Identifying the Pathways for Meaning Circulation using Text Network Analysis

Dmitry Paranyushkin

Version December 2011

Published in Nodus Labs

Received 26 October 2011, Accepted 15 November 2011, Published 11 December 2011

Dmitry Paranyushkin
Nodus Labs

3, Am Flutgraben,
12435, Berlin, Germany

www.noduslabs.com
info@noduslabs.com

Identifying the Pathways for Meaning Circulation using Text Network Analysis

Dmitry Paranyushkin
Nodus Labs, Berlin, Germany, e-mail: dmitry@noduslabs.com

Abstract

In this work we propose a method and algorithm for identifying the pathways for meaning circulation within a text. This is done by visualizing normalized textual data as a graph and deriving the key metrics for the concepts and for the text as a whole using network analysis. The resulting data and graph representation are then used to detect the key concepts, which function as junctions for meaning circulation within a text, contextual clusters comprised of word communities (themes), as well as the most often used pathways for meaning circulation. We then discuss several practical applications of our method ranging from automatic recovery of hidden agendas within a text and intertextual navigation graph-interfaces, to enhancing reading and writing, quick text summarization, as well as group sentiment profiling and text diagramming. We also make a quick overview of the existing computer-assisted text analysis (and, specifically, network text analysis), and text visualization methods in order to position our research in relation to the other available approaches.

Keywords: network text analysis, text, network, meaning, narrative, discourse, language understanding, semantics, structure, system, semantic networks, context, cognition, interpretation, graph, diagram, visualization, interface, reading, writing, image

1. Introduction

Any text can be represented as a network. At a basic level, the words, or the concepts are the nodes, and their relations are the edges of the network. Once a text is represented as a network, a wide range of tools from network and graph analysis can be used to perform quantitive analysis and categorization of textual data, detect communities of closely related concepts, identify the most influential concepts that produce meaning, and perform comparative analysis of several texts.

In this paper we will introduce a method for text network analysis that allows one to visualize a readable graph from any text, identify its structural properties along with some quantitive metrics, detect central concepts present within the text, and, finally, identify the most influential pathways for the production of meaning within the text, which we call *the pathways for meaning circulation*. This method can have a variety of practical applications: from increasing the speed and quality of text cognition, to getting a better insight into the hidden agendas present within a text and better understanding of its narrative structure. The fields where it can be applied range from media monitoring to comparative literary analysis to creative writing.

The use of diagrammatic approach to better understand text and analyze its narrative structures is not new. There has been a lot of research on this subject. For instance, “Interacting Plans” (Bruce et al., 1978) focused on uncovering social interaction structures from semantic structures of texts using visual and graphic analysis. “Plot Units and Narrative Summarization” (Lehnert, 1981) focused on identifying so-called “plot units” within a text. These plot units were graphical representations of “affect states”, which were not focussing on complex emotional reactions or states, but, rather, on gross distinctions between “negative” and “positive” events and “mental events” with zero emotionality. “The Role of Affect in Narrative” (Dyer, 1983) established affect - and goal-seeking behavior - as a moving force behind narrative structures.
Later work, most notably “Coding Choices for Textual Analysis” (Carley, 1993) focused on using map analysis to extract the main concepts from the texts and relations between them. In their research on spatial analysis of text documents (Wise et al., 1995) proposes spatial visualization techniques for various texts based on their similarity, reducing the workload when performing text analysis.
In “Knowledge Graphs and Network Text Analysis” (Popping, 2003) the author advocates the use of schemes, and specifically knowledge graphs, to represent texts in order to gain a better understanding of textual data and to tackle the dynamic nature of knowledge. Her other work “Computer-Assisted Text Analysis” (Popping, 2000) has a good overview of network text analysis techniques that use graphs to represent text visually and graph analysis tools in order to obtain quantitive data for later analysis.
Later research, for instance “Diagramming Narratives” (Ryan, 2007) proposes to think of diagram as an heuristic device, which can represent a narrative in a spatial, temporal, and mental plane. There is also work by the scientists involved in computer game creation (Loewe et al 2009), where formal representations of narrative structures are used to detect similarities between different stories.

Thus, there’s a long history of diagrammatically representing textual data as graphs and applying network text analysis in order to gain a better understanding of the text’s meaning and structure.

Many of the approaches presented above focused on semantic relations between the words when representing texts as networks. While this is definitely helpful in gaining a better understanding of text, such approach is also adding an extra layer of ontologies and complexity on top of the textual data. The decisions regarding which concepts are related together are based on their affective affinity (Lehnert, 1981 and Dyer, 1983), their causal relations (Bruce, 1978), chronological sequence (Loewe, 2009), and semantic analysis (Van Atteveldt, 2008). All these approaches introduce a strong subjective (and even cultural) bias into the structure of the resulting text graphs. When the basis for connectivity is an external system of rules and logic, the resulting structure will be a result of negotiation between the text itself, the representational system used, and these external systems that define the basis for connectivity.

An interesting relation between the “meaning” and text network analysis can be found in a paper on meaning as sociological concept (Leydesdorff, 2011) where he writes: “Meaning is generated in a system when different pieces of information are related as messages to one another, for example, as words in sentences (Hesse, 1980; Law & Lodge, 1984). The information is then positioned in a network with an emerging (and continuously reconstructed) structure. This positioning can be done by an individual who – as a system of reference – can provide personal meaning to the events; but meaning can also be provided at the supra-individual level, for example, in a discourse. In the latter case, meaning is discursive, and its dynamics can therefore be expected to be different from those of psychological meaning.”

The difference in our approach is that we propose to avoid as much subjective and cultural influence as possible during the process of translating the textual data into a text network and visualizing it into a graph. We will only use the proximity of concepts and the density of their connections to encode the relations between the words, not their meanings or affective relations. We will then apply various graph visualization techniques, community detection mechanisms, and quantitative metrics in order to get a better insight into resulting structures without imposing any other external semantic structures on top. The resulting visual representation of text will thus be a translation of textual data into a graph where we attempt to avoid any filtering, generalization and distortion that may result from interfering into that process with an external ontology. After the textual network is visually represented as a graph of interconnected concepts it is finally open to interpretation by the observer.

Our approach can inform the existing methods of finding structure within text that employ graphical models and topic modeling: latent semantic analysis or LSA (Landauer et al 1998), pLSA (Hofmann 1999), Pachinco allocation (Li & McCallum 2006), latent dirichlet allocation or LDA (Blei et al 2003), and relational topic models (Chang & Blei 2010). These methods are based on retrieving the topics from text by identifying the clusters of co-occurrent words within them. This data can then be used to classify similar documents, improve text indexing and retrieval methods, and to identify evolution of certain topics overs a period of time within a specific text corpus (Blei 2004). Combined with hyperlinking data between texts it can also be used to find similar texts, find relations between different topics, or to predict citations based on the presence of similar topics in a text corpus (Chang & Blei 2010).

The difference of our method is that it doesn't only take into account the probabilistic co-occurrence of words in a text to identify the topics, but also the structural properties of the text itself, using graph analysis methods along with qualitative and quantitive metrics.

While it is yet to be seen how this approach specifically compares to semantic network analysis, we believe that it will definitely be useful for the researchers in this field. Italo Calvino once said that reading is “a way of exercising the potentialities contained in the system of signs” (Calvino, 1987). Therefore what we want to achieve with our approach is to simply propose a different way of reading a text through representing it as an image. And while the specific algorithm behind the image formation is described, it is the actual interpretation of the external observer that interests us the most and the possibility to fold the temporal aspect of a text onto a two-dimensional plane – thus giving a global overview of local chronological phenomena. Following Victor Shklovsky’s words when talking about a plot and a story, “Energy of delusion [...] means a search for truth in its multiplicity. This is the truth that must not be the only one, it must not be simple. Truth that changes; it recreates itself through a repetitive pattern.” (Shklovsky, 2007). The pathways for meaning circulation that we attempt to discover through our methodology are these repetitive patterns derived from the text’s structure, using their connectivity and the intensity of interactions between them as the only criteria for their belonging together. It is then through interpretation and comparative analysis that their semantic relations can be established, but we want to leave this out of the picture to allow the text to speak for itself.

“Time prevents everything from being given at once” (Bergson, 2002). Visualizing text as a network we remove the variable of time and let the history of the text appear through the diagram.

2. Data Extraction Methodology

In order to demonstrate the proposed methodology for identifying the pathways for meaning circulation within a text, we will use the introduction above as an example.

The first step is to remove the most frequently used words from the text that participate in binding the text together, but do not specifically relate to the content (see Appendix A). These are the articles, conjunctions, auxiliary verbs, and some frequently used words, which do not directly affect the context. The choice of the latter should be very considerate, because, for instance, the word “becoming” could be very important in Gilles Deleuze’s writing, but carries much less contextual significance in most newspaper articles. Moreover, it’s better to use the same delete list for several texts when doing comparative analysis in order to ensure that the differences in these delete lists do not affect the possible divergences between the different texts analyzed. Applying delete list helps to remove unnecessary content from the text and reduce the amount of noise. It also makes it more likely that the distribution of words’ frequency will not necessarily follow Zipf’s power law distribution (Zipf, 1935), as most of the simple, short words are removed. This allows one to more easily detect abnormal distribution patterns within some texts, which may be particularly useful for their comparative analysis.

The second step is to stem the remaining words in the text, in order to transform them to their appropriate morphemes. The two major algorithms for this process are the Krovets Stemmer - also called KStemmer - (Krovets, 1993) and the Porter Stemmer (Porter, 1980). We will use KStemmer as it’s less aggressive than Popper Stemmer. The stemming algorithm allows to simplify the further processing of the text and make it much more efficient. Stemming can be seen as clustering related words around a certain morpheme (Krovets, 1993), thus enabling one to reduce the complexity of the resulting network and provide a clearer starting point for further analysis.

One important detail here is that both stemming and word deletion may affect named entities. For instance, the word “Lynch” in the proper name “David Lynch” will be considered as “lynch” morpheme, and thus David Lynch will mistakenly be closer to the cluster of words such as “lynched” or “lynching”. It is important at this stage to identify what the purpose of the analysis is. If the task is to uncover the key structural properties of a text, then the aspect above may be not so important. However, if the goal is, for instance, to uncover the social network (Diesner & Carley, 2004) or a network of named entities that underly the text and integrate it into the resulting visualization (which may be useful when analyzing the interviews for example), it could be necessary to modify the stemming algorithm, so it wouldn’t affect these important semantic units. Such modification could be done using Thompson Reuters Calais system, which provides an API that detects and categorizes semantic units within the text, which could then be appropriately marked, so that the stemming algorithm doesn’t apply to them.

The third step is to further normalize the text by transforming all capital letters to lowercase (thus, avoiding that the same word is seen as two different ones by the encoding software), remove unnecessary spaces, remove the symbols and punctuation (parenthesis, dashes, dots, commas, semicolons, colons and other auxiliary signs), and numbers (unless their presence in the text crucially affects the meaning and the context).

The resulting text would look like this:

text represent network basic level word concept node relation edge network text represent network wide range tool network graph analysis perform quantitive analysis categorization textual data detect community closely relate concept identify influential concept produce mean perform comparative analysis text

paper introduce method text network analysis visualize readable graph text identify structural property quantitive metric detect central concept present text finally identify influential pathway production mean text call loop mean circulation method variety practical applications increase speed quality text cognition insight hide agenda present text understand narrative structure field apply range media monitor comparative literary analysis creative write

diagrammatic approach understand text analyze narrative structure lot research subject instance interacting plan bruce newman focus uncover social interaction structure semantic structure text visual graphic analysis plot unit narrative summarization lehnert focus identify socal plot unit text plot unit graphical representation affect state focus complex emotional reaction state   gross distinction negative positive event mental event emotionality the role affect narrative dyer establish affect goalseeking behavior move force narrative structure
work notably coding choice textual analysis carley focus map analysis extract main concept text relation spatial analysis text document wise thoma crow propose spatial visualization technique text base similarity reduce workload perform text analysis
knowledge graph network text analysis pop author advocate scheme specifically knowledge graph represent text order gain understand textual data tackle dynamic nature knowledge work computerassisted text analysis pop good overview network text analysis technique graph represent text visually graph analysis tool order obtain quantitive data analysis
research instance diagram narrative ryan propose diagram heuristic device represent narrative spatial temporal mental plane work scientist involve computer game creation loewe al formal representation narrative structure detect similarity story

there long history diagrammatically represent textual data graph apply network text analysis order gain understand text mean structure

approach present focus semantic relation word represent text network helpful gain understand text approach ad extra layer ontology complexity top textual data decision concept relate base affective affinity lehnert dyer causal relation bruce chronological sequence loewe semantic analysis van atteveldt approach introduce strong subjective cultural bias structure result text graph basis connectivity external system rule logic result structure result negotiation text representational system external system define basis connectivity

interest relation meaning text network analysis find meaning sociological concept leydesdorff writes
meaning generate system piece information relate message word sentence hesse law lodge information position network emerge continuously reconstruct structure position individual system reference provide personal mean event mean provide supraindividual level discourse case mean discursive dynamic expect psychological mean

difference approach propose avoid subjective cultural influence process translate textual data text network visualize graph proximity concept density connection encode relation word mean affective relation apply graph visualization technique community detection mechanism quantitative metric order insight result structure impose external semantic structure top result visual representation text translation textual data graph attempt avoid filter generalization distortion result interfere process external ontology textual network visually represent graph interconnect concept finally open interpretation observer approach specifically compare semantic network analysis researcher field italo calvino read a exercise potentiality contain system sign calvino achieve approach simply propose read text represent image specific algorithm image formation actual interpretation external observer interest possibility fold temporal aspect text twodimensional plane give global overview local chronological phenomena victor shklovsky word talk plot story energy delusion mean search truth multiplicity truth simple truth recreate repetitive pattern shklovsky loop mean circulation attempt discover methodology repetitive pattern derive text structure connectivity intensity interaction criteria belong interpretation comparative analysis semantic relation establish leave picture text speak

time prevent bergson visualize text network remove variable time history text diagram

The next step is to convert this text into the graph data, which could later be used in order to represent it visually as a graph using various software, for example, a popular graph visualization and analysis tool Gephi (Bastian et al., 2009). One of the more common formats is GraphML (Brandes et al, 2002). It is a type of XML file for graph data. In our case the structure is such that the words (or the nodes) are listed first, and then their connections (or edges) are listed after along with the special metrics, which measure the weight for each edge. The graph type is undirected.

It's important to save all intermediary data in common formats, so that the data could later be used by other researchers who are using different software (Brandes, 2002). Also, transparent and open formats allow a higher traceability of the algorithm and possible modifications to the process in case the researcher is interested in a different approach or in modifying some detail parameters. The steps below, for example, were found to be particularly useful for our local research situation, however, if the parameters need to be amended the clear format of the text file above and the XML structure of the resulting file will allow one to quickly identify the algorithm in order to modify it if necessary.

It is also possible to use the algorithms described below to record the data directly into a relational or graph database, such as the popular MySQL or Java-based Neo4j. Using a graph database can be more efficient when it's needed to query this data online and retrieve it much faster, especially for text network databases (Vicknair et al., 2010).

During our research we found that the best results for encoding textual data into a graph structure are achieved with a two-pass process described below.

First, the normalized text above is scanned using a 2-word gap. For each word, if it appears the first time in the text, it's recorded as a new node with the id that equals the name of the node. When two words appear within the gap, the algorithm first checks if the pair exists already. If the pair does not exist yet, a new connection (an edge) is recorded where the first word is the source and the second word is the target, the weight equals 1. If the pair exists already, the weight of the corresponding edge is incremented by 1. This way we trace the narrative and create a concept graph from the text. Each connection is based on the words' proximity to each other. The more frequent the combination of words, the higher is the weight of connection between them. When the scanner reaches the end of the paragraph, it jumps to the next one in order to avoid that the last word from the previous paragraph is linked to the first word from the next one. This helps us to somewhat translate the spatial structure of the text into the graph. A modification of this scanning algorithm can allow it to make connections between different paragraphs, so that the last word of a paragraph is connected to the first word of the next one. This will create a more interconnected graph and the decision as to which particular version of the algorithm to use should depend on the importance of text paragraph structure for the researcher.

The second pass uses a 5-word gap and follows a similar procedure. For each combination of 5 words, starting from the beginning of the text, the algorithm first checks whether each word pair exists. If it does already (as a result of the 2-word gap pass before), the weight of the connection (or the edge) between the pair is incremented by one. If it does not exist, the new pair is recorded as the new edge (the weight equals 1), where the source is the word to the left of the gap and the target is the word to the right of the gap. The words adjacent to the words in the beginning and in the end of each paragraph will have a slightly less intense connection (as the 5-word gap starts at the first word of a paragraph and terminates when it reaches the last word of the paragraph, then jumping to the next paragraph and starting again from the first word). Such approach allows us to accommodate further for the spatial structure already utilized within the text. It also allows us to increase the intensity of connections between the words that are more proximate to each other. If the first 2-word gap scan is sketching a general structure of the text intensifying repetitions of adjacent words within the text and outlining its paragraph structure, then the 5-word gap scan is a kind of zooming in tool into the local areas of the text, which allows us to intensify the local clusters of meaning overlaying them on the general structure created before.

It is also possible to add additional parameters to this data. For example, each edge can be recorded with a timestamp, so that a dynamical graph can be generated instead of a static one. That would allow to observe the formation of text as a graph in time and provide very interesting insights on its structural properties. While this is a direction for our further research, we will focus on static graphs for the purposes of this particular paper.

After the resulting data is recorded in XML-compatible format (e.g. GraphML) or into the database, it can be retrieved to be visualized directly or using one of graph visualization and analysis software, for instance, Gephi.

3. Text Graph Visualization

If the data above is directly represented as a graph, the nodes will be aligned randomly in a two-dimensional space and such image will not give a clear idea about the text's structure. In order to produce a more readable representation of the text, we will apply an Force Atlas algorithm (Jacomy, 2009), which is itself derived from force-layout algorithm for graph clustering (Noack, 2007). This algorithm pushes the most connected nodes (hubs) away from each other, while aligning the nodes that are connected to the hubs in clusters around them. This provides a much more readable representation of the graph.

Figures 1 and 2 below show a graph representation of the 2-word gap text network and the 5-word gap text network respectively.

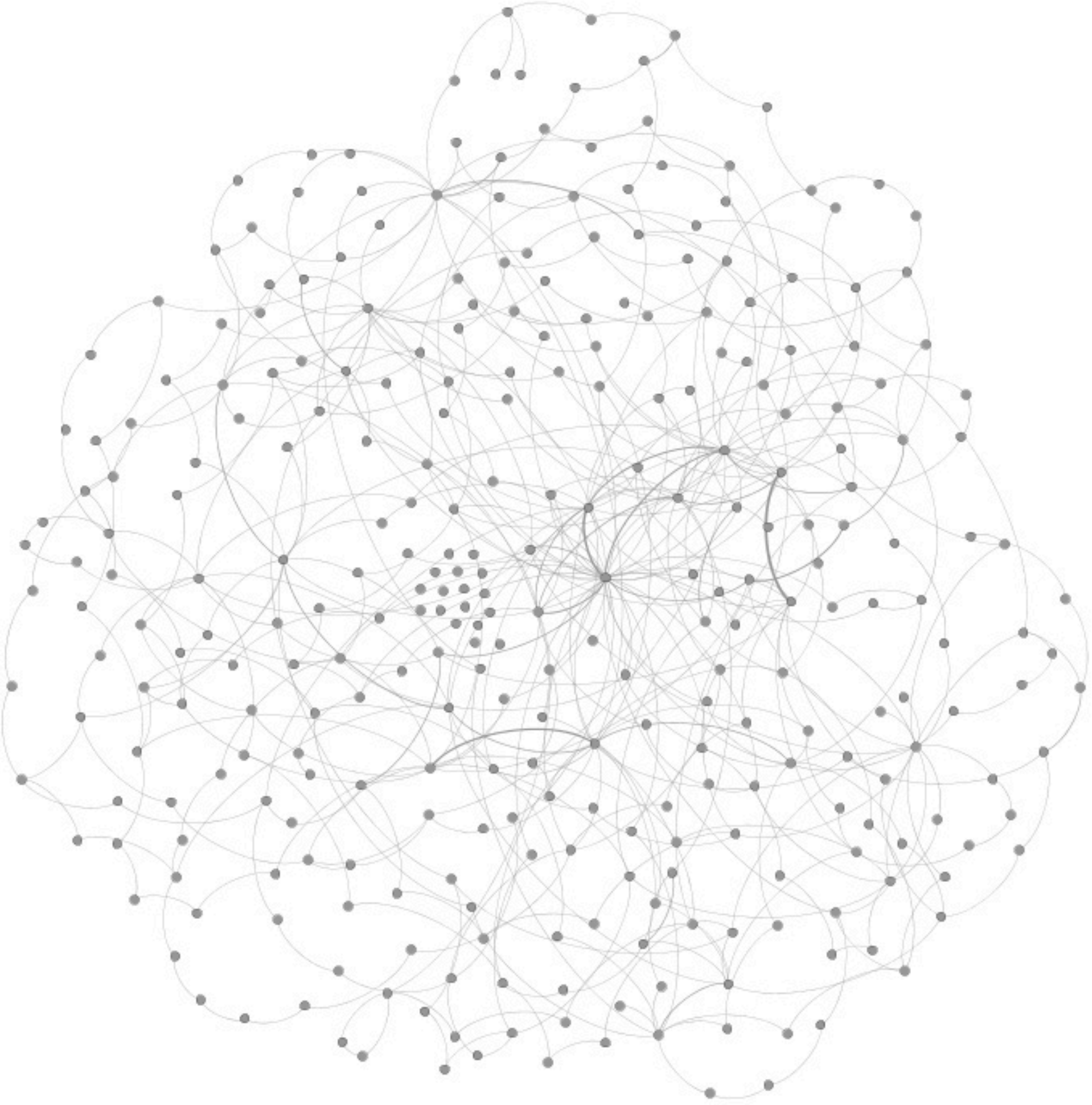

Figure 1: Force-atlas layout of 2-word gap graph

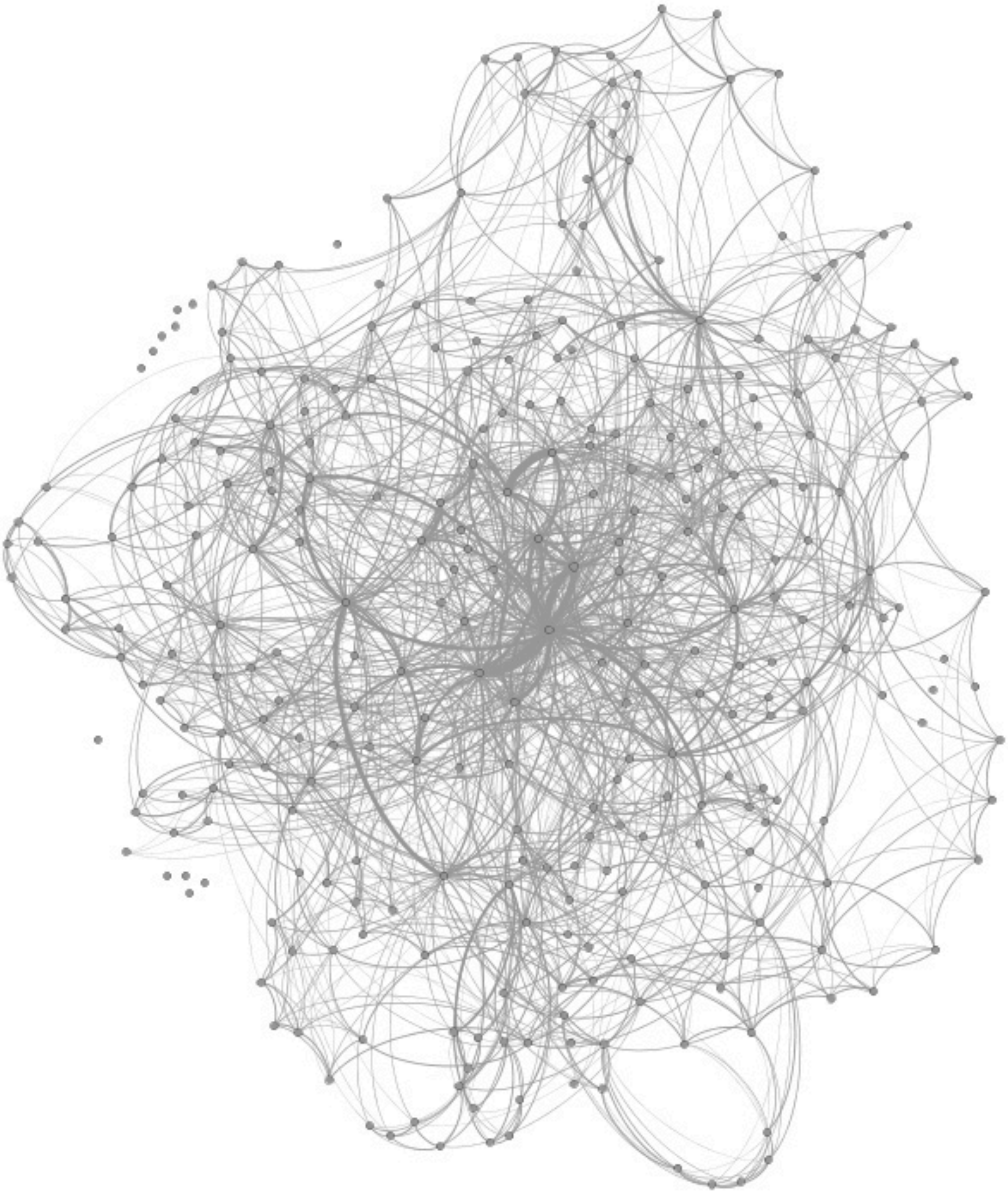

Figure 2: Force-atlas layout of 5-word gap graph

We show these intermediary graphs in order to illustrate the points from the previous chapter. It can clearly be seen that while the first text scan produces a general interconnected structure, the second scan with a bigger word gap makes it much easier to produce a more meaningful visualization, as it emphasizes not only the most frequently mentioned concepts, but also takes into account their local contextual relevance.

Figure 3 below shows the graph that results from putting together the both networks above: the 2-word gap and the 5-word gap one.

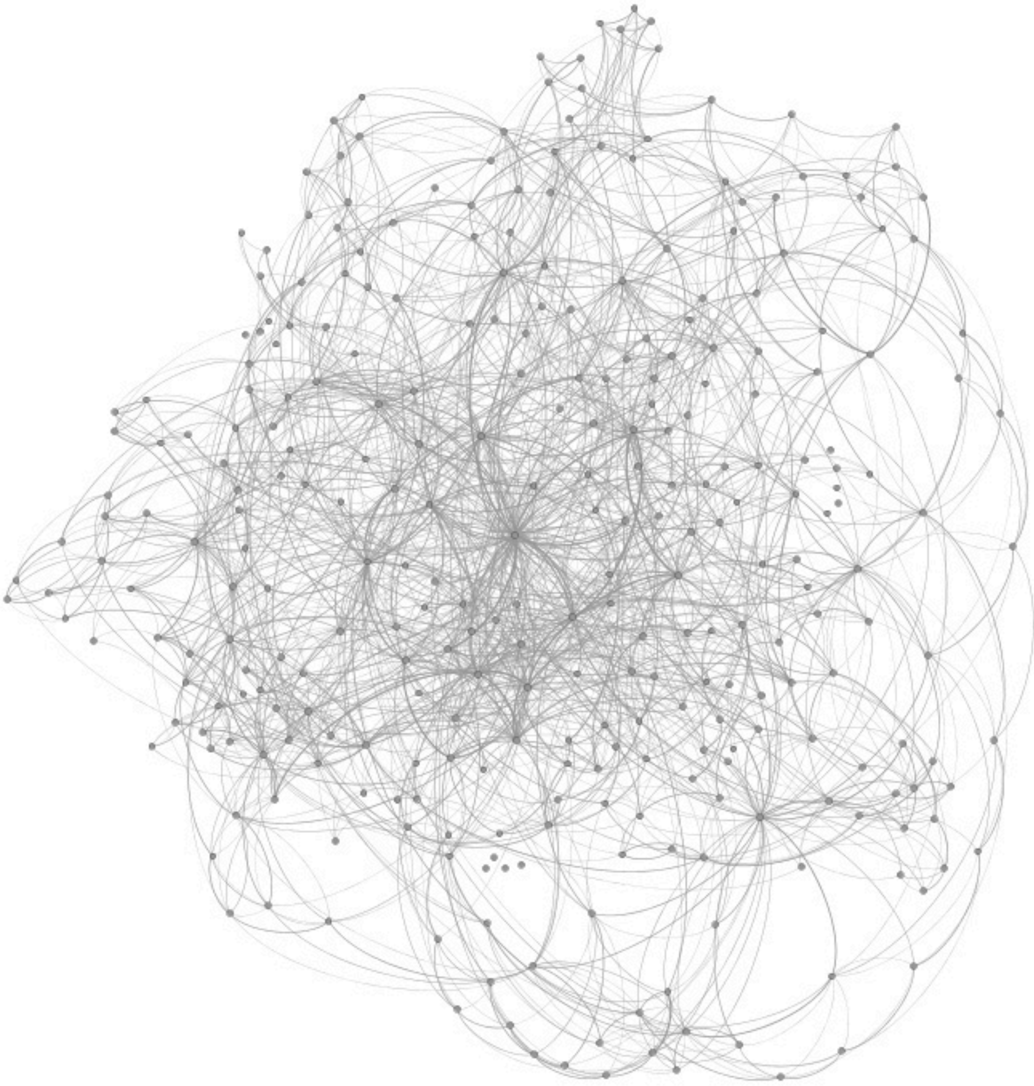

Figure 3: Force-atlas layout of both 2-word gap and 5-word gap text graphs

The visualizations are made using Gephi software and the parameters we use for the Force Atlas layout are the following (Table 1):

Table 1:

Repulsion strength:	10000
Attraction strength:	10
Maximum displacement:	10
Austostabilization strength:	80.0
Austrabilization sensitivity:	0.2
Gravity:	400

In order to provide a more meaningful image, we will range the sizes of the nodes according to their betweenness centrality. Betweenness centrality measure for each node indicates how often it appears between any two random nodes in the network. The higher it is, the more influential is the node because it functions as a junction for communication within the network (Freeman, 1977; Brandes, 2001). Betweenness centrality is different from the node's degree (or the number of edges it has). For instance, it's possible that a node is connected to a lot of other nodes within a certain cluster (high degree), but has few connections to the other clusters in the network. It will then be influential within its cluster, but will have less influence than a node, which has fewer connections, but links different communities together. This is different from tag clouds or other text network visualizations, which often use the node's frequency (or degree in our case) to emphasize the hubs in the network (Kaser, 2007).

Betweenness centrality shows the variety of contexts where the word appears, while high degree shows the variety of words next to which the word appears. In our approach we emphasize the word with the highest betweenness centrality, because they are the most important junctions for meaning circulation within the network. and in that way it's
After running some basic metrics for this network (see Appendix B) and then ranging both the nodes and their edges by their betweenness centrality we obtain the following graph (Figure 4):

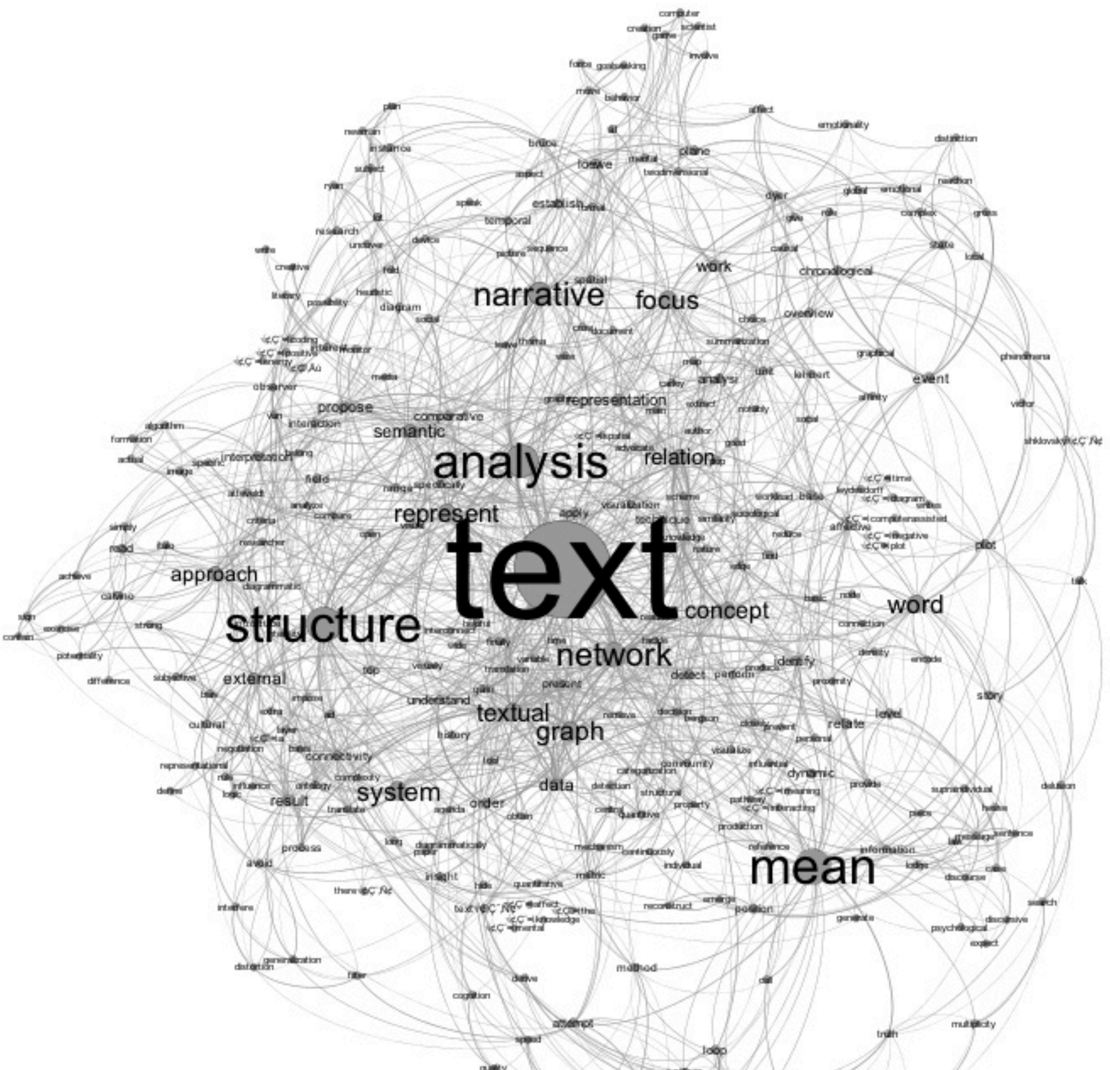

Figure 4: Force-atlas layout, emphasis on the nodes with a higher BC

This representation already shows us that the central junctions for meaning circulation within the introductory part of this article are:

text - structure - mean - analysis - narrative - network

The data below provides a clearer idea about these concepts:

Table 2:

Label	Betweenness Centrality	Degree
analysis	3795.93348067254	74
mean	4018.56122275846	54
narrative	2311.92367170237	49
network	2258.33733816798	65
structure	4051.30522587258	65
text	13149.4837285632	138

As we can see, “text” is the central term in this text and it’s also adjacent to the most words in the network.

The term “analysis” has a high degree, but lower betweenness centrality than the words “structure” and “mean”, which have a lower degree measure. This indicates that the this word “analysis” is an important local hub that binds together a cluster of terms that form a specific context, but it’s not as central as other important terms to the text as a whole.
The term “network” also has the lowest betweenness centrality, but a relatively high degree among the most influential words in the text. This could indicate another contextual cluster around that term within the text.
The term “mean” (morpheme of “meaning”), on the contrary, has a relatively low degree for its high betweenness centrality. That could indicate that just like the term “text” it is used more to connect different contextual clusters together rather than define a certain context within the text.

Thus, so far, we have the two terms “text” and “mean” (or “meaning”), which are the central junctions for meaning circulation within the text, and the terms “analysis” and “network” as the major hubs of two distinct contextual clusters. Finally, the words “structure” and “narrative” function both as the local hubs for contextual clusters and as important junctions for meaning circulation within the whole text. The text could therefore be seen as a tension field between “network” and “analysis” on one side, and “structure” and “narrative” on the other side – mediated by the terms “meaning” and “text”.

The next step is to detect the community structure, in order to be able to see the contextual clusters within the text more precisely. We will use the community detection mechanism (Fortunato, 2010; Blondel, 2008) based on modularity, where the nodes that are mode densely connected together than with the rest of the network are considered to belong to the same community.

The image below shows the community structure for this text – each community has a particular color (Figure 5 and Figure 6 below).

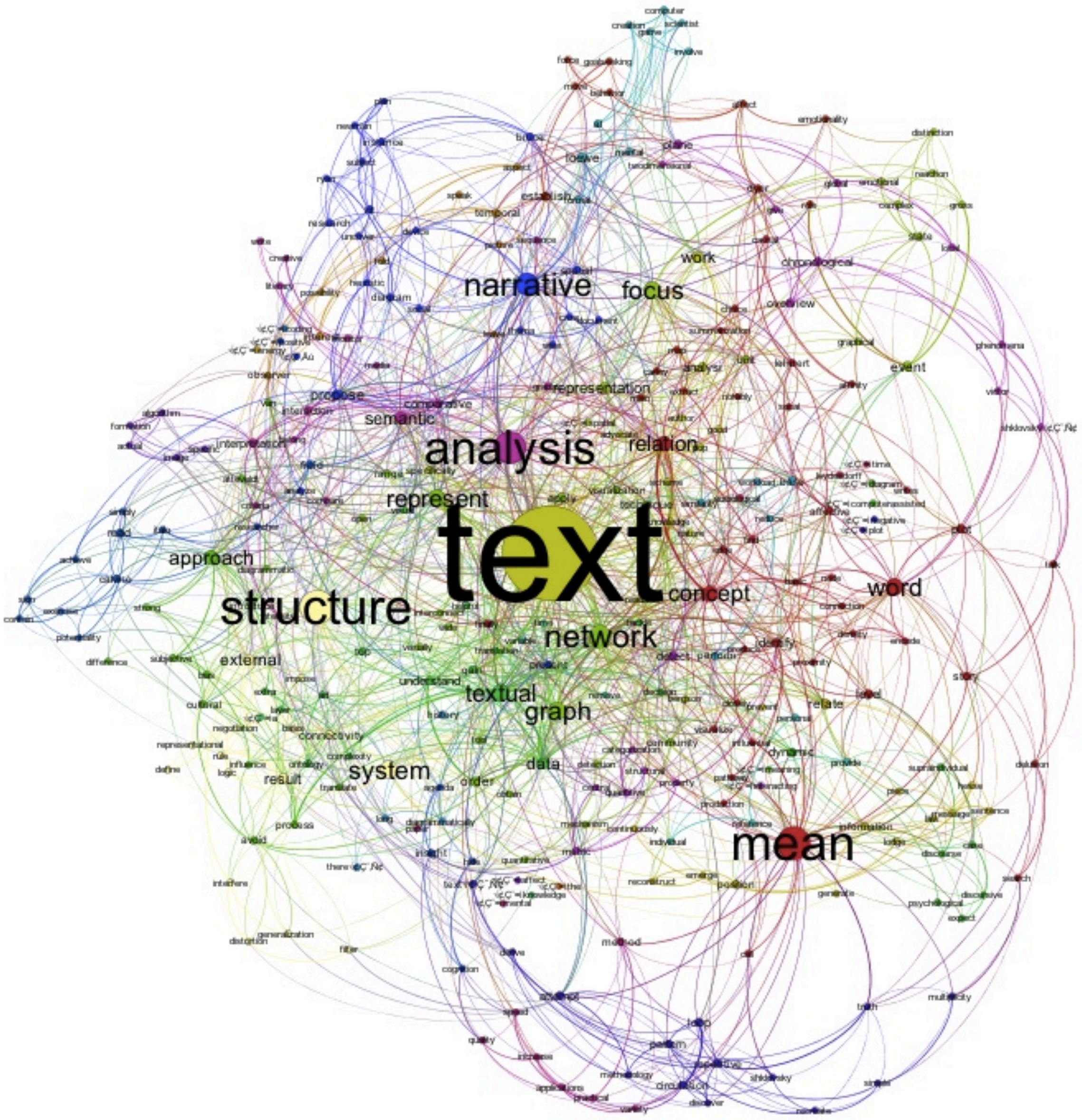

Figure 5: Community structure of the text network

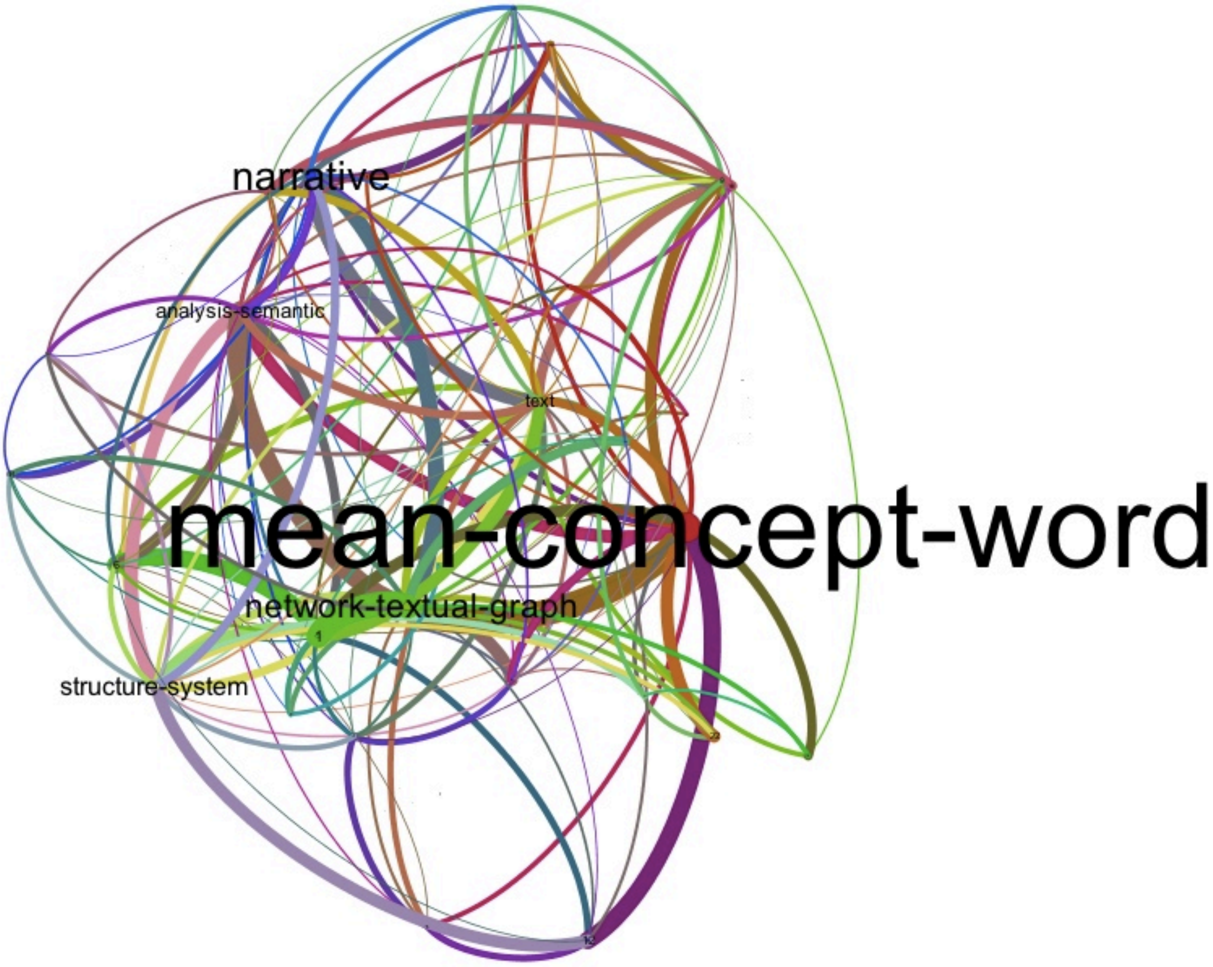

Figure 6: Community-only view

We can clearly see from these images that there are three main contextual clusters within this text. The largest community (context) is comprised of 10.4% of the total nodes and contains the words “mean”, “concept” and “word”. The second largest community (6.8% of the nodes) is the cluster around “narrative” at the top left part of the graph. The third largest community (6.2% of the nodes) is the cluster around “network”, “textual” and “graph”. The fourth largest community (5.2%) is the cluster around “structure” and “system”. The community around the words “text” and “analysis-semantic” are not that large, but the nodes that are contained within them are quite central, which means that these terms function more as junctions rather than hubs for contextual clusters.

4. Data and Graph Interpretation

Putting together the data calculated for the text graph, we can now attempt to identify the pathways for meaning circulation within this text. It is important to provide the global metrics for the graph (Figure 5) before making analysis in order to also have a quantitive insight on its structural properties.

Nodes (Words):	45
Edges (Connections):	241
Average path length:	2.5
Average degree:	5.9
Distance:	5
Graph density:	0.039
Modularity:	0.397
Connected components:	18
Clustering:	0.588

Average path length indicates the number of steps one needs to make on average in the graph in order to connect two randomly selected nodes (Newman, 2010). The lower the number, the more interconnected is the text network, meaning that certain combinations of words occur quite often within the text and that several central concepts appear quite often within contextual clusters, thus contributing to their connectivity. In case of this graph 2.5 is relatively close to the possible minimum (which is 0) and quite far from the possible maximum (which could equal 42 / 4 = 10.5 if all the words were unique), so we can safely say that this particular text graph is quite interconnected.

Graph distance is the longest path between the nodes that exists in the network (Newman, 2010). The maximum with the 2- and 5-word gap we're using equals the number of nodes divided by 4, which is 10.5 in this case. The lowest is 0. High distance value could indicate that there are deviations in the text, which do not have so much to do with the central concepts – for instance, the use of metaphors or elaborate storytelling. In case of this text the distance has average measure that is not so far from the average path length, so it can be said that the whole text (and even its periphery) is quite well connected to the central concepts and to the main contextual clusters.

Average degree is obtained by dividing the total number of edges by the number of nodes (Newman, 2010), showing how many connections (on average) each word has to other unique words in the text. The higher the number, the more there are frequent words within the text and the more diverse and elaborate is the text itself. A lower number could indicate the presence of many repetitions within the text. A lower number could also be obtained if the algorithm used for text scanning does not make connections between the paragraphs and the paragraphs are short (thus indicating dispersed paragraph structure within a text). For this text network we use a 5-word gap scanning, so the minimum possible degree (given that each paragraph contains at least 5 morphemes) is 4. Maximum degree that exists within the text is 138 (see Table 2 above). The distribution of nodes is close to exponential distribution following a power law (see Figure 7 where the X axis is the degree and the Y axis is the number of nodes), most of the nodes (at least 150) have between 5 and 10 edges and only a few have more. So it can be safely said that this text's connectivity is relatively medium and a few but significant number of concepts function as the central, more frequently used ones.

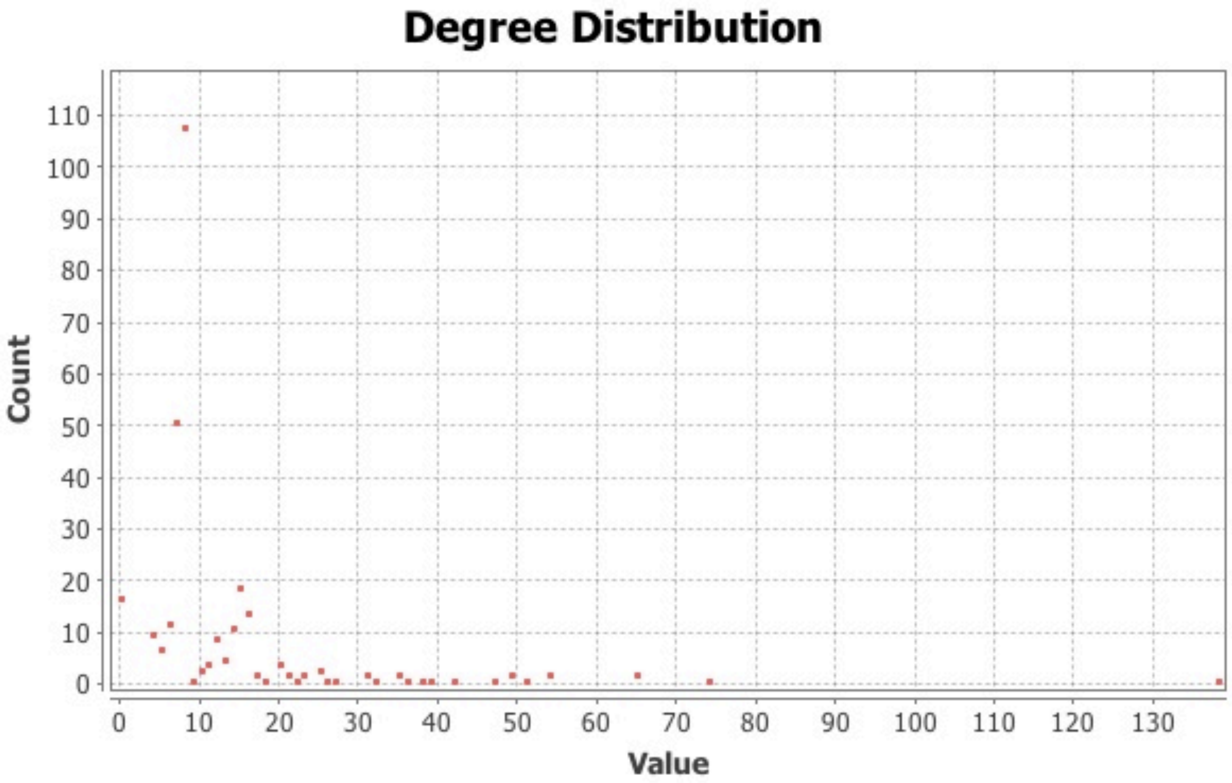

Figure 7: Node degree distribution

The modularity measure that is more than 0.4 (Blondel, 2008) indicates that the partition produced by the modularity algorithm can be used in order to detect distinct communities within the network. It indicates that there are nodes in the network that are more densely connected between each other than with the rest of the network and that their density is noticeably higher than the graph's average.

Now, taking these considerations about the structure of the text into account, we can interpret the data and its visual graph representation (Figures 5 and 6), identify the contextual clusters, and finally identify the pathways of meaning circulation within the text.

First, the *central concepts*, the words (nodes) with the *highest betweenness centrality* (which we also call the *junctions for meaning circulation*) are

text - structure - mean - analysis - narrative - network

We can see from the community structure (Figure 6) that each of these terms is a central node in its own "community", forming what we call *contextual clusters* around them.

Both these central terms and their contextual clusters are the backbone for meaning circulation within the text. Each time this text is read it is most likely that they will play the most important role in establishing the meaning for the text and its interpretation.

The nodes "text" and "mean" (as well as their respective communities) play a more conducive role in the text than other concepts because the relation of their influence to the

total number of connections they have is the highest. These terms and their respective contextual clusters act as *mediators* within the discursive field of the textual graph.
The meaning is generated through the dialectics between the remaining major contextual clusters (represented by their respective terms), which are quite distinct from each other: "narrative" and "network textual graph" as well as "structure-system" and "analysis-semantic". The resulting field of tension (Asvatsaturov, 2007) or dispositif (Deleuze, 1989) is a sum total of these forces that are present at the same time and they form the major loop of meaning circulation within the text.

We could now identify dispositif of the text we are seeing as an interaction between two major themes (contextual clusters): "narrative" on one hand, and "network textual graph" on another. This interaction is mediated through the term "text" and the contextual cluster "meaning-concept-word-relation".

In other words (and that is only one possible interpretation), the text is presenting us with a question of how a narrative could be represented as a network through deconstructing the text into meaningful concept-word relations. The concepts of systems and structure as well as semantic analysis are evoked as the tools that can be used to interpret the resulting textual graph and to offer a different way of reading the text.

There could be many other readings of the resulting structure, of course. The emphasis in our research, however, is on identifying the backbone of meaning circulation, so that it can be visually clear what the main themes of the text in fact are.
When we read a text, we follow a narrative guided by the author, rules of grammar, logic, and common sense. When we read a network, we follow the affirmative drive of contingent associative flow.

"A qualitative multiplicity is not an aggregate of parts with an apparent unity constituted by the relation of separate numerical or physical existents (the Galilean world of purely external relations) but an event, an actual occasion of experience" (Whitehead, 1938)

It is this occasion of experience that text network visualization allows: after all, it is just one more possibility to read any text again and to get new insights about its structure.

5. Conclusion and Future Work

We have here presented a formal approach to identifying the pathways for meaning circulation in textual data. To summarize, the steps are the following:

1) Process the initial textual data in order to convert it into the kind of data that can be represented as a graph, so that network analysis tools can be applied to it and that certain metrics can be calculated;

2) Represent this data visually as a graph and identify
   a) the most ***influential nodes*** (words) that function as ***junctions*** for meaning circulation;
   b) the ***contextual clusters*** or distinct word ***communities*** (or ***themes***) that are present within the text;
   c) the main quantitive properties of the graph as a whole and of the most influential nodes;

3) Using the data from 2) explore the relations between the contextual clusters (communities) and the role of junctions in linking these clusters together. This will identify the ***pathways for meaning circulation*** within the text

4) Find ***alternative pathways*** through which these communities connect bypassing the main loop for meaning circulation – these will usually exemplify the main agenda, but in different terms;

5) Interpret the data.

The method we propose has several practical implications.

First, our approach can help one **identify the text's main agenda** very quickly. The purpose is similar to that of identifying the plot units within text for narrative summarization (Lehnert, 1981) or to uncovering the text's dispositif (Deleuze, 1989). However, our approach is different in the way that the connectivity is calculated based on the graph's properties rather than on affective or semantic relations between the terms. This allows us to postpone the intrusion of subjective cognitive processes into the fabric of the text, which can be especially useful for comparative analysis of several texts or for avoiding overlaying semantic, affective, and ideological layers over the textual structure.

Second, our approach helps to **unlock the potentialities present within the text** (Calvino, 1987) by presenting a text in a non-linear fashion, opening it up for interpretations that are not so readily available via standard sequential reading. It allows the text to speak in its multiplicity or to use Bakhtin's term multiglossia (Bakhtin, 1981).

Third, our approach allows to see the **plot structure within a story** much easier. In this way it's somewhat related to formalist analysis (Shklovsky, 2007) as it allows one to see the form of a text much clearer. It also allows us to see psychological and affective variations in text and uncover the underlying psychological structure for the text's narrative (Rudnev, 2000).

Fourth, our approach can be useful for **group sentiment analysis**. It is possible to conduct interviews with the group members and bring the resulting graphs together in order to reveal the key concepts that bring the group together as well as the peripheral concepts that indicate each member's particular area of expertise and interest. Some attempts in this direction have been done before (for example, Diesner, 2004), however our approach does not require semantic processing of the resulting text and it focuses more on revealing the potential areas for group collaboration rather than uncovering its social structure through text analysis. Besides, it's much less computation-heavy.

Fifth, our approach can be useful for **comparative analysis of different texts** or to detect a certain similarity between different types of textual data. Given that the same processing algorithm is applied to different texts the resulting metrics could be used to categorize these texts according to their structural properties or to automatically detect their formal similarities and dissimilarities.

Sixth, our approach is a way to **visually represent a text as a Gestalt.** This can be especially useful for writers, editors, and copywriters. A text represented as a Gestalt or a diagram allows for a more holistic perception of its interconnectedness (Wertheimer, 1997) and opening up more possibilities for interpretation (Ryan, 2007).

Seventh, our approach allows to create a whole range of **navigational, archival and search tools for textual data** if the resulting visual representations are translated into easy-to-use interfaces. The difference from the existing tag clouds that serve this function (Kaser, 2007) is that the focus is made on interconnectivity between the different concepts and texts, as well as the contextual data, rather than on the terms' frequency of occurrence.

Finally, there are possibilities for **dynamic analysis of textual data** where the resulting graph is visualized in a spatio-temporal frame, allowing one to observe the formation of meaning that is more closely related to the process of reading or listening. This can be especially useful for creating interfaces for navigating related content (the graph representing only the part of the text that is currently read) or for creating a live audio-visual cross-content navigational system where speech recognition mechanisms could provide an immediate graphical representation of the speech and show how it relates to other audio-visual content both within the same context and outside of it.

Appendix A

A list of stopwords used for text modification:

a, a's, able, about, above, according, accordingly, across, actually, after, afterwards, again, against, ain't, all, allow, allows, almost, alone, along, already, also, although, always, am, among, amongst, an, and, another, any, anybody, anyhow, anyone, anything, anyway, anyways, anywhere, apart, appear, appreciate, appropriate, are, aren't, around, as, aside, ask, asking, associated, at, available, away, awfully, be, became, because, become, becomes, becoming, been, before, beforehand, behind, being, believe, below, beside, besides, best, better, between, beyond, both, brief, but, by, c'mon, c's, came, can, can't, cannot, cant, cause, causes, certain, certainly, changes, clearly, co, com, come, comes, concerning, consequently, consider, considering, contain, containing, contains, corresponding, could, couldn't, course, currently, definitely, described, despite, did, didn't, different, do, does, doesn't, doing, don't, done, down, downwards, during, each, edu, eg, eight, either, else, elsewhere, enough, entirely, especially, et, etc, even, ever, every, everybody, everyone, everything, everywhere, ex, exactly, example, except, far, few, fifth, first, five, followed, following, follows, for, former, formerly, forth, four, from, further, furthermore, get, gets, getting, given, gives, go, goes, going, gone, got, gotten, greetings, had, hadn't, happens, hardly, has, hasn't, have, haven't, having, he, he's, hello, help, hence, her, here, here's, hereafter, hereby, herein, hereupon, hers, herself, hi, him, himself, his, hither, hopefully, how, howbeit, however, , i, i'd, i'll, i'm, i've, ie, if, ignored, immediate, in, inasmuch, inc, indeed, indicate, indicated, indicates, inner, insofar, instead, into, inward, is, isn't, it, it'd, it'll, it's, its, itself, just, keep, keeps, kept, know, knows, known, last, lately, later, latter, latterly, least, less, lest, let, let's, like, liked, likely, little, look, looking, looks, ltd, mainly, many, may, maybe, me, mean, meanwhile, merely, might, mine, more, moreover, most, mostly, much, must, my, myself, name, namely, nd, near, nearly, necessary, need, needs, neither, never, nevertheless, new, next, nine, no, nobody, non, none, noone, nor, normally, not, nothing, novel, now, nowhere, obviously, of, off, often, oh, ok, okay, old, on, once, one, ones, only, onto, or, other, others, otherwise, ought, our, ours, ourselves, out, outside, over, overall, own, particular, particularly, per, perhaps, placed, please, plus, possible, presumably, probably, provides, que, quite, qv, rather, rd, re, really, reasonably, regarding, regardless, regards, relatively, respectively, right, said, same, saw, say, saying, says, second, secondly, see, seeing, seem, seemed, seeming, seems, seen, self, selves, sensible, sent, serious, seriously, seven, several, shall, she, should, shouldn't, since, six, so, some, somebody, somehow, someone, something, sometime, sometimes, somewhat, somewhere, soon, sorry, specified, specify, specifying, still, s, sub, such, sup, sure, t's, take, taken, tell, tends, th, than, thank, thanks, thanx, that, that's, thats, the, their, theirs, them, themselves, then, thence, there, there's, thereafter, thereby, therefore, therein, theres, thereupon, these, they, they'd, they'll, they're, they've, think, third, this, thorough, thoroughly, those, though, three, through, throughout, thru, thus, to, together, too, took, toward, towards, tried, tries, truly, try, trying, twice, two, un, under, unfortunately, unless, unlikely, until, unto, up, upon, us, use, used, useful, uses, using, usually, value, various, very, via, viz, vs, want, wants, was, wasn't, way, we, we'd, we'll, we're, we've, welcome, well, went, were, weren't, what, what's, whatever, when, whence, whenever, where, where's, whereafter, whereas, whereby, wherein, whereupon, wherever, whether, which, while, whither, who, who's, whoever, whole, whom, whose, why, will, willing, wish, with, within, without, won't, wonder, would, would, wouldn't, yes, yet, you, you'd, you'll, you're, you've, your, yours, yourself, yourselves, zero

Appendix B:

Data table for network textual data

Label	**Betweenness Centrality**	**Degree**
analysis	3795.93348067254	74
approach	1099.57632490571	38
concept	1376.92248622351	51
data	710.66286544905	39
dynamic	323.476497293674	14
event	453.081493372521	14
external	735.228174789833	31
focus	1587.27835035067	35
graph	1627.36553278749	54
identify	354.642930420184	26
level	382.275733471347	15
loewe	372.910451466497	16
mean	4018.56122275846	54
narrative	2311.92367170237	49
network	2258.33733816798	65
order	392.077647539612	25
plot	328.356068507915	15
propose	524.097981741999	27
relate	421.673692511309	21
relation	1211.34678647754	42
represent	1490.02452614555	47
representation	635.30909225652	22
result	562.411474518848	31
semantic	837.221737993167	36
structure	4051.30522587258	65
system	1627.89395686862	32
text	13149.4837285632	138
textual	1501.31450722844	49
understand	302.798453505203	25
word	1550.70795210806	35
work	676.446245489501	20

References

Asvatsaturov, A. (2007). *Phenomenology of Text: Game and Repression. Moscow: New Literary Review, XLII*

Bakhtin, M.M. (1981). The Dialogic Imagination: Four Essays by M.M. Bakhtin. *Austin: University of Texas Press.*

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

Bergson, H. (2002). The Possible and the Real. *Continuum.*

Blei, D. Ng, A. Jordan, M. (2003) Latent Dirichlet Allocation. *In Journal of Machine Learning Research, 1/03, 993-1022*

Blei, D. (2004). Probabilistic Models of Text and Images. *Disseration, University of California, Berkeley*

Blondel, V.; Guillaume, J-L; Lambiotte, R; Lefebvre, E; (2008). Fast Unfolding of Communities in Large Networks. *In Journal of Statistical Mechanics: Theory and Experiment, Volume 2008*

Brandes, A. (2001). Faster Algorithm for Betweenness Centrality. *In Journal of Mathematical Sociology 25(2):163-177*

Brandes, U.; Eiglsperger, M.; Herman. I., Himsolt, M.; and Marshall M.S. (2002). GraphML Progress Report: Structural Layer Proposal. *Proc. 9th Intl. Symp. Graph Drawing (GD '01), LNCS 2265, pp. 501-512. Springer-Verlag,*

Bruce, D; Newman, D. (1978). Interacting Plans. In *Cognitive Science, Volume 2, Issue 3, July-September: 195-233*

Calvino, I. (1987). The Uses of Literature: Essays. *Mariner Books.*

Carley, K. (1993). Coding Choices for Textual Analysis: A Comparison of Content Analysis and Map Analysis. In *Social Methodology, Vol. 23: 75-126*

Chang, J. & Blei, D. (2010). Hierarchical Relational Models for Document Networks. *Annals of Applied Statistics, Vol. 4, No. 1, 124–150*

Deleuze, G. (1989). Qu’est-ce qu’un dispositif. *In Michel Foucault Philosophe: Rencontre internationale, Paris: Seuil, 185-95.*

Diesner, J; Carley, K. (2004). Using Network Text Analysis to Detect the Organizational Structure of Covert Networks, *Center for Computational Analysis of Social and Organizational Systems (CASOS), Institute for Software Research International (ISRI) School of Computer Science Carnegie Mellon University*.

Dyer, G. (1983). The Role of Affect in Narratives. *In Cognitive Science 7: 211-242,* Fortunato, S. (2010). Community Detection in Graphs. *In Complex Networks and Systems Lagrange Laboratory, Torino.*

Freeman, L. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry Vol. 40, No. 1 (Mar., 1977): 35-41*

Gephi. An Open Source Software for Exploring and Manipulating Networks. Gephi Consortsium. *URL:* *http://www.gephi.org*

Hesse, M. (1980). Revolutions and Reconstructions in the Philosophy of Science. *London: Harvester Press.*

Hofmann, G. (1999). Probabilistic latent semantic indexing. *In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*

Jacomy, M. (2009). Force-Atlas Graph Layout Algorithm. *URL:* *http://gephi.org/2011/forceatlas2-the-new-version-of-our-home-brew-layout/*

Kaser, O. & Lemire, D. (2007). Tag-Cloud Drawing: Algorithms for Cloud Visualization. In *Proceedings of Tagging and Metadata for Social Information Organization (WWW 2007)*

Krovetz, R. (1993). Viewing Morphology as an Inference Process. *SIGIR '93 Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*

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

Law, J; Lodge, P; (1984). Science for Social Scientists. *London: Macmillan.*

Lehrnert, W. (1981). Plot Units and Narrative Summarization. In *Cognitive Science, Volume 5, Issue 4: 293–331*

Leydesdorff, L. (2011). ‘Meaning’ as a sociological concept: A review of the modeling, mapping and simulation of the communication of knowledge and meaning. In *Social Science Information 50(3–4) 391–413*

Li, W. & McCallum, A. (2006). Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations. *In Proceedings of the 23rd International Conference on Machine Learning*

Loewe, B. (2010). Comparing Formal Frameworks of Narrative Structure. In *Association for the Advancement of Artificial Intelligence,*

Loewe, B.; Pacuit, E.; and Saraf, S. (2009). Identifying the structure of a narrative via an agent-based logic of preferences and beliefs: Formalizations of episodes from CSI: Crime Scene Investigation. In *Duvigneau, M., and Moldt, D., eds., Proceedings of the Fifth International Workshop on Modelling of Objects, Components and Agents. MOCA’09, FBI-HH-B-290/09, 45–63.*

MySQL. An Open Source Database. *URL:* *http://www.mysql.com*

Neo4j. An Open Source NoSQL Graph Database. *URL:* *http://www.neo4j.org*

Newman, M. (2010). Networks: An Introduction. *Oxford: Oxford University Press*

Noack, A. (2007). Energy Models for Graph Clustering. In *Journal of Graph Algorithms and Applications, vol 11, no 2: 453-480*

Shklovsky, V; (2007). Energy of Delusion: A Book on Plot. *Dalkey Archive Press.*

Popping, R. (2003). Knowledge Graphs and Network Text Analysis. In *Social Science Information*

Popping, R. (2000). Computer-assisted Text Analysis. *SAGE*.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program: electronic library and information systems, Vol. 14 Iss: 3, pp.130 - 137*

Rudnev, V.P. (2000). Away from Reality: Study of Text Philosophy. *Moscow: Agraph*

Ryan, M-L. (2007). Diagramming Narrative. In *Semiotica 165–1/4: 11–40.*

ThisIsLike. An online mnemonic network. Nodus Labs. *URL:* *http://www.thisislike.com*

Van Atteveldt, W. (2008). Semantic Network Analysis: Techniques for Extracting, Representing, and Querying Media Content. *Booksurge LLC*

Vicknair, C.; Macias, M.; Zhao, Z.; Nan, X.; Chen, Y.; and Wilkins, D; (2010). A comparison of a graph database and a relational database: a data provenance perspective. *In Proceedings of the 48th Annual Southeast Regional Conference (ACM SE '10). ACM, New York, NY, USA, , Article 42 , 6 pages. DOI=10.1145/1900008.1900067 http://doi.acm.org/10.1145/1900008.1900067*

Wertheimer, M. (1997). Gestalt Theory. *New York: Gestalt Journal Press*

Whitehead, A.N. (1938). Modes of Thought. *New York: Macmillan*

Wise, J; Thomas, J; Pennock, K.; Lantrip, D.; Pottier, M.; Schur, A.; and Crow, V. (1995). Visualizing the non-visual: spatial analysis and interaction with information from text documents. From *Information Visualization 1995 Proceedings: 51-58,* 1995

Zipf, G. K. (1935). The psycho-biology of language. *Oxford, England: Houghton, Mifflin. ix, 336 pp.*