Semantic Density mapping: A discussion of meaning in William Blake's *Songs of Innocence and Experience*

Elina Koristashevskaya

Submitted in fulfilment of the requirements for the Degree of Master of Research in English Language

School of Critical Studies

College of Arts

University of Glasgow

September 2013

Abstract:

This project attempts to bring together the tremendous amount of data made available through the publication of the *Historical Thesaurus of the Oxford English Dictionary* (eds. Kay, Roberts, Samuels and Wotherspoon 2009), and the recent developments in digital humanities of 'mapping' or 'visually displaying'[1] literary corpus data. Utilising the Access *HT-OED* database and 'Gephi' digital software, the first section of this thesis is devoted to establishing the methodology behind this approach. Crucial to achieving this was the concept of 'Semantic Density', a property of a literary text determined by the analysis of lexemes in the text, following the semantic taxonomy of the *HT-OED.* This will be illustrated with a proof-of-concept analysis and visualisations based on the work of one poet from the Romantic period, William Blake's *Songs of Innocence and Experience* (1789/1794). In the later sections, these 'maps' will be used alongside a more traditional critical reading of the texts, with the intention of providing a robust framework for the application of digital visualisations in literary studies. The primary goal of this project, therefore, is to present a tool to inform critical analysis which blends together modern digital humanities, and traditional literary studies.

[1] See: Moretti (2005), Hope and Witmore (2004;2007)

Table of Contents

List of Tables

List of Figures

Acknowledgement

I would like to thank my supervisor, Jeremy Smith, for his support and encouragement during this project. I would also like to thank Marc Alexander, for providing additional support and valuable resources which made this project possible.

For their interest and encouragement, I would like to thank Professor Nigel Fabb at the University of Strathclyde, and Heather Froelich, his 2nd year PhD candidate.

Finally I must give my thanks to my partner, Eachann Gillies, for his sympathy and understanding and Duncan Pottinger, for listening to all of my ideas and poking holes in them.

Declaration

I declare that, except where explicit reference is made to the contribution of others, that this thesis is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Signature _____

Printed Name _____

Chapter 1 - Introduction

1.1 Introduction

1.1.1 The *Historical Thesaurus of the Oxford English Dictionary* (eds. Kay, Roberts, Samuels and Wotherspoon 2009) is a unique resource for the analysis of the English language. Encompassing the complete second edition of the *Oxford English Dictionary* (*OED),* and additional Old English vocabulary, the *HT-OED* displays each term organised chronologically through 'hierarchically structured conceptual fields' (Kay 2012: 41). Despite the relatively recent publication, the *HT-OED* is already being explored by academics from both literary and linguistic backgrounds[2] as a tool for the analysis of language. Such was the intention of the creators of the *HT-OED*, the project being originally born out of Michael Samuels' 'perceived gap in the materials available for studying the history of the English language, and especially the reasons for vocabulary change' (Kay 2012: 42).

1.1.2 The *HT-OED* was developed over a period of five decades, during which time both technological developments and, consequently, academic practice continued apace. In particular, new digitalised methods of corpus analysis began to breach the same gap as the one identified by Samuels in 1965. As noted by one of the earlier pioneers of digital corpus analysis, John Sinclair, with instant access to digital corpora the ability to examine text in a 'systematic manner' allowed 'access to a quality of evidence that [had] not been available before' (Sinclair 1991: 4). In-keeping with this progress, the *HT-OED* has been integrated into the *OED* online, and plans are currently in motion at the University of Glasgow for an 'integrated online repository' using the Enroller project (Kay and Alexander 2010; Kay 2012). Despite this, there is as yet no comprehensive tool for utilising *HT-OED* data for digital text analysis, and this project marks an attempt to address this void by using existing tools for digital corpus analysis.

1.1.3 The goal of this project is to present a new way of engaging with the *HT-OED*, in-keeping with the current developments in digital humanities, but not seeking to replace or replicate the future goals of the *HT-OED* team. Working on the hypothesis that semantic properties of a text can be discussed through electronic analysis and classification, this thesis serves as a proof-of-concept for a holistic study of literary texts. At its core, this hypothesis relies on the well-

[2] A selected bibliography can be found on the Historical Thesaurus of the Oxford English Dictionary website http://historicalthesaurus.arts.gla.ac.uk/webtheshtml/homepage.html

established foundation of electronic corpus analysis in literary linguistics, and strives to blend these methods with traditional critical theory for a modernised approach to critical studies.

1.1.4 Corpus linguistics has been increasingly developed to cope with the demands of literary analysis, and has over the last two decades grown into a rich field of study[3]. For this project, work by Franco Moretti (2005) and Michael Witmore (2004; 2007; 2011) is identified as particularly important, but several other studies on Semantic Network Analysis (Krippendorff, 2004; Van Atteveldt 2008; Roberts 1997; Popping 2000) are valuable for the manner with which they engage with large corpora and digital representation. While the intended outcome of this project differs from the goals of these authors, their work is credited for helping to establish the validity of this project. In particular, Moretti's (2005) work on 'distant reading' engages with several themes that are present in this thesis, and will be discussed in greater detail in the literature review.

1.2 Semantic Density

1.2.1 Similar to existing forms of Semantic Network Analysis, this projects follows the path of first representing the content of the data as a network in an effort to address the research question, rather than 'directly coding the messages', and then querying the representation to answer the research question (Van Atteveldt: 4). This project departs from the work of previous authors by introducing the concept of 'semantic density' (SD) to cope with the data obtained from the *HT-OED*. Outlined briefly, SD is a property of a text that is delimited by the semantic categories of the *HT-OED*[4], where each lexical term has a statistical relationship with the semantic categories, and the other lexical terms in the text. For instance, a text may include several words from the semantic field of 01.02 Life, e.g. *bird*, *tree*, *green*, etc. Such a text has a specific property of semantic density with regard to the field 01.02 Life. This density will either be high or low, depending on the number of collocates present within the text that also fall within the field of 01.02 Life. Texts may contain two or more semantic fields with a high semantic density, often resulting from the polysemous characteristics of many words (including metaphor).

1.2.2 To illustrate this, it is possible to look at two sub-categories of the *HT-OED*, 01.04.09 Colour and 01.02.04 Plants. A text may, for instance, include the word *green* alongside *hill*, *leaf*,

[3] See: Sinclair (1991;2004)
[4] For the purpose of reference, categories of the *HT-OED* are listed alongside their hierarchical number.

grass etc., but also alongside *tinted*, *red*, *coloured* etc. Such a text would have a SD property in both 01.04.09 Colour and 01.02.04 Plants, which could be measured by how frequently these collocates appear in the text. When a text is being read, collocates are frequently used to determine the appropriate connotation or denotation of a polysemous word, while collocates from multiple interpretations frequently establish the use of metaphor. Therefore, in a text where a polysemous word is mentioned with predominant collocates from only one semantic field, as in '*green* coloured wallpaper', SD can be used to display this relationship. In the aforementioned example, the sentence will have a higher semantic density count of for the field 01.04.09 Colour than 01.02.04 Plants. Thus, it is possible to infer the denotation of this instance of *green* based on SD.

1.2.3 Of course, real examples are rarely so clear cut, and it would be highly unusual for a longer text to have such a clearly defined SD count. What this example represents, however, is the possibility of scanning large texts for SD counts in a fast and efficient way, which can then be represented through large visualisations of the text as a whole, defined for the purpose of this project as 'semantic density mapping'. The purpose of identifying the visualisations as 'maps', instead of simply referring to them as networks relates to the information that they are trying to portray. These networks don't simply describe the relationship between the words and the semantic categories, but rather visualise a property of the original text, and offer a way of 'reading' the text at network level.

1.2.4 Returning to the previous example of a text where the semantic field of 01.04.09 Colour is represented by multiple collocates, and 01.02.04 Plants by very few, the visualisation will be representative of this, indicating the predominant theme of the text. SD is a response to existing work being carried out by corpus linguists, which moves beyond the lexical items of the text into a form of visual representation that combines lexical choice with pre-defined semantic categorisation. Reading corpus data through the filter of semantic density allows for increased visibility and accessibility in highlighting semantic patterns in literary texts. Intended initially as a tool to complement and re-evaluate existing critical work, it could also be used to discover new patterns in old texts.

1.3 Historical Thesaurus of the Oxford English Dictionary

1.3.1 This project was born out of the desire to utilise the *HT-OED* in critical literary analysis,

which in turn serves to inform the methodology in two fundamental ways. Firstly, as illustrated above, the hierarchical semantic categorisation of the *HT-OED* is used for the SD analysis. The *HT-OED* is expertly suited to this as it encompasses within its complex taxonomy both 'single notions', which are 'expressed as synonym groups', and 'related notions', which can 'encompass as much of the lexical field within which the particular group of lexical items is embedded as the researcher wishes to pursue'(Kay 2010: 42). This project makes use of both phenomena for the purpose of SD mapping. It will therefore be necessary to explore the categorisation itself, as the theoretical approach behind this project relies on the coherency of these categories. At this stage, however, it is possible to state that the categories function as the 'tags' of groups of lexical items, which in turn are used in the visual representation of the corpora.

1.3.2 The second key significance of using the *HT-OED* for this project, is the ability to analyse a word's meaning at a specific point in time. By cross referencing the data obtained from the semantic analysis of the corpora with the meaning's recorded date of usage in the *HT-OED*, it is possible to not just identify the semantic categories that the words used by the authors fall into, but also filter the data to display only those meanings that were in use during the author's lifetime. To make use of this, the data taken from the *HT-OED* only recorded words which were cited within fifty years of the publication of the original text. The use of the *HT-OED* for this function has begun to be tested by linguists, taking for example Jeremy Smith's exploratory study of medical vocabulary in the work of John Keats (2006). Despite the synchronic approach of both Smith's work and this paper, it is possible to see how this methodology could be adapted for a diachronic analysis, highlighting for example the dominant semantic fields of a literary period, or of one author's work during their lifetime. In this manner, the *HT-OED* allows for a more accurate description of semantic distribution within a text than the traditional 'Dictionary Approach' (Krippendorff, 2004, p.283). In his original treatise for the creation of the *HT-OED*, Samuels argued that what was missing from contemporary tools for studying semantic change was the ability to see 'words in the context of others of similar meaning' (Kay 2010: 42). To this end, this project hopes to utilise the framework created by Samuels and his team to achieve this goal in relation to literary texts.

1.3.3 Principal to this is the unique taxonomy that was created for the *HT-OED*. The multi-level semantic categorisation was conceived by the authors for the purpose of 'semantic contextualisation' of lexical items (Kay 2010: 42). At the highest level, the *HT-OED* is organised in a 'series of broad conceptual fields' (Kay 2010: 43), which are 01 The World, 02 The Mind

and 03 Society. For the purpose of classification, this level is referred to as the 'first level', and is then split further into 'second level' categories such as 01.03 Physical sensibility, 02.02 Emotion, and so forth. While the early stages of the project used the categories of the 1962 edition of Roget's *Thesaurus of English Words and Phrases* (Dutch 1962) as a 'preliminary filing system', these were largely abandoned as the project progressed, in favour of the extensive 12-place hierarchically numbered taxonomy which is used in the *HT-OED* today (Kay 2010: 44-52).

1.3.4 For the purpose of this project, only three of those levels were utilised in the network analysis. Due to the large size of the literary corpora, and the exploratory nature of this proposal, it was necessary to limit the amount of data for processing. Each word entry (later referred to as a 'node), was only processed up to the third level within the *HT-OED* taxonomy**.** As the data was originally obtained by cross-referencing a lemmatised version of the text with the *HT-OED* 'Access' database, the resulting table of entries had to be cut to the third level category. An example of this can be seen for one of the entries for the word *lamb* in Table 1 and 2 below:

Occur	Word	Part	Group	Sub	Heading	MajHead	AppS	AppE
18	lamb	n	01.02.08.01.05.05.	08.	(.lamb)	Mutton	1620	2000

Table 1 [5]- Original output from HT-OED Access database

Occur	Word	Part	Group	MajHead
18	lamb	n	01.02.08	Mutton

Table 2 - Modified entry for *lamb* record

1.3.5 As seen above in Table 2, the MajHead definition was also kept alongside each record, and was later utilised in the network graphs. The title 'MajHead' is taken from the *HT-OED* Access database as the shorthand for the main sequence headings which appear after the designated category number, and is adopted for this project. An example of where the MajHead would appear in the *HT-OED* can be seen below, in this instance, for the word *Mutton*:

5 Occurrence marks the number of times the lemma appeared in the text, the Part marks the word class, AppS is the date the word is first recorded and AppE is the last recorded date, with 2000 marking words that were still in use when the *HT-OED* was published.

‘01.02.08.01.05.05 (*n.) Mutton*
mutton c1290- · sheep-meat/sheepmeat 1975-
01 *quality* muttoniness 1882 **02** *carcass of* […]’
(eds. Kay, Roberts, Samuels and Wotherspoon 2009: 335)

1.3.6 The MajHead added an extra level between the word and the third level semantic group, acting as a proxy definition, or otherwise suggesting towards the specific connotation or denotation of each word. This resulted in a more readable network, which identified specific meanings within the broader semantic categories. An example of this can be seen in Table 3 below.

1.3.7 From the MajHeads visible in Table 3, it is possible distinguish between the definitions for the word *sleep* which fall into the category 01.03.01 Sleeping and Waking. Although the MajHead is not the same as a definition, acting instead as a more specific semantic group which the word belongs to, it offers a way of organising the words by meaning without having to display the full multi-level taxonomy.

1.3.8 Coding each word in this way allowed for both a broad view of the text using the higher level semantic categories, and a closer analysis of each possible usage based on the MajHeads. Of course, cutting the heading at the third level (Table 2) distributes the meaning of the specific word within the broader semantic category. Returning to Table 1 and 2, this is displayed as a specific word within the broader semantic category. Returning to Table 1 and 2, this is displayed by the word *lamb* being counted towards the SD of 01.02.08 (Table 2) instead of 01.02.08.01.05.05 (Table 1). This, however, is the goal of the project; a broader and more distant view of the text using the dominant semantic fields. By focusing only on the higher tier of categories, each semantic field has the potential to reach a higher SD than focusing at, for example, the 6th or 7th level of the *HT-OED* taxonomy**.** As this project relies on visual representation of these categories, having more distinct categories instead of countless minor ones is more suitable for analysing the broader themes of the corpus.

Occur	Word	Part	Group	MajHead
2	sleep	vi	01.02.04	Age/be defined by cyclical growth periods
2	sleep	vi	01.03.01	Sleep
2	sleep	vi	01.03.01	Go to bed/retire to rest
2	sleep	vi	01.05.05	Be inactive
2	sleep	vi	01.05.05	Be quiet/tranquil

Table 3 - Example of entries for the word *sleep*

1.4 Gephi

1.4.1 Gephi is an open source[6] software package for visualisation and manipulation of data networks. It was chosen for this project for a number of reasons, the dominant one being its ability to cope with a very large number of source nodes and edges. The 'nodes' for this project, as mentioned previously, represent each individual lemma entry in the network, and are visually displayed as a round dot in the network. In addition to lemma nodes, each semantic category at the third and second level (eg. 01.01.02 and 01.01) had a node entry to represent them, their titles capitalised to set them apart from their counterparts. The third type of node used in this network was the MajHead node that determined each denotation of the lemma node, and was marked with asterisks at each side. The 'edges' represent the connections between one node and another, and are displayed as a line between the two. For this project, the 'connection' dictated the relationship between the lemma node and the MajHead, the MajHead with the third level semantic category, and the third level category with the second (Figure 1). The reason for using all three types of nodes was the result of a limitation within Gephi, as discussed below, but resulted in a large number of entries for the networking software to cope with. Despite the aforementioned limitation, Gephi was expertly capable of handling the large amount of data necessary to this project, and was the clear choice amongst rival software. In addition to this, Gephi came pre-packaged with a number of tools for network analysis, of which the Weighted Degree and

[6] Available to download at: https://gephi.org/

Betweenness Centrality algorithms were used for this project. Furthermore, Gephi has a large online user community[7] which helped with troubleshooting, and a number of free plug-ins have been created to expand its capabilities.

1.4.2 For this project, the plug-ins used were OpenOrd[8], Noverlap[9] and Sigmajs Exporter[10]. OpenOrd is a layout algorithm which displays the nodes in clusters for clearer visibility, while Noverlap adjusts the nodes within the OpenOrd layout to prevent overlap and label confusion. Sigmajs Exporter was used for creating .html export files using JavaScript code. The resulting files can be opened using a web browser[11], showing the full interactive network which can be searched and navigated using the zoom and span functions. All of these plug-ins were adjustable, so several templates had to be created within Gephi to standardise the output across different data networks.

1.4.3 An example of an edge file created for this project can be seen in Table 4[12] above. The format read by the Gephi import function is Comma-Separated Values (CSV), in this case with each record separated by a semicolon. The table does not have titles or 'labels' as these are only necessary for the Node CSV file, and are automatically attributed using the 'ID' within Gephi. For the nodes that represent semantic categories, the ID was set to the corresponding category number within the *HT-OED*, but all other nodes and edges had a randomly generated number series, as seen above with 60001 to 60025 and so on. This is a necessity for Gephi, as each entry must have a unique ID. The edge weight had to be adjusted by two decimal points to avoid unnecessary bulk**.**

1.4.4 It is necessary to note that Gephi software was not without its limits. A particular issue had to be overcome as a result of the program's lack of support for multiple edges between nodes. As shown in Table 3, the word *sleep* fell into the category 01.01.05 Water twice, once with the MajHead 'Be inactive' and once with 'Be quiet/tranquil'. Originally the data was to be presented using the MajHead as the label for the 'edge' or the connection between the word node and the Semantic Category node, but this would have required multiple connections (edges) between the

---

[7] Accessible at: http://forum.gephi.org/
[8] Available to download at:https://marketplace.gephi.org/plugin/openord-layout/ or through the Gephi plugins tab.
[9] Avaliable to download at: https://marketplace.gephi.org/plugin/noverlap/ or through the Gephi Plugins window.
[10] Available to download at: https://marketplace.gephi.org/plugin/sigmajs-exporter/ or through the Gephi plugins tab.
[11] Currently, SigmaJs only supports the Mozilla Firefox web-browser for files which are not hosted online. As this is the case for the digital networks created for this project, a README text file is included with each relevant Appendix with instructions on how to open these files.
[12] Complete versions of the files can be found in the Appendix 7-10 folders.

same two nodes. Neither Gephi, nor any similar open-source software was compatible with this design, so instead, the word node was connected by an edge to the MajHead, and then to the corresponding semantic category. An example of this connection for the word *lamb* can be seen in Figure 1 below:

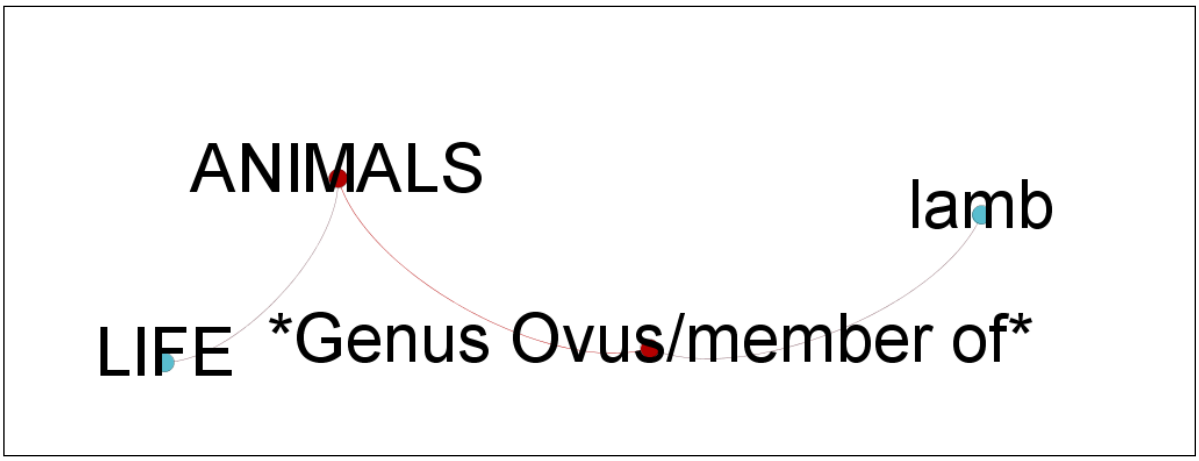


Figure 1 - Example visualisation within Gephi for the word *lamb*

1.4.5 Figure 1 follows the path of one entry for the word *lamb*, which goes from the word node, to the MajHead, then to the third level semantic category and finally to the second level semantic category of 01.02 Life. In the complete semantic networks, the second level categories aggregate into the corresponding first level headings, but for the purpose of this visualisation, a simplified format was used.[13] Using this chain it was possible to encode a clear level of semantic distinction for each node without overburdening the already complex network. For future analyses, it would be possible to include more or less information as needed, while maintaining the same overall degrees and semantic density results.

1.5 Original proof-of-concept

1.5.1 In order to test this theory, an initial proof-of-concept study was carried out, using a machine-readable corpus of William Blake's *Songs of Innocence and Experience* (1789/1794) to analyse the range of possible words which could have been used by Blake to realise a given

[13] One final correction had to be made to the data for it to be used in Gephi, and that is the removal of all commas in the Semantic categories and MajHeadings, so that the comma-delimited graphs were not corrupted. These were replaced with a '/'.

concept. For that project, the *HT-OED* was only cross-referenced with a list of the ten most frequent lexemes from each set of poems. This limit was imposed on the data as the lemmas were cross-referenced manually with the *HT-OED.*

1.5.2 With the resulting data, the SD distribution was displayed using a treemap visualisation showing the difference between *Songs of Innocence* (1789) (*SoI)* and *Songs of Experience* (1794) (*SoE).* This analysis determined in a preliminary way the particular semantic densities characteristic of each set. The data derived from the top ten lexical items alone, however, proved to be too limited to carry out a thorough analysis of the author's style. It was, nevertheless, possible to discern from it the overall viability of a future project by harnessing a derived methodology on a larger scale, which is attempted in this thesis.

1.6 Songs of Innocence and Experience

1.6.1 Before continuing, it is necessary to account for the decision to use William Blake's *Songs of Innocence and Experience* as the literary text for this project. As mentioned previously, his work was already used for the original proof-of-concept study, and was retained for this project. The reason for this choice, as it was for the original pilot study, stems from existing critical work on the *Songs*.

1.6.2 It is widely acknowledged that the *Songs* display distinct and socially motivated themes, veiled in the child-like nursery rhyme form (Bottrall 1970; Bronowski 1954). *Songs of Innocence* (1789) was originally published as a book for children, and Blake continued to market it as such even after the publication of *Songs of Experience* (1794) which more visibly showcased mature themes (Bottrall 1970: 13). Posthumous interest in Blake's work (Yeats 1961 [1897]) led to a resurgence in critical analysis of his work, and now the 'critical exegesis has laid bare, even in these seemingly direct little poems, complexities of meaning undreamed of by Blake's earlier admirers' (Bottrall 1970:11). The *Songs*, therefore, appeal to this analysis in two ways: they engage readers on multiple levels, and they can be split into two collections with contrasting themes, suited to a comparative analysis.

1.6.3 The latter of these assessments, as summarised by Bronowski, raises a further boon the *Songs* add to this analysis:

> ‘The happy world of the *Songs of Innocence* had been a state of mind. The unhappy world of the *Songs of Experience* is the contrary state of mind, though that contrary has been thrust upon the mind of the hypocrite. The symbol of innocence has been the child. The symbol of experience, mazy and manifold as the hypocrite, and as fascinating, is the father.’
>
> (Bronowski 1954: 166-167)

Bronowski’s mention of symbols in the *Songs* is particularly suited to showcasing the benefits of Semantic Density mapping. For this technique to be useful in critical analysis of literary texts, it would have to be capable of picking up on symbolism in the text. This will be explored further in Chapter 4 and 5, with the discussion of results.

1.6.4 One further benefit of choosing an author from the Romantic period pertains to the reduction of all possible meanings by the recorded date of usage in relation to the text. While the language used by Blake and his contemporaries naturally deviates from modern English, the casual reader would likely feel confident in anticipating specific connotations of the poet’s words. By referring to the *HT-OED*, however, it is clear that several meanings that were present during the Romantic period have since become obsolete. Of course, this knowledge is not new within Linguistic and Literary academic circles. Working from the assumption that being aware of these retired definitions could illuminate something new about the work of John Keats, Jeremy Smith utilised the *HT-OED* for precisely this purpose[14] (Smith, 2006). This project hopes to emulate this method of discovery, but on a larger scale, through the digital networks of the poet’s works.

1.7 Revised Claim

1.7.1 This project continues from the original proof-of-concept, expanding into an analysis of every lexical item in Blake’s *Songs of Innocence and Experience*. Access to the *HT-OED* Access database allowed for the expansion of the size of the corpora, which would have not been possible if each entry had to be manually recorded (the *SoI* corpus cross referenced with the database, for example, returns over 13,000 *HT*-OED entries).

1.7.2 Originally, the expansion was intended to include the work of an additional author from the Romantic period, which would serve to open up a comparison-driven study. When taking this

[14] Amongst his discoveries was the meaning of the word *touch* in reference to a gynaecological examination in Keats time. Combined with Keats medical background, Smith was able to make a positive claim for a re-evaluation of the word in *Endymion (*Smith 2006).

into consideration, however, the scope of the project had to be adjusted significantly, and the decision was made to keep this proof-of-concept focused on the work of only one author. It was still possible to discuss the *Songs* as two separate units, and for the purpose of the network analysis they were converted into two separate corpora, one with the collected poems from *Songs of Innocence* and the other with the poems from *Songs of Experience.*[15] This project, serving as a further proof-of-concept for SD mapping, will utilise both corpora as a trial for future application to the work of multiple authors and literary periods. Although both corpora come from the same poet and time period in this project, the critical analysis will address the capabilities of SD mapping in identifying the idiosyncrasies of each corpus. This thesis will therefore serve as an investigation of the methodology behind Semantic Density analysis, and the overall viability of this approach.

1.8 Roadmap

1.8.1 The following chapter of this thesis provides a literature review, the purpose of which is to position this project within an existing body of work in digital humanities. In particular, the background to corpus linguistics will be established through a discussion of the work of John Sinclair (1991; 2003), who is noted for his achievements in regulating both the theory and the methodology of corpus analysis. An outline of existing methods and approaches to Semantic Network analysis will be presented through the work of Klaus Krippendorf (2004) and Van Atteveldt (2008). To take this project closer to literary analysis, a discussion of the current work by Franco Moretti (2005), Michael Witmore and Jonathan Hope (2004; 2007) and their colleagues[16], will follow, with particular attention afforded to the 'distant reading' concept conceived of by Moretti (2005).

1.8.2 The third chapter will outline in further detail the concept of Semantic Density in relation to existing techniques, and will explore the theory behind SD mapping. Expanding on existing work by the aforementioned linguistics, this section will showcase the application of the *HT-OED* in corpus analysis, and how this can be used to infer the semantic properties of a text. This section will also include an outline of the project methodology, and a discussion of Gephi algorithm and analysis results.

---

[15] See: Appendix 7 and 8.
[16] See: Allison, S., et al. (2011). "Quantitative Formalism: an Experiment." (Pamphlet) In: Literary Lab 1.

1.8.3 Chapter 4 will examine the data obtained from the corpus analysis, and *HT-OED* tagging of the lexical items in both texts. Here, the theory of SD mapping will be put into practice, with visualisations obtained from the analysis of the corpora. Four separate data sets were created for this purpose, one each for the *SoI* and *SoE* collections, and smaller networks for one poem from each collection: 'The Lamb' and 'The Tyger'. This section will test the methodology for the analysis, and will observe the use of the *HT-OED* Access database and the corpus linguistics AntConc tool. As this project is an expansion of a previous proof-of-concept, some of the data gathered for that study will be used here. The widening of the corpus data to encompass all lexical items from the chosen texts, however, will showcase a broader analysis of the literary texts. For this purpose, Gephi software will be used to display the SD analysis data. This section will also form the foundation for the critical analysis of the author's work.

1.8.4 The following chapter will address the use of semantic density mapping as well as semantic networks in literary analysis. Contrary to the work of Franco Moretti (2005), this project will address the effectiveness of a 'distant reading' analysis in combination rather than as a replacement for a more traditional close reading of a text. Here, existing critical work on the *Songs* will be examined side by side with the SD visualisations, in the hope of establishing a new way of conducting literary criticism.

1.8.5 In the sixth chapter, these results will be discussed in relation to future applications and research. As this project is intended to establish a working framework of analysis, it will be possible to apply this model to different texts and literary periods. In addition to this, the imposed limitations of the word count for this thesis dictate that several sections have to be left for future exploration. Of these, one of the most prominent areas of future research is the relationship between the cognitive associations formed by readers, and the semantic mapping using the *HT-OED*. This section will therefore conclude with a brief outline of implications for future research.

1.8.6 Finally, chapter 6 will summarise and conclude the paper, returning to the original hypothesis and highlighting any unexpected or illuminating results. This project is ambitious in both scope and theoretical implication, so any deviation from the expected results will guide necessary developments in the future of SD mapping.

Chapter 2 - Literature review

2.1 Corpus linguistics

2.1.1 This project originated as the result of the increased interest and possible uses of the *HT-OED* in literary analysis, and only through trial and error developed into a digital corpus analysis project. As a result, it was necessary to place the notion of SD mapping within an already established body of work. The principles of corpus creation and processing came from the work of John Sinclair (1991; 2004)**,** and the Birmingham school of corpus linguistics. Despite the fact that Sinclair's most prominent work on the subject, *Corpus, Concordance, Collocation* (1991) is now more than two decades old, the robust framework and methodology presented for corpus creation and analysis was endlessly helpful. Of particular interest to this project, however, was the question raised by Sinclair during his development of the theoretical approach to corpus linguistics: can **'**discrete units of text, such as words, […] be reliably associated with units of meaning?' (Sinclair 1991: 3)**.** This project hopes to answer this by combining digital corpus analysis with the semantic categorisation of the *HT-OED*.

2.1.2 It is important to note that Sinclair's opinions on corpus linguistics is not without criticism, in particular his advocation of minimal annotation has given rise to competing theories that promote broader engagement with the corpus data[17]. Consequently, as this project relies on a second dimension to the corpus data, namely the semantic categorisation based on the *HT-OED*, it in many ways frustrates Sinclair's core principles. His stance, however, that 'the ability to examine large text corpora in a systematic manner allows access to a quality of evidence that has not been available before' (Sinclair 1991: 4) is one that forms the basis for this investigation.

2.2 Content Analysis

2.2.1 John Sinclair's work, as mentioned above, was the foundation for the corpus analysis methods used for this project. The techniques that were used to manage the resulting data were borrowed from another field within language studies: Content Analysis. As mentioned previously, 'mapping' texts using the semantic categories of the *HT-OED* shares aspects of both the Semantic Network approach and the Dictionary approach, both of which are methods within

[17] See: Wallis (2007)

the wider field of electronic Content Analysis (Krippendorff 2004; Van Atteveldt 2008; Roberts 1997). In brief, Semantic Network Analysis, or Network Analysis seek to represent language as a network of 'concepts and pairwise relations between them' (Carley 1997: 81), resulting in a web-like visualisation. The Dictionary approach involves grouping words within a text by 'shared meanings' and tagging them with pre-determined notional categories (Krippendorff 2004: 284-285). As summarised by Van Atteveldt, 'in the social sciences, Content Analysis is the general name for the methodology and techniques to analyse the content of (media) messages' (Van Atteveldt 2008: 3). It is important here to note the use of 'social sciences', as the work on automatic Content Analysis is almost exclusively framed within this discipline.

2.2.2 In spite of the similarities between Content Analysis methods and those detailed in this thesis, the grounding of the technique within the Social Sciences discipline resulted in the majority of the research for this project to have been conducted before coming in contact with the approach. Applying the methodology retrospectively to Semantic Density networks, however, has proven to be favourable. One possible cause for this is offered by Van Atteveldt who stated that:

> 'Content Analysis and linguistic analysis should be seen as complementary rather than competing: linguists are interested in unravelling the structure and meaning of language, and Content Analysts are interested in answering social science questions, possibly using the structure and meaning exposed by linguists'
>
> (Van Atteveldt 2008: 5).

2.2.3 Diverging from Van Atteveldt's stance that Content Analysis is suited primarily to answering social science questions (albeit doing so without competing with linguistics), this project attempts to utilise Content Analysis from a literary-linguistic perspective. To a degree, this project is an attempt to adapt the paradigm for use in literary analysis. The end goal, however, is to move beyond existing methods of Content Analysis through the Semantic Density approach. Consequently, this thesis will address the ways in which SD can account for some of the issues raised by traditional Content Analysis.

2.2.4 Van Atteveldt argued that 'a recurrent problem in searching in a text is that of synonyms' and similarly sought answers to this problem in the 'lists of sets of synonyms' available in thesauri (Van Atteveldt 2008: 48). Referring to two thesauri specifically, Roget's Thesaurus (Kirkpatrick 1998) and WordNet (Miller 1990; Fellbaum 1998), Van Atteveldt acknowledged the

application of thesaurus resources in Semantic Network Analysis. His interest in them, however, did not extend to the semantic taxonomies used within the thesauri, choosing to focus instead on the ability to scan a text for synonyms, and disambiguating words using Part-of-Speech (POS)[18] tagging (Van Atteveldt 2008: 48). Offering as an example that 'safe as a noun (a money safe) and as an adjective (a safe house) have different meanings.' Van Atteveldt chose not address the implications of this distinction in his analysis (Van Atteveldt 2008: 48). This is particularly interesting when coupled with Van Atteveldt's concerns over 'standard ways to clearly define the meaning of nodes in a network and how they relate to the more abstract concepts' (Van Atteveldt 2008: 5), and indicates a gap in current materials for Content Analysis. This project is an attempt to address these issues by first defining broad semantic groups of nodes using the *HT-OED*, and then referring to the Semantic Density to determine the most likely node meanings.

2.2.5 To illustrate the sentiment above, it is possible to look at the path for Semantic Network analysis, as diagrammed by Van Atteveldt in his book:

'Text -> Extraction -> Network Representation -> Query -> Answer'

(Van Atteveldt 2008: 4,205)

This project offers an additional step between 'Extraction' and 'Network Representation': Semantic classification and density analysis.

2.2.6 The Dictionary approach to Content Analysis, as outlined by Krippendorff, involved using the dictionary taxonomy for representing text 'on different levels of abstraction' (Krippendorff 2004: 283). Offering the example of Sedelow's (1967) work as a 'convincing demonstration that analysts need to compare texts not in terms of the character strings they contain but in terms of their categories of meanings', he recounted the example of her work on Sokolovsky's *Military Strategy*, which found that two respectable translations of the text 'differed in nearly 3,000 words' (Krippendorff 2004: 283). He inferred from this that 'text comparisons based on character strings can be rather shallow', and that if done well, the Dictionary approach can serve 'as a theory of how readers re-articulate given texts in simpler terms' (Krippendorff 2004: 283-284). His argument in favour of the Dictionary approach can also be applied to SD analysis, which operates from a similar foundation. Even closer to this was Sedelow's original observation which proposed 'applying ordinary dictionary and thesaurus entries to the given text and obtaining

[18] In corpus analysis, POS tagging refers to identifying the lexical class of the word using adjacent words.

frequencies, not of actual character strings, but of work families that would account for texts on the basis of common meanings' (Krippendorff 2004: 284). Taking this into account, it is possible to think of the methodology for this project as the next step in the progression of the Dictionary Approach to Content Analysis.

2.2.7 Despite the similarities in handling the data, Krippendorff's account for the use of frameworks differs from the one proposed by this project. His stance that 'in content analysis, semantic networks are of particular interest because they preserve relationships between textual units' (Krippendorff 2004: 294) is in keeping with the core foundation of this project. What is missing from Krippendorff's commentary, however, is the application of these resources to literary texts. Therefore, while certain concerns shared by Krippendorff were key to the methodology behind this project (in particular, that the results are 'reliable', 'replicable' and 'valid' (Krippendorff 2004: 18)), the second part of the thesis marks a departure from the scientific approach into relatively subjective critical analysis.

2.2.8 From the above survey of Content Analysis, it was possible to draw several conclusions. Firstly, that Semantic Density analysis shares a common ancestry with Dictionary and Semantic Network approaches. Secondly, that these areas of research, like this project, were concerned with the application of digital resources to texts for the purpose of statistical analysis. Lastly, that, despite the similarities between the approaches**,** this thesis takes into account several factors which were not considered in the original methods. This includes the use of the complex taxonomy of the *HT-OED*, and the time-bound factor of the denotations and connotations considered for the networks. In addition to this, this project is concerned with literary texts, which brings with it a different range of concerns, such as genre and style which were lacking in previous approaches. Furthermore, the networks attempted by this project offer an interactive approach to 'reading' the text through the network itself. This aspect of the project draws closer to the work on 'distant reading' and will be discussed in the following section. Finally, this project presents these networks not in isolation, but as a tool which can be used for critical analysis of literary texts, alongside traditional methods rather than as a replacement for them. Whether it will function as well as previous methods, or offer any new discoveries will be determined through the discussion of the results (Chapter 4), but the rich background of work already conducted on this topic serves to strengthen its foundation.

2.3 Distant Reading

2.3.1 In investigating more recent advances in corpus linguistics, two studies stood out as paramount to this project. The work of Jonathan Hope and Michael Witmore in the analysis of genre in Shakespeare's dramatic work (2004; 2007) and Franco Moretti's[19] work on 'distant reading' (2000) and further 'reduction and abstraction' (2007: 3) in *Graphs, Maps, Trees* as well as later collaborative work (Allison,Witmore, Moretti, et al. 2011). Although the primary concern of the authors in the above texts was that of specific literary features, such as genre or historical and geographical narrative analysis, they all chose to employ the use of quantitative analysis in their work, instead of more traditional approaches. This, as described by Moretti (2010: 28-29) served to distance the reader from the text, allowing the 'focus on units much smaller or much larger than the text' itself, which in turn became 'a condition of knowledge'. In this regard, 'distant reading' is similar to Semantic Network and Dictionary analysis, as it removes the reader from the source and allows for a perspective that was inaccessible at text level.

2.3.2 Hope and Witmore's (2004; 2007) study of genre within Shakespeare's work was accomplished with the help of Docuscope, a digital tool for corpus-based rhetorical analysis. Their intention was to allow the computer to attempt to calculate the classification itself, which could then be used to make an informed judgement on what features are most prominent in this classification process. Despite the varying success of Hope and Witmore's Shakespeare project, their research opened up yet more paths for SD analysis. In particular, this project has in common with theirs the notion that 'computer visualisation' can allow access to 'whole texts' (Hope and Witmore 2004). This is valid approach in both the field of corpus analysis, where the language is taken as a whole and is unedited by human perception, and Semantic Network Analysis, where visual representations of a text can signify that text as a whole. This project will attempt to represent the *Songs* in this way, and to show how it is possible to draw conclusions about the text based on the information contained in these visual representations.

2.3.3 A similar digital humanities study, and one more closely tied to this this project is the shared work of Witmore and Moretti et al in 2011, in which the authors conducted a series of tests to determine if quantitative analyses could be used to distinguish between different genres and authors. Amongst their results, the authors found that in using digital corpus analysis tools, they were able to discover 'imperceptible linguistic patters that provide an unmistakable stylistic

[19] The work of Whitmore and Moretti is combined in the article 'Quantitative Formalism: an Experiment' (2011).

'signature' (Allison, Witmore, Moretti, et al 2011: 14). This project is similarly concerned with distinguishing a property of the text using quantitative analysis, albeit one that is displayed by notional rather than lexical hierarchy.

2.3.4 To this end, it is possible to take Moretti's approach to map-making in the direction that was not explored by the author himself. Discussing Franco Moretti, his friend and colleague Steven Berlin Johnson noted that the theorist believed that 'the future of literary criticism was going to lie in map-making' (2011: 81). Although for Moretti, these maps were taken from geography, and served to visually represent narrative space (unlike the 'maps' created from an SD analyses), Moretti's notion that visual representation of literary data was the future of critical analysis is one shared by this project. The core belief that Moretti bases all three sections of his *Graphs, Maps, Trees* on is that 'a field this large cannot be understood by stitching together separate bits of knowledge about individual cases, because it *isn't* a sum of individual cases: it's a collective system, that should be grasped as such, as a whole' (2007: 4). In this essay, this notion will be reflected in the 'distant' analysis of SD mapping, but the goal of this research, unlike Moretti's, is not to replace close reading in critical analysis, but rather to provide a robust system of digital corpus analysis that can then be used to complement a close reading of a literary text.

Chapter 3 – Methodology

3.1 Weighted Degree

3.1.1 The primary goal of this project was to highlight new tools for the analysis of literature. For the visual aspect, the digital networks created with Gephi aim to provide an interactive experience for 'distantly reading' texts. Behind the networks, however, lie the complex algorithms which calculate the connections between nodes to present a meaningful representation of the data. The simplest of these is the 'Weighted Degree', which calculates the 'degree', or number of edges coming into a node as well as the weight attributed to each edge, thus combining the degree with the weight of the connected edges for prioritising nodes in the network.

3.1.2 The 'weight' in the case of this project was taken from the number of occurrences of each lemmatized word in the corpora. Returning to Table 2, the weight for the 01.02.08 […] Mutton record for the word *lamb* was 18, as this is the number of times the lemma appeared in the *SoI* corpus. In the network, the edge between the *lamb* node and the MajHead Mutton has a weight of 18, as does the edge from Mutton to 01.02.08 Food and Drink. The edge between 01.02.08 Food and Drink and 01.02 Life is the sum of each MajHead weight coming into the 01.02.08 node. The decision behind setting the edge weight as the occurrence of the lemma arose from the desire to highlight lemmas that appeared more frequently within the text, and their associated semantic categories. The followed logic was that words which appeared more frequently within the text had a greater impact on the distribution of semantic categories in relation to the text, and should contribute towards a higher Semantic Density.

3.1.3 Another example of this can be taken from Table 3 for the word *sleep*. With an edge weight of 2, *sleep* does not contribute a large amount to the Semantic Density of its nesting categories. However, as it appears twice within 01.03.01 Sleeping and Waking, this category gains a Semantic Density of 4 in relation to *SoI*. The Weighted Degree doesn't just affect the Semantic Category nodes, but the MajHeads and the lemma nodes themselves. As the edges used for this analysis were set to 'undirected', with no designated 'source' and 'destination' node, the edge weight benefits all connected nodes within the network. This was designed so that the lemma nodes which appear most frequently within the corpora have visual prominence within the Semantic Network. Despite the focus of this analysis on Semantic Density within the

frameworks, the 'distant reading' approach serves to highlight to the reader the words that the poet chose to use most frequently. In this manner, Semantic Density mapping engages with the stylistic choices of the author.

3.1.4 The final category level taken into account within this network analysis, are the three upper-level categories of the *HT-OED*, 01 The World, 02 The Mind and 03 Society. These will have the highest incoming edge weight of all the nodes, as they connect to every second level category on in the network. The hypothesis presented in relation to these categories, was that they will vary significantly depending on the content of the corpus. Having these three overarching categories for reference, however, could indicate the semantic leanings of a particular text. Below, Figures 2 shows cropped visualisations of the upper-level semantic nodes in *SoI*, when selected within the Sigmajs export.

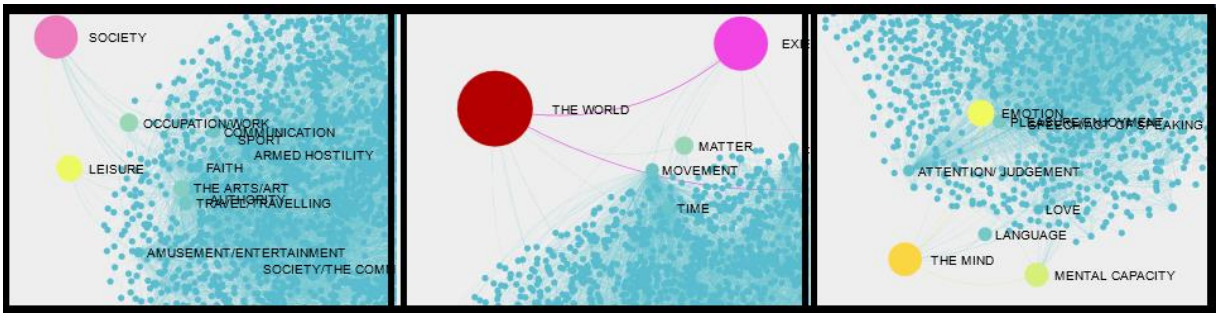


Figure 2 – Cropped images of the three upper-level semantic category nodes, taken from the same screenshot of the *SoI* Weighted Degree network[20].

3.1.5 As visible from Figure 2 above, the three upper-level nodes vary significantly in size[21]. 02 The Mind is the smallest of the three, and 01 The World is by far the largest. This Figure does not, on its own, show anything significant about the text. To determine whether the size of the upper-level categories plays a significant role in Semantic Density analysis, it is necessary to discuss the nature of these categories, and their relationship to their nested nodes. This discussion takes place in Chapter 4, alongside similarly high-level Treemap analyses of the *Songs*.

3.1.6 Running the Weighted Degree algorithm in Gephi, in addition to providing a scale for the nodes within the network, outputs a graph representation of the node's distribution. The nodes are distributed along a count and value axis, with count representing the combined number of degrees

[20] See: Appendix 11.
[21] As it was necessary to capture all three within the same resolution, the screenshot for this image was taken with the zoom set at distant. The full screenshot can be found in Appendices at the end of the paper, under Screenshot

entering a node and value representing the highest value of a particular degree. In the graphs below, Figure 3 shows that *SoI* has a higher count distribution than *SoE* in Figure 4. The nodes represented at the higher levels in *SoI* have a larger amount of incoming edges than those in *SoE*. As previously stated, the edges represent the connections between words and semantic categories, so a higher overall connection number implies a more prominent category or word than a lower one. Following this logic indicates that *SoI* has themes that are stronger at the highest level than *SoE*, with the edges in *SoE* being more evenly distributed into different nodes. In short, this data should translate into more visible, coherent themes being present in *SoI* than *SoE*.

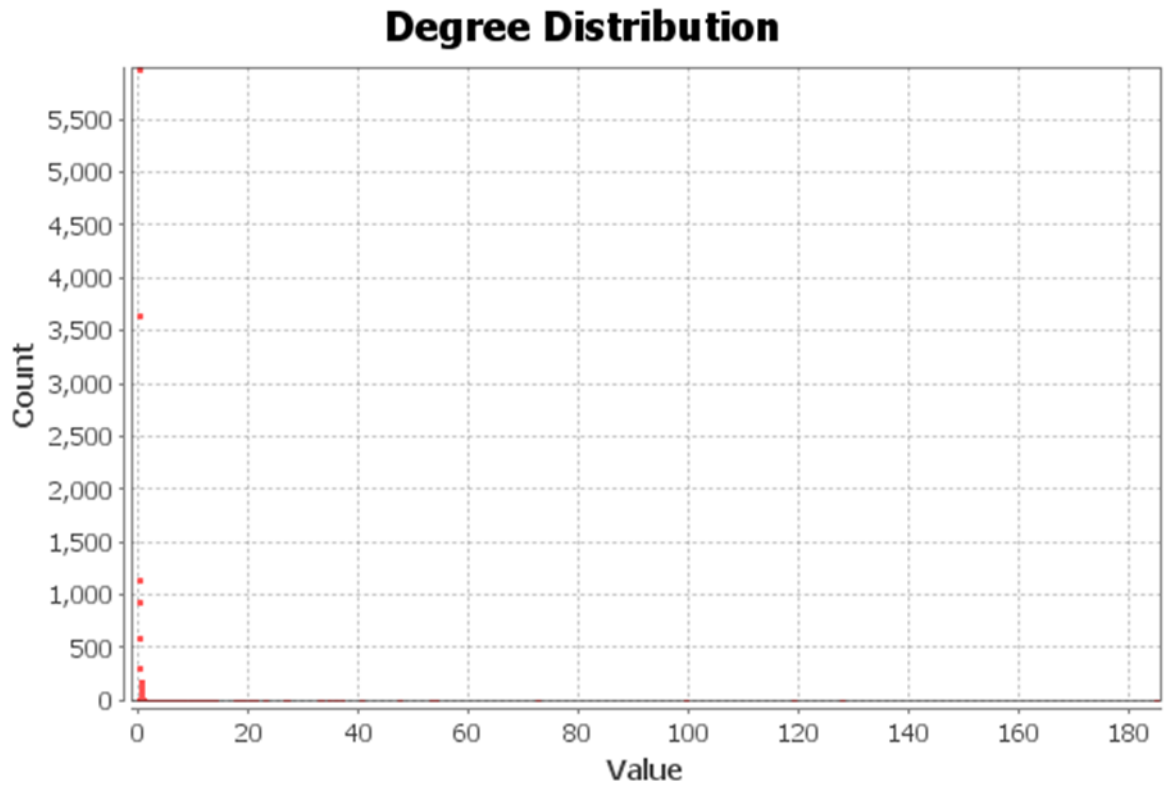


Figure 3 - *SoI* Weighted Degree graph.

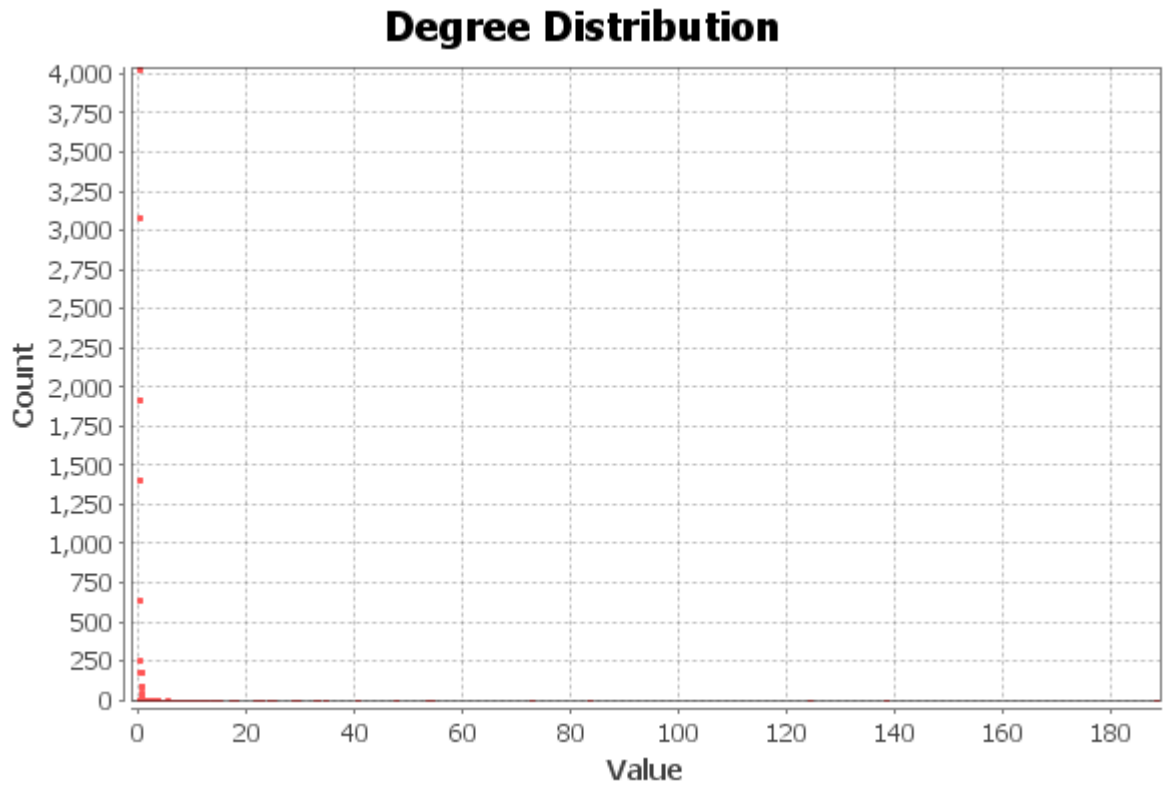


Figure 4 - *SoE* Weighted Degree graph.

3.1.7 Studying these graphs does not necessarily garner different or new results when compared to a reading of the original text, rather, the benefits from the digital computation of the weighted degree allows for almost instantaneous information, and thus is more suited to handling large corpora. Additionally, the data presented in Figures 3 and 4 above does not display the node titles represented by the markers on the graph, instead offering only an overview of the data. Running this diagnostic through Gephi, however, allows for the sorting of node and label size and colour based on the Weighted Degree, as seen in Figure 2, and will be examined in greater detail in Chapter 4. Gephi offers multiple diagnostic tools, all of which, when run, allow for the manipulation of the networks based on their results. For this project, two of these tools were used, the Weighted degree analysis as described above, and the Betweenness Centrality algorithm described below.

3.2 Betweenness Centrality

3.2.1 The second algorithm used for scaling the networks for this project was Betweenness Centrality. The purpose of using two different algorithms was to determine which would be best

suited to this project. As this is a proof-of-concept it was necessary to try multiple approaches for reaching the end goal of a coherent and cohesive Semantic Network.

3.2.2 Betweenness Centrality is another method for measuring a node's 'centrality' within the network. Weighted Degree above measured the centrality of each node by the number of other nodes connected to it, and the weight of each connection. Centrality is important in network analysis as it highlights the 'most active' nodes, which 'have the most ties to other actors in the network graph' (Wasserman & Faust, 1994, p. 178). Betweenness Centrality measures 'the share of times that a node [needs another] node […](whose centrality is being measured) in order to reach [a third node] via the shortest path' (Borgatti 2005: 60). This type of algorithm depends heavily on the amount of edges between each node, as this is the primary method of measuring centrality. A node which connects the largest number of nodes is seen as the most prominent.

3.2.3 Originally, this was seen as suitable for SD mapping, because the nodes that showed the largest number of connections to other nodes indicated a high semantic relevance to the corpus. An example of a Betweenness Centrality network can be seen in Screenshot 2 in Appendices. In this network, Semantic Category nodes are displayed as more popular than their lemma node counterparts. This was a positive result for Semantic Density mapping, as it allowed for a focus on the categories which have the highest number of words present in the corpora. Unfortunately, because Betweenness Centrality does not take into account the edge weight when processing the network connections, it does not fulfil the full demands of SD mapping. It is possible that this capability will be developed in the future, which would make this algorithm useful for SD analysis, but for this project, the following networks were all created using the Weighted Degree algorithm.

3.3 Methodology challenges

3.3.1 Before continuing with the results from the analysis, it is necessary to account for some of the issues that presented themselves during the design stage of the analysis. Rather than commenting on the results themselves, this section outlines some of the challenges that had to be overcome, and others that were set aside for the next stage of this methodological approach.

3.3.2 Some issues with the methodology have already been mentioned, namely the incompatibility of Betweenness Centrality with the goals of SD analysis and the inability to display multiple edges between nodes. The latter of these issued caused a problem with the

readability of the network. When a node is selected in the networks created through Gephi, all of the nodes connected to it are selected as well, and the rest of the network fades from view. An example of how this appears visually can be seen in Figure 5 below:

Figure 5 – Example of node selection for the category LOVE in the full *SoI* network[22].

3.3.3 As the network was originally intended to show only the connections between the word node and the semantic category node, it would have been possible to instantly see all of the semantic categories that the word falls into. The MajHead was to be used as a label for the edge that connected the word node to the category node, and would be visible by either selecting it or choosing to display edge titles in Gephi options. Unfortunately, as mentioned above, this would have required some nodes to have more than one edge connecting them, which is a feature not yet available in open source software, so the networks were created with the MajHead as a connecting node between the word and category nodes. As a result, as demonstrated in Figure 5 above, selecting the category node Love displays only the MajHeads within that category that define relevant definitions of the lemma nodes in the corpora. Love, as a third level category is also connected to the second level category Emotion. Selecting Emotion within the same network

[22] See: Appendix 11.

would display all of the third level categories that nest into it, and the upper-level category of The Mind, as seen in Figure 6 below:

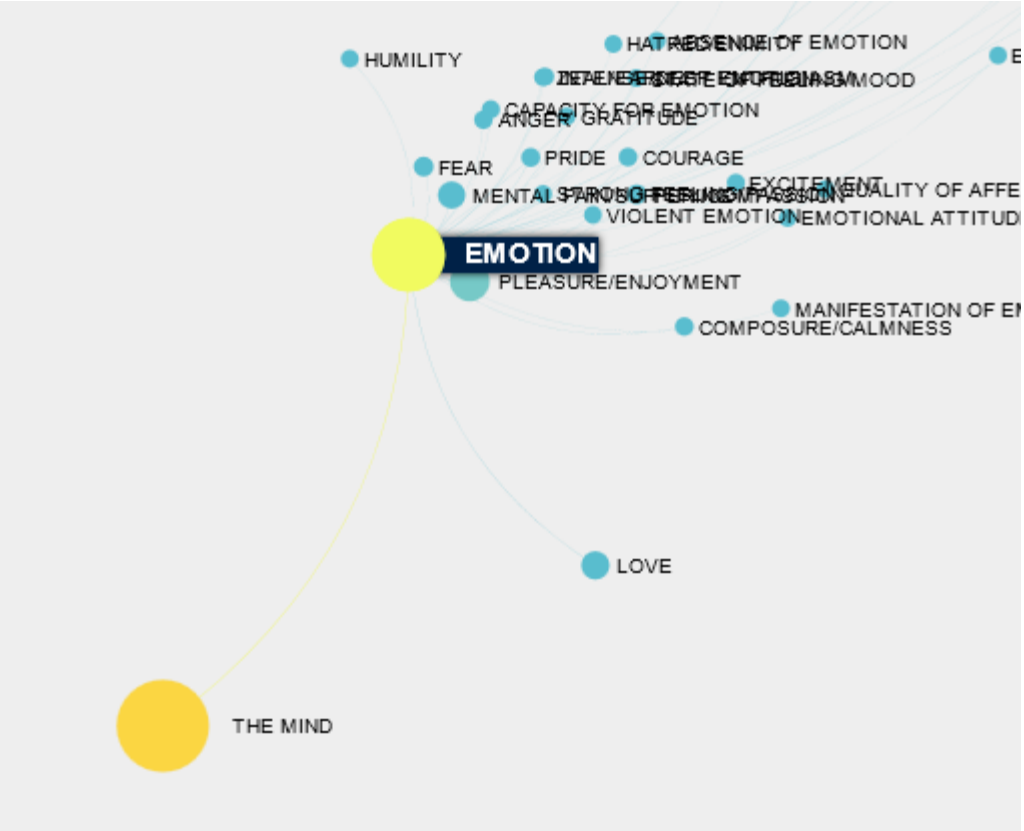


Figure 6 – Example of node selection for the category Emotion in the full *SoI* network[23].

3.3.4 In Figure 6 above, the titles of some nodes overlap, making them difficult to make out. It is possible to see them more clearly by zooming into the network or highlighting them with the cursor. Unfortunately, this issue is prevalent amongst all graphs of this size, and it not avoidable without making the networks too sparse to be coherent. It is, however, possible to view the original node Love from Figure 5, now unselected, as well as other nodes connected to the second level category Emotion.

3.3.5 In overcoming the multiple edge issue, the visual coherency of the networks suffered, and future SD projects have to resolve this to become more reader-accessible. Unfortunately, for this project it was not possible to find a viable alternative, and the MajHead fix had to be put in place. This did not invalidate the calculation of Semantic Density within the networks, as the edge

[23] See: Appendix 11.

weights were kept the same for the word to MajHead to semantic category connections thus affecting each other in the same way as a direct connection.

3.3.6 A third issue resulted from the content analysis part of the project. As mentioned above, one of the stages that a corpus can go through before being used for a discussion of the text is part-of-speech (POS) tagging. An attempt was made to tag the corpora used for this project in the same manner, but several issues prevented this from going ahead.

3.3.7 Due to the size of the corpora, the tagging process had to be automated to be viable for use. The programming language Python was used to attempt this, with a Natural Language Toolkit[24] (NLTK) module that is available online. A lack of familiarity with the language led to this process having limited success. Though it was possible to apply POS tags to the corpus, this was only achieved with one line at a time, a process that would have taken a similar amount of time to manually tagging the whole corpus. Additionally, unusual spelling and pre-modern words in Blake's text were not capable of being processed in this way, and would have to be manually tagged.

3.3.8 A further issue with POS tagging for a Semantic Density analysis of the text, it that of using the tagged corpus with the *HT-OED* database. As the corpus had to be lemmatised to be referenced with the *HT-OED*, it was not suitable for POS tagging in this state. If POS tagging was completed before lemmatising the text, then each tag would have to be re-attributed to the lemma form after. It would also be necessary to keep count of the number of times a word was used as the same part of speech in the corpus, so that the edge weight could be attributed based on this rather than just the basic frequency with which it appeared in the text. For example, if *green* appeared in the text four times as an adjective and two times as a noun, the edge weight between *green* and 01.04.09 Colour would be ten, as it appears within this category once as an adjective and three times as a noun.

3.3.9 As a result of these limitations and conversions, POS tagging was not completed for this project, and would have to be adjusted for future use of SD analysis. The downside of using a corpus that has not been parsed or tagged in any way is that there is no way to determine which usage listed in the *HT-OED* could be deemed relevant to this text. This is a known issue, and one that alters the results of this analysis. The networks displayed below do not distinguish between different parts of speech, displaying all possible variants with recorded usage. This had to be

[24] Available at http://nltk.org/

ignored for the purpose of this project, though it is important to note that for future SD analysis, POS tagging would have to be included to provide valid results.

3.3.10 The final issue facing this project was the aforementioned 'non-standard' spellings in Blake's work. To circumvent this, the corpora were created from a digital file of Project Gutenberg [25]collection of Blake's work, in which these are standardised. In this version, 'Tyger', becomes *tiger*, allowing for the word to be referenced with the *HT-OED*. This could be mistaken for a disregard of the importance variant spellings hold in literary texts. It is important to note that this is not the case, with the original spelling being used for the literary analysis. Because this project is concerned primarily with semantic value, however, unusual spelling variation was seen as rarely affecting the overall semantic category that a word can belong to. There are obviously exceptions to this, and the ability of these networks to distinguish and cope with these would have to be addressed in future projects.

[25] Accessible at http://www.gutenberg.org/files/1934/1934-0.txt

Chapter 4 - Results

4.1 Treemaps

4.1.1 Although this project relies predominantly on the networks created through Gephi to visualise Semantic Density results, this is not the only method available for this. One of the simplest ways to display the SD data is using treemap diagrams. Though lacking in detail, these maps can be useful when comparing data from multiple corpora. In the pilot study for this project, treemap representations of the top ten lexical words for SoI and SoE proved to be too limited in scope to provide any conclusive analysis. Using the Access database for this project increased the amount of data available for mapping, so a similar treemap analysis was conducted to determine any variation between the two texts at this level.

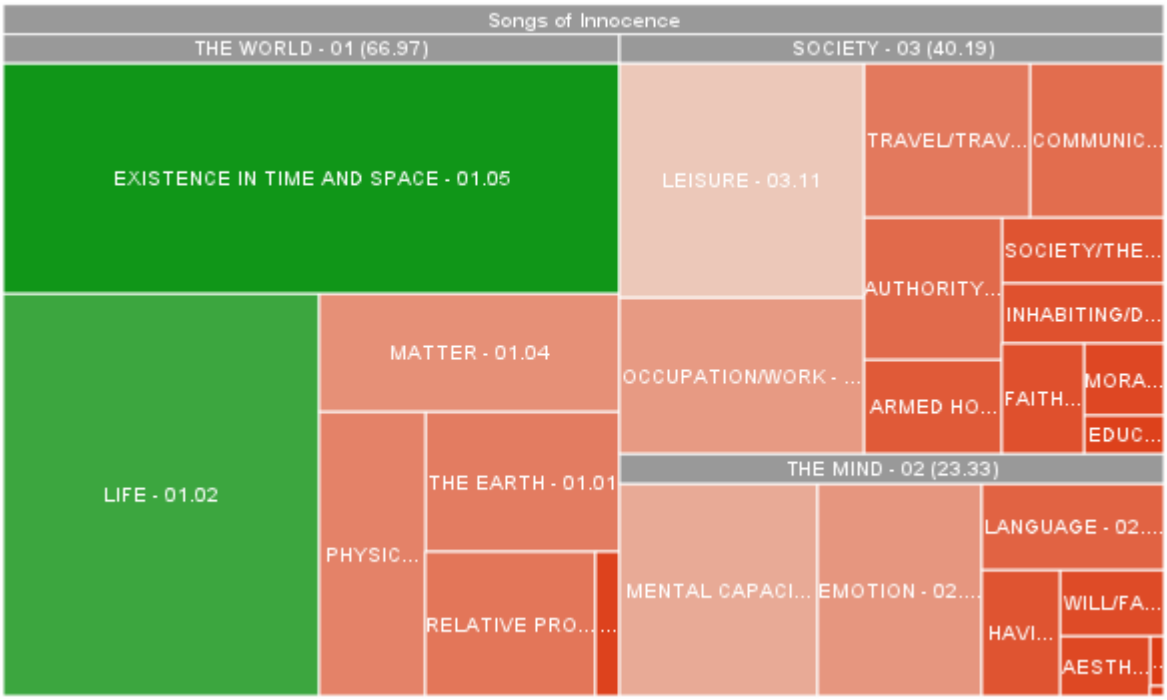


Figure 7 - Treemap *SoI*

4.1.2 The treemaps in Figure 4 and 5 display SD properties of the corpora to the second level of *HT-OED* taxonomy. The data compiled for the treemap diagram served primarily to highlight similarity between the two texts, rather than any obvious divergence. Interestingly, a similar discovery was made using just the top ten most frequent lexical words in the original proof of concept.

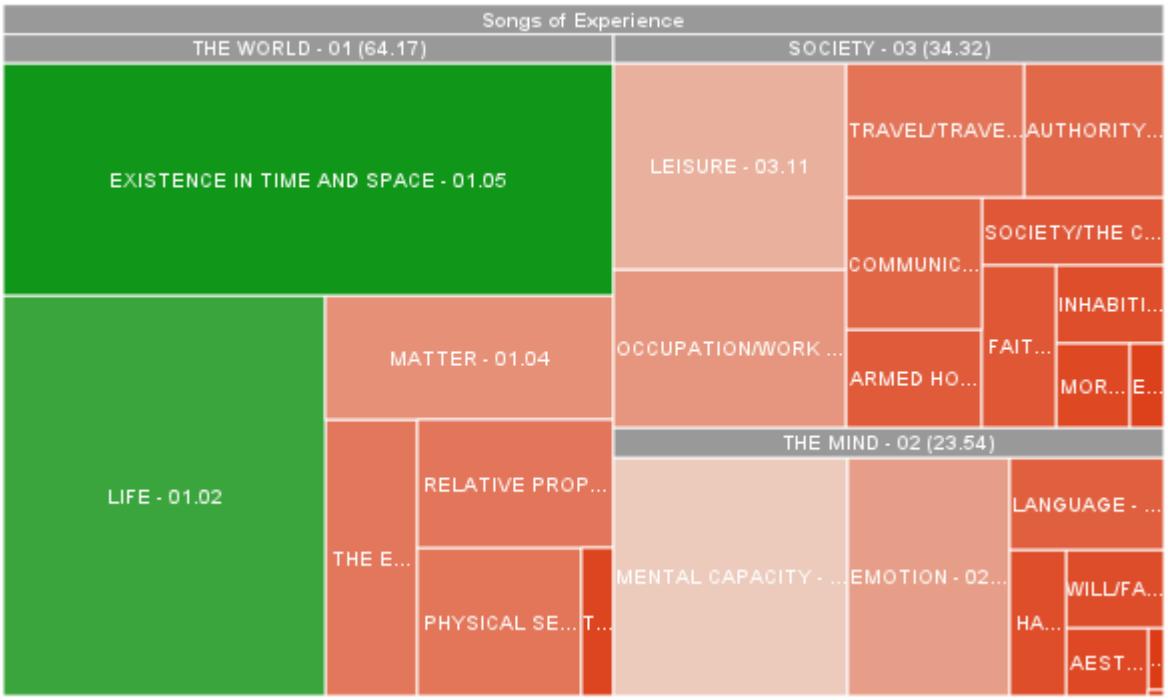


Figure 8 - Treemap *SoE*

4.1.3 Visually, it is obvious that certain categories (Life, Existence in Time and Space) are dominant in both texts, as are, to varying degrees below them Mental Capacity, Matter, Leisure, Occupation/Work and Emotion. An important question was raised by this discovery: are certain categories favoured due to the ubiquity of the lexical items that fall into them, or is authorial style responsible for this congruity?

4.1.4 From the outlines provided by the authors of the *HT-OED*, the answer pointed towards the former conclusion. The categories which were had the highest SD factor fell into the first section, 01 The World, which was defined by the authors as:

> 'the most readily observable phenomena of the universe, the land, sea, sky, etc., followed by living beings, their characteristics and physical needs.'
>
> (Kay, et al. 2009: xvii).

4.1.5 With this in mind, it was unlikely that any coherent text involving characters and a physical environment would have a higher SD property in either of the two remaining sections, 02 The mind, and 03 Society. To find themselves in the latter two categories of the *HT-OED*, the author would first have to depict their world using the language from the first.

4.1.6 After this conclusion, it is tempting to disregard treemap diagrams as static phenomena, serving to describe the process of characterisation in a literary text, and incapable of uncovering the author's idiosyncrasies. This is not unjustified, as the purpose of the treemap diagram is to highlight the highest frequencies in greatest detail, confining those less prominent to a proportionally smaller area of the graph. For this reason, the networks used for the visual analysis section of this thesis were produced using *Gephi* software, which allows for a more in-depth visualisation of the text, capable of clearly displaying all categories, large and small.

4.1.7 The original data used for the treemap analysis, however, can still be useful for the discussion of the lesser-featured categories. The full graph of the side-by-side comparison of the data used for the treemaps above can be seen in Appendix 2. A shortened version is visible in Table 4 below:

Category	**SoI**	**SoE**
THE WORLD - 01	66.97	64.17
THE MIND - 02	23.33	23.54
SOCIETY - 03	40.19	34.32
THE EARTH - 01.01	5.06	4.26
LIFE - 01.02	23.3	21.34
EXISTENCE IN TIME AND SPACE - 01.05	25.86	23.52
THE SUPERNATURAL - 01.07	0.71	0.83
EMOTION - 02.02	7	6.91
PHILOSOPHY - 02.03	0.15	0.2
WILL/FACULTY OF WILL - 02.05	1.39	1.39
REFUSAL/DENIAL - 02.06	0.03	0.02
ARMED HOSTILITY - 03.03	2.47	2.32
MORALITY - 03.05	1.11	1.09
EDUCATION - 03.06	0.61	0.53
COUNT	265.54	244.06

Table 4 – Shortened version of the table showing the comparison of the data used for the treemap analysis.

4.1.8 By examining Table 4, it is possible to identify the semantic categories which were used the least by Blake in his *Songs.* Very little is said, for example, of 01.07 The Supernatural, 02.03 Philosophy, 02.06 Refusal/Denial, and 03.06 Education. A reader who is already familiar with the *Songs*' dominant themes would perhaps be surprised to find these categories so poorly represented. 03.06 Education, for example, could be expected to reveal a higher 'semantic density' in relation to the *Songs of Innocence,* intended as they were, to be read to a young

audience (Bottrall 1970: 13). The answer to this, perhaps, can be discussed in relation to Gardner's analysis of Blake's original illustration for the title page of *SoI* (Figure 6):

> 'Blake's illustration is saying more to the imagination than meets the eye, preparing us for the uniqueness of his technique. The nurse and children are at once unrealistically yet naturally undisturbed by the breeze which blows about them. With the chair and the book, they inhabit a circle of composure in a landscape of vivid movement. Blake is painting, not simply the 'scene' but the idea derived from it. One hint we have to accept from this title-page is that the exuberance of nature and the education of children are complementary, and will delightfully and advantageously co-exist in *Songs of Innocence*'
>
> (Gardner 1986: 16)

This coexistence could explain the low SD of 03.06 Education if it is seen to be connected to a similar Density in 01.01 The Earth, a semantic category which contains the highest number of nature-associated imagery.

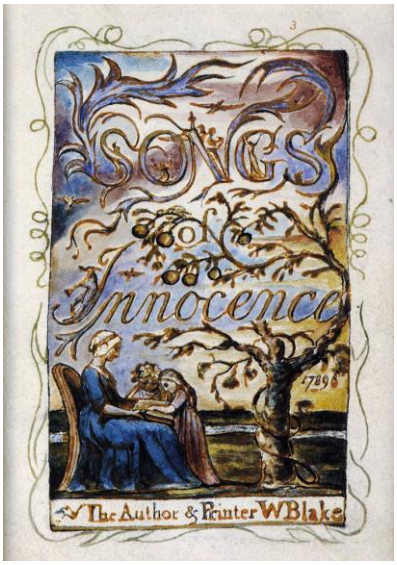


Figure 9 – Blake's illustration for the title-page of *SoI* [26]

4.1.9 Discussing the corpus in terms of such high-level categories is, of course, purely hypothetical. Yes, 03.06 Education has a semantic density of 0.61 in relation to the *Songs of Innocence*, but which words and which specific definitions contribute to this cannot be discovered from this data alone. What this data can do, however, is to prompt further examination. Can the connection between 03.06 Education and 01.01 The Earth be justified in the text? Why, in a set of tales for children, is the Semantic Density of 03.03 Armed Hostility four times higher than 03.06 Education?

[26] Taken from http://www.wikipaintings.org/en/william-blake/songs-of-innocence-1825#close. Blake's artwork is now in the public domain and as such can be freely distributed.

4.1.10 These questions cannot be answered with treemaps. For this reason, it was necessary to find an alternative way of displaying Semantic Density connections. The result was the creation of complex Semantic Networks using Gephi.

4.2 Gephi Results

4.2.1 When examining the networks for the first time, the most salient categories are the upper-level semantic categories, as would be expected. These hold the same proportion as shown in Table 4, with 01 The World as the upper-level category with the highest SD count for both corpora[27]. Below this are the second and third level semantic categories, again in with the same distribution as the treemap data. To maximise visibility, the label threshold for the networks was set to 4, displaying only the labels for the largest nodes when viewing the network zoomed out. By zooming in slightly, more labels appear, and zooming in fully displays the labels for every node. To begin viewing labels past those available through the treemap analysis, it was necessary to zoom in once into the graph. As mentioned before, the Weighted Degree measures the number of weighted edges coming into a node. As a result, the largest nodes belong to the semantic categories, as they have the highest number of weighted edges coming in and out of them. This is not the case for Betweenness Centrality[28], where some of the largest nodes belonged to the lemmas that had the highest number of possible definitions within the *HT-OED*.

4.2.2 For the purpose of this project, the discussion of results had to be limited to showcasing the capabilities of SD analysis. To guide this inquiry, it is possible to return to the questions raised in section *4.1.9*, thus allowing for a cursory analysis of SD potential. The first of these to be attempted returns to 03.06 Education and the connection between the category 01.01 The Earth. To keep this part of the project concise, the following section will concentrate on *SoI.*

4.2.3 Looking first at 03.06 Education in *SoI* (Figure 10), it is possible to see all of the third level categories which fall into Education in the corpus:

[27] See: Appendix 11 and 13 for full networks.
[28] See: Appendix 12.

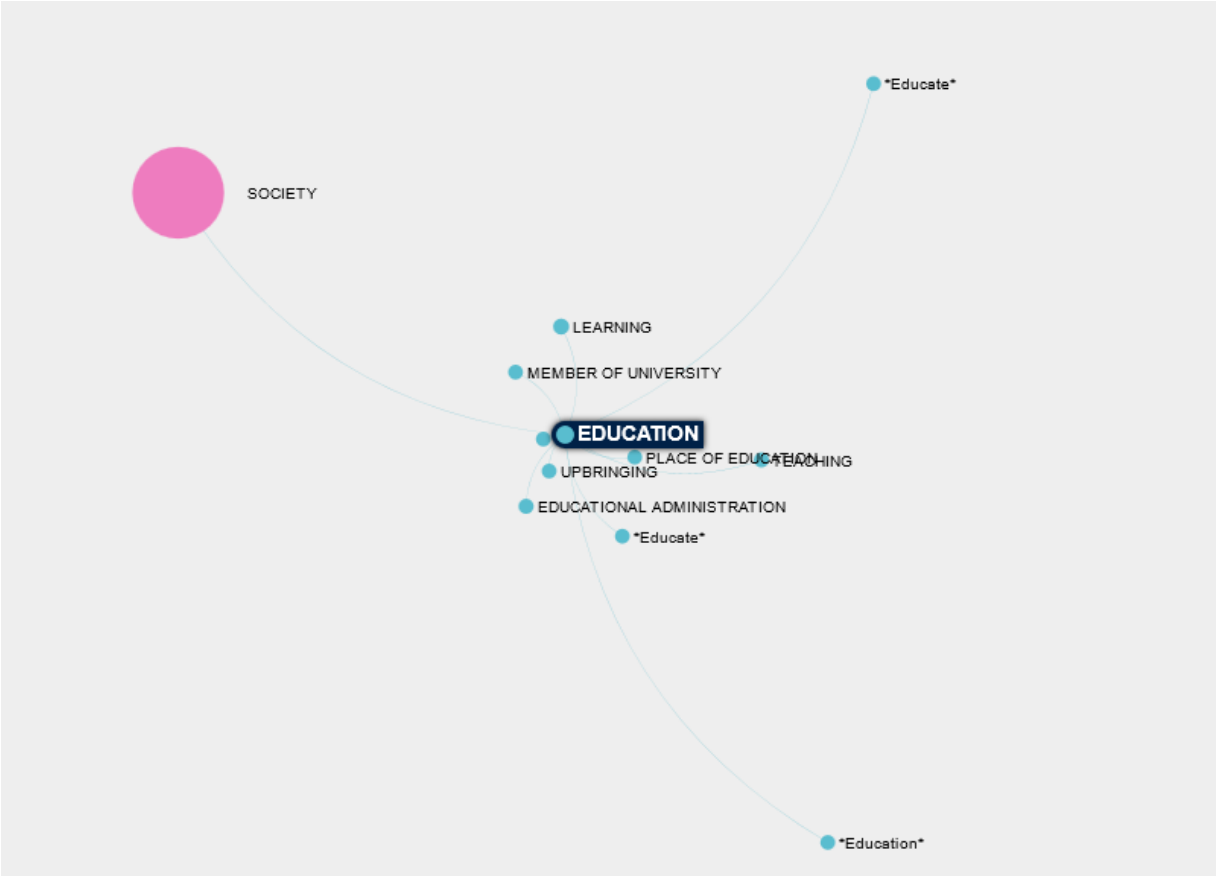


Figure 10 – 03.06 Education in *SoI*.

As visible from Figure 10, the categories which fall into 03.06 Education do not vary significantly in density (with the exception of the upper-level 03 Society). It is possible to dismiss some instances of low semantic density as the result of specific or technical categories, as is the case perhaps with 03.06.04 Member of University in figure 10, into which only three lemmas in the corpus could possibly fall: *father*, *man* and *gown.* More general categories, such as 03.06.01 Upbringing and 03.06.03 Learning have a higher number of edges, but are still relatively small within the network.

4.2.4 A study of category 01.01 The Earth is shown to contain primarily those semantic fields which can be used for natural imagery (Figure 12). Following Gardner's original assessment that nature and education coexist within the poems (Gardner 1986: 16), it would be tempting to attribute the higher SD in 01.01 Earth as a counterbalance to the lower SD in 03.06 Education. This assertion, however, would be difficult to justify without a complementary analysis of texts aimed at a similar audience, and for this reason cannot be determined within the remits of this project.

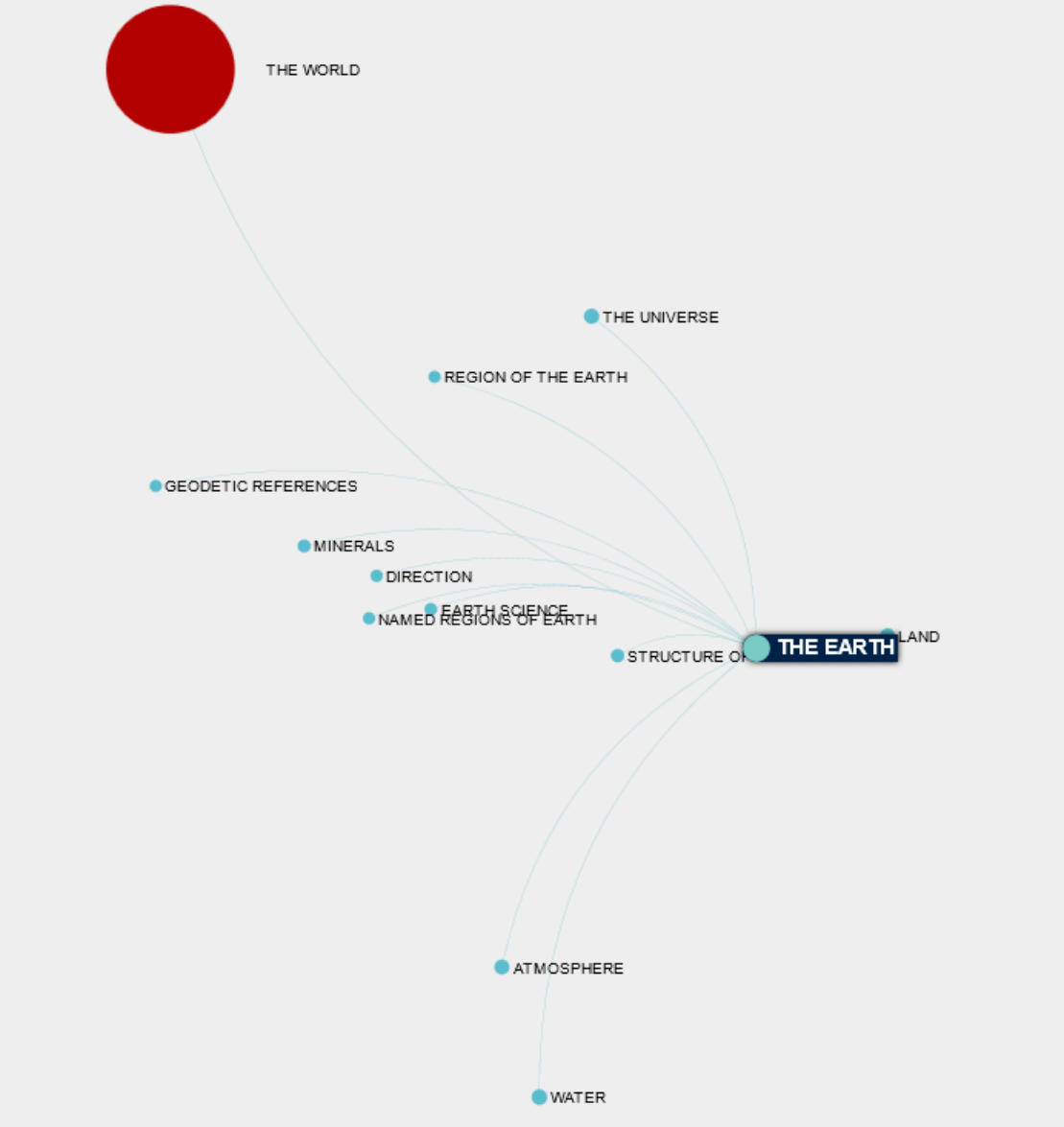


Figure 11 – 01.01 The Earth in *SoI*.

4.2.5 The second question posed during the discussion of the treemap analysis was 'why, in a set of tales for children, is the Semantic Density of 03.03 Armed Hostility four times higher than 03.06 Education?'. This unusual SD proportion cast doubt over the viability of SD analysis and had to be discussed in greater detail to determine the cause for this phenomenon.

4.2.6 The current limitations of the Gephi visualisations, as outlined in section *1.4.4* create an obstacle for using the networks for a closer analysis of the categories. To determine which words fall into the category 03.03, for example, it would be necessary to select each MajHead node

connected to the 03.03 Armed Hostility node, and take a note of each word one at a time. This is a time-consuming process, and frustrates the notion of quickly scanning large texts from a distance. Possible solutions to this issue are discussed in sections *6.2.2.* and *6.2.3*, but could not be implemented for use in this phase of the project.

4.2.7 In order to discuss the high SD count in 03.03 in detail, it was necessary to return to the original Access data to determine the lexical items from the *SoI* and *SoE* corpora that fall into this category:

> *ancient, annoy, arm, armed, arrow, back, ban, bard, bare, barrel, bear, beat, bed, blind, blue, body, bore, break, chain, chase, clasp, close, coat, cock, company, crow, cry, curtain, day, desert, distress, dress, drop, duty, fall, feed, feel, field, filled, fire, flute, foe, foot, form, former, friend, gorge, guard, hammer, head, heavy, horn, iron, jack, join, keep, lamb, lay, left, lie, light, live, lost, lot, maiden, make, man, mark, mine, mouth, move, naked, peace, plain, play, pole, put, raise, receive, red, reed, relief, reply, rest, return, ring, rise, round, run, secret, self, set, shoulder, sight, sit, spear, spoil, spring, stand, star, stray, stroke, sweep, time, turn, two, virgin, weak, well, white, wing*

This is a remarkable number of words to fall into such a specific category. Looking at the *HT-OED* entries for these more closely, the dates indicate that many of the definitions associated with 03.03 were only in use during a short period of time around Blake's lifetime (1757 – 1827). The following list provides some examples of these entries, alongside the word Heading [29]and MajHead:

> *Stand* – (..colours of regiment) – Insignia – 1746-1794
>
> *Lamb* – Soldier of spec. force/unit – Soldier of special force/unit – 1744-1849
>
> *Flute* – (.pistol) – Small-arm – 1842-1842
>
> *Mouth* – (..muzzle) – Parts & fittings of fire-arms – 1587-1802
>
> *Relief* – (..gradual widening of bore) – Parts & fittings of fire-arms – 1824-1858

[29] The number of dots within the brackets of the Heading indicates the sub-category level of the item.

In total, 40 of the definitions are no longer in use, with the last recoded usage shortly before or after the publication of the *Songs of Innocence and of Experience* (1794), and 90 of the possible definitions are still listed as in use in the *HT-OED* database. The proliferation of applying military connotations to certain words could have been a consequence of the time, as stated by Bronowski, 'William Blake lived in the most violent age of English history.' (Bronowski 1954:185). Blake's political awareness could have led to his use of military terminology during a time of social unrest. While it is unlikely that the *Songs of Innocence* are a veiled attempt at rallying the readers to take arms, it is not impossible, considering the struggle with censorship faced by Blake's contemporaries (Bronowski 1954: 187). Concealing his intentions in a collection of poems for children would be a creative way to avoid persecution.

4.2.8 To fully discuss whether the SD analysis truly picked up something unusual about the text, it is necessary to put these definitions into context, to determine whether they could have been used to otherwise conceal a message from the general public, and through the loss of meaning over time, the modern reader. For this purpose, a section of 'The Chimney-Sweeper' from the *Songs of Innocence* collection was chosen, as it had a large proportion of words which could fall into 03.03:

> There's little Tom Dacre, who cried when his head,
> That curled like a lamb's back, was shaved; so I said,
> 'Hush, Tom! never mind it, for, when your head's bare,
> You know that the soot cannot spoil your white hair.'
>
> (Blake, 1789)

Following the process of the methodology, the grammatical words are excluded leaving the lexical items. When lemmatized, they are:

> *little, cry, head, curl, lamb, back, shave, say, hush, mind, bare, know, soot, spoil, white, hair*

Of these, *cry, head, lamb, back, bare,* and *white* can be found in 03.03.

*4.2.*9 As this part of the analysis involved such a small portion of the corpus, it was possible to manually POS-tag the text, so only the corresponding *HT-OED* entries were considered. In doing this, the true extent of how an un-tagged analysis could corrupt the semantic mapping of a text was discovered. The 03.03 Armed Hostility entry for *cry* referred to a noun, but *cry* was used as a verb in the excerpt above, and the entry for *white* was in reference to a noun which does not

apply to the adjective use of the word in the poem.

4.2.10 The issues indicated above suggest that the networks could not be considered fully representative of the text unless a POS-tagged corpus is used for the analysis. These errors were discovered when assessing a minor portion of the corpus, and it follows that the issue would be prevalent throughout the dataset. It would therefore be necessary to resolve this in the methodology for future projects. Some of the entries did correspond to the correct parts of speech, however, and are listed below:

Head – noun – (.head) – Sharp weapon – 1400-Current

Head – noun – (.other parts of carriage) – Gun carriage – 1823-1823

Head – noun – (.head of arrow) – Arrow – 1400-Current

Lamb – noun – Soldier of spec. force/unit – 1744-1849

Back – noun – (.rear) – Part of army by position – 1597-1737

Bare – adjective – (.drawn) – Cut/thrust of sword – 1604-1855

In comparing these entries with the passage, no obvious pattern emerges which would indicate that these definitions could have been intended by Blake. If the semantic networks created in Gephi allowed for a direct edge between the lemma node and the category heading, it could perhaps be possible to discern alternative meanings which would fit with these definitions in the text, but that resource was not available for this project, and manual extraction of this information would be time-consuming. While conducting this type of analysis could prove futile, it would be interesting to determine if a connection can be made between 03.03 Armed Hostility, and the *Songs of Innocence and Experience*, and something which could be explored further in future projects.

4.2.11 The sections above explore some of the ways in which SD analysis could be used for studying literary texts. Despite using only the second and third level categories in Figures 10 and 11, it was possible to discuss the relationship between 01.01 The Earth and 03.06 Education. Understanding the social factors affecting language during Blake's lifetime allowed for a cursory discussion of how SD analysis can be used for exploring these discrepancies. Using the *HT-OED* for dating the words in this way could lead to new interpretations of texts that have been overlooked in modern analyses. The proliferation of entries within 03.03 Armed Hostility during the short period of Blake's lifetime highlighted the potential application of semantic network mapping as a language discovery tool. Of course, the discussion of the results above is suited to a

proof-of-concept rather than an application of SD analysis to literary works. The methodology is still evolving, but the viability of this framework as a powerful analytical tool is increasingly evident.

Chapter 5 - Critical Analysis: 'The Lamb' and 'The Tyger'

5.1 The Poems

5.1.1 One of the issues faced within Chapter 4 was that of visual clarity. Nodes had to be selected individually in Figure 10 and 11 so that they could be discussed. When looking at the full networks, even in the web-browser where it is possible to zoom in and out, the number of nodes poses a challenge in studying the results. One possible solution to this would be to overcome the need for the MajHead node, thereby cutting the number of nodes almost in half. As this was not possible for this project, two smaller networks were created for two individual poems in the *Songs*. These were 'The Lamb' from *Songs of Innocence* and 'The Tyger' from *Songs of Experience*.

5.1.2 These poems were chosen for their distinct quality of corresponding with one another, while being commonly believed to be symbolically opposite. In his critique of the Songs, Alexander Gilchrist (1970: 59) describes 'The Lamb' as a 'sweet hymn of tender infantine sentiment appropriate to that perennial image of meekness; to which the fierce eloquence of 'The Tyger'… is an antitype'. This sentiment is shared by several critics[30], so it was naturally interesting to see if SD mapping could highlight the same themes raised by these critics.

5.1.3 It was already established how SD analysis of notional categories can inform general assumptions about the semantic properties of a text. Having scrutinised the data in the quantitative analysis, it was now necessary to test these hypotheses through a closer look at the text. The purpose of this is to determine if through the use of electronic text analysis and the *HT-OED,* it would be possible to draw any conclusions at level of the poem, in addition to the corpus as a whole. Additionally, this section will establish the way in which SD analysis can be used alongside traditional literary criticism of the text.

5.2 The Analysis

5.2.1 Instead of splitting the poems up over two sections, this analysis will discuss both of them side-by-side. The networks used for informing this analysis can be found in Appendices 14 and

[30] See: Bronowski (1954); Gardner (1987), Leader (1981).

15. Focusing at first on the similarities in SD between the two poems, two tables were created displaying the SD distribution for the corpora (Appendix 5 and 6). Although this information is visible in the networks seen in Screenshots 4 and 5, without being able to navigate the network in the browser it is difficult to discern the details of SD distribution.

5.2.2 The categories with the highest SD in both 'The Lamb' and 'The Tyger' are 01.05.05 Action/Operation and 01.02.08 Food and Drink. The first of these is not surprising, as it contains most standard action verbs and would be found as having a high SD in relation to most texts that describe something that is happening. 01.02.08 Food and Drink, on the other hand, poses a slight problem in relation to the texts. As this data is based on a Weighted Degree analysis, 01.02.08 Food and Drink becomes inflated due to the variety of MajHeads each term can have in relation to the lemma, and the large number of semantic fields that fall into this category. One possible way to overcome this in the future would be to adapt the Betweenness Centrality algorithm to function with the SD analysis. This could raise the categories which are more appropriate to the text as a whole above those that have a high SD because of the large number of subcategories. Whether this would have the desired effect cannot be determined at this stage, because the functions are not compatible.

5.2.3 The SD relationship between 01.02.08 Food and Drink and the poems is not completely baseless. 'The Lamb' depicts the acts of care and nurture, within which 01.02.08 Food and Drink play a key role. 'The Tyger' contains more visceral imagery, and the devouring nature of fire, which uses terminology from 01.02.08 Food and Drink. As a result, even with the potential for SD inflation, the network does not fail to pick up on these themes, rather it relates them to their literal counterparts as well as the figurative ones. In further projects working with SD, it would be interesting to discuss the relationship between reader interpretation of figurative and literal meanings, and how this effects the effectiveness of SD analysis. For now, however, the high density of 01.05.05 Action/Operation and 01.02.08 Food and Drink has proven favourable to the analysis.

5.2.4 Looking again at the tables in Appendix 5 and 6 (and corresponding Screenshots 4 and 5), the top 10 categories with the highest SD count for each poem will be discussed (Table 5):

‘The Lamb’		‘The Tyger’	
Category	**SD**	**Category**	**SD**
01.05.05	123	01.05.05	119
01.02.08	81	01.02.08	114
03.11.03	64	01.05.07	101
02.08.03	**58**	**01.02.05**	**96**
03.10.13	**53**	**01.05.08**	**88**
02.01.15	52	**01.02.06**	**77**
01.05.07	34	**02.02.30**	**68**
03.09.00	**33**	03.11.03	65
02.07.12	**32**	**01.04.03**	**64**
03.11.01	**30**	02.01.15	57

Table 5 – Top 10 categories with the highest SD for ‘The Lamb’ and ‘The Tyger’

In table 5 above, the categories which appear only in one poem or the other have been highlighted. For ‘The Lamb’, these categories were: 02.08.03 Speech/act of speaking, 03.10.13 Trade and commerce, 03.09.00 Travel/travelling, 02.07.12 Giving, and 03.11.01 Amusement/entertainment. For ‘The Tyger’, these were: 01.02.05 The Body, 01.05.08 Movement, 01.02.06 Animals, 02.02.30 Fear, and 01.04.03 Properties of materials.

5.2.5 While it is possible to discuss the merits of each one of these categories in relation to the text, the limits imposed on this analysis predicate a more focused approach. Of particular interest here are the categories 02.07.12 Giving and 02.02.30 Fear. The way in which they outline the contrary positions of the poems complements Damon’s stance on their symbolic nature: ‘Blake used the Lamb as a symbol of Innocence (*SoI*, “The Lamb”), and of God’s love, as contrasted with his wrath (*SoE*, “The Tyger”)’ (Damon 1988: 232). He further qualified this in a later text by stating that ‘the problem of *The Tyger* is, quite simply, how to reconcile the Forgiveness of Sins (the Lamb) with the Punishment of Sins (the Tyger)’ (Damon 1924: 277). These notions of ‘innocence’ and ‘wrath’ as well as ‘forgiveness’ and ‘punishment’ are emulated in the distribution of SD for 02.07.12 Giving and 02.02.30 Fear. Here it is possible to see definitive evidence that SD analysis is capable of highlighting dominant themes within a text, and can be suitable for a comparative analysis of two separate texts.

5.2.6 One further aspect of SD analysis in relation to Blake's poetry must be considered before concluding this brief analysis. That is the question of the 'Tyger'. Although the *HT-OED* is capable of coping with old-English words, it is not designed to work with non-standard spellings. *Tyger*, of course, is not a word in the *HT-OED*, and had to be changed to *tiger* for the SD analysis. As Damon, however, (1988: 413) makes clear in *A Blake Dictionary*, it is not without specific meaning or connotation for Blake:

> 'The TYGER is Wrath (MHH 9:5; FZ ii: 35[31]). He is the Wrath of the Heart, for his position is East. He is one of a quaternary (MHH 8:14; J 98:43): Lion (north), Tyger (east), Horse (south), and Elephant (west). He is the fallen Luvah, when Love has turned to Hate; he is Orc (revolution).
>
> "Wrath is a fire," wrote Spenser (FQ II. Iv.35), utilising a common symbol; and fire in Blake is often associated with the Tyger. But though at least once he is "burning bright" (SoE, "The Tyger" 1) …'
>
> (Damon 1988: 413)

Can such a complete understanding of the *tyger* be achieved by changing the spelling to *tiger* for use with the *HT-OED*? The answer to this question is no, as the *HT-OED* is not designed to be used for a study of an author's individual symbolism. Though one could envisage how a concordance of Blake's entire works could provide some of the connections made by Damon (perhaps it could even clarify exactly how many times the *tyger* is associated with fire), but to develop the intertextural connections with Spencer's work is still a task only made possible by manual analysis and interpretation. It would not be possible to mimic these findings electronically even if a corpus large enough to span several literary periods was created, as it would be necessary to manually determine semantic compatibility. SD analysis cannot replace the decades Damon devoted to his analysis of Blake's imagery, which resulted in two published works on this subject alone (1988; 1924). The *HT-OED* cannot answer for intertextuality or pastiche, because it is not a literary dictionary, but rather the companion piece to the English language. It can, therefore, inform the reader about the language that the author used, and alongside a study of the author's artistic licence, help to construct a sensible opinion of the author's style.

5.2.7 The techniques and issues outlined in this chapter serve as an assessment of the methodology. The conclusions formed as part of this analysis are not intended to inform the reader about Blake, but rather show how SD analysis could be used for critical discussion of an

[31]The abbreviations here are Damon's own. MHH stands for 'The Marriage of Heaven and Hell', FZ refers to 'The Gour Zoas' and FQ to the 'Faerie Queene'.

author's work. Significant work on the methodology would be necessary before SD mapping can be used in a legitimate analysis of authorial style. This, however, marks a starting point for future SD mapping projects.

Chapter 6 – Discoveries, Limitations, Future Research and Conclusion

6.1 Discoveries

6.1.2 By working through the SD mapping methodology, it was possible to begin forming conclusions about the semantic properties of Blake's *Songs of Innocence and Experience.* This proof-of-concept made it clear that there were different themes present in both collections, but also highlighted the similarities within the corpora that unite the collections. The validity of these results has been discussed throughout, with questions being raised as to the likelihood of a high semantic density with 01.01 The World being a result of the ubiquity of lexical items that fall within it. Despite concluding that a high SD within this field is unavoidable, the extent to which an author's text makes use of the establishing themes within 01.01 The World could help to distinguish their personal style.

6.1.3 Discussing the categories with the least SD in relation to the text, as highlighted by the treemaps, spurred on an examination of the relationship between semantic categories within the text. It was determined that to compare SD counts for the semantic categories of the *HT-OED*, it was necessary to have access to a reference corpus to fully express their relationship within the text. While it was possible to formulate assumptions about the relationships, for example the 'coexistence' of 03.06 Education and 01.01 The Earth, as suggested by Gardner (1986), the low SD of 03.03 could instead be the result of the category size and technical content. It would be a priority for future studies to incorporate a reference corpus for this purpose.

6.1.4 By analysing the high density of 03.03 Armed Hostility, an interesting discovery was made about the English language during Blake's time. Attempts to relate the *HT-OED* entries to the text, however, proved to be a challenge without extending the capabilities of the networks themselves. The limitations of this proof-of-concept prohibited a conclusive analysis of this relationship, but suggested another use for SD mapping in literary texts, prompting a discussion of unusual patterns within the categories themselves.

6.1.5 The data used for this project could only take the analysis so far, and future developments could alter the way in which SD mapping promotes discoveries in reading literary texts. The current framework has already proven capable of encouraging discussion of authorial style. Additionally, in Chapter 5, the differences between the top 10 categories with the highest SD value had proven complementary to existing criticism of the poet's work. This suggests a

legitimate future for SD mapping in critical analysis.

6.2 Limitations

6.2.1 As outlined in section 3.3, this project was faced with a number of challenges. In particular, the lack of parsing and tagging that the data went through served to undermine any conclusions that the analysis could project. In future projects handling SD mapping, this issue would have to be resolved as a point of key importance.

6.2.2 The inability to display multiple edges through Gephi posed a further issue, in this case with the intelligibility of the networks. One possible solution that could be explored for this, before multiple edge networks become possible, is the implementation of an arbitrary connection between the lemma node and the semantic category node, in addition to those formed through the MajHead node. This would display the lemma node when the category node is highlighted. The edge weight would have to be offset to prevent inconsistency with the node centrality, but this potential solution demands further investigation.

6.2.3 One further solution to the visibility issue could be found in networking software that supports loosely connected edges. Selecting a node in this network would highlight not just the directly connected nodes, but also all of the nodes connected to them. As of writing this thesis, no open source software was found that supported this functionality, but the concept is frequently discussed in online communities on digital data representation.

6.2.4 Another limitation, resulting from the constrained word count of this thesis, was the decision to use the work of only one poet for SD analysis. Without contrasting data from another source, if is not possible to measure the true weight of the semantic networks beyond the conceptual stage. It would be necessary to expand the scope of this analysis in future trials to fully explore SD mapping capabilities.

6.3 Future Research

6.3.1 Once the framework for a synchronic analysis has been fully established, it would be possible to expand the methodology into a study of diachronic variation in the distribution of SD across literary texts. A large scale study of a corpus that encompasses multiple authors and more than one literary period would be useful in highlighting trends that are unique to a specific period.

6.3.2 One possible use for SD mapping, once the methodology becomes complete, would be to make the networks available online for readers from a variety of backgrounds. Automating aspects of the process could allow for users to input their own texts to be SD mapped. As this project relies on copyrighted data from the *HT-OED*, necessary permissions would have to be obtained before this becomes possible.

6.3.3 Another concern that could be addressed in future projects, as mentioned briefly in earlier chapters, is that of cognitive associations formed by readers, and their relationship with SD. There is already some precedent for this study. Van Atteveldt raised some interesting arguments for grounding Semantic Network Analysis in cognitive psychology (Van Atteveldt 2008: 70). Working from the 'associative network model of human memory' as outlined by Collins and Quillian (1969), Van Atteveldt extrapolated the theory that 'the concepts in semantic memory are represented as nodes in a complex hierarchical network [within which] each concept […] is directly related to other concepts' (Van Atteveldt 2008: 70). As mentioned briefly in section 1.7 Roadmap, the cognitive concerns for the viability of Semantic Network Analysis, and by the same token SD Analysis were too broad to approach within the confines of this project. However, it is encouraging that similar concerns were shared by academics working on SNA projects, and that these proved favourable to the concept.

6.4 Conclusion

6.4.1 Though it was possible to determine several ways in which SD data could serve as a trigger for an informed notion of the semantic properties across a body of the poet's work, it would be necessary to resolve a number of outstanding issues with the methodology before continuing with SD mapping. Once this framework is established, it would be possible to utilise the data in literary critical analysis. Two key uses of this data were highlighted in the pilot study. The first allows for an investigation of the author's style, as highlighted by the *HT-OED* semantic fields which most frequently feature the highest density in relation to their work, and could be used in a comparative study of several texts, authors or literary periods.

6.4.2 The second relates to the use of SD mapping in addition to traditional literary criticism. By combining these two approaches, it is possible to construct a discussion of the text that has a visual reference point. It is also possible that SD mapping could be used to as a prompt to literary criticism, with focus on specific themes being derived from SD distribution.

6.4.3 The ultimate goal of SD mapping is to provide a user-friendly visual representation of an author's work, a single literary piece or the collected works from a specific period. The methodology for this analysis is still being developed, but the results obtained through this trial indicate that Semantic Density mapping could be used to identify an author's literary fingerprint. This project was an attempt to showcase SD approaches, as well as recent developments in digital tools which can be used for the analysis of literary texts. Continuing advancements within this field point to an exciting future for digital approaches to literary criticism, within which SD mapping will hope to find its place.

Word count: 17,532.

Appendices

Appendix 1 - Excerpt from a *SoE* edge file for categories 01.01 - 01.02.11[32].

Source;Target;Type;Label;Id;Weight
01.01;01;Undirected;;60001;11.4
01.01.01;01.01;Undirected;;60002;0.02
01.01.02;01.01;Undirected;;60003;0
01.01.03;01.01;Undirected;;60004;0.07
01.01.04;01.01;Undirected;;60005;1.91
01.01.05;01.01;Undirected;;60006;2.63
01.01.06;01.01;Undirected;;60007;0.18
01.01.07;01.01;Undirected;;60008;0.990000000000001
01.01.08;01.01;Undirected;;60009;0.64
01.01.09;01.01;Undirected;;60010;0.5
01.01.10;01.01;Undirected;;60011;2.43
01.01.11;01.01;Undirected;;60012;2.03
01.02;01;Undirected;;60013;59.4399999999997
01.02.00;01.02;Undirected;;60014;2.19
01.02.01;01.02;Undirected;;60015;4.91
01.02.02;01.02;Undirected;;60016;0.8
01.02.03;01.02;Undirected;;60017;1.33
01.02.04;01.02;Undirected;;60018;4.79
01.02.05;01.02;Undirected;;60019;6.13
01.02.06;01.02;Undirected;;60020;9.01000000000001
01.02.07;01.02;Undirected;;60021;5.88
01.02.08;01.02;Undirected;;60022;17.2099999999996
01.02.09;01.02;Undirected;;60023;2.29
01.02.10;01.02;Undirected;;60024;3.59
01.02.11;01.02;Undirected;;60025;1.31

[32] Complete version can be seen in Appendix 8.

Appendix 2 - Full list of data used for Treemap diagrams.

Category	**SoI**	**SoE**
THE WORLD - 01	66.97	64.17
THE MIND - 02	23.33	23.54
SOCIETY - 03	40.19	34.32
THE EARTH - 01.01	5.06	4.26
LIFE - 01.02	23.3	21.34
PHYSICAL SENSIBILITY - 01.03	5.5	4.11
MATTER - 01.04	6.51	5.92
EXISTENCE IN TIME AND SPACE - 01.05	25.86	23.52
RELATIVE PROPERTIES - 01.06	4.59	4.19
THE SUPERNATURAL - 01.07	0.71	0.83
MENTAL CAPACITY - 02.01	8.44	9.96
EMOTION - 02.02	7	6.91
PHILOSOPHY - 02.03	0.15	0.2
AESTHETICS - 02.04	1.12	1.05
WILL/FACULTY OF WILL - 02.05	1.39	1.39
REFUSAL/DENIAL - 02.06	0.03	0.02
HAVING/POSSESSION - 02.07	2.01	1.44
LANGUAGE - 02.08	3.19	2.57
SOCIETY/THE COMMUNITY - 03.01	2.01	2.09
INHABITING/DWELLING - 03.02	1.85	1.47
ARMED HOSTILITY - 03.03	2.47	2.32
AUTHORITY - 03.04	3.7	3.25
MORALITY - 03.05	1.11	1.09
EDUCATION - 03.06	0.61	0.53
FAITH - 03.07	1.73	2.05
COMMUNICATION - 03.08	3.89	3.05
TRAVEL/TRAVELLING - 03.09	4.81	4.05
OCCUPATION/WORK - 03.10	7.28	6.33
LEISURE - 03.11	10.73	8.09
COUNT	265.54	244.06

Appendix 5 – ‘The Lamb’ SD distribution

‘The Lamb’							
Category	**SD**	**Category**	**SD**	**Category**	**SD**	**Category**	**SD**
01.05.05	123	03.11.04	13	03.08.07	6	01.03.04	2
01.02.08	81	01.02.03	12	01.06.01	5	02.02.20	2
03.11.03	64	02.01.18	12	02.01.04	5	02.02.27	2
02.08.03	58	03.07.03	12	02.01.13	5	02.02.30	2
03.10.13	53	03.09.04	12	02.02.25	5	02.02.31	2
02.01.15	52	01.02.01	11	02.03.23	5	02.07.	2
01.05.07	34	02.01.17	11	03.04.06	5	03.03.16	2
03.09.00	33	03.05.01	11	03.06.05	5	03.04.05	2
02.07.12	32	01.04.07	10	01.03.05	4	03.05.04	2
03.11.01	30	01.04.08	10	01.03.06	4	03.10.	2
01.02.06	29	02.01.07	10	01.04.09	4	03.10.09	2
01.03.08	29	02.08.04	10	02.01.06	4	03.11.05	2
01.05.02	29	03.01.01	10	02.01.10	4	01.02.02	1
02.01.12	29	03.04.09	10	02.02.05	4	01.02.09	1
02.02.22	27	03.10.12	10	02.05.06	4	01.02.11	1
01.05.03	24	02.01.14	9	02.07.05	4	01.03.01	1
02.02.19	24	02.05.04	9	02.07.14	4	01.04.03	1
01.05.06	23	01.06.07	8	03.03.07	4	02.02.03	1
01.05.08	23	02.05.01	8	03.03.09	4	02.02.17	1
01.04.04	22	02.07.08	8	03.04.12	4	02.02.21	1
01.06.02	21	03.03.15	8	03.08.02	4	02.04.01	1
02.08.06	21	03.03.17	8	03.09.03	4	02.05.02	1
01.01.05	20	01.01.04	7	03.10.04	4	02.05.05	1
02.01.16	20	01.01.11	7	03.10.11	4	03.01.04	1
01.02.00	19	01.02.10	7	01.03.02	3	03.01.07	1
03.08.05	19	01.03.00	7	01.04.06	3	03.03.01	1
01.02.07	18	02.01.08	7	01.04.10	3	03.03.02	1
01.06.04	17	03.01.06	7	01.07.04	3	03.03.06	1
03.04.13	17	03.04.07	7	02.01.09	3	03.04.	1
01.06.06	16	03.06.03	7	02.02.18	3	03.04.03	1
03.07.02	16	03.11.02	7	02.04.05	3	03.04.08	1
01.05.01	15	01.02.05	6	02.07.11	3	03.05.	1
02.08.05	15	01.04.05	6	02.08.07	3	03.06.01	1
03.02.07	15	01.06.05	6	03.04.01	3	03.08.01	1
03.05.05	15	02.02.08	6	03.04.10	3	03.09.05	1
03.08.04	15	02.02.29	6	03.08.14	3	03.10.07	1
03.10.06	15	02.07.13	6	01.01.08	2		
01.02.04	14	03.07.00	6	01.01.10	2		
03.04.02	14	03.08.06	6	01.03.03	2		

Appendix 6 – ‘The Tyger’ SD distribution

‘The Tyger’									
Category	**SD**	**Category**	**SD**	**Category**	**SD**	**Category**	**SD**	**Category**	**SD**
01.05.05	119	01.03.08	15	03.09.05	7	02.02.29	3	03.05.01	1
01.02.08	114	02.02.22	15	02.08.06	7	02.01.13	3	02.07.06	1
01.05.07	101	03.10.12	15	02.07.12	6	03.07.01	3	02.02.16	1
01.02.05	96	03.04.13	14	02.02.14	6	03.03.07	3	02.02.28	1
01.05.08	88	03.08.04	14	02.02.21	6	03.01.07	3	03.11.02	1
01.02.06	77	02.01.08	14	02.08.07	6	03.03.12	3	03.03.18	1
02.02.30	68	01.02.09	14	02.01.04	6	03.03.06	3	01.04.10	1
03.11.03	65	01.03.03	13	03.04.06	6	01.02.11	3	03.03.01	1
01.04.03	64	02.08.03	13	02.01.16	6	02.02.03	3	03.10.10	1
02.01.15	57	03.01.01	13	02.01.06	6	01.03.02	2	03.03	1
01.02.01	54	03.05.05	13	03.07.02	6	03.02.03	2	01.05.04	1
01.02.04	49	03.09.01	12	03.10.04	6	03.03.13	2	03.01.06	1
01.05.06	49	02.02.31	12	01.03.04	5	03.06.01	2	03.08.03	1
03.11.04	43	01.01.05	12	03.08.05	5	03.08.12	2		
03.10.06	42	01.06.03	12	02.01.10	5	02.02	2		
01.05.02	39	01.07.04	12	01.01.07	5	03.05	2		
03.10.11	39	02.01.07	11	01.04.05	5	02.03.23	2		
03.09.04	38	01.01.04	11	03.10.01	5	02.07.05	2		
03.11.01	32	01.02.07	10	01.04.07	5	03.04.02	2		
03.09.00	29	01.06.01	10	03.06.03	5	03.04.07	2		
03.10.13	27	03.03.17	10	02.07.11	4	03.06.05	2		
01.04.08	27	03.09.02	10	02.08.05	4	03.08.06	2		
01.05.03	24	02.01.18	9	03.04	4	03.02.02	2		
02.01.17	24	03.04.12	9	03.07.00	4	01.03.01	2		
03.03.16	24	02.04.05	9	02.05.06	4	03.08.14	2		
02.02.19	23	01.02.00	9	03.10.08	4	01.04.02	2		
02.01.14	19	03.10.09	8	02.03.27	4	01.01.11	2		
01.03.07	19	02.04.07	8	02.07.14	4	03.03.05	2		
01.04.09	18	01.02.02	8	03.06.02	4	03.03.15	2		
01.06.04	18	02.02.17	8	02.02.05	4	01.04.04	2		
01.06.06	17	01.01.10	8	02.02.08	4	03.10	1		
01.02.10	17	02.05.04	8	02.02.12	4	03.08.11	1		
03.02.07	17	01.05.01	8	01.02.03	4	01.01.09	1		
02.01.12	17	01.06.02	8	02.07.08	4	03.06.06	1		
01.06.05	15	02.07.13	8	03.03.09	4	03.08.10	1		
03.01.04	15	03.08.07	7	02.05.03	4	01.03.00	1		
03.04.09	15	01.07.03	7	03.03.03	4	03.08.13	1		
02.02.15	15	03.08.08	7	03.11.05	4	02.01.02	1		
01.06.07	15	02.02.20	7	03.09.03	4	03.04.03	1		

List of Appendices on attached CD:

Appendix 7 - Songs of Innocence Nodes and Edges

Appendix 8 - Songs of Experience Nodes and Edges

Appendix 9 - ‘The Lamb’ Nodes and Edges

Appendix 10 - ‘The Tyger’ Nodes and Edges

Appendix 11 - SoI Weighted Degree Network

Appendix 12 - SoI Betweenness Centrality Network

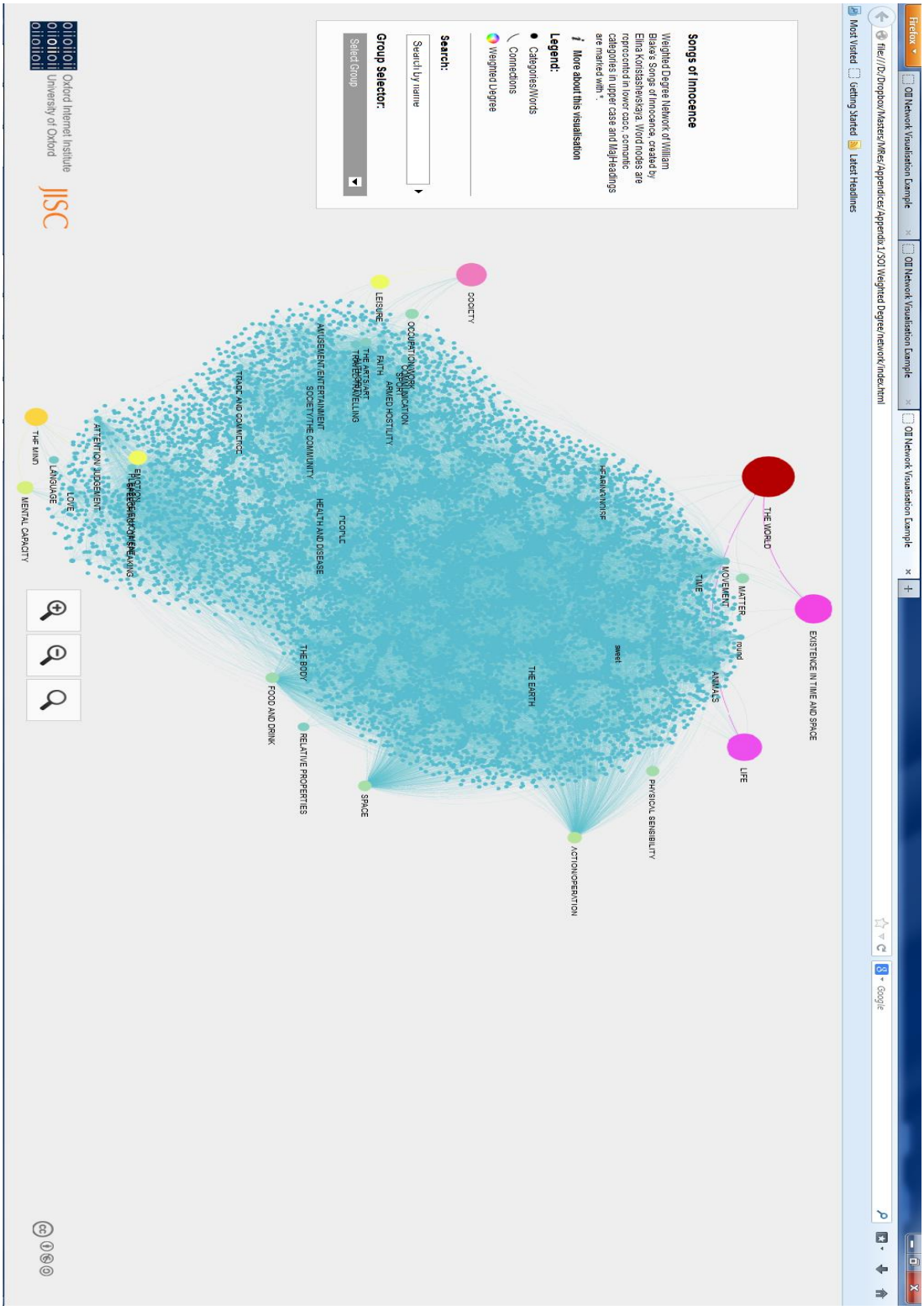
Appendix 13 - SoE Weighted Degree Network

Appendix 14 - ‘The Lamb’ Weighted Degree Network

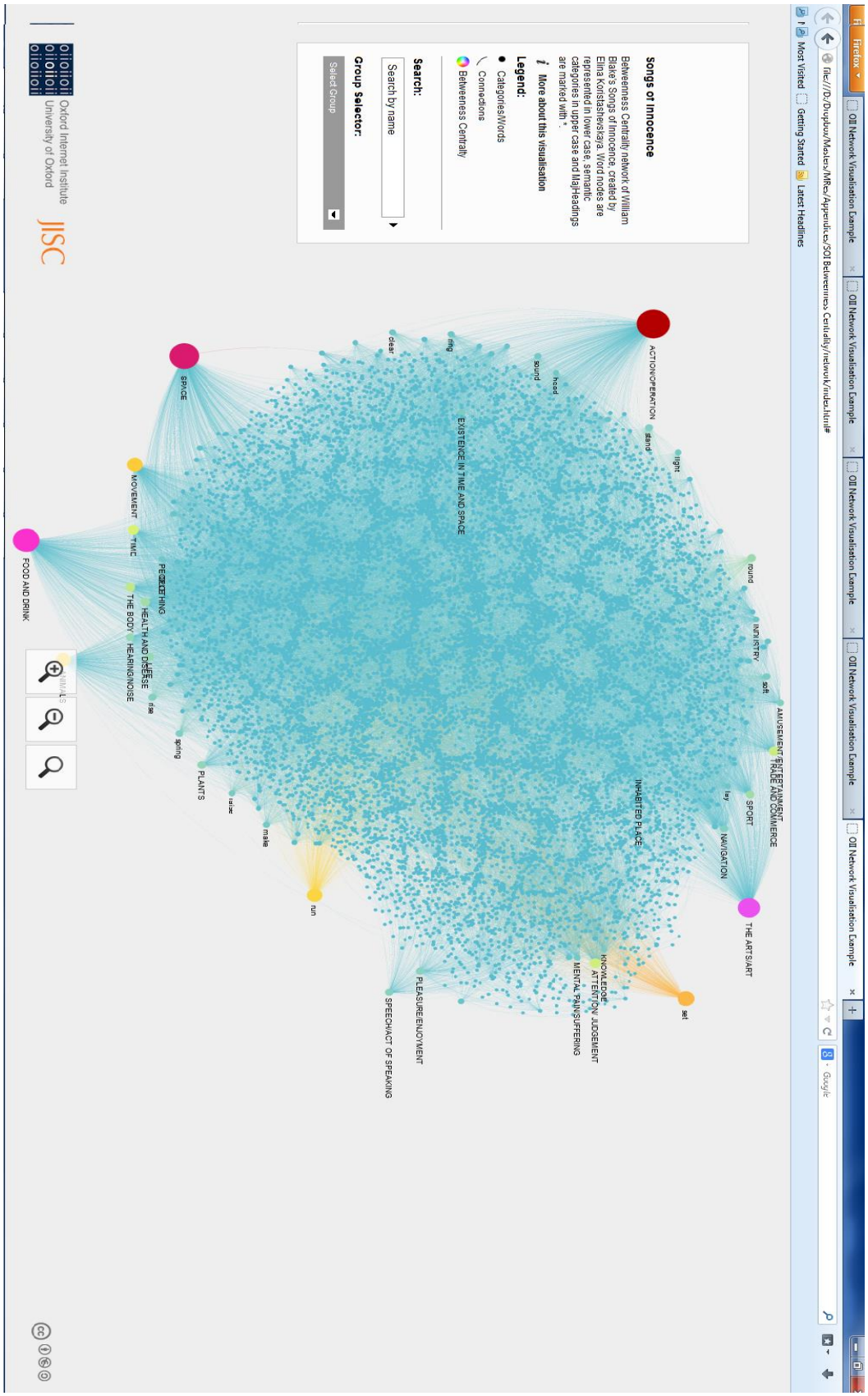
Appendix 15 - ‘The Tyger’ Weighted Degree Network

Screenshots:

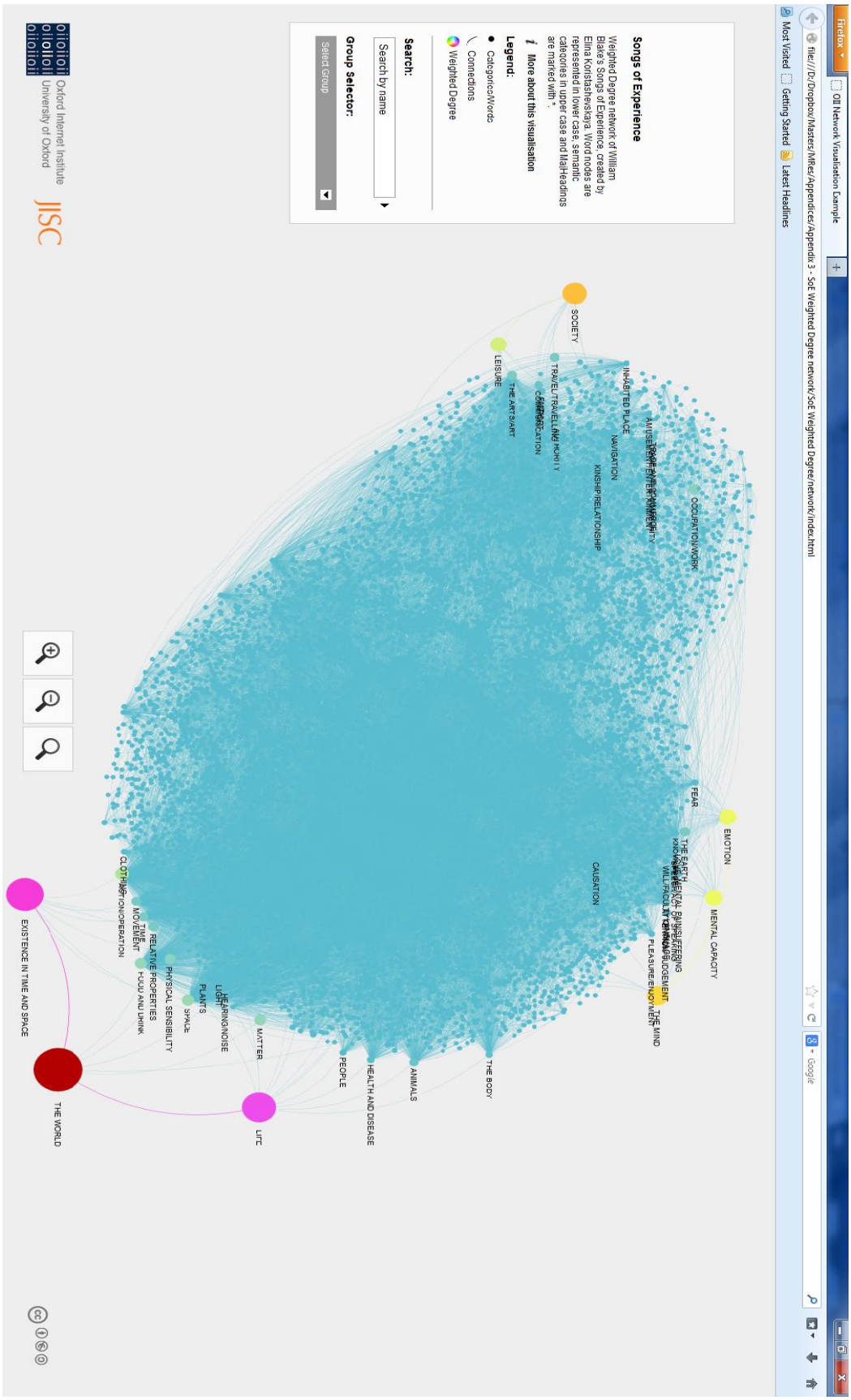
Screenshot 1 – *SoI* Weighted Degree



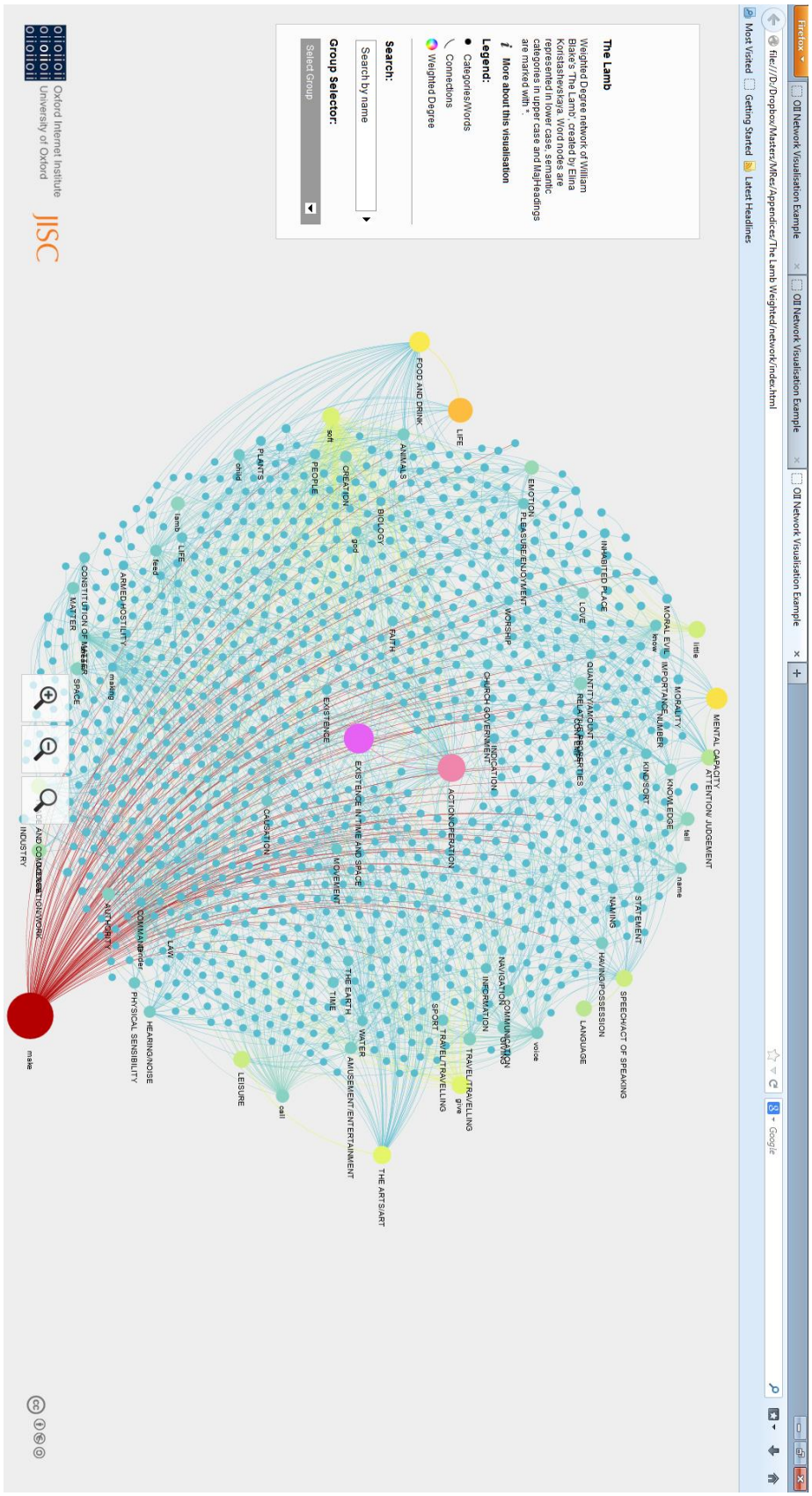
Screenshot 2 – *Sol* Betweenness Centrality



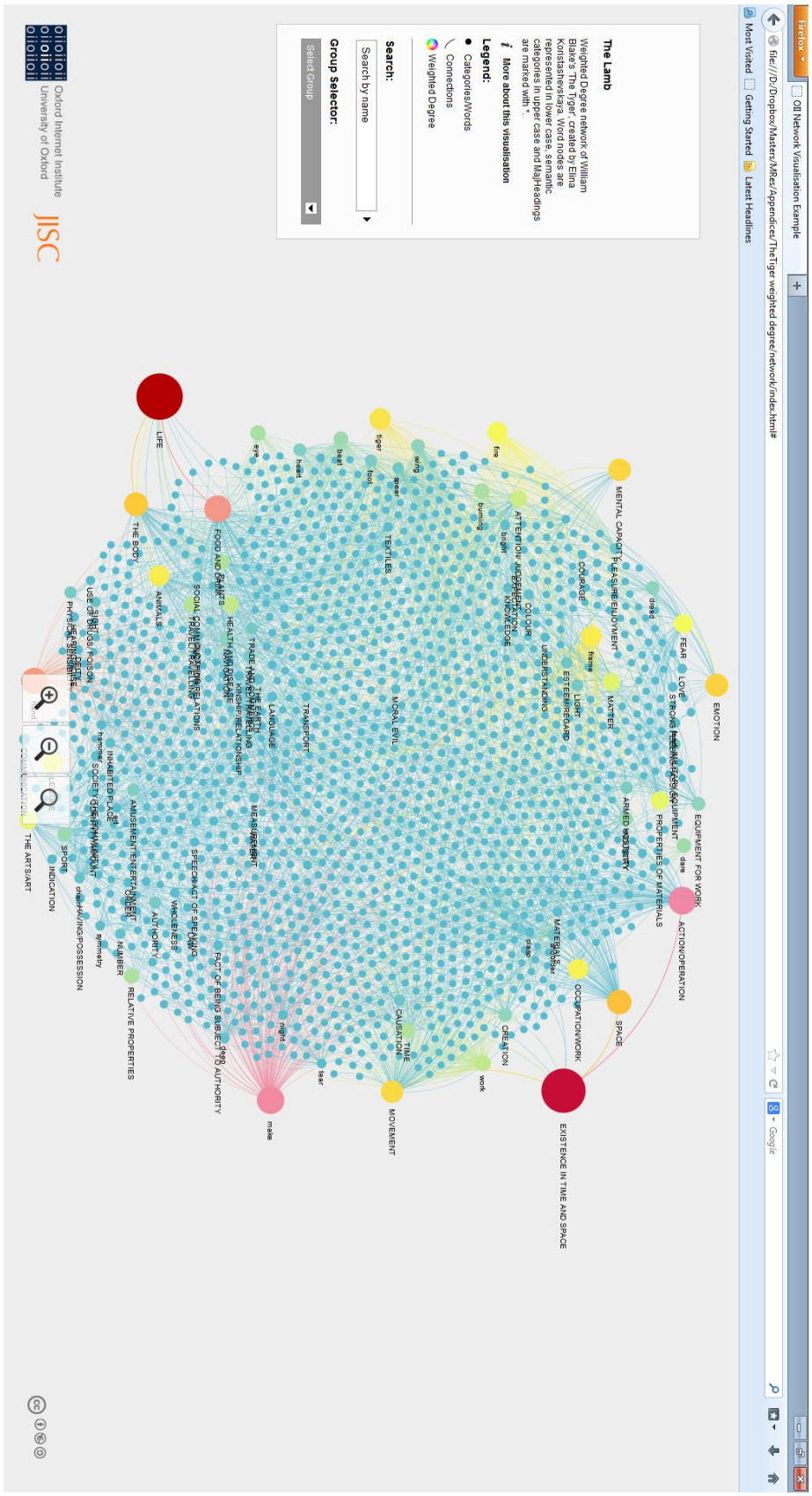
Screenshot 3 - *SoE* Weighted Degree



Screenshot 4 – ‘The Lamb’ Weighted Degree



Screenshot 5 – ‘The Tyger’ Weighted Degree



References

Bibliography

Allison, S., Heuser, R., Jockers, M., Moretti, F. and Witmore, M. (2011). "Quantitative Formalism: an Experiment." (Pamphlet) In: Stanford Literary Lab 1.

Borgatti, S.P. (2005). Centrality and Network Flow. *Social Networks*, 27(1), 55-71.

Bronowski, J. (1954). *William Blake: A Man Without a Mask*. Harmondsworth, Middlesex: Penguin Books Ltd.

Carley, Kathleen M. (1997). Network Text Analysis: The Network Positions of Concepts. In Carl W. Roberts (Ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts* (pp. 79-100). Mahwah, NJ: Lawrence Erlbaum.

Ceri, C. (2007). Liguistic Resources, Development, and Evaluation of Text and Speech Systems. *Text, Speech, Language and Technology*. 37: 221-261.

Collins, A. M. and Quillian, M. R. (1969). Retrieval Time from Semantic Memory. *Journal of Verbal Learning and Verbal Behavior*, 8: 240–248.

Damon, S.F. & Eaves, M. (1988). *A Blake Dictionary: The Ideas and Symbols of William Blake*. New England: Brown University Press.

Damon, S.F. (1924). *William Blake: His Philosophy and Symbols*, London: Constable and Company Ltd.

Dutch, R. A., & Roget, P. M. (1962). *The Original Roget's Thesaurus of English Words and Phrases* (New ed.). New York: St. Martin's Press.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA.

Gardner, S. (1986). *Blake's Innocence and Experience Retraced.* London: Athlone Press.

Gardner, S. (1987). *Some Notes on Blake's Songs of Innocence and of Experience*. Essex: University of Essex Press.

Gilchrist, A. (1970). 'Comments and Critiques 1863-1907'. *In:* Bottrall, M. (ed.) *William Blake: Songs of Innocence and Experience.* Bristol: Macmillan.

Hope, J. and Witmore, (2007), M., "Shakespeare by the Numbers: On the Linguistic Texture of

the Late Plays" in *Early Modern Tragicomedy*, eds. Subha Mukherji and Raphael Lyne. London: Boydell and Brewer, 133-53.

Hope, J. and Witmore, M., (2004): "The Very Large Textual Object: A Prosthetic Reading of Shakespeare," *Early Modern Literary Studies*, 9.3 Jan. 6.1-36.

Kay, C., Roberts, J., Samuels, M. & Wotherspoon, I. (2009). Historical Thesaurus of the Oxford English Dictionary: With additional material from A Thesaurus of Old English, OUP Oxford.

Kay, C., and Alexander, M. (2010) *Life After the Historical Thesaurus of the Oxford English Dictionary*. Dictionaries: Journal of the Dictionary Society of North America, 31 . pp. 107-112.

Kay, C. (2012). 'Current Methods in Historical Semantics'. *In:* Allan, K. & Robinson, J. A. (eds.) *Topics in English Linguistics.* Berlin ; Boston, Mass.: De Gruyter Mouton.

Keynes, G. (ed.) (1969). *Blake: Complete Writings With Variant Readings*, New York: Oxford University Press.

Kirkpatrick, B. (1998). *Roget's Thesaurus of English Words and Phrases*. Penguin, Harmondsworth, England.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Leader, Z. (1981). *Reading Blake's Songs*. London: Routledge.

Miller, G. (1990). *Wordnet: An On-line Lexical Database.* International Journal of Lexicography (Special Issue), 3:235–312.

Moretti, F. (2005). Graphs, Maps, Trees: Abstract Models for a Literary History. London: Verso.

Popping, R. (2000).*Computer-Assisted Text Analysis*. Sage, Newbury Park / London.

Roberts, C. W., editor (1997).Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts. Lawrence Erlbaum, Mahwah, NJ.

Sedelow, S. Y. (1967). *Stylistic analysis*. Santa Monica, CA: SDC.

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. (2004). Trust the Text: Language, Corpus and Discourse. London: Routledge.

Smith, J. (2006). 'Notes on the medical vocabulary of John Keats'. *In:* Caie, G. H., C;

Wotherspoon, I (ed.) *The Power of Words: Essays in Lexicography, Lexicology and Semantics*. New York: Rodopi.

Van Atteveldt, W. (2008). Semantic Network Analysis Techniques for Extracting, Representing,and Querying Media Content. BookSurge Publishers, Charleston SC.

Wasserman, S., and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press.

Yeats, W.B. (1961)[1897]. 'William Blake and the Imagination'. *Essays and Introductions*. London: Macmillan and Co Ltd.

Accessed Online:

Anthony, L. (2011). AntConc (Version 3.2.4) [Computer Software]. Tokyo, Japan: Waseda University. [Online] Available: http://www.antlab.sci.waseda.ac.jp/ [28 August 2013].

Bastian M., Heymann S., Jacomy M. (2009). *Gephi: An Open Source Software for Exploring and Manipulating Networks***.** International AAAI Conference on Weblogs and Social Media. [Online], Avalilable: https://gephi.org/ [28 August 2013].

Bird, S., Loper, E., and Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc. [Online], Available: http://nltk.org/ [28 August 2013].

Blake, W. [1789/1794] (1901). *Songs of Innocence and Songs of Experience*, [Online], Avaliable: http://www.gutenberg.org/files/1934/1934-0.txt [28 August 2013].

Blake, W. (1789). *Songs of Innocence*. Etching. Metropolitan Museum of Art, New York City. [Online], Available: http://www.wikipaintings.org/en/william-blake/songs-of-innocence-1825#close [28 August 2013].

Hale, S. (2013). *Sigmajs Exporter*, [Gephi Plug-in], [Online], Available: https://marketplace.gephi.org/plugin/sigmajs-exporter/ [28 August 2013].

Jacomy, M. (2013). *Noverlap*. [Gephi Plug-in], [Online], Available: https://marketplace.gephi.org/plugin/noverlap/ [28 August 2013].

Martin, S., Brown, W. M., Klavans, R. and Boyack, K. (2011). *OpenOrd: An Open-Source Toolbox for Large Graph Layout*. SPIE Conference on Visualization and Data Analysis (VDA). [Gephi Plug-in], [Online], Available:

https://marketplace.gephi.org/plugin/openord-layout/ [28 August 2013].

Wallis, S. (2007). 'Annotation, Retrieval and Experimentation', in Meurman-Solin, A. & Nurmi, A.A. (ed.) *Annotating Variation and Change*. Helsinki: Varieng, [Online], Available: http://www.helsinki.fi/varieng/series/volumes/01/wallis/ [28 August 2013].