The Third Way: Discovery Beyond Search and Browse in *Letters of 1916*

Susan Schreibman
susan.schreibman@nuim.ie
Maynooth University, Ireland

Sara Kerr
sarajkerr@icloud.com
Maynooth University, Ireland

Shane McGarry
shane.mcGarry@nuim.ie
Maynooth University, Ireland

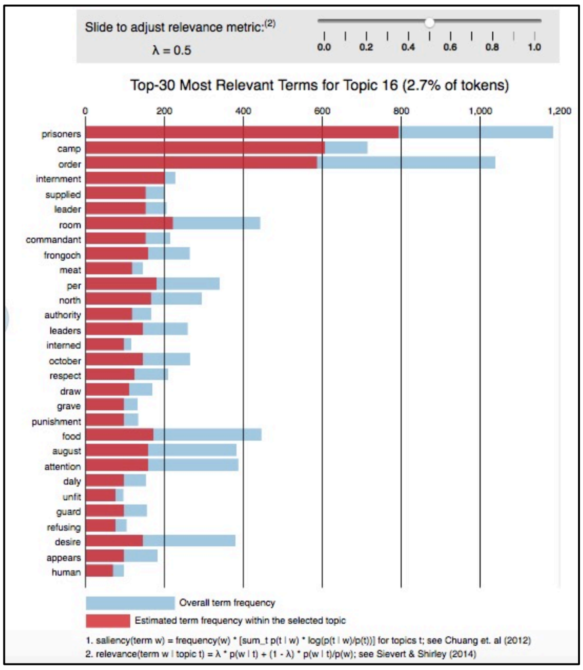
Letters of 1916 is Ireland's first digital public humanities project. Launched in 2013, it has become one of the key corpora in providing new insights and understandings of the 1916 period. The project collects and transcribes through crowdsourcing epistolary documents from October 1915 to November 1916 with the goal of creating a window onto a year in the life of the nation. In the middle of the project's collection period is the Easter Rising (24-29 April 1916), arguably one of the most important events in Irish history as it sets in motion Irish independence from Great Britain in 1921.

The year 1916 was chosen not only because of the centrality of the Easter Rising, but because of Irish participation in the Great War and the historical significance for this throughout the following century in the construction of Irish identity. The collection also reestablishes the role of women in their participation in the Great War as well as the Rising and its aftermath. In addition, epistolary documents were chosen as a record of the everyday and the quotidian, providing a window onto a social history that has too often been repressed or ignored.

With a collection of over 3,500 letters contributed by 54 families and 32 institutions (with new letters being added continually) the corpus is too large to read in its entirety. While search and browse functionality in the project's 'Explore' Database allows users to restrict results to more manageable subsets, the complexity of the letter form, with its frequently meandering content, makes many letters difficult to categorise using a tightly restricted set of keywords. Full text searching also misses many possible letters of interest on a particular topic due to the broad register and idiosyncratic use of language utilised by a wide variety of correspondents (Altman, 88).

A solution to these issues has been to explore the use of alternative methods of analysis and discovery, including topic modelling, vector space modelling with t-SNE vector reduction, and semantic network analysis. These methods provide alternative ways to explore a corpus of this size: too large to be read in its entirety via close reading, yet not big enough to qualify as big data. Rather, this type of collection, not out of reach of many DH projects, might be typified as one of the middle distance in which visualisations can serve as a series of lenses through which areas of interest can be identified for further research.

Topic modelling, although computationally expensive and requiring pre-processing, provides a detailed overview of the corpus and its themes. It highlights the complexity and variety of the letters' content, as well as suggesting areas for further analysis. Combining the full text of the letters with the editorially-assigned keywords has proved a powerful combination in providing a bird's-eye view of the corpus by theme.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 1. Murder at Portobello Barracks

Figures 1 and 2 show topics 14 and 16 from a model where only the standard stopwords were removed. Topic 14 relates to the murder of Sheehy Skeffington, while Topic 16 focuses on internment. The visualisation was created using 'LDAvis', where the most distinct terms in the topic can be viewed by adjusting the relevance metric λ to 0.5.

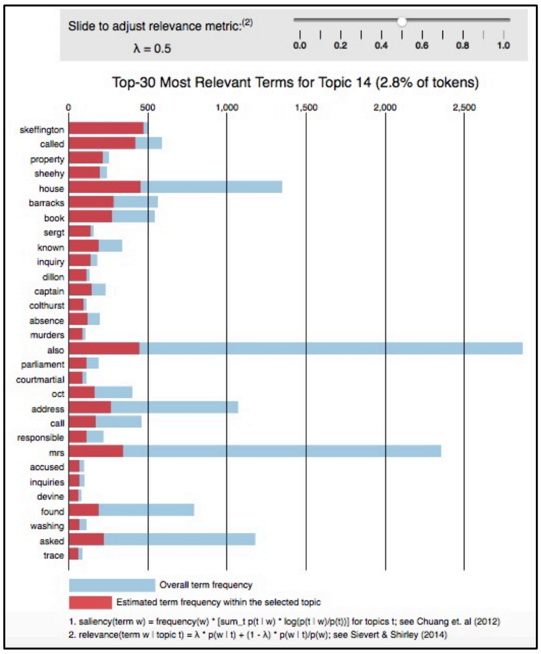


Figure 2. Topic 16 - Internment at Frongoch

Vector space modelling, using the R package 'wordVectors' (Schmidt and Li) which is based on the 'word2vec' algorithm (Mikolov et al.), is effective for a corpus where the researcher is familiar with the broad topics and wishes to zone in on specific aspects. Rather than asking 'which topics are in this corpus?', it allows the researcher to ask 'what does the corpus tell us about this topic?', revealing syntactic and semantic relationships. This type of analysis requires less pre-processing than topic modelling and does not exclude words with a low frequency.

The resulting vector space model can be interrogated, for example by extracting the words nearest to the word vector. For example, the vector for 'rising' (eg the Irish Rising) results in: 'outbreak', 'scene', 'leaders', 'theatre', and 'hostility'. While we may expect the term 'theatre' to refer to a theatre of war, close reading of the six references to 'theatre' in the letters reveals that two refer to an operating theatre while the remaining four are regarding theatre as a place of entertainment. Vector rejection can be used to exclude particular word meanings in cases of polysemy, thus allowing the researcher to specify which meaning they wish to search for. In the above example a researcher interested in entertainment could exclude the alternative meanings by rejecting those words which are related to theatre in the sense of 'hospital' and 'operation'. This would create a vector with the meaning 'theatre as a place of entertainment' enabling a more focused search.

The words nearest the desired search term can be visualised through vector space reduction to two dimensions. We use the Barnes-Hut implementation of t-SNE (t-distributed stochastic neighbour embedding). The visualisation of the 500 words nearest to the vector for 'rising' includes a cluster of words: 'portobello', 'murders', 'accused', 'colthurst' and 'dickson'. This refers to the murders of Sheehy Skeffington, Dickson and MacIntyre at Portobello Barracks on the orders of British officer Bowen-Colthurst.

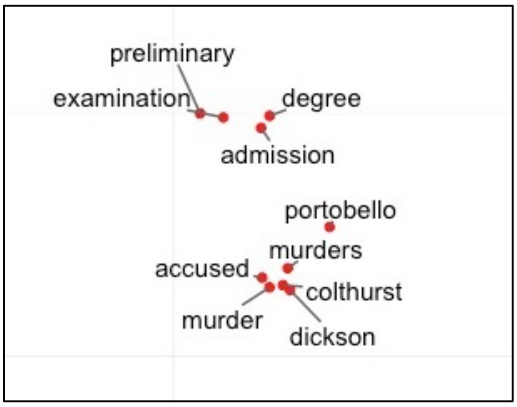


Figure 3. The 'Portobello' cluster

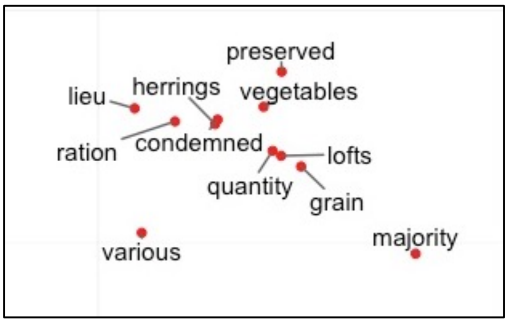
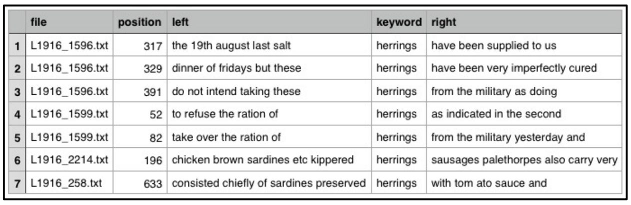


Figure 4. The 'herrings' cluster

A second cluster containing the words 'herrings', 'condemned' and 'ration' is less clear. A key word in context analysis demonstrates one of the strengths of the vector space model, the word 'herrings' only appears seven times in the corpus, but it sheds significant light onto the lives of those in internment. Figure 5 illustrates that the first five references to "herrings' is in the context of the poor quality of rations provided to the Irish prisoners in Frongoch, a prisoner of war camp where some 3500 Irish men were sent after the British put down the Rising.



	file	position	left	keyword	right
1	L1916_1596.txt	317	the 19th august last salt	herrings	have been supplied to us
2	L1916_1596.txt	329	dinner of fridays but these	herrings	have been very imperfectly cured
3	L1916_1596.txt	391	do not intend taking these	herrings	from the military as doing
4	L1916_1599.txt	52	to refuse the ration of	herrings	as indicated in the second
5	L1916_1599.txt	82	take over the ration of	herrings	from the military yesterday and
6	L1916_2214.txt	196	chicken brown sardines etc kippered	herrings	sausages palethorpes also carry very
7	L1916_258.txt	633	consisted chiefly of sardines preserved	herrings	with tom ato sauce and

Figure 5. 'herrings' Key Word in Context

A third analysis builds upon the vector space model by creating a semantic network based on the cosine similarity of terms. The network is visualised using the 'visNetwork' R package (Almende and

Thieurmel) which can be viewed as an interactive HTML file. This analysis enables the exploration of multiple dimensions of the model. Here terms can be clustered and connected with other terms providing a more detailed representation of the semantic space. This network can suggest a broader and more nuanced range of themes. Each of the analyses, and associated visualisations, provide a transformation of the text, disrupting expectations and providing new avenues for exploration (Clement).

These visualisations are, however, but the first step down the path of the third way. In order to engage the parahippocampal area of the brain, which is responsible for drawing context and meaning, we seek to create a more immersive experience for the reader (Bouchard et al). Thus a layer of interactivity is being explored to enhance these visualisations. While the visualisations themselves provide interesting insights into the corpus, they are, much like a traditional search and browse, rather static implementations in that they are pre-selected and curated. The project is thus exploring how users can be provided with the ability to customise these visualisations through the "slicing" of additional metadata in order to draw comparisons that may not be readily apparent.

However, this type of interactivity is not without its drawbacks. Both topic modelling and vector space visualisations require significant processing power in order to generate the base visualisations and the demands on internal memory of the processing machine are high as a result. Thus, creating these types of visualisations in an "on demand" environment is not feasible without a hardware investment that is beyond the reach of most of DH projects. Thus other options—such as caching, indexing, NoSQL databases, or various other data-related optimisation techniques—must be used in order to develop technical solutions that allow for interactivity while supporting a relatively low cost hardware solution.

These visualisations have begun to offer tantalising new insights into the corpus, providing a third way beyond search and browse. This paper will explore both the visualisations themselves, their strengths and weaknesses within the context of a corpus of the middle distance, as well as the novel readings they enable. The paper will conclude by discussing how interactive visualisations such as these can augment traditional modalities of interaction through a rich toolset for research and exploration.

Bibliography

Almende, B. V., and Thieurmel, B. (2016) *VisNetwork: Network Visualization Using "vis.js" Library*. https://CRAN.R-project.org/package=visNetwork. R package version 1.0.2.

Altman, J.G.. (1982) *Epistolarity: Approaches to a Form.* Columbus: Ohio State University Press. 1982.

Bouchard, S., et al. (2015) "The Meaning of Being There is Related to a Specific Activation in the Brain Located in the Parahypocampus." 12th Annual International Workshop on Presence. November 2009. PDF. 11 July 2015.

Clement, T. (2013) "Text Analysis, Data Mining and Visualisations in Literary Scholarship." *MLA Commons | Literary Studies in the Digital Age*, Oct. 2013, https://dlsanthology.commons.mla.org/text-analysis-data-mining-and-visualizations-in-literary-scholarship/

Mikolov, Tomas et al. "Efficient Estimation of Word Representations in Vector Space." Proceedings of the International Conference on Learning Representations (ICLR 2013) (2013): 1–12.

Schmidt, B., and Li, J. (2015). *WordVectors: Tools for Creating and Analyzing Vector-Space Models of Texts*. R package version 1.3.