When Computer Science Met Austen and Edgeworth

Sara J. Kerr, Maynooth University

Introduction

Jesse Rosenthal states in the introduction to the 2017 special issue of *Genre*: 'data is a big deal right now. We cannot talk about data and the novel without recognizing the particular importance that the question of data has in literary studies' (2017, p. 4). This paper is positioned at the intersection of Literary Studies and Computer Science. It explores the application of computer based analysis to novels from the long eighteenth century (an historical period between approximately 1640 to 1830) and, specifically, examines the insights that are gained by using these tools to compare novels by Jane Austen and Maria Edgeworth. It also considers the challenges these methods may present for Humanities scholars, and the benefits of combining computational approaches with close reading.

The title of this paper comes from the film 'When Harry Met Sally...' (1989). The line at the heart of the film proposes that 'men and women can't be friends because the sex part always gets in the way', before ultimately demonstrating that, for Harry and Sally, combining sex with friendship leads to a positive relationship. This analogy echoes some of the arguments against the use of digital analysis in literary studies, or, to rephrase it 'literary studies and computer science can't be friends because the tech part always gets in the way', but it also suggests a possible way forward.

Computational Approaches to Analysing Texts

There has been some resistance to the idea of applying computational techniques to literary texts, although they have a history dating back to at least Father Busa's 1949 collaboration with IBM on the works of Saint Thomas Aquinas. Scholars such as Katie Trumpener (2009) and Stanley Fish (2012) have argued vociferously against the use of digital techniques, claiming that they are unnecessary, and that simply reading more will help to develop more questions. Others consider the potential benefits the digital has to offer, but warn that 'distant reading may actually blunt our critical faculties, inviting us to inadvertently adopt biased views of literature under the mask of objectivity' (Ascari, 2014, p. 3).

In her article, 'The Achievement of Scholarly Authority for Women: Trends in the Interpretation of Eighteenth-Century Fiction', Toni Bowers comments that:

> Widespread interest in a new subject or method often builds from an initially slow response to new research that only gradually comes to influence other scholars' work. Pioneering scholars republish previously out-of-print primary texts; deploy innovative interpretative methods toward unlikely textual subjects; or produce provocative rubrics for previously overlooked or dismissed categories of writing, writers, textual production, or readers. Where there has accrued what we might call a critical mass of this kind of groundbreaking research - enough to suggest a significant body of previously obscured work and to demonstrate the value of recovering and reading it and

> to suggest appropriate methods for interpreting it - scholarly focus shifts, first to recognize the existence of the new object or method of study, then to take it fully on board. (2009, p. 52)

I would argue that we are at the point where the ‘critical mass’ has built to the point where digital techniques are an accepted method, but have not yet reached full acceptance within the Literary Studies community. At present, the arguments regarding what these techniques may show, and what their value is, have not been satisfactorily answered for those for whom this type of analysis seems alien and clinical. There is a tension which exists between the proponents of digital tools and the more traditional style of close reading which needs to be addressed. The perception of digital techniques is often that it stems from a desire to reduce texts to data, stripping them of their context and humanity.

Stephen Marche wrote in an article for the LA Review of Books that, ‘Literature cannot meaningfully be treated as data. The problem is essential rather than superficial: literature is not data. Literature is the opposite of data’. He went on to say, ‘The process of turning literature into data removes distinction itself. It removes taste. It removes all the refinement from criticism. It removes the history of the reception of works’ (2012, no pagination). Although this statement no doubt includes some element of hyperbole, it reflects a number of concerns raised by literary scholars regarding the nature of computational techniques. However, the sentiment expressed by Marche also seems very similar to an article titled ‘Against Theory’, which appeared 30 years earlier (Knapp & Michaels, 1982, pp. 741-742):

> The theoretical impulse, as we have described it, always involves the attempt to separate things that should not be separated…Our thesis has been that no one can reach a position outside practice, that theorists should stop trying, and that the theoretical enterprise should therefore come to an end.

Perhaps things have not changed quite so much after all?

In reality, the digital offers us a different perspective, an alternative way of reading and interpreting texts, in a similar way to the advent of critical theory. These new lenses, through which we can read and interpret a corpus of texts, seek to augment our understanding of the texts, not replace previous understandings. To borrow from Clifford Geertz’s concept of culture as ‘webs of significance’ (1994, p. 214) which become a context ‘within which [social events, behaviours, institutions, or processes] can be intelligibly—that is, thickly—described’ (1994, p. 220), new ways of reading aim to increase what we can say about a text or corpus of texts rather than merely reduce them to a series of sterile numbers. What these methods do enable is the ability to step back from the tight focus of close reading and to consider the texts from another angle. They also offer the opportunity of making textual analysis a little more replicable, allowing researchers to repeat a particular analysis, and view the source of an interpretation. The challenge these techniques seek to address is how to quantify and visualise these ‘webs of significance’ between text, context and meaning.

The rapid development of computer technology, including improvements in storage capacity and processing power, over the past 20 years and more has enabled scholars to create and analyse large datasets. Large-scale digitisation projects, for example the Google Books Library Project, allows access to large corpora, some far larger than a single scholar could read in a lifetime, and have necessitated the development of new tools and methods of computerised reading of texts, automatically, and in some cases, unsupervised, which mean that it is now possible to explore literary texts using a variety of 'distant' methods.

The advent of distant and scaled reading techniques has explored the question of how to present texts in a manner which 'defamiliarize…making them unrecognizable in a way…that helps scholars identify features they might not otherwise have seen' (Clement, 2013, no pagination). Martin Mueller refers to this type of scaled reading as 'DATA' or 'digitally assisted text analysis' (2012, no pagination). What scholars in Digital Humanities are also keen to highlight is that solutions may be found through a combination of techniques (Allison et al., 2011; Mueller, 2012; Clement, 2013; Jockers, 2013). A large proportion of the recent research carried out in this area has focused on non-fiction texts, especially those generated by social media interactions. However, there has been a relatively small, but increasing, interest in applying these techniques to literary texts.

When commenting on large-scale quantitative literary studies, David Brewer states that they 'remind us of just how broad and varied the literary field of the past actually was, and what a small fraction of it receives scholarly attention of any sort' (2011, p. 161). He goes on to say that it is 'accompanied by a sobering reminder that our customary modes of investigation are simply not up to the task of really grasping this broadened field and its forms' (2011, p. 161). Large-scale studies seek to explore and understand the broad sweep of literary history, the development of the novel and its genres over time for example, but this comes 'at a cost. In order to be countable (and so graphable or otherwise capable of being traced over time), texts have to be treated as if they were comparable units' (Brewer, 2011, p. 162). This is something that Moretti acknowledges in his description of distant reading: 'Distant reading: where distance … is a condition of knowledge: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes - or genres and systems ... If we want to understand the system in its entirely, we must accept losing something' (Moretti, 2013, pp. 48-19).

In computational analysis, the choice of tool is often a matter of scale. If we want to understand world literature (which Moretti is discussing when he first uses the term 'distant reading') it is not possible to read everything, even if there were sufficient time. As Moretti says: 'That's the point: world literature is not an object, it's a problem, and a problem that asks for a new critical method: and no one has ever found a method by just reading more texts […] they need a leap, a wager - a hypothesis, to get started' (2013, p. 46).

However, while the large-scale study by necessity needs to strip back the texts to certain key metrics, this is not the case when considering a medium sized corpus such as the one explored in this paper. The 'middle distance' is a profitable area for exploration as it enables the texts it be considered in their own right, as well as part of a broader corpus. To shed light on texts which fall outside the traditional canon, and compare them with canonical texts, allows us to understand more about the texts, authors, and the contexts in which they were written. It is dangerous to assume that texts which became accepted as part of the canon were the more popular, or more accomplished. The expansion of the traditional scope of texts from the long eighteenth century has led to the inclusion of those which were 'othered' because of nationality or political perspective.

A number of studies which take advantage of the digitisation of eighteenth and nineteenth century texts have their genesis in work carried out in the field of corpus linguistics. The interest in applying computational techniques to the works of Austen, in particular, originate in corpus linguistics (Burrows, 1986; Burrows, 1987; DeForest & Johnson, 2001; Starcke, 2006; Fischer-Starcke, 2010). Austen's popularity in the academic community, accessibility, and the lack of copyright attached to her novels, have made her a logical choice when carrying out this type of research.

My research considers the political nature of women novelists publishing between 1800 and 1820. The corpus explored in this paper consists of six novels by Jane Austen and eight novels by Maria Edgeworth. Jane Austen and Maria Edgeworth published the majority of their novels between 1800 and 1820. This was a period of time marked by social and political upheaval in Europe and beyond; revolutions in France and America and a series of rebellions in England and Ireland caused many to question the status quo, where a person's position in the world was largely defined by an accident of birth. As a result, structures of power and regulation were examined, formally or informally, in many of the texts written during this period.

In their novels, Austen and Edgeworth examine the domestic, social, and political world they live in, albeit it from differing social and political perspectives. They argue for greater freedoms for women and those who live outside the traditional hierarchy of the aristocracy and the landed gentry. These are not the silent and domestic voices so often, and erroneously, associated with women writers from this period.

Traditional close reading by necessity focuses on the detailed analysis of small sections of text; it must be selective in the examples chosen to support the argument being presented. While it is possible to construct a convincing argument regarding the political beliefs of Austen and Edgeworth, supported by extensive quotations from their novels, it is equally possible to construct an opposing argument using the same novels. 'It sounds impossible, but Jane Austen has been and remains a figure at the vanguard of reinforcing tradition *and* promoting social change. In early 1900s London, when elite men were drinking, singing, and calling Austen an apolitical author in their private clubs...suffragists were marching through the streets outside with her name emblazoned on a banner' (Looser, 2017, p. 3). Thus, Beth Tobin's 'apologist for the landed classes' (1990, p. 229) is also the Austen who, in Audrey Bilger's *Laughing Feminism: Subversive Comedy in Frances Burney, Maria Edgeworth and Jane Austen*, 'indict[s] the masculine culture that produces such figures' as John Thorpe in *Northanger Abbey* (2002, p. 132). In searching for insight into the traces of a novelist's political views, we need to look for more subtle patterns within and across the texts. In effect, we are looking for an understanding which goes beyond the individual novel, and in Moretti's words 'close reading will not do it' (2013, p. 48). One group of methods which can be used in this way are vector space models.

Vector space models have their origins in frequency-based information retrieval systems developed for computers. The underlying structures created to enable computers to extract information from texts have increasingly been leveraged by literary scholars to explore meaning circulation within and between texts. A vector space model is a matrix type structure used by computers to make sense of texts and to extract information. The vector space model represents documents, or smaller text elements, as points in space which reveal their semantic and syntactic relationships. The two most commonly used types are the ‘term-document matrix’ and the ‘word-context matrix’.

The term-document structure was originally used for automatic computer indexing (Salton et al., 1975). It relates to the bag of words hypothesis which states that ‘the frequencies of words in a document tend to indicate the relevance of a document to a query’ (Turney & Pantel, 2010, p. 153). The aim of the term-document structure is to reveal the similarity between documents. Blei et al.’s ‘Latent Dirichlet Allocation’ algorithm (2003), which is commonly used for topic modelling, built on previous work on Latent Semantic Analysis, creating a method which has been applied effectively to literary texts (Deerwester et al., 1990; Papadimitriou et al., 1997; Hofmann, 1999). Topic modelling has been a popular method to explore large corpora, both literary and archival (Blei, 2012; Mohr & Bogdanov, 2013; Buurma, 2015; Hengchen et al., 2016).

A relative newcomer to literary analysis, the word-context structure, which includes word vectors, uses the distributional hypothesis that ‘words in similar contexts tend to have similar meanings’ (Turney & Pantel, 2010, p. 143). The ‘word2vec’ algorithm is one of the most commonly used algorithms for creating word vectors. Originally created by Tomas Mikolov and his colleagues at Google, the algorithm takes in a corpus of texts and represents words as points in a multi-dimensional space, and word meanings and relationships between words are encoded as distances and paths in that space, through the creation of an artificial neural network (2013). The created model is a simple, shallow neural network which encodes not only syntactic but also semantic relationships between words. Like topic modelling, word vectors aim to reveal the underlying structure of a text or corpus of texts. However, unlike topic modelling, they allow us to ask ‘what does this corpus say about this theme?’. The ‘word-context matrix’, or word vector model, will be the focus for the case study below.

Case Study: Independence in Austen and Edgeworth

A word vector model was created for each of the authors using the ‘wordVectors’ package for R (Schmidt & Li, 2015). The multiple dimensions in the vector space model were reduced to something readable by a human (2 or 3 dimensions) using ‘Rtsne’, and plotted using ‘ggplot2’ and ‘ggrepel’ (Krijthe, 2015; Wickham, 2009; Slowikowski, 2016).

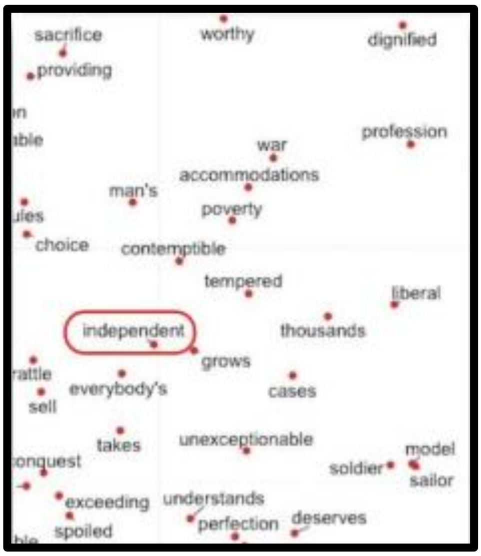


Figure 1: Independence in Austen

Figure 1 shows a section of the plot for the 500 words nearest to ‘independence’. The red points show the local relationship between the words and allow the semantic space to be explored. We can see a number of possible sub-themes: means of independence including ‘profession’ and ‘war’, as well as ‘soldier’ and ‘sailor’; benefits of independence, such as ‘choice’, and perhaps an indication of who benefits—‘man’.

There is also a selection of terms (‘sacrifice’, ‘spoiled’, ‘contemptible’) which suggest that Austen’s view of independence was not entirely positive. Some of these terms suggest the negative impact independence could have on the recipient, as well as the impact of the unfair distribution of wealth, challenging Beth Tobin’s view of Austen as ‘an apologist for the landed classes’ (1990, p. 229).

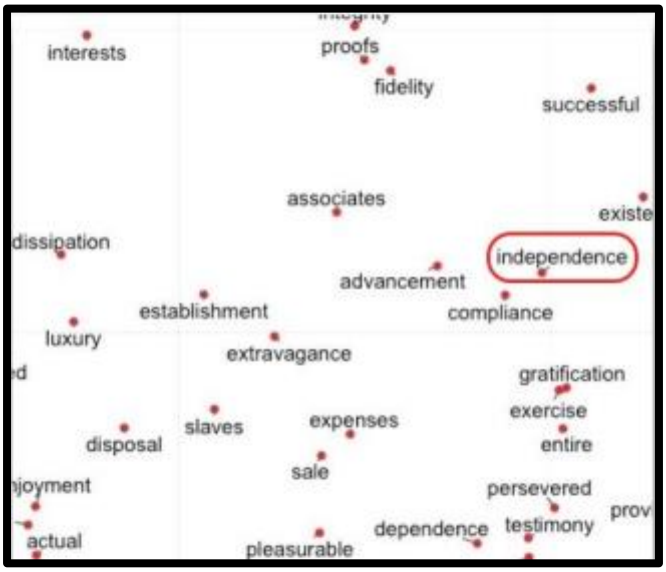
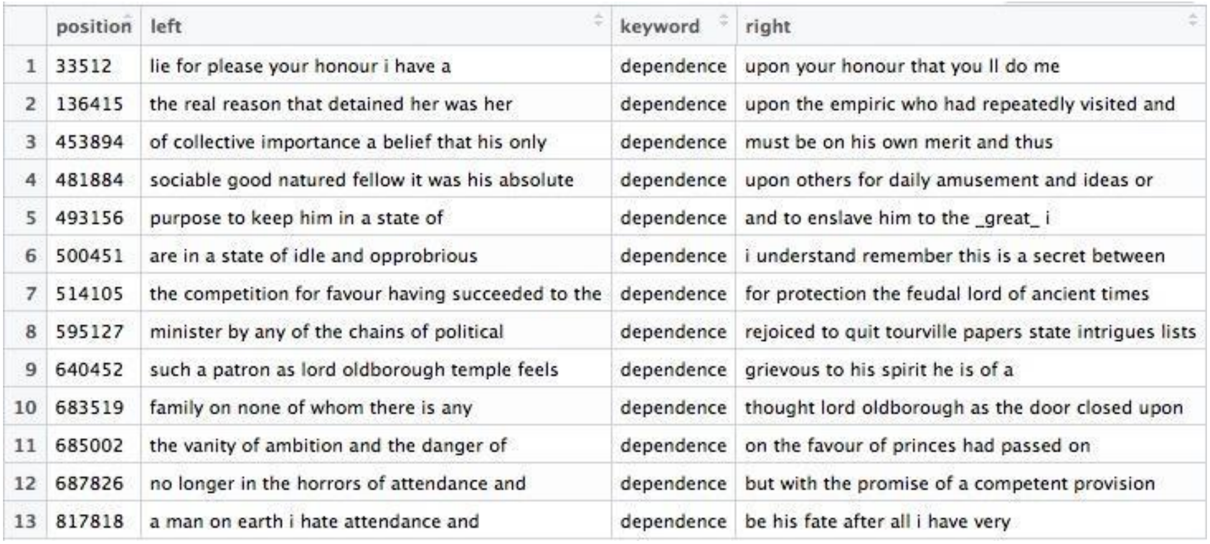


Figure 2: Independence in Edgeworth

A section from the model for Edgeworth’s novels (Fig. 2) suggests similarities to Austen - the use of ‘advancement’ implying independence through work rather than fortune. Like Austen, Edgeworth also highlights the negative impact of independence. Those without their own independence must comply with the desires of others. Exploring the context of ‘dependence’ using Key Word in Context (KWIC) highlights this further (Fig. 3), for example, in *Patronage*, ‘dependence’ is described as being ‘grievous to’ the ‘spirit’ (Edgeworth, 1814, p. 267).



	position	left	keyword	right
1	33512	lie for please your honour i have a	dependence	upon your honour that you ll do me
2	136415	the real reason that detained her was her	dependence	upon the empiric who had repeatedly visited and
3	453894	of collective importance a belief that his only	dependence	must be on his own merit and thus
4	481884	sociable good natured fellow it was his absolute	dependence	upon others for daily amusement and ideas or
5	493156	purpose to keep him in a state of	dependence	and to enslave him to the _great_ i
6	500451	are in a state of idle and opprobrious	dependence	i understand remember this is a secret between
7	514105	the competition for favour having succeeded to the	dependence	for protection the feudal lord of ancient times
8	595127	minister by any of the chains of political	dependence	rejoiced to quit tourville papers state intrigues lists
9	640452	such a patron as lord oldborough temple feels	dependence	grievous to his spirit he is of a
10	683519	family on none of whom there is any	dependence	thought lord oldborough as the door closed upon
11	685002	the vanity of ambition and the danger of	dependence	on the favour of princes had passed on
12	687826	no longer in the horrors of attendance and	dependence	but with the promise of a competent provision
13	817818	a man on earth i hate attendance and	dependence	be his fate after all i have very

Figure 3: KWIC ‘dependence’ in Edgeworth

There is a discourse of implied criticism of traditional power structures running through the novels of both authors, a criticism which may be hard to identify through close reading alone. Far from reinforcing the belief that the upper classes had a hereditary right to financial security and property ownership, Austen and Edgeworth present a world in which ‘new’ money and professions are a much more positive driving force behind the success of the country than those of the traditional aristocracy. The aristocratic characters are often presented as ‘spoilt’, ‘dissipated’ or prejudiced, in some cases both physically and morally diseased. Edgeworth’s Lady Delacour in *Belinda* is one such example, the physical disease of her breast acts as a metaphor for the destruction caused by her lack of morality and the excesses of London life. Sir Walter Elliot in Austen’s *Persuasion* is another, contrasted, to his detriment, with the self-made Admiral Croft and Captain Wentworth.

Once a vector space model has been created, it can be explored in more detail by calculating the cosine similarity between words and creating a semantic network from the results. Cosine similarity (Fig. 4) generates a metric that says how related two vectors are by measuring the angle between them. The value will be between -1 and 1, 1 being totally similar and -1 being totally dissimilar (Perone, 2013).

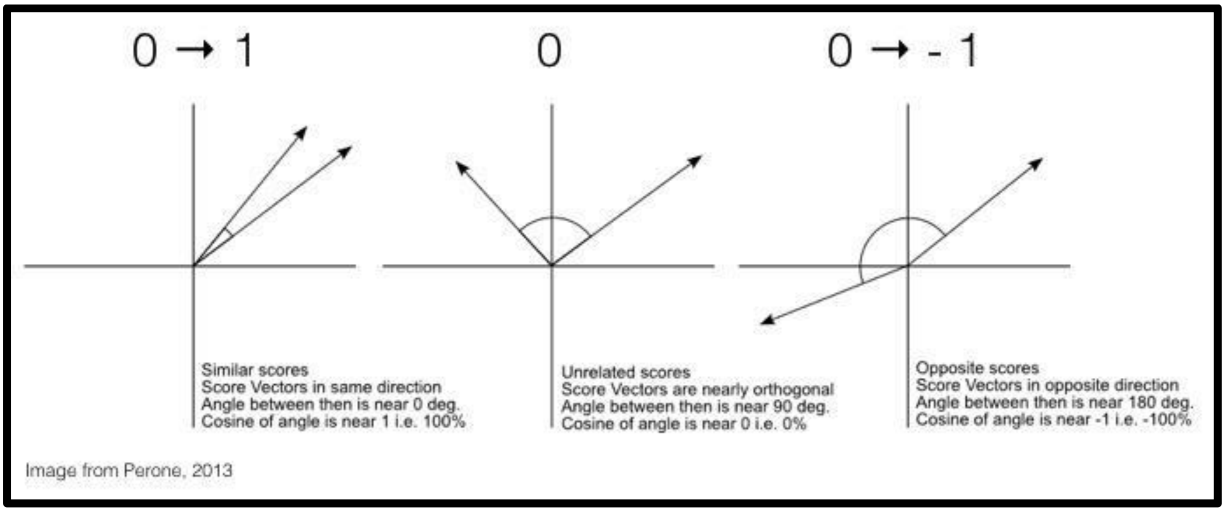

0 → 1
0
0 → - 1
Similar scores
Score Vectors in same direction
Angle between then is near 0 deg.
Cosine of angle is near 1 i.e. 100%
Unrelated scores
Score Vectors are nearly orthogonal
Angle between then is near 90 deg.
Cosine of angle is near 0 i.e. 0%
Opposite scores
Score Vectors in opposite direction
Angle between then is near 180 deg.
Cosine of angle is near -1 i.e. -100%

Image from Perone, 2013

Figure 4: Explanation of Cosine Similarity

Creating a network enables not only links between words to be viewed, but also provides a method of finding 'meaningful groups' through 'community detection' (Heuser, 2016).

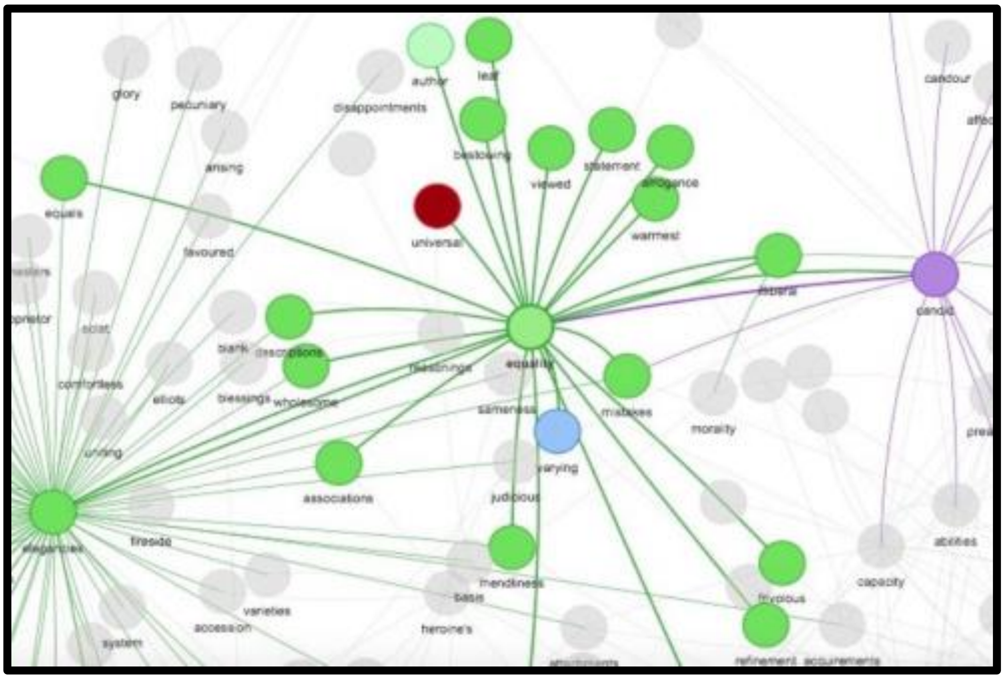


Figure 5: Semantic Network of Independence in Austen

Figure 5 was created using the 'visNetwork' R package and represents a section of the relationships between the 100 words nearest to an 'independence/independent' vector and the whole Austen corpus (Almende & Thieurmel, 2016). The words with a cosine similarity of 0.55 or more are shown. The colours represent communities of words. The original network is an HTML file which allows nodes to be moved or highlighted and for the user to zoom in and out.

‘Equality’, the central node in Figure 5, was a potentially controversial term in the early nineteenth century. While the revolutionary ideals of equality proved a popular topic of discourse in the early 1790s, expressed, for example in Mary Wollstonecraft’s *A Vindication of the Rights of Women*, the French Revolution’s Reign of Terror dramatically shifted public opinion. Supporting these ideals was viewed as seditious and a threat to the nation, even in fictional form:

> The revolutionary ideas of France have already made but too great a progress in the hearts of men in all countries, and even in the very centre of every capital. If every crime be crowned with reward in France, every individual may hope that the subversion of order in his own country will procure him a situation, if not honourable, at least honoured. IT IS NOT BONAPARTE THAT AT PRESENT FORMS THE DANGER OF EUROPE...IT IS THE NEW OPINIONS. (Anon, 1815)

Authors openly advocating equality were viciously attacked by conservative reviewers. Therefore, it is not surprising that writers were cautious in expressing these ideas.

Although Austen is often presented as a conservative writer, the semantic network shows links between ‘equality’, ‘wholesome’ and ‘friendliness’ implying that Austen is favourable towards the concept. The node ‘varying’ provides a link from ‘equality’ to ‘heroine’. Further linked words suggest there is a criticism of the lack of equality, created by societal forces, that many of Austen’s female characters initially face.

In network theory, a giant component is a group of interconnected nodes which accounts for a large proportion of the nodes in the network as a whole. In comparison with the Austen network, the giant component in the Edgeworth network has a greater number of interconnections, suggestive of the relative complexity and density of Edgeworth’s language. A large number of the nodes belong to the same community. This includes: ‘profession’, ‘abilities’, ‘connexions’, ‘talent’, and ‘occupation’ reinforcing the suggestion from the vector space reduction that independence through employment is an important theme in Edgeworth’s novels. The perceived rise of the middle classes during this period and the challenge they represented to the established social hierarchy was a concern for many of those who gained power and wealth through inheritance. The vestige of this concern exists to this day, especially in the UK with the distinction between ‘old’ and ‘new’ money.

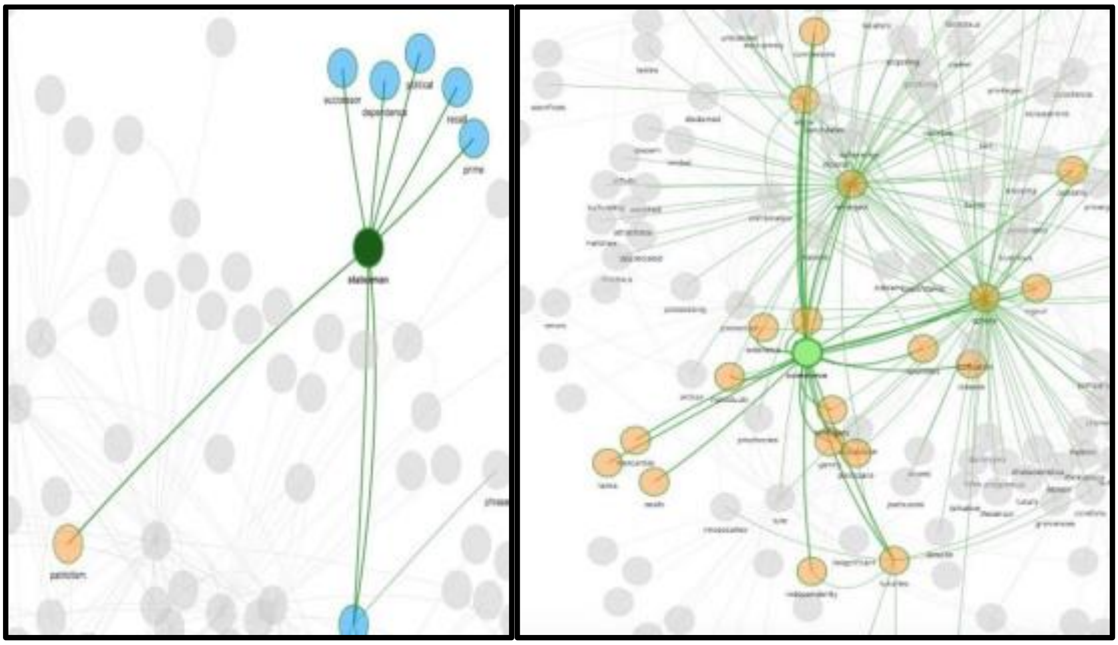


Figure 6: 'Statesman' Cluster and 'Commerce' Cluster in Edgeworth

Traditionally, women were viewed as belonging to the private or domestic sphere, a domain which excluded the 'masculine' topics of politics, economics and history. The challenge of defining the public and private spheres is acknowledged by Susie Steinbach who explains that this 'was not a rigid set of rules internalized as natural and adhered to unquestioningly. Rather, separate spheres were in the process of being constructed, rife with internal contradictions, and frequently challenged (both overtly and covertly)' (Steinbach, 2012, p. 830). However, in contrast to Austen, Edgeworth openly explores both politics, as seen in the group surrounding 'statesman', and economics, as seen in the cluster surrounding 'commerce', in her novels (Fig. 6). This led some of her contemporary reviewers to criticize her depiction of things they felt, as woman, she could have no experience of (Anon, 1814).

Practical Challenges

Utilising digital tools in the analysis of novels is not without its challenges. While the novels of Jane Austen have been digitised in a wide variety of formats, the same cannot be said for Edgeworth's less popular novels. The reach of the English literary canon still has influence over what is valued and studied, and, although digitisation projects have made many previously unstudied texts available, the physical quality of these texts can prove a hurdle which cannot be easily overcome.

Beyond the quality of the text to be analysed, the choice, availability, and appropriateness of the tools and methods used may also prove challenging. A corpus, for example like the Austen-Edgeworth corpus used here, may contain 14 novels and approximately 1.8 million words but still be considered too small for some tools. Topic modelling's need for a large corpus was the original motivation for this study's focus on the use of word vectors. The scale and output of the chosen methods also requires additional statistical, mathematical, and programming skills—skills not frequently a part of literary studies. It is this uncertainty and unfamiliarity which is often the greatest barrier to Humanities scholars using these methods.

Yet, beyond the technical challenges raised by the use of computational analysis, there is also the familiar. While a computer program can reveal interesting aspects of a novel or other text, it is only with the application of contextual details, associated with traditional close reading, that the significance of the findings become clear. The computer has no awareness of the text it explores or the results it produces, the scholarly skills of close reading are required to interpret these results.

Conclusion

On the surface, Austen and Edgeworth appear to reinforce the traditional and restricted life of the woman in the long eighteenth century. However, through a closer analysis of their novels, they show the desire and ability to criticise the social norms of the times they are living in. Their criticism encompasses the clergy, the titled aristocracy, and the treatment of unmarried women. In effect, Austen and Edgeworth simultaneously construct and deconstruct the early 19th century image of womanhood. A computational analysis helps to provide empirical evidence of these ideas which are skilfully woven through the novels, helping to support existing interpretations arrived at through close reading, as well as challenging others.

It is not the purpose of Digital Humanities to replace traditional tools and methods, it merely provides us with an ever broader range of tools to choose from. Not all research projects are ‘big data’ projects, and not all research projects will require the use of digital tools. However, these open access tools help make research more accessible, more replicable and perhaps more objective than close reading alone (Omar, 2010).

Bibliography

Allison, S. et al. (2011). *Quantitative Formalism: An Experiment*. Stanford Literary Lab [online]. Available at: https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf [Accessed 22nd March, 2015].

Almende, B. V. and Thieurmel, B. (2016). *visNetwork: Network Visualization using “vis.js” Library*. R package version 1.0.2 [software]. Available at: https://CRAN.R-project.org/package=visNetwork.

Anon. (1814). Patronage by Miss Edgeworth. *The British Critic and Quarterly Theological Review*. Volume 1. pp. 159-173.

Anon. (1815). Letter to Times Newspaper, 6 June, 1815. *Cobbett’s Weekly Political Register*, 27(23), pp. 705-706.

Ascari, M. (2014). The Dangers of Distant Reading: Reassessing Moretti’s Approach to Literary Genres. *Genre*. Volume 47, Issue 1. pp. 1-19.

Bilger, A. (2002). *Laughing Feminism: Subversive Comedy in Frances Burney, Maria Edgeworth and Jane Austen (Humor in Life & Letters)*. Detroit, Michigan: Wayne State University Press.

Blei, D. M. (2012). Probabilistic Topic Modeling. *Communications of the ACM*. Volume 55. pp. 77-84. Available at: https://cacm.acm.org/magazines/2012/4/147361-probabilistic-topic-models/fulltext [Accessed 15th February, 2016].

Blei, D. M., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*. Volume 3, Issues 4-5. pp. 993-1022.

Bowers, T. (2009). The Achievement of Scholarly Authority for Women: Trends in the Interpretation of Eighteenth-Century Fiction. *The Eighteenth Century*. Volume 50, Issue 1. pp. 51-71. Available at: http://muse.jhu.edu/content/crossref/journals/the_eighteenth_century/v050/50.1.bowers.html [Accessed 19th June, 2013].

Brewer, D. A. (2011). Counting, Resonance, and Form, A Speculative Manifesto (with Notes). *Eighteenth-Century Fiction*. Volume 24, Issue 2. pp. 161-170.

Burrows, J. F. (1986). Modal verbs and moral principles: an aspect of Jane Austen’s style. *Literary and Linguistic Computing*. Volume 1, Issue 1. pp. 9-23. Available at: http://llc.oxfordjournals.org/content/1/1/9.short [Accessed 19th December, 2014].

—. (1987). Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style. *Literary and Linguistic Computing*. Volume 2, Issue 2. pp. 61-70. Available at: http://llc.oxfordjournals.org/cgi/doi/10.1093/llc/2.2.61 [Accessed 3rd September, 2013].

Buurma, R. S. (2015). The fictionality of topic modeling: Machine reading Anthony Trollope’s Barsetshire series. *Big Data & Society*. Volume 2, Issue 2. p. 1-6. Available at: http://journals.sagepub.com/doi/pdf/10.1177/2053951715610591 [Accessed 15th February, 2017].

Clement, T. (2013). Text Analysis, Data Mining, and Visualizations in Literary Scholarship. *Literary Studies in the Digital Age: An Evolving Anthology*. MLA Commons [online]. Available at: https://dlsanthology.mla.hcommons.org/text-analysis-data-mining-and-visualizations-in-literary-scholarship/ [Accessed 3rd April, 2015].

Deerwester, S., Dumais, S. T., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*. Volume 41, Issue 6. pp. 391-407. Available at: http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf [Accessed 15th February, 2017].

DeForest, M., and Johnson, E. (2001). The density of Latinate words in the speeches of Jane Austen’s characters. *Literary and linguistic computing*. Volume 16, Issue 4. pp. 389-401. Available at: http://llc.oxfordjournals.org/content/16/4/389.short [Accessed 19th December, 2014].

Edgeworth, M. (1814). *Patronage*. Volume 3. J. London: Johnson and Co.

Fischer-Starcke, B. (2010). *Corpus Linguistics in Literary Analysis: Jane Austen and Her Contemporaries*. London: Continuum.

Fish, S. (2012). Mind Your P's and B's: The Digital Humanities and Interpretation. *New York Times Opinionator Blog* [online]. Available at: http://opinionator.blogs.nytimes.com/2012/01/23/mind-your-ps-and-bs-the-digital-humanities-and-interpretation/?_php=true&_type=blogs&_r=0 [Accessed 20th October, 2014].

Geertz, C. (1994). Thick description: Toward an interpretive theory of culture. In: M. Martin, and L. C. McIntyre, eds. *Readings in the Philosophy of Social Science*. Cambridge, Massachusetts: The MIT Press. pp. 213-231.

Hengchen, S. et al. (2016). Exploring archives with probabilistic models: Topic Modelling for the valorisation of digitised archives of the European Commission. In: *IEEE Proceedings of the First Workshop on Computational Archival Science*. Washington DC: IEEE. Available from: http://ieeexplore.ieee.org/document/7840981/ [Accessed 3rd April, 2017].

Heuser, R. (2016). Word vectors in the eighteenth century, episode 4: Semantic networks. *Adventures of the Virtual* [online]. Available at: http://ryanheuser.org/word-vectors-4/ [Accessed 26th September, 2016].

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. In: *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, UAI 1999*. San Francisco, CA: Morgan Kaufman. pp. 289-296. Available from: https://arxiv.org/pdf/1301.6705.pdf [Accessed 15th February, 2017].

Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana, USA: University of Illinois Press.

Knapp, S., and Michaels, W. B. (1982). Against Theory. *Critical Inquiry*. Volume 8, Issue 4. pp. 723-742.

Krijthe, J. H. (2015). *Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation*. R package version 0.13[software]. Available at: https://github.com/jkrijthe/Rtsne.

Looser, D. (2017). *The Making of Jane Austen*. Baltimore, Maryland: Johns Hopkins University Press.

Marche, S. (2012). Literature is not Data: Against Digital Humanities. *The Los Angeles Review of Books* [online]. Available at: http://lareviewofbooks.org/article/literature-is-not-data-against-digital-humanities [Accessed 17th February, 2015].

Mikolov, T. et al. (2013). Efficient Estimation of Word Representations in Vector Space. In: *Proceedings of the International Conference on Learning Representations (ICLR 2013)*. pp. 1–12. arXiv preprint arXiv:1301.3781. Available at: http://arxiv.org/pdf/1301.3781v3.pdf [Accessed 8th April, 2016].

Mohr, J. W., and Bogdanov, P. (2013). Introduction—Topic models: What they are and why they matter. *Poetics*. Volume 41, Issue 6. pp. 545-569. Available at: http://www.sciencedirect.com/science/article/pii/S0304422X13000685 [Accessed 12th November, 2014].

Moretti, F. (2013). *Distant Reading*. London: Verso.

Mueller, M. (2012). Scalable Reading. *Scalable Reading Blog* [online]. Available at: https://scalablereading.northwestern.edu [Accessed 20th October, 2014].

Omar, A. A. (2010). Addressing Subjectivity and Replicability in Thematic Classification of Literary Texts: Using Cluster Analysis to Derive Taxonomies of Thematic Concepts in the Thomas Hardy's Prose Fiction. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*. Volume 1, Issue 2. pp. 1-14.

Papadimitriou, C. H. et al., (2000). Latent Semantic Indexing: A Probabilistic Analysis. *Journal of Computer and System Sciences*. Volume 61, Issue 2. pp. 217–235. Available at: http://www.sciencedirect.com/science/article/pii/S0022000000917112 [Accessed 22nd February, 2017].

Perone, C. S. (2013). Machine Learning: Cosine Similarity for Vector Space Models (Part III). *Terra Incognita* [online]. Available at: http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/ [Accessed 14th September, 2016].

Rachum, A. (2016). Knowledge Debt. *Amir Rachum's Blog* [online]. Available at: http://amir.rachum.com/blog/2016/09/15/knowledge-debt/ [Accessed 20th September, 2016].

Reiner, R. (1989). *When Harry Met Sally…*. USA: MGM.

Rosenthal, J. (2017). Introduction: "Narrative against Data". *Genre*. Volume 50, Issue 1. pp. 1-18. Available at: http://genre.dukejournals.org/lookup/doi/10.1215/00166928-3761312 [Accessed 23rd March, 2017].

Salton, G., Wong, A., and Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Information Retrieval and Language Processing*. Volume 18, Issue 11. pp. 613-620.

Schmidt, B., and Li, J. (2015). *wordVectors: Tools for creating and analyzing vector-space models of texts*. R package version 2.0 [software]. Available at: http://github.com/bmschmidt/wordVectors.

Slowikowski, K. (2016). *ggrepel: Repulsive Text and Label Geoms for "ggplot2".* R package version 0.6.5 [software]. Available at: https://cran.r-project.org/package=ggrepel.

Starcke, B. (2006). The phraseology of Jane Austen's Persuasion: Phraseological units as carriers of meaning. *ICAME journal*. Volume 30. pp. 87-104.

Steinbach, S. (2012). Can We Still Use "Separate Spheres"? British History 25 Years After Family Fortunes. *History Compass*. Volume 10/11. pp. 826-837.

Tobin, B. F. (1990). The Moral and Political Economy of Property in Austen's *Emma*. *Eighteenth-Century Fiction*. Volume 2, Issue 3. pp. 229-254. Available at: http://muse.jhu.edu/content/crossref/journals/eighteenth_century_fiction/v002/2.3.tobin.html [Accessed 19th June, 2013].

Trumpener, K. (2009). Paratext and Genre System: A Response to Franco Moretti. *Critical Inquiry*. Volume 36, Issue 1. pp. 159-171.

Turney, P. D., and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*. Volume 37. pp. 141-188.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.