Using Text Network Analysis in Corpus Studies — A Comparative Study on the 2010 TU-154 Polish Air Force Presidential Plane Crash Newspaper Coverage

Marta Gruszecka and Michal Pikusa

***Abstract*—This paper presents the results of a comparative study of corpora gathered from newspaper articles on the 2010 Polish Air Force TU-154 crash in Smolensk, Russia. We investigated the main concepts around which the narrative structure of the articles was built using the text network analysis. For the analysis we gathered articles from two mainstream Polish newspapers, Gazeta Wyborcza and Nasz Dziennik, and transformed the texts from the resulting corpora into corresponding graphs. Each graph consisted of nodes based on the words included in the text, and edges between the nodes based on the proximity relations between the words. Resulting graphs were then filtered to show the most important nodes, which acted as junctions for meaning circulation in the texts. After comparing the graphs from both corpora we found that the articles from Gazeta Wyborcza focused more on the aspects of Polish-Russian dialog, while articles from Nasz Dziennik focused on the national aspects of the tragedy. These findings show that the text network analysis can be successfully used as an alternative to other statistical corpus methods for discourse analysis.**

***Index Terms*—2010 polish air force TU-154 crash, quantitative corpus analysis, Smolensk catastrophe, text network analysis.**

I. Introduction

The present study examines a body of data compiled from newspaper articles covering the 2010 crash of the Polish Air Force TU-154 in order to extract the main concepts around which the articles' narratives are built. On the 10th of April 2010 96 delegates, including the President of Poland Lech Kaczynski, his wife, former President Ryszard Kaczorowski, many military and government officials, prominent members of Polish clergy, as well as family members of the Katyn Massacre victims and other figures prominently present in the Polish public sphere boarded the Polish Air Force TU-154 plane heading to Katyn. The delegates were to participate in a commemorative ceremony remembering the Katyn Forest massacre of 1940, an event still painfully present in the collective memory of the Polish nation.

In 1940 in the Katyn area (including the grounds surrounding Smolensk), approximately 21,000 Polish prisoners of war, mostly Polish officers, were brutally murdered by the Soviet Army and buried in the surrounding forests [1]. The massacre was held in secrecy until the discovery of mass graves made by German officers in 1943. Although found on Russian territory, the government of Russia blamed the crime on Nazi Germany and closed the investigation. For decades to come the families of Polish officers who had gone missing in the Katyn area were left with no explanation of what happened to their relatives. Moreover, the families of the victims faced tremendous difficulties when trying to investigate the case on their own or draw public attention to the tragedy [1]. Allen Paul [2] describes the silence surrounding the Katyn Forest massacre as a "complex symbol of Polish suffering at the hands of Stalin". Although Russia finally acknowledged its blame for the crime in 1990, the Katyn Forest massacre remains a troubling point in Polish history.

The 2010 TU-154 plane crash in Smolensk was not only a blow to the Polish nation as a catastrophe that claimed the lives of its President and a sizable portion of prominent governmental officials, but it also ignited a heated discussion concerning the crash in the context of the Katyn Forest massacre, including some members of the public sphere putting forward theories of Russia's involvement in the catastrophe.

That is why in this study we examine the newspaper articles about the crash and its aftermath, and with the help of text network analysis we extract the pathways of meaning circulation, which help us identify the main concepts that contribute to the meaning of the texts. Using the text network analysis, we visualize a readable graph from the text, thus gaining a better understanding of its hidden agendas and narrative structure. As any text can be easily transformed into a graph, with words acting as nodes and relationships between them acting as edges, the text network analysis provides a fair tool of comparative analysis of textual data.

Even though the graph approach to text analysis is not new, most of the studies introduce subjective bias into the resulting graphs, as they rely on relationships between concepts based on casual relations [3], affective affinity [4], [5], semantic relations [6], and chronological sequence [7]. In order to avoid such a bias, we use the words' proximity as a base for a relationship between them. In this way we extract the meaning of the texts by identifying the clusters of co-occurent words within them. In the present study, we use the text network analysis of texts about the Smolensk catastrophe from two influential Polish newspapers in order to extract the concepts around which the texts revolve.

Manuscript submitted December 12, 2013; revised January 20, 2014.

The authors are with the Faculty of English, Adam Mickiewicz University, Poznan, Poland (e-mail: mgruszecka@wa.amu.edu.pl, mpikusa@wa.amu.edu.pl).

II. Methodology

A. Data Collection

The corpora compiled for the purposes of this study are a compilation of all articles related to the TU 154 Presidential plane crash in Smolensk that appeared in two national Polish newspapers, *Gazeta Wyborcza* (a liberal title, estimated around 319,000 issues, critical towards Kaczynski's administration) and *Nasz Dziennik* (a conservative title advocating for patriotism or religion, supportive of Kaczynski's political program, estimated circulation around 250,000 issues). The data were taken from printed editions of both newspapers, specifically from the general news sections (domestic news and world news, unless the issue was not divided into sections). Moreover, the data were taken from national editions of the newspapers, regional editions were excluded. All the articles selected for each corpus appeared in press during the time period of three weeks from the tragedy. The choice of the time period, though targeted at analyzing the initial reaction to the tragedy, was arbitrary. An article was tagged as Smolensk plane crash related and included in the corpus if it contained the word *Smolensk* (treated as a morpheme), TU-154 or the words *catastrophe* or *tragedy* in collocation with *Smolensk* or *Katyn* (treated as a morpheme). The corpus compiled from *Gazeta Wyborcza* contains c. 120,000 words, taken from 170 articles selected as Smolensk plane crash related, while the one compiled from *Nasz Dziennik* contains c. 140,000 words taken from 246 articles tagged as Smolensk related.

B. Text Processing

Before each of the corpora could be visualized as a graph, a few preprocessing steps had to be taken.

The first step was to normalize the text of the corpus by removing all punctuation and symbols, numbers, and unnecessary spaces, so that the resulting text was a continuous series of words. Then, all capital letters were converted into lowercase, in order to avoid the later identification of two occurrences of the same word as two different words. Finally, all stopwords, i.e. the words that bind the text together but do not contribute to the meaning, were removed. As the stopwords are the most frequently used words, removing them helped reduce the amount of noise in the corpus and made detection of abnormal distribution of words easier, as the text no longer followed the Zipf's power law [8].

The next step was to transform all the words in the remaining corpus to their base forms through lemmatization. As Polish is a highly inflectional language, lemmatization helps group together different inflected forms, thus reducing the complexity of the resulting network. Lemmatization was done using PSI-TOOLKIT software [9].

C. Graph Construction

In order to convert the text into the graph data, an algorithm was devised to make a graph file readable by a graph visualization and analysis tool Gephi [10].

First, the normalized text of the corpus was scanned using a 2-word gap. Each of the words, if it appeared for the first time in the text, was recorded as a new node in the network. When two words appeared within the gap, the algorithm first checked whether the pair already existed. If it did not, an edge (connection) between the words was established, with the first word being the source word and the second word being the target word, and the weight of the edge was set to 1. If the pair of words already existed, the weight was incremented by 1. This way each connection was based on words' proximity to each other, and the more frequent the combination of two words, the higher the weight of the connection between them. After the scanning of the whole corpus, a graph file with unique nodes, edges between them, and weights was created.

D. Graph Visualization and Parameter Calculation

When the graph data were loaded into Gephi, the nodes were aligned randomly in two-dimensional space. As such representation does not give a clear idea about the text structure, a Force Atlas algorithm [11] was applied to the graph. This pushes the most connected nodes (hubs) away from each other and aligns the nodes that are connected to those hubs in clusters around them. This way a more readable representation was made.

After aligning the nodes, we calculated the betweenness centrality and ranged the size of the nodes according to this measure. Betweenness centrality measures how often a certain node in the network lies on the shortest path between two random nodes in the network. The higher it is, the more influential is the node, as it functions as a junction for communication within the network [12]. This way it shows the variety of contexts where the word appears, as the words with highest betweenness centrality are the most important junctions for meaning circulation [13]. Along with betweenness centrality, we also calculated the degree, i.e. the number of nodes that are connected to each word. This helped determine whether the words with highest betweenness centrality acted as important junctions globally, in which case the degree is relatively high, or locally, in which case the degree is relatively low. Then, we calculated Pearson's correlation coefficient to check if there was a correlation between the measures of betweenness centrality and degree, as a statistically significant value would point to a steady trend in either the global direction (positive r value) or the local direction (negative r value).

The last step was to detect community structure in order to make contextual clusters more visible. Community detection mechanism [14] was used, in which the nodes that were more densely connected together than with the rest of the network were considered a part of the same community. After applying the mechanism, we filtered out the nodes from the graph, leaving only the contextual communities that contributed to meaning circulation. Random colors were assigned to each community to make the distinction easier. The size of the words on the graph relates to their betweenness centrality, and the size of the edges relates to the weight of the connections between them.

III. Results

The analysis of the graph based on the corpus from Gazeta Wyborcza revealed 23 communities revolving around the

words with the highest betweenness centrality. These words are, ordered from the ones with the highest betweenness centrality to the ones with the lowest, polski (EN. Polish), mówić (EN. talk), prezydent (EN. president), człowiek (EN. person), katastrofa (EN. catastrophe), mieć (EN. have), samolot (EN. plane), być (EN. be), pilot (EN. pilot), rosyjski (EN. Russian), tragedia (EN. tragedy), czas (EN. time), raz (EN. time), chcieć (EN. want), smoleński (EN. Smolensk, adj.), polityk (EN. politician), polityczny (EN. political), państwo (EN. country), zostać (EN. become), ofiara (EN. victim), osoba (EN. person), miejsce (EN. place), and powiedzieć (EN. say). There was a significant positive correlation between the measures of betweenness centrality and degree, as the Pearson's correlation coefficient r was .98 ($p < .001$). The average degree of the graph was 11.3, and the average path length was 1.31. The graph is presented in Fig. 1.

The analysis of the graph based on the corpus from Nasz Dziennik revealed 21 communities centered on the words with the highest betweenness centrality. These words are, ordered from the ones with the highest betweenness centrality to the ones with the lowest, polski (EN. Polish), katastrofa (EN. catastrophe), człowiek (EN. person), mieć (EN. have), być (EN. be), prezydent (EN. president), osoba (EN. person), narodowy (EN. national), tragedia (EN. tragedy), chcieć (EN. want), zostać (EN. become), czas (EN. time), śmierć (EN. death), sprawa (EN. issue), miejsce (EN. place), wielki (EN. great), powiedzieć (EN. say), samolot (EN. plane), móc (EN. can), życie (EN. life), mówić (EN. talk). Even though there was a significant positive correlation between the measures of betweenness centrality and degree, as the Pearson's correlation coefficient r was .88 ($p < .001$), two words, zostać and samolot, contributed to the lower r the most by skewing the trend in the direction of local connectivity in the case of zostać, and in the direction of global connectivity in the case of samolot. The average degree of the graph was 8.2 and the average shortest path was 1.3. The graph is presented in Fig. 2.

IV. Discussion

As can be seen in the graphs both corpora built their narratives around the national (Polish) dimension of the Smolensk catastrophe, although each newspaper chooses a significantly different focus on the issue. While in both cases one of the most prominent nodes is polski (EN. Polish), in the corpus compiled from Gazeta Wyborcza polski (EN. Polish) appeared mostly in the context of words signifying the victims of the crash, as can be inferred from the nodes closely surrounding it. Hence, in Gazeta Wyborcza the focus appears to be on the victims of the Smolensk catastrophe as a personal loss for the Polish nation. In the case of Nasz Dziennik, however, the nodes surrounding polski (EN. Polish) suggest a socio-political focus on the significance of the plane crash on Poland. Here, the focus is shifted from the victims of the catastrophe to the tragic event itself. In fact, the node ofiara (EN. victim) is absent in the graph compiled from Nasz Dziennik whatsoever.

Moreover, the nodes raz (EN. time) and chcieć (EN. want) further suggest that the texts in Gazeta Wyborcza may present the 2010 tragedy as an opportunity and willingness for political dialog provoked by another time when the Katyn forest became a place of tragic death for Polish citizens. In contrast, while chcieć (EN. want) in Gazeta Wyborcza suggests a will to take action, móc (EN. can) in Nasz Dziennik can suggest either the ability (or lack of ability) to take action in investigating the crash. In Gazeta Wyborcza the political aspect of the catastrophe is concerned mostly with dialog through mówić (EN. talk), while in Nasz Dziennik one of the two nodes standing for dialog, powiedzieć (EN. say) is not in a close proximity to any other concepts, whereas mówić (EN. talk) is in close proximity to prezydent (EN. president). It appears that in Nasz Dziennik verbs connected with speech are used for reporting purposes, while in Gazeta Wyborcza they may also imply political dialog.

Another interesting observation can be made when looking at the node prezydent (EN. president), as in Gazeta Wyborcza this node is in close proximity to the node rosyjski (EN. Russian), which is not present in the graph made from Nasz Dziennik. This suggests that the texts from Gazeta Wyborcza indeed focused more on the dialog between the two nations which resulted from this tragic event, as also signified by the prominent edge connecting polski (EN. Polish) and rosyjski (EN. Russian). The same observation can be made, when analysing the nodes polityk (EN. politician), polityczny (EN. political), and państwo (EN. country), which are not present in the graph compiled from Nasz Dziennik. While the aforementioned nodes indicate a political focus of the texts and draw an image of Poland as a civic society, the nodes found exclusively in the graph from Nasz Dziennik suggest a different situation. The nodes narodowy (EN. national), śmierć (EN. death), życie (EN. life) and wielki (EN. great) contribute to a poetic, martyrological narration of Poland as a nation, not so much a political state.

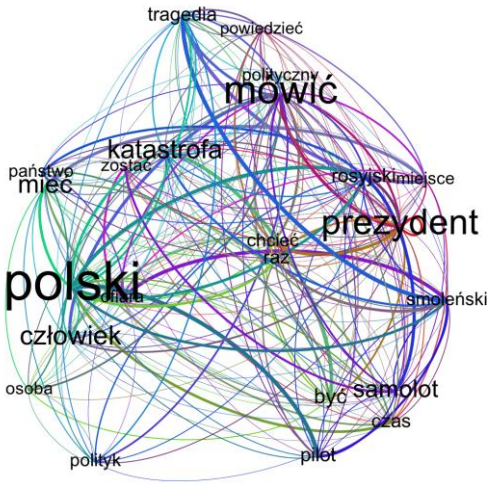

Fig. 1. A graph showing distinct words, around which the meaning circulates in the texts from *Gazeta Wyborcza*. The size of the nodes depicts the importance of the word as a meaning junction, and the size of the edges depicts the strength of the connection between the words.

What is more, the graphs suggest that in Gazeta Wyborcza the TU-154 crash functions under the concept of the Smolensk tragedy (as signified by the strong edge between tragedia [EN. tragedy] and smolenski [EN. Smolensk, adj].), while in Nasz Dziennik the main concept that the texts focus

on is the Polish tragedy, as there is a strong edge between the two nodes, in the absence of Smolenski in the graph (EN. Smolensk, adj.).

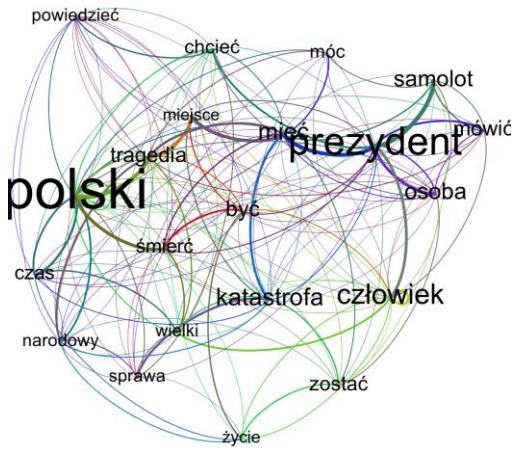

Fig. 2. A graph showing distinct words, around which the meaning circulates in the texts from *Nasz Dziennik*. The size of the nodes depicts the importance of the word as a meaning junction, and the size of the edges depicts the strength of the connection between the words.

V. Conclusions

In conclusion, it is clearly visible that the text network analysis can yield new possibilities in quantitative corpus analysis and make the data easily interpretable due to readable visualizations. In the case of the data analyzed in this study, the text network analysis showed that there is in fact a significant difference between the conceptual structure of the narratives in two corpora compiled from the same type of newspaper articles (general news sections). In the first corpus we observed a general focus on the Polish-Russian dialog concerning the Smolensk tragedy, treating the victims as a personal loss for their country. In the second corpus, however, the focus appears to be on the national aspect of the tragedy. Even though it is still a quantitative study, the method used provides quick and clear insight into the contextual surroundings of the most significant concepts found in the data.

References

[1] M. Kalinowska, "Monuments of memory: Defensive mechanisms of the collective psyche and their manifestation in the memorialization process," *Journal of Analytical Psychology*, vol. 57, pp. 425-444, 2012.
[2] A. Paul, *Katyń: The untold story of Stalin's Polish Massacre*, Charles Scribner's Sons, New York, 1991.
[3] D. Bruce and D. Newman, "Interacting plans," *Cognitive Science*, vol. 2, no. 3, pp. 195-233, 1978.
[4] W. Lehrnert, "Plot Units and narrative summarization," *Cognitive Science*, vol. 5, no. 4, pp. 293-331, 1981.
[5] G. Dyer, "The role of affect in narratives," *Cognitive Science*, vol. 7, pp. 211-242, 1983.
[6] W. V. Atteveldt, *Semantic Network Analysis: Techniques for Extracting, Representing, and Querying Media Content*, Booksurge LLC, 2008.
[7] B. Loewe, E. Pacuit, and S. Saraf, "Identifying the structure of a narrative via an agent-based logic of preferences and beliefs: formalizations of episodes from CSI: Crime scene investigation," in *Proc. the Fifth International Workshop on Modeling of Objects, Components and Agents*, pp. 45-63, 2009.
[8] G. K. Zipf, *The Psycho-Biology of Language*, Oxford, England: Houghton, Mifflin. ix., 1935
[9] F. Graliński, K. Jassem, and M. J. Dowmunt, *PSI-Toolkit: Natural Language Processing Pipeline. Computational Linguistics-Applications*, Heidelberg: Springer, 2012.
[10] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: open source software for exploring and manipulating networks," *Association for the Advancement of Artificial Intelligence*, pp. 361-362, 2009.
[11] M. Jacomy. (2009). Force-Atlas Graph Layout Algorithm. [Online]. Available: http://www.gephi.org/2011/forceatlas2-the-new-version-of-our-home-brew-layout/
[12] A. Brandes, "Faster algorithm for betweenness centrality," *Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163-177, 2001.
[13] P. Dmitry, *Identifying the Pathways for Meaning Circulation Using Text Network Analysis*, Nodus Labs: Berlin, 2011.
[14] S. Fortunato, *Community Detection in Graphs. Complex Networks and Systems*, Lagrange Laboratory, Torino, 2010.

Marta Gruszecka lives in Poznan, Poland and she is a Ph.D. student at the Faculty of English, Adam Mickiewicz University. In 2010, she obtained her master's degree in english from School of English, Adam Mickiewicz University. Her research interests include collective and cultural trauma and political discourse analysis.

Michal Pikusa lives in Poznan, Poland and he is a Ph.D. student at the Faculty of English, Adam Mickiewicz University. In 2011, he obtained his master's degree in english from School of English, Adam Mickiewicz University. His research interests include neurolinguistics and computational methods in linguistics.