparative analysis, which now comprises 19 instead of 24 nodes.

First, a correlation analysis was run to evaluate the associations among nodal frequencies, LMRC activation, TNA *BC* and TNA *D*. As depicted in Table 5, bivariate correlations were .92 to .97 ($p < 01$) demonstrating a strong positive relationship among the LMRC and TNA computational measures thus supporting concurrent validity.

Table 5
Bivariate Correlations of Nodal Measures

Nodal Measure	Frequencies	Activations	Betweenness Centrality	Degree

Frequencies	1.00			
Activations	.96	1.00		
Betweenness	.94	.92	1.00	
Centrality				
Degree	.97	.96	.96	1.00

Note: All correlations are significant at < .01.

Next, a principal components analysis with Varimax rotation revealed a single component solution for frequencies, activations, *BC*, and *D* with an Eigenvalue of 3.86. Loadings were .97 - .99 explaining 96.57% of nodal variance. This result suggests that these measures tap into the same construct—nodal influence within the tale.

Last, a Spearman rho association test was run on the nodal thematic communities followed by Cohen Kappa agreement assessment. The Spearman *rho* correlation was .65 ($p = .002$) and the Kappa was .62. Both of these measures fall short of the common accepted minimum of .80 (Krippendorff, 2013; Taylor & Watkinson, 2007).

Results

Hypothesis 1: Nodal frequencies, LMRC-based activations, TNA *BC*, and TNA *D* will be highly correlated. This hypothesis was confirmed. The correlations among these measures were .92 to .97 ($p < .01$). This finding strongly suggests that the TNA and network principles approach closely captures content analysis from the receiver perspective as conceptualized by the LMRC.

Hypothesis 2: The LMRC Connection Matrix results and principal components analysis of the text network analysis ascribed nodal communities will not be different. The Spearman rho was .66 ($p = .002$) and Cohen Kappa was .62. These measures reveal a level of low moderate/weak association and agreement respectively between the communities. The LMRC focuses on the number of relationships (*D*) between words, but not on the strength (*BC*) of those relationships as is the case when measures of *BC* are considered. Thus, as might be expected the connections describe some of the relationship, however, the strength is not accounted for.

Research Question 1: What benefits do the methodologies have over each other? In the case of LMRC, researchers can select specific nodes for analysis. Moreover, various types of weighted values can be assigned to activation such as explicit mentions, causal, and enabling conditions. Next, based on this study, the resulting thematic communities predicated on the cross-products of node activation per reading cycle and principal components analysis have face validity. In short, after unitization and the establishment of validity sufficiency, the procedure becomes computational and automated thus eliminating the need for further inter-coder reliability. On the other hand, TNA offers some advantages over LMRC, unitization is not required, and thus substantial time is saved. Second, related to the prior point, subjectivity is reduced to final interpretation of thematic communities because intersubjective unitization is eliminated. In TNA; there is no unitization. It is a computation process predicated on network analysis principles. Third, nodal influence is operationalized based on the relationships among nodes rather than the cross-product of frequency strength per reading cycle. Last,

TNA provides a graphical representation of nodal influence in the context of thematic communities. This features makes complex text analyses easier to comprehend.

Discussion

Table 6 summarizes the key terms in this content analysis concurrent validity study. As these comparative results reveal, both methods' selection of nodes and their relative influence in *A Knight's Tale* are consistent as indicated by the high correlations among frequencies, activations, *BC*, and *D*. LMRC bases its influence on co-occurring nodes and their weighted influence per reading cycle as determined by coders. On the other hand, TNA operationalizes nodal influence based on the impact of a node on other nodes such that the node is required to establish a relationship among two or more other nodes to form communities. Based on this study, both methods are highly correlated and may very well capture this influence.

Table 6
Key Methodological Terms

Term	Description
BC	Betweenness Centrality represents how often a word appears on the shortest paths between two words in the network.
D	Degrees are the number of links (or co-occurrences) a word possesses; these are the number of direct connections to other nodes.
LMRC	The Landscape Model of Reading Comprehension is a framework for understanding and conducting content analysis based on the reader's perspective.
TNA	Text Network Analysis is an unsupervised content analysis method based on network principles.

However, there is some divergence in the assignment of thematic communities. This is not unexpected because of the different basis for ascribing nodes to groups. The LMRC calculates a bi-nodal "connection" score based on the cross-product of weighted co-occurring activation scores within a reading cycle (i. e. one sentence per cycle). This study took this procedure further by running a covariation principal components analysis and assigned nodes to communities based on factor loadings. These assignments were supported by strong Cronbach Alpha reliability scores. Keep in mind, however, that nodes with high connected scores do not necessarily indicate that they represent the same construct; their loadings indicate that they are among a grouping of high cross-product nodes. On the other hand, TNA tabulated nodal communities are based on the strength of their *BC*. In other words, they are grouped predicated on how closely they connect to each other and provide meaning between nodes, as well as across and within thematic communities. Perhaps a larger text corpus would have clearly revealed similar communities. TNA allows nodes that link communities to be identified. In the case of the LMRC, activations may very well be closely connected as in the case of *BC*, but not necessarily so. The Cohen Kappa suggests this relationship. TNA is ideal for big data CA and for identifying themes and messages not clearly evident.

One limitation of this study is the small text corpus. Previous research (Vieira & Grantham, in review) revealed more potential of the TNA method. This study was limited

to text that was previous published in an LMRC article. It was not possible to run the data in the LMRC application owing to technical and operating system issues.

Perhaps future research would compare human coding and TNA as well as integrating all of the steps (i. e. data scrubbing and generation of XML file) into a seamless automated process. From there, a host of larger scale studies could be conducted. For example, applications could include social media communication among individuals incorporating participants' demographic variables in order to identify segments and social influencers associated with specific content themes and to what extent themes drive the discourse. Another area for TNA is corporate social responsibility. The CSR reports of major corporations in strategic industries could be analyzed and compared against performance measures. In short, there are many potential applications using TNA and using TNA in conjunction with other modes of data collection.

Conclusions

TNA is an innovative research tool. First, it does not require human coding and intercoding reliability measures. It has potential to streamline CA especially today with the advent of big data. TNA can be automated to run large volume CA studies in cross-sectional, longitudinal, or real-time analytics-based research. Second, the graphical representation of the text provides an easily understandable depiction of key words, text themes, and their interrelationships.

Third, unlike other computational models, the two scan process provides a layer of depth resulting in high resolution theme-based communities. Fourth, TNA offers CA from the reader's perspective. In an era of instant online and recorded communication where content is user generated via Web 2.0 and driven by such dynamics as crowdsourcing in social media, TNA is a valuable tool to make sense of the massive amounts of user created content. Rather that analyzing the message sent, TNA focuses on the message received, which aids in discovering whether a message is effective on the reader. With the growth of Web 2.0, it stand to reason that this approach would be an appropriate alternative to many existing systems.

Fifth, receiver-centered CA allows communicators to craft effective communications through message pre-testing to assure that the intended message is the received message.

Last, this approach can easily be deployed across disciplines on the Internet and on many organizational Intranets to study communication themes, social opinions, social and political trends, product/service preferences, employee motivations, and many other topics. It can be used to identify opportunities and challenges, as well as inform decision-making. Using social media, TNA can identify opinion leaders and those content themes that drive individuals to certain opinion leaders.

In sum, this study reveals concurrent validity between the LMRC and TNA in establishing influential nodes. There is also some evidence for concurrent validity between the nodal assignment of thematic communities. Findings suggest that TNA provides more enriching CA information including a graphical function, removes the inter-coder unitization assessment function, is not labor intensive, and can be utilized for inductive and deductive study designs on large as well as smaller text corpus.

References

Bastian, M., Heymann, S., & Jacomy, M. (2009). *Gephi: An Open Source Software for Exploring and Manipulating Networks*. Association for the Advancement of Artificial Intelligence. https://gephi.org/ .

Boik, R. J. (2013). Model-based principal components of correlation matrices. *Journal of Multivariate Analysis, 116*, 310-331.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, P10008. DOI: 10.1088/1742-5468/2008/10/P10008.

Cheung, M., & Brandes, B. J. (2011). Enhancing treatment outcomes for male adolescents with sexual behavior problems: Interactions and interventions. *Journal of Family Violence, 26*, 387 -401.

Dennis, S. (2004). An unsupervised method for the extraction of propositional information from text. *Proceedings of the National Academy of Sciences, 101*, 5206-5213.

Durbin, M. A., Earwood, J., & Golden, R. M. (2000). Hidden Markov models for coding story recall data. *In Proceedings of the 22nd Annual Cognitive Science Society Conference,* pp. 113-118. Mahwah, N.J.: Erlbaum.

Freeman, L. C. (1979). Centrality in Social Networks Conceptual Clarification. *Social Networks 1*, 215-239.

Freeman, L. C. (2000). Visualizing Social Networks. *Journal of Social Structure 1*(1). Retrieved from http://www.cmu.edu/joss/content/articles/volume1/Freeman.html .

Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics, 28*, 245-288.

Graneheim U. H., & Lundman B. (2004). Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. *Nurse Education Today 24*, 105-112.

Goldman, S. R., Golden, R. M., & Van den Broek, P. (2007). Why are computational models of text comprehension useful? In F. Schmalhoefer & C. A. Perfetti (Eds.), *Higher Level Language Processes in the Brain: Inference and Comprehension Processes,* (pp. 27-51). Mahwah, N.J.: Erlbaum.

Hays, D. (1960). *Automatic content analysis*. Santa Monica, C.A.: Rand Corporation.

Hays, D. (1969). Linguistic foundations for a theory of content analysis. In G. Gerbner, O. R. Holsti, K. Krippendorff, W. J. Paisley, & P. J. Stone (Eds.), *The analysis of communication content: Developments in scientific theories and computer techniques* (pp. 57 – 67). New York City, N.Y.: John Wiley.

Izquierdo, L. R., & Hanneman, R. A. (2006). *Introduction to the formal analysis of social networks using mathematica*. Retrieved from http://luis.izquierdo.name.

Kleinnijenhuis, J., De Ridder, J. A., & Rietberg, E. M. (1997). Reasoning in economic discourse: An application of the network approach to the Dutch press. In C. W. Roberts (Ed.), *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts* (pp. 191 – 207). Mahwah, N.J.: Lawrence Erlbaum.

Krippendorff, K. (2013). *Content analysis an introduction to its methodology*. Los Angeles, C.A.: Sage Publishing.
Krovetz, R. (1993). *Viewing Morphology as an Inference Process*. SIGIR 1993 Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval.
Kyngas, H., & Vanhanen, L. (1999). Content analysis as a research method. *Hoitotiede, 11*, 3-12.
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104*, 211-240.
Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.
McNamara, D. S. (2011). Computational methods to extract meaning from text and advance theories of human cognition. *Topics in Cognitive Science 3*, 3-17.
Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, C.A.: Sage.
Osgood, C. E. (1959). The representation model and relevant research methods. In Ithiel de Sola Pool (Ed.), *Trends in content analysis* (pp. 33-88). Urbana, I.L.: University of Illinois Press.
Paranyushkin, D. (2013). Addresses to the Federal Assembly of the Russian Federation by Russian presidents, 2008–2012: Comparative analysis. *Russian Journal of Communication, 5*(3), 265-274.
Paranyushkin, D. (2012). *Visualization of text's polysingularity using network analysis*. Berlin, Germany: Nodus Labs.
Paranyushkin, D. (2011). *Identifying the pathways for meaning circulation using text network analysis*. Berlin, Germany: Nodus Labs.
Potter, W. J.,& Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research, 27*, 258-284.
Rapp, D. N.,& van den Broek, P. (2005). Dynamic text comprehension: An integrative view of reading. *Current Directions in Psychological Science, 14*, 276-279.
Taylor, J, & Watkinson, D. (2007). Indexing reliability for condition survey data. *The Conservator, 30*, 49 -62.
Tzeng, Y. (2007). Memory for narrative texts: How do parts of the Landscape Model work. *Chinese Journal of Psychology, 49*(3), 245-269.
Tzeng, Y., van den Broek, P., Kendeou, P., & Lee, C. (2005). The computational implementation of the Landscape Model: Modeling inferential processes and memory representation of text comprehension. *Behavioral Research Methods, Instruments & Computers, 37*, 277-286.
van den Broek, P. (2010). Using texts in science education: Cognitive processes and knowledge representation. *Science, 328*, 453-456. DOI: 10.1126/science.1182594.
van den Broek, P., Rapp, D. N., & Kendeou, P. (2005). Integrating memory-based and constructionist approaches in accounts of reading comprehension. *Discourse Processes, 39*, 299–316.
van den Broek, P, Risden, K., Fletcher, C. R., & Thurlow, R. (1996). A "landscape" view of reading: Fluctuating patterns of activations and the construction of a stable

memory representation. In. B. K. Britton & A. C. Graesser (Eds.), *Models of Understanding Text* (pp. 165-187). Mahwah, N.J.: Erlbaum.
Vieira, E. T., Jr., & Grantham, S. (in review). A Comparison of ExxonMobil's CEO 2002 and 2012 Corporate Citizen Report Letters Using a New Content Analysis Methodology. *Social Science Research*.
Weber, R. P. (1990). *Basic content analysis*. Beverly Hills, C.A.: Sage.
Yeari, M., & van den Broek. (2011). A cognitive account of discourse understanding and discourse interpretation: The Landscape Model of reading. *Discourse Studies, 13*(5) 635-643.

AUTHOR DETAILS

Edward T. Vieira, Jr., earned his Ph.D. from the University of Connecticut. Currently, he is an Associate Professor and Director of Research at the School of Management at Simmons College in Boston, M. A.

Susan Grantham earned her Ph.D. from the University of Florida. Currently, she is a Professor and Director of the School of Communication at the University of Hartford in West Hartford, C. T.