A computer-assisted analytical technique addressing a topic of relevance to the proposed dissertation.

Nina Lapot

UCC MADAH

117110040

DH6014 Digital Skills for Research Postgraduates in the Humanities and Social Science

Dr. James O'Sullivan

April 20, 2018

It has been estimated that over 14 million people passed through the hundreds of camps that comprised the Soviet Gulag labour system from the years 1930 to 1960. Much literature has been dedicated to the many facets of this phenomenon and infamous works such as Solzhenitsyn's *Gulag Archipelago* and *One Day in the Life of Ivan Denisovich* are the predominant instances of the genre.

For the purposes of this assignment, it was decided to analyse more far reaching narratives beyond those that are traditionally associated with the genre. Although the gulag system is widely regarded as having being constituted solely of labour camps, it was in fact much wider in scope, also encompassing settlements and open villages.

In order to specify my aims for this assignment, a breakdown of the two texts chosen is deemed necessary. Both volumes cover survivors' testimonies in the approximate years around World War Two and the analysis herein is specific to this time frame.

Anne Applebaum's *Gulag Voices* contains thirteen, self-authored, accounts from individuals who survived the Gulag system and the book is loosely structured to track the path from arrest to release. The volume amounts to approximately 47,000 words. The authors are literate and well educated and the background of the writers reflects somewhat the diversity of the camps themselves, counting amongst them an American writer and some renowned literary scholars. However, as Applebaum herself relates, and warns, in the preface, some important aspects of the Gulags are not represented by those portrayed in the collection. "Although the majority of the Gulag's prisoners, particularly in the early days, were peasants and uneducated

workers, their experiences do not feature here, for the simple reason they could not write." (Applebaum 2011, 131)

Gulag Voices - Oral Histories of Soviet Incarceration and Exile by Jehanne M. Gheith and Katherine R. Jolluck gives voice to that group omitted in the Applebaum volume. Standing at approximately 88,000 words, the book is comprised of sixteen oral histories of individuals and is predominantly concerned with the dekulakization process, a plan which originally targeted richer peasants but quickly became indiscriminate in this aim. The interviewees are, for the most part, less educated and articulate than those in the Applebaum volume. The interviews therein were largely gathered by Gheith and Jolluck between 1998 and 2006 and so the participants are older and the narratives less focused than in *Gulag Voices.*

Two computer assisted analytical techniques, Mallet and Voyant, will be employed to analyse these collections of literature. The aim of this analysis will be to decipher the manner in which these different social strata within the gulag system perceived their lives. Insight will be garnered into the different experiences of just *two* of the different groups in the system at the time and a research narrative built out of the predominant findings.

Matthew L. Jockers states that "as a result of a disciplinary habit of thinking small, the traditionally minded scholar…often fails to recognize the potentials for analysis that an electronic processing of texts enables," (Jockers 2013, 17) while Stanford scholar Franco Moretti refers to the use of technology that enables us to get a 'birds eye view of literature' as 'distant reading.' Large scale digitization of texts has enabled this new potential for a wider scope of analysis and insight and allows the individual to bring interpretation to a huge manner of texts that was never before attainable. "Big data have been a major catalyst," according to Jockers, and

"the questions we may now ask were previously unconceivable, and to answer these questions requires a new methodology, a new way of thinking about our object of study." (4)

Jockers suggests a “blended approach ...a unification of the macro and micro scales that promises a new, enhanced and better understanding of the literary record.” (Jockers 2013, 26) This element of human interpretation remains crucial and allows the scholar to build theories that may have before remained hidden when engaged in a method of close reading.

MALLET (MAchine Learning for LanguagE Toolkit) is a tool used for the application of machine learning and topic modeling. Robert K. Nelson (2011), director of the Digital Scholarship Lab, explains that “the real potential of topic modeling . . . isn’t at the level of the individual document. Topic modeling, instead, allows us to step back from individual documents and look at larger patterns among all the documents, to practice not close but distant reading, to borrow [Moretti's] memorable phrase.” Scott Weingart (2011) warns that topic modelling is “powerful, widely applicable, easy to use, and difficult to understand — a dangerous combination.” The user adopts a position of responsibility in subjectively interpreting the results it emits.

Voyant is an accessible tool with no configuration required. When the user engages with the tool, they are engaging with a broader sense of the text in question. Predominantly, the aim is to identify any patterns or outliers that may have gone unnoticed previously by using analytic functions such as collocation and concordance. The user applies their reasoning and subjectivity upon this objective computer assisted analysis.

Matthew L. Jockers (2013) has purported that the “quality of the final model (which is to say the coherence and usefulness of the topics) is largely dependent upon preprocessing.” Both texts for this assignment were sourced in e-book format and converted with Calibre to text

format. These texts were then cleaned of any unnecessary ‘noise’ such as chapter headings and captions. While Applebaum’s book proved straightforward in process due to its format of self-authored accounts, the second text proved less intuitive in preparation for analysis. The volume is comprised of interactions between interviewer and interviewee. A subjective decision was made to omit the interjections of the interviewer and the resulting focus is on the words of the interviewee solely. This decision was made as it is the aim of this assignment to assess emotional language and personal preoccupations of the interviewee in relation to their experiences.

Upon entering my first text, *Gulag Voices*, into Mallet, clusters of words were returned that enabled me to construct insight and meaning into the collection. The program returned a list of 10 topics generated by analysis of the word and word sequences in the text. The Topic Modelling tool offers default values for the number of topics, iterations (how many loops it will do before narrowing down on a topic) and topic proportion threshold (if it is set to.05 the tool will only pick those which occur more than .05 times in the text). I increased the proportion threshold due to the nature of my smaller data set, influenced by similar decisions made in the *Hannah More Project*. (Coleman et al. 2013) Several topic reveals with different advanced settings were conducted before settling on the following breakdown: topics: 10, iterations: 200, topic words printed: 10, and topic proportion threshold: 0.10. It is relevant to note that the list of topics are not organised by frequency, although the words within that topic are.

	A	B	C	D	E	F	G	H
1	topicId	words..						
2	1	camp work men bread found barracks group cooler called cut						
3	2	life children world war rest end russia home short half						
4	3	soviet made years good turned young political family arrested stalin						
5	4	back guards guard hard don ll small woman worked hands						
6	5	camps gulag women zaliva working moscow russian kind place voice						
7	6	eyes head face put looked light things stepan child chief						
8	7	prisoners man night left morning visit house wife allowed room						
9	8	day began prison days food felt heard knew side thought						
10	9	time cell water cold floor didn sleep interrogation line asked						
11	10	people long prisoner free year stood ten tarasiuk forest front						
12								

Figure 1: CSV results for Mallet analysis on Gulag Voices

By making certain assumptions on what the clusters imply, it may be discerned that the memoirs collected in the Applebaum collection are largely preoccupied with themes such as: punishment regime (topic 1), separation from family (topic 3), family visits (topic 7), the passing of time (topic 8), and interrogation and labour (topics 9 and 10). Negative and dark sentiments of unrest, separation and control are evoked. Immediately evident is wording associated with guards and punishment. This highlights the fact that, in the gulag system, being an ‘intellectual’ or ‘political’ was seen as a crime worse than that of criminal, and the analysis on Applebaum’s book demonstrates an awareness on the part of the subjects of punishment methods associated with the system (cooler, sleep, interrogation) and names of guards (Zaliva, Stepan and Tarasuik) highlighting the power the guards held over life and death for these individuals.

Equipped with this feedback from Mallet, I entered subjectively determined relevant data into Voyant to further my hypothesis and draw out further reasoning. A collocational analysis of the term Tarasiuk in Voyant reflects the indications given in Mallet. The name collocates with terms such as pellagra, rations, tyrannous and upset. A further collocational analysis on the term

pellagra reveals it to be coupled with Tarasiuk, rations, given and work, in descending order, highlighting again the power represented by the guards in controlling the lives of the prisoners in their hold over portion control and rationing of food. A similar search using the word bread shows it to collocate with: grams, ounces and day. The prisoners daily workload was measured and they were allocated a certain amount of rationing in correspondence with this, reflected in this collocate result.

An isolation of the name Tarasiuk, placed within the Context pane in Voyant, reflects the aura of control and fear that his name summoned for those in the memoirs.

the head and explain that	tar…	was the worst bastard of
to his character. In fact,	tar…	represented the most extreme and
bore the hallmark of Colonel	tar…	. Now Tarasiuk was in charge
hallmark of Colonel Tarasiuk. Now	tar…	was in charge of our
grew pale with fear when	tar…	appeared. Those who could work

Figure 2: Contextual analysis of Tarasiuk in Voyant

Upon entering *Oral Histories* into Mallet in the same manner, the results are distinct from that of *Gulag Voices* in certain manners.

	A	B	C	D	E	F	G	H	I
1	topicId	words..							
2	1	worked family party village years children director timber front son							
3	2	father big husband thought day kind wrote arrested lot asked							
4	3	gave don left ll couldn afraid woods person completely apartment							
5	4	died wasn thing forest remember bread ve give volynka weren							
6	5	people put lived prison order german kolkhoz ten river home							
7	6	called brought settlement year made living wood horse cut days							
8	7	time back war long feel library half house exiled money							
9	8	germans working started built young happen army began horses russians							
10	9	didn stalin mother wanted things happened wouldn good power special							
11	10	work commandant office labor place bring live time camp guys							
12									

Figure 3: CSV results for Mallet analysis on Oral Histories

Themes such as family identity (topic 1), labour in the settlement (topic 6), exile (topic 7), rebuilding (topic 8) and importance of work (topic 10) feature in the results. Given that *Oral Histories* is largely concerned with relating the stories of peasants in the gulag system these results correspond to the narrative of their background and the process of dekulakization.

The analytical results of *Oral Histories* speak of more family orientated sentiment, indicating that rather than it being a solitary experience, these peasants adapted to this new regime in a more perfunctory manner and constructed a society around their confinement. The textual analysis reveals words associated with family grouping and identity: family, children, husband, son. This theme of family is not as prevalent in the analysis of *Gulag Voices* and reflects the case that in the *Oral Histories* collection we are privy to accounts of entire family groups being exiled *en masse* as opposed to the solitary detainment of individuals that is characteristic of Applebaum's collection.

Their experience of exile relates a total resettlement of life in that terms such as village, apartment, home, library and office occur, disclosing that their experience was not that of being physically imprisoned by four walls, but more an open village structure. The word Kolkhoz

makes an appearance, which refers to a collective farm from which the villagers were not allowed leave.

In regards to the sentiment expressed by the Mallet analysis of *Oral Histories,* much literature written in retrospect of the cases of these individuals details that as their home and lands were reshaped, so too was their identity. The Foreign Policy Research Institute relates that the process of dekulakization was successful in its aim, "all resistance was broken." (Satter 2007) The textual analysis of *Oral Histories* reveals a scarcity of words that may be associated with anger or animosity. *Oral Histories* recounts the manner in which one particular interviewee, Balashina, "rather than reject the values of the system that had repressed her, (she) internalized them." (Gheith et al. 2011, 24) An analysis of the recounts told in *Oral Histories* indicates a relegation of details of personal life and portrays these details as footnotes to a life narrative where "work served as both the instrument and measure of normality." (Kotkin 1997, 202)

A further analysis with Voyant extrapolates deeper meaning into the results extracted from Mallet. A collocational search of the word forest through Voyant places the term with emotionally neutral words such as work, cut and timber. In stark contrast, the same word when focused on in a Voyant analysis of *Gulag Voices* results in negative terminology being paired with forest: Tarasiuk, rations, survive, and highlights the diverse attitudes to labour between the two groups of individuals.

Many accounts refer to dekulakization as a type of internal colonization and this identity was also internalized by many of the oppressed. It is evident that those who had been resettled seemed to accept their fate with a quiet air of inevitability, whereby those incarcerated for political crimes and separated from family elicit more an air of unrest. Many accounts corroborate this resigned nature of those resettled, relating at how, when the time came for many

who had been resettled for freedom, they remained in the areas they had been resettled, and often continued with similar labour. “We get a significantly different view of the attitudes towards the Soviet regime of those who were persecuted by it (in *Oral Histories*); for most their internalization of Soviet values survived their frequently horrific experiences, and it is remarkable how little bitterness is expressed by most of the interviewees.” (Young 2011)

In contrast to the identification of certain guards and of specific punishments being meted out in the Applebaum collection, *Oral Histories* is remarkable in that it does not distinguish names of oppressors nor punishments. This indicates that its subjects resided predominantly in resettlements, where the punishment *was* their daily life, and their oppressor was often the barren landscape, which required guard nor barbed wire to prevent escape such was its remoteness and brutality. The lack of bitterness reported by Sarah Young may be seen to indicate the broken spirit of the resettled. From the analysis it may be read that less combative themes are evident in *Oral Histories* when compared to *Gulag Voices*, displaying that the spirits of the peasants were captured and indoctrinated by the process of dekulakization.

Miriam Posner gently warns that “topic modeling is not a way of revealing any objective “truth” about a text; instead, it’s a way of deriving a certain kind of meaning — which still needs to be interpreted and interrogated.” (Posner 2012) It has been the aim of this assignment to apply my subjective reasoning to a minute portion of a huge canon. The tools engaged with in this assignment have helped confirm hypotheses from previous close readings and raised further questions for future exploration.

References

Applebaum, Anne, ed. 2011. *Gulag Voices: An Anthology*. New Haven: Yale University Press.

Coleman, Sinead, Melissa Filbeck, Mary Gifford, Stephanie Harper, Lizette Hernandez, Nazanin Keynejad, Lida Perez, Jason Sirkin and Lillian Thiemens, eds. 2013. "The Hannah More Project. Computational Analysis, Author Attribution, and the Cheap Repository Tracts of the 18th Century." *Scalar*. Accessed April 08, 2018. http://scalar.usc.edu/works/the-hannah-more-project/topic-modeling-process.

Jehanne M. Gheith and Jolluck, K., eds. 2011. *Gulag Voices - Oral Histories of Soviet Incarceration and Exile*. New York: Palgrave MacMillan US.

Jockers, Matthew, L. 2013. *Macroanalysis Digital Methods and Literary History*. Illinois, University of Illinois.

Jockers, Matthew, L. 2013. "'Secret recipe' for Topic Modelling Themes." *Matthew L. Jockers*. Accessed April 01, 2018. http://www.matthewjockers.net/2013/04/12/secret-recipe-for-topic-modeling-themes/.

Kotkin, Stephen. 1997. *Magnetic Mountain. Stalinism as a Civilisation*. London: University of California Press.

Nelson, Robert. 2011. “Mining the Dispatch.” *Internet Archive Wayback Machine*. Accessed 10 April, 2018. http://web.archive.org/web/20110823214412/http://dsl.richmond.edu:80/dispatch/pages/intro.

Posner, Miriam. 2012. “Very Basic Strategies for interpreting results from the Topic Modelling Tool.” *Miriam Posner*. Accessed April 10, 2018. http://miriamposner.com/blog/very-basic-strategies-for-interpreting-results-from-the-topic-modeling-tool/.

Satter, David. 2007. “The Soviet Gulag.” *Foreign Policy Research Institute*. Accessed April 17, 2018. https://www.fpri.org/article/2007/06/the-soviet-gulag/.

Underwood, Ted. 2012. “Topic Modelling made just simple enough.” *The Stone and the Shell*. Accessed April 02, 2018. https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/.

Weingart, Scott. 2011. “Topic Modelling and Network Analysis.” *Scottbot*. Accessed April 01, 2018. http://www.scottbot.net/HIAL/index.html@p=221.html.

Young, Sarah. 2011. “Gulag Voices: Two Books.” *Sarah J. Young*. Accessed April 02, 2018. http://sarahjyoung.com/site/2011/09/05/gulag-voices-two-books/.

-