Decoding British Empiricism: A Distant Reading of Locke, Berkeley, and Hume

by

Jason Richard Bradshaw

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Arts
in
Philosophy

Department of Digital Humanities
University of Alberta

Abstract

In a world inundated with information that is becoming increasingly more digital by the day there is value in unique academic disciplines that can make sense of this new landscape—interdisciplinary fields like the Digital Humanities (DH). DH scholars have the ability to breach the divide between the traditional physical objects of study found within academia and the digital artifacts that are quickly becoming present reality. The field can provide a degree of clarity in these uncertain times and has already given insight into many of the texts that circulate throughout the modern world. Textual analysis techniques have allowed researchers to probe the depths of literature and the masses of user generated content from popular social media sites. Large amounts of text to analyze are by no means a difficult thing to come by in the age of information. However, this interdisciplinary field has often neglected an ancient academic practice with a wealth of textual content and a substantial amount of close reading—Philosophy. This alone is grounds for the following research questions: *Why is philosophical source material underrepresented in DH, how can methods be established that yield important results from classical philosophy,* and *what can textual analysis methods reveal about classical philosophical literature?*

Using established textual analysis techniques, this research introduces a sampling of classical philosophical literature to the 21st century. By analyzing the work of three of the founding philosophers of the school of British empirical thought (John Locke, David Hume, and George Berkeley), it is shown that the addition of a distant reading to the discussion can provide a deeper understanding of this source material than a close reading alone. These quantitative text analysis methods can either prove or disprove some of the assumptions made by academics in the field, as well as open up new avenues of research.

This has been accomplished by running the texts through the following three techniques: a Lexical Richness Index (LRI), Topic Modelling, and a Multiple Correspondence Analysis (MCA).

The results speak to the capabilities of using such techniques on philosophical sources, but the method established herein does much more than just that. It shows that philosophy cannot only be treated as a source for distant reading, but also as a framework. Philosophical concepts readily lend themselves to DH analyses, and this is certainly true of the British empiricists. This school of philosophy helped to form the foundations of rigorous, hard scientific investigation, much like how DH distant readers apply computational analysis to their own source materials. Additionally, the analysis takes the shape of a tool for further research into this area. All of the code has been made available in a GitHub repository for future researchers to download and repurpose for their own projects. The British empiricist corpus, as well as the results, are also freely available for those who would continue to pursue how these two disciplines function together.

Dedication

For Debbie, Sherry, and Josh.

The family that I lost much too soon.

Acknowledgements

This work would not have been possible without the generous research grant provided by the Social Sciences and Humanities Research Council (SSHRC). The funding allowed me to continue my graduate education and make the time that I required to examine these sources in depth. I am also indebted to my supervisory committee for the work that they did to make this all possible: Dr. Harvey Quamen, for his positive outlook and unending support; Dr. Ingo Brigandt, for his philosophical perspective and taking the time from sabbatical to attend virtually from Oslo; Dr. Sean Gouglas, for chairing the whole thing and remaining completely optimistic even when I, perhaps, was not; and Dr. Geoffrey Rockwell, for being not only a major source of inspiration for this endeavour and friend, but also my biggest proponent. Finally, I would like to acknowledge all of the work done by Digital Humanities grad supervisor Nicola DiNicola in preparing me for this moment and for what is to come in the future.

Contents

List of Tables

List of Figures

DECODING EMPIRICISM

"If we take in our hand any volume; of divinity or school metaphysics, for instance; let us ask, Does it contain any abstract reasoning concerning quantity or number? No. Does it contain any experimental reasoning concerning matter of fact and existence? No. Commit it then to the flames: for it can contain nothing but sophistry and illusion."
David Hume, *An Enquiry Concerning Human Understanding*, p. 217.

Introduction

According to Genesis, the Tower of Babel was constructed by the remnants of humanity after the Great Flood. These people shared a few things in common. They were all driven to build a tower that could reach to the heavens, were the last of their kind, and spoke a single language. Naturally, the God of the Old Testament did not take kindly to such human hubris: "And the Lord said, Behold, the people is one, and they all have one language; and this they begin to do: and now nothing will be restrained from them, which they have imagined to do."[1] After making this proclamation God swiftly moved in, 'confounded' the speech of this last vestige of humanity and proceeded to scatter them across the land. For millennia, this was the narrative behind the origin of human languages. That is until rational thinkers of the enlightenment began to question the dogma. Most notably, the philosopher John Locke gave a very different account of the formation of language in *An Essay Concerning Human Understanding*. Locke saw words as being nothing more than "pure convention, contrived and shaped by humans through the centuries and differing through time and space [. . .] because of the uniqueness and diversity of human experience itself."[2] Words are simply arbitrary, malleable configurations that humans

[1] Genesis, [11:6].
[2] Edward Gray, *New World Babel: Languages and Nations in Early America,* (New Jersey: Princeton University Press, 2016), 85.

assign to external and internal objects to make sense of the world around them. The Tower of Babel mythos may be a fiction, but it can be argued that humanity currently finds itself in an analogous situation. Information technologies have united much of the modern world and the construction of a new Tower of Babel has begun—with data and information functioning as a shared language. Likewise, the potential for being confounded remains.

Data, text, and media proliferate and flourish in the age of the internet. At times it feels like the possibilities opened up by advancements in information technology are limitless, but this has also resulted in an information landscape that can be overwhelming and confusing. It has driven a need for unique academic disciplines to make sense of the current situation. Interdisciplinary fields like the Digital Humanities (DH). In an era where the lines between digital and analog media become increasingly blurred, there is value in being able to cross the barriers that separate disciplines. DH scholars have the ability to breach the divide between the traditional physical objects of study found within academia and the digital artifacts that are quickly becoming present reality. DH can provide a degree of clarity in these uncertain times. For instance, the field has yielded insight into many of the texts that circulate throughout the modern world. Textual analysis techniques have allowed researchers to probe the depths of literature and the masses of user generated content from popular social media sites. Large amounts of text to analyze are not a difficult thing to come by in the age of information. However, it can be argued that DH is currently experiencing a lack of projects related to an ancient academic practice with a wealth of textual content, a substantial amount of close reading, and a historical precedent within the DH field—philosophy. This alone is grounds for exploring the following research questions: *Why is philosophical source material underrepresented in DH, how can methods be* *established that yield important results from classical philosophy,* and *what can textual analysis methods reveal about classical philosophical literature?* John Locke deconstructed the biblical Tower of Babel with his empirical philosophy and so it is only appropriate that he, and the other British empiricist philosophers, form the corpus from which to launch the investigation. The digital Tower of Babel is capable of uniting the confounded speech of humanity and finding patterns in the unique language of different authors. One can only hope that the God of the Old Testament does not catch on and bring his wrath down upon humankind once again.

Using established textual analysis techniques, this research introduces a sampling of classic philosophical literature to the 21st century. By analyzing the work of three of the founding philosophers of the school of British empiricism (John Locke, George Berkeley, and David Hume) it will be shown that the addition of distant reading, *sensu* Franco Morretti, can provide a deeper understanding of this source material than a close reading alone. These quantitative text analysis methods can either prove or disprove some of the assumptions made by academics in the field, as well as open up new avenues of research. Moreover, the techniques form the basis for a methodological approach to analysing philosophical source material that can be applied to other corpora. This has been accomplished by running the texts through the following three techniques:

1. Lexical Richness Index
2. Topic Modelling
3. Multiple Correspondence Analysis

Before any analysis can begin it is necessary to establish this work within the broader fields of the disciplines it borrows from. Therefore, the frameworks from DH and philosophy that this research utilizes and complements will first be discussed. Following this, the contemporary and historical relationship of philosophy and DH will be explored. Next, the analysis proper will begin with an examination of the corpus, methodology, and techniques to be used. Finally, findings of the research and their associated visualizations will be analyzed in depth and recommendations for future research will be made.

Framework

Distant Reading

Two influential distant reading researchers who have successfully applied a similar mixed methods approach are Franco Moretti and Matthew Jockers.[3] Moretti, an Italian literary scholar currently based out of the Swiss Federal Institute of Technology, sparked interest in using digital tools to analyze the history of literature with the release of his 2005 book, *Graphs, Maps and Trees.* In the examination, Moretti presents a quantitative analysis of the rise of the novel.[4] He defines the concept of distant reading as an abstraction and reduction of a text, but with the gained advantage of delivering "*a specific form of knowledge*. Fewer elements, hence a sharper sense of their overall interconnection."[5] The approach is useful when presented with a large volume of text that could never be read by one individual in a lifetime. It also provides a means of academic inquiry that is distinct from close reading. Through the use of computationally generated line graphs, Geographical Information Systems (GIS), and histograms, Moretti paints a unique picture of historical literature analysis using digital tools. The visualizations add another layer to

[3] It should be noted that *distant reading* and *text analysis* are used interchangeably throughout.

[4] This was to be an expansion upon his earlier work on the connections between literature and space in *Atlas of the European Novel, 1800–1900* (1998).

[5] Franco Moretti, *Graphs, Maps, and Trees: Abstract Models for a Literary History*, (New York: Verso, 2005), 1.

what was most often an interpretative and primarily textual endeavour. This approach provided a different kind of data than literature analysts were accustomed to, one that attempts to be independent of interpretation.[6] Moretti expands upon this concept of distant reading in his 2013 book of the same name, *Distant Reading*. In this follow up he offers a deeper digital analysis of world literature by narrowing down and refining those previous techniques from *Graphs, Maps, and Trees.* Techniques that were typically reserved for the hard sciences.

Another DH researcher worthy of investigation is Matthew Jockers. A colleague of Moretti's, the two founded the Stanford Literary Lab in 2010.[7] This collegiality is apparent in Jockers' book, *Macronalysis,* when he emphasizes where the value of distant reading lies by making direct reference to Moretti's own pursuits:

> Moretti's intent is not to vanquish traditional literary scholarship by employing the howitzer of distant reading, and microanalysis is not a competitor pitted against close reading. Both the theory and the methodology are aimed at the discovery and delivery of evidence. This evidence is different from what is derived through close reading, but it is evidence, important evidence. At times the new evidence will confirm what we have already gathered through anecdotal study [. . .] At other times, the evidence will alter our sense of what we thought we knew. Either way the result is a more accurate picture of our subject.[8]

Moretti's methods and analysis are somewhat hard to locate in his work, while Jocker's technical approach to text analysis clearly presents his process. Indeed, a number of the techniques that have been chosen for this research come directly from suggestions made by Jockers in his book. He evaluates a variety of corpora and documents the text analysis

[6] Moretti, *Graphs, Maps, and Trees*, 9.

[7] See: https://litlab.stanford.edu/.

[8] Matthew Lee Jockers, *Macroanalysis: Digital Methods and Literary History*, (Chicago: University of Illinois Press, 2013), 90.

process from beginning to end, including additional tips on best practices. Jockers walks the reader through his methods and analysis of topic models, LRIs, text classification, MCA, and a multitude of others, complete with links to tools and examples for the would-be researcher to conduct their own analysis.

While both Moretti and Jockers did uncover intriguing results by using distant reading techniques on their respective corpora, this was not the ultimate goal of the projects. They were attempting to show the validity of using quantitative methods in a traditionally interpretative arena. For Moretti, his introduction of models and theoretical analysis to the field allowed for a means to "*concretely change the way we work*: [. . .] they allow us to enlarge the literary field, and re-design it in a better way, replacing the old, useless distinctions (high and low; canon and archive; this or that national literature . . .) with new temporal, spatial, and morphological distinctions."[9] He was injecting something new into a discipline that had long relied on mostly subjective interpretation. Moretti was catching the literary analysis field up with the rest of the increasingly digitized world by introducing verifiable and repeatable text analysis techniques. Jockers was on a similar mission as that of his contemporary. The DH researcher posits that the "humanities computing/digital humanities revolution has now begun, and big data have been a major catalyst. The questions we may now ask were previously inconceivable, and to answer these questions requires a new methodology, a new way of thinking about our object of study."[10] Following this line of thought, not only should digital text analysis techniques be considered valid in literary analysis, they may even be necessitated given the nature of text

[9] Moretti, *Graphs, Maps, and Trees*, 91.
[10] Jockers, *Macroanalysis*, 2-3.

in the 21st century. This leads to three important general findings on distant reading championed by the researchers:

1. *Analysis of literature has reached an oversaturation point. Computational means are now necessitated to carry out research on massive volumes of work.*

This is the point previously made by Jockers. Interpretative close reading of material was the method most commonly employed by researchers for centuries. It was not until very recently that the means have become available to carry out computational analyses on massive corpora. Now that they are it may even be the researcher's responsibility to couple these digital techniques with interpretive analysis. That may seem like a bold statement, but it is something that Moretti himself alludes to when thinking about the large corpora studied in literary analysis, "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."[11] This is not due to any shortcoming on the researcher's part. They are, after all, only human and as a human the processing power required to assimilate the knowledge from thousands upon thousands of documents is not available.

2. *Distant reading stands to uncover information that close reading cannot.*

The human mind, due to the aforementioned cognitive limitations, may miss connections in large corpora that distant reading easily uncovers. A computer operates much differently than the human brain. Not many researchers will sit around counting each and every use of the word 'and' from the collected works of Shakespeare. This is impractical and, from the perspective of a flesh and blood analyst, most likely of minimal

[11] Moretti, *Graphs, Maps, and Trees*, 4.

value. However, the computer can make short work of such tasks and find connections in the mundane details that would have otherwise been overlooked, disregarded or left unconsidered by its human counterpart. By no means are humans inferior to machines in the realm of literature analysis; it is only that their strength lie in a different method—interpretation. As noted by Jockers, "though the computer cannot perfectly replicate human synthesis and intuition, it can take us a long way down this road and certainly quite a bit further along than what the human mind can process."[12] This suggests that neither a close nor distant reading approach are superior on their own, but that they complement each other.

3. *Distant reading is not intended to supplant close reading in academic discourse, but to enhance it.*

Distant reading and the technology it utilizes are not here to replace academics and their traditions (any time soon), but to function as an aid to further research. The techniques *inform* interpretation, they do not provide it. The computer counting up every occurrence of a word is freeing up time for the researcher to ponder the outcome of the analysis. As put by Jockers, "the two scales of analysis work in tandem and inform each other. Human interpretation of the "data," whether it be mined at the macro or micro scale, remains essential."[13] Furthermore, the world is quickly entering an age where the two techniques may become mutually exclusive. The data now available to researchers is massive and close reading on its own is no longer practical. On the other hand, viewing literature only from a distance increases the likelihood of missing something important. In

[12] Jockers, *Macroanalysis*, 34-35.

[13] Ibid., 54.

the words of Jockers, "the two scales of analysis, therefore, should and need to coexist."[14] These points made by Jockers and Moretti are important for any researcher planning to undertake a distant reading project using digital tools, and they will certainly inform the research found in this project. Arrogantly claiming that one method of discovery trumps the other can lead only to faulty conclusions and failure.

As it was with the DH authors thus far examined, the outcome of this research can be considered twofold. The project certainly stands to uncover and confirm some of the realities surrounding the literature, but, as was stated earlier on, philosophical literature and projects are underrepresented in contemporary DH research. This creates a need to first establish the techniques as a reliable source of analysis when it comes to philosophical literature. Providing substantial and intriguing findings from the works of the British empiricists may even be considered a beneficial by-product. Showing that text analysis is an appropriate and accurate means to study philosophy is a major underlying impetus of the project. This does not lessen the fact that given a source such as the written works of the British empiricists, one that has been picked over by philosophers for centuries, that distant reading stands to uncover new information by having the gained advantage of testability against numerous close readings done by experts in the field.

Both Jockers and Moretti have given their own definitions of distant reading in their pioneering DH research. Moretti viewed the distance as a way to derive a very specific form of knowledge that provided data rather than interpretation, while Jockers, who was for the most part in agreement with his colleague, placed an additional emphasis on the need of such techniques in a world that is becoming more digital by the day. It would be beneficial

[14] Jockers, *Macroanalysis*, 16.

at this point to explicitly state what the concept of distant reading, in regards to this undertaking, entails: distant reading is a digital, computational method that can enhance and supplement subjective close reading through the use of textual data and visualizations that can then be analyzed. The outlined framework of the DH distant readers and simple definition of the term will help to guide and ground this research.

British Empiricism

Empiricism is a somewhat nebulous term that can mean different things depending upon who is asked. The word that might first come to mind is *empirical*. There is empirical evidence, empirical data, and empirical laws. The word invokes images of white-coated scientists, beakers, and labs, which is fair given the term's common modern use. However, *empiricism*, specifically of the British kind, is a different entity entirely. The British empiricists chosen for this project, though certainly not the only individuals exploring these concepts, have been lumped together given their similar ideas, significant contributions, and close geographical and temporal proximity.[15] Hume, Locke, and Berkeley were all philosophers of the Enlightenment writing in and around the 18th century. They are also representative of the three nations of the British Isles. Hume was from Scotland, Locke from England, and Berkeley from Ireland.[16] Geography and time period aside, these philosophers also had similar visions on how philosophy should be conducted. As laid out by Margaret Atherton, "Locke, Berkeley, and Hume share a common genetic account of the contents of the understanding: There are no mental contents that

[15] Other empiricists that are not part of the study include: Francis Bacon, Thomas Hobbes, Pierre Gassendi, Robert Boyle, and Isaac Newton (Laurence Carlin, *The Empiricists: A Guide for the Perplexed,* London: Continuum, 2009, 2).

[16] Margaret Atherton, *The Empiricists: Critical Essays On Locke*, Berkeley, and Hume, (Lanham: Rowman & Littlefield Publishers, 1999), vii.

cannot be derived from sensation or reflection. This common approach can reasonably be called Empiricism."[17] An even more succinct account of this uniting idea of the empiricists is provided by Laurence Carlin in the first line of *The Empiricists: A Guide for the Perplexed*, "*Empiricism* is the view that sensory experience is the source of human ideas (or concepts) and/or human knowledge."[18] One can now begin to trace the roots of *empirical* back to *empiricism*. It takes no stretch of the imagination to see Carlin's definition of empiricism reflected in empirical science, a science that is based upon the use of observation to derive findings from repeatable experimentation. These aforementioned similarities have been grounds for grouping Hume, Locke, and Berkeley together under the flag of British empiricism, but this overlooks the distinct contributions made by these philosophers to the field. They should instead be viewed as individuals carrying out their own projects "with motivations that differ significantly one from another", but with a shared "general view in common that can fairly be called 'Empiricist'."[19] With this in mind, a brief examination of these individuals and their unique approaches to empiricism is in order.

John Locke was born in the summer of 1632 in Somerset, England. He began his education at Westminster school before moving on to Christ Church, Oxford University where he was subjected to "a large dose of Scholastic Aristotelianism."[20] Locke, like the other empiricists, was opposed to the innatism laid out centuries before by Aristotle and Plato that was still widely taught during this time period.[21] This opposition found its expression in Locke's oft-quoted concept of the human mind being a *tabula rasa* (blank

[17] Atherton, *The Empiricists: Critical Essays*, viii.

[18] Carlin, *The Empiricists: A Guide*, 1.

[19] Atherton, *The Empiricists: Critical Essays*, vii.

[20] Carlin, *The Empiricists: A Guide*, 78.

[21] Innatism being "the view that the mind comes to the world already endowed with certain ideas, truths, or 'principles'" (Carlin, *The Empiricists: A Guide*, 83).

slate) at birth that is then imprinted with experiences from the external world.[22] For Locke, the experience that fills a blank human mind comes in two varieties: *sensation* and *reflection*. Sensations come to the mind by perceiving the qualities of bodies that are external, while reflections happen internally and provide insight into the happenings of the mind itself.[23]

These two types of experience described by Locke yield *ideas*, which the philosopher spent a great deal of time exploring. Ideas, for Locke, can be either *simple* or *complex.* Simple ideas come into a mind through sensation and reflection, while complex ideas are combinations of several simple ideas. The simple ideas, coming to the mind through experience, are received passively. They are the data that the brain receives from the outside world through the sensory organs. In Locke's own words, "whatsoever is so constituted in Nature, as to be able, by affecting our Senses, to cause any perception in the Mind, doth thereby produce in the Understanding a simple *Idea*."[24] While simple ideas of sensation may be received passively, one can also reflect internally to produce them. Processes such as reasoning, judgement, and remembrance are simple ideas that are derived by the mind observing "its own Actions about those *Ideas* it has."[25] These simple ideas become the building blocks for complex ideas, which come in three different types once simple ideas are combined: complex ideas of *substance* are created by combining sensible qualities of objects in the external world in order to perceive distinct objects; complex ideas of *relation* do not really combine simple ideas, but rather consider them

[22] John Locke, *An Essay Concerning Human Understanding,* ed. Pauline Phemister, (New York: Oxford University Press, 2008), xii.
[23] Carlin, *The Empiricists: A Guide*, 86-87.
[24] Locke, *An Essay Concerning Human Understanding*, 73.
[25] Ibid., 69.

separately and as a means to contrast one thing to another; complex ideas of *abstraction* are created when one takes several simple ideas and abstracts from them a category, or a "mode."[26]

Simple ideas can be related to two distinct types of quality, *primary qualities* and *secondary qualities*. Locke makes a firm distinction between idea and quality. Qualities are the things that exist in bodies in the external world, while ideas are the imperfect representations of those qualities in the human mind.[27] Primary qualities are in bodies no matter what changes may occur to it; things like solidity, extension, figure, motion or rest and number.[28] Secondary qualities, however, may be separable from the bodies that they are attributed to. Things like color, taste, and sound produce ideas in the mind that are, to Locke, mere arrangements of primary qualities that may "cause one to attribute a redness to that body, or a pleasant aroma."[29] This is a familiar sentiment echoed to this day. Who is to say that one person's 'red' is another's 'red'? Color is caused by differing wavelengths of light entering the eye and it is difficult to determine that every human is objectively perceiving the same colors.

It is clear that each of these concepts lead into each other and are dependant upon the prior. By creating a hierarchy of experience, ideas, and qualities, Locke was able to bypass innatism. More than that, he essentially created an epistemological theory that is supported by a scientific approach and "the new philosophy of experimentation."[30] Though not the first or only enlightenment thinker to advocate for what are now considered

[26] Carlin, *The Empiricists: A Guide*, 88.
[27] Ibid., 89.
[28] Locke, *An Essay Concerning Human Understanding*, 75-76.
[29] Carlin, *The Empiricists: A Guide*, 90.
[30] Ibid., 81.

empiricist ideals Locke provided the foundations for this mode of enquiry and, as will be shown, points of contention for some.

George Berkeley was born in March of 1685, close to Thomastown in County Kilkenny, Ireland. He received his formal education at Trinity College in Dublin and finished his studies at the young age of 19. Berkeley became a fellow at the University and shortly after was ordained a priest of the Anglican Church. His best known works (*Essay Towards a New Theory of Vision, A Treatise Concerning the Principles of Human Knowledge,* and *Three Dialogues Between Hylas and Philonous*) were all published before Berkeley hit 29 years of age.[31] Any discussion of Berkeley requires first establishing his philosophy's underlying *idealism*. Like Locke, Berkeley believed that all human knowledge of the natural world came through the senses. However, Berkeley worried that Locke's distinction between the realm of ideas and a physical reality external to the body could lead to scepticism directed towards empiricist ideals.[32] To counter this potential future attack on empiricism, Berkeley did what he believed to be the most sensible thing. He brought everything from the natural world into the sphere of human knowledge by denying the existence of anything external to the human mind and its contents. For Berkeley, "the only things that exist are minds and ideas."[33] However, it should be noted that Berkeley also makes room for God in his formulation of immaterial idealism. God "is the immediate producer of all ideas," and he does so "in a law-like fashion in order that humans can explain things, and use nature to their purposes."[34] This was a bold move on Berkeley's part and one that has received a fair

[31] Carlin, *The Empiricists: A Guide*, 128-129.

[32] Berkeley envisioned this scepticism manifesting itself in relation to Locke's *representational theory of perception.* (Ibid., 143).

[33] Atherton, *The Empiricists: Critical Essays*, xiii.

[34] Carlin, *The Empiricists: A Guide*, 145 - 146.

amount of criticism. It seems odd for an empiricist of the enlightenment to avoid scientific reasoning for an immaterial idealism, but Berkeley's philosophical project needs to be considered in light of this underlying attempt to avoid the scepticism that he foresaw. Keeping this in mind, an exploration of this unique empiricist's work is in order.

Berkeley did still believe that ordinary objects exist, but that these objects are nothing more than collections of *ideas*. Where Locke saw ideas as being representative of physical objects that inhere in a material substrate, Berkeley distinguished himself by asserting that all objects and their qualities are nothing more than bundles of ideas in the human mind.[35] Put another way, Locke believed that these objects existed independent of the human mind, while Berkeley viewed everything as being mind-dependent. This idealism is best exemplified by Berkeley's famous phrase, "*esse est percipi,* or *to be is to be perceived*."[36] Ideas are not the only thing to exist in Berkeley's empiricism, the philosopher also made room for the existence of souls. While ideas may be considered passive objects of knowledge, "there is likewise something which knows or perceives them, and exercises divers operations, as willing, imagining, remembering, about them. This perceiving, active being is what I call *mind, spirit, soul,* or *myself*."[37] This extension from idea to perceiver is logical and straightforward enough. In a world devoid of materiality, minds (the only thing to exist besides ideas) are a necessary requirement for anything to exist at all. To bolster his case for a reality based upon immaterial idealism, the philosopher put forth two arguments in his defence.

[35] Carlin, *The Empiricists: A Guide*, 131.

[36] Ibid., 151.

[37] George Berkeley, *A Treatise Concerning the Principles of Human Knowledge*, ed. Thomas J. McCormack, (New York: Dover Publications, 2003), 30.

The first of these arguments is the argument against *Strangely Prevailing Opinion.* The opinion Berkeley speaks of is the human inclination to believe that "houses, mountains, rivers, and in a word all sensible objects have an existence natural or real, distinct from their being perceived by the understanding", which seems strange to the philosopher "for what are the aforementioned objects but the things we perceive by sense, and what do we perceive besides our own ideas or sensations."[38] Carlin summarizes the argument superbly in the form of a syllogism:

P: "We perceive sensible objects such as houses, mountains and rivers."

P: "We only perceive sensible ideas, and ideas cannot exist outside of a mind."

C: "Thus, sensible objects such as houses, mountains and rivers *just are* sensible ideas and cannot exist outside of a mind."[39]

This is Berkeley's first attempt at establishing his philosophy of immaterial idealism. How can one be sure that any external objects exist when all that resides in the mind are representations or ideas of them?

For those in doubt of the logical soundness of his first argument, Berkeley believed that he delivered his coup de grâce to the sceptics with his *Master Argument*. The philosopher challenges those still in doubt to look "into your own thoughts, and so trying whether you can conceive it possible for a sound, or figure, or motion, or colour to exist without the mind or unperceived. This easy trial may perhaps make you see that what you contend for is a downright contradiction."[40] In essence, the Master Argument posits that perception is key to the existence of anything external to the mind. How might something

[38] Berkeley, *Principles*, 31.
[39] Carlin, *The Empiricists: A Guide*, 134.
[40] Berkeley, *Principles*, 41.

exist if not seen, heard, felt or tasted? Berkeley anticipated a negative reaction to this statement and completes his argument thusly, "there is no difficulty in it; but what is all this, I beseech you, more than framing in your mind certain ideas which you call books and trees, and the same time omitting to frame the idea of any one that may perceive them? But do not you yourself perceive or think of them all the while?"[41] This is where Berkeley believed the strength of his Master Argument to lie. The simple act of trying to conjure up images of mind-independent objects necessitates its perception within the mind. One cannot ever conceive of anything external to it (save a God that implants ideas) because no perception takes place outside of the mind. Berkeley's immaterial idealism may stand apart from the empiricism of other enlightenment thinkers, but it does still adhere to its main tenet: all knowledge is derived through sensory experience that informs ideas in the mind.

The final philosopher of the British empiricist trifecta is David Hume, born in Edinburgh, Scotland in the year 1711. Hume began his educational career at Edinburgh University and then moved to La Flèche, France where he harassed the local Jesuits from the Jesuit College with his many arguments against religious belief. Unlike the empiricists so far examined, Hume was a staunch religious sceptic and the absence of appeals to a god for explanation distinguishes his philosophy.[42] Another way that Hume deviated from the other empiricists was by re-establishing some of the terminology surrounding human knowledge. He grouped all contents of the mind under the term *perception*, and held that there were two types of things contained within it—*impressions* and *ideas*. For Hume, impressions are delivered to the mind through sensory experience. They are the passive

[41] Berkeley, *Principles*, 42.

[42] Carlin, *The Empiricists: A Guide*, 155-156.

information coming in from the outside world. However, he did state that the process could happen internally as well. Hume calls upon his predecessor Locke by defining *inward impressions* as being products of *reflection*, while *outward impressions* are caused by *sensation.* Hume considered impressions to be 'lively', but that their counterpart, ideas, were 'less lively' and objects of mere recollection. Ideas are any functions of the mind that are not impressions, they are, however, "'copies' of earlier impressions."[43] This interplay between impressions and ideas led Hume to a conclusion on how it is that experience can create knowledge—the *Copy thesis*.

As its name implies, the Copy thesis refers to the origin of ideas from impressions. Ideas are less lively than impressions because they are merely copies of prior impressions. Take pain for instance. One can easily distinguish the difference between experiencing pain (a lively impression) and simply thinking about pain (the less lively idea).[44] Hume saw this as clearly pointing to the fact that all human knowledge is derived from experience. If ideas are copies of impressions then everything within the human mind can trace its origins back to some sensory event. Another reason Hume believed in the Copy thesis and the role of epistemic experience was due to the activity in the minds of disabled individuals:

> If it happen, from a defect of the organ, that a man is not susceptible of any species of sensation, we always find that he is as little susceptible of the correspondent ideas. A blind man can form no notion of colours; a deaf man of sounds. Restore either of them that sense in which he is deficient; by opening this new inlet for his sensations, you also open an inlet for the ideas; and he finds no difficulty in conceiving these objects.[45]

[43] Carlin, *The Empiricists: A Guide*, 158.
[44] Atherton, *The Empiricists: Critical Essays*, xvi.
[45] Hume, *Human Understanding*, 15.

Without the required hardware, these poor souls are unable to receive the impressions that would then inform the subsequent ideas. These foundational components of the Copy thesis gave Hume grounds for an empirical philosophy based upon experience, impressions, and sensation. From here, the philosopher was able to flesh out how the human mind is capable of cognizance and complex thought.

Yet another defining feature of Hume in comparison to the other empiricists was his account of the relation of ideas. Hume believed that "an examination of the contents of our own minds reveals that we naturally relate certain ideas together according to certain principles."[46] The empiricist isolated and described three such principles: *resemblance*, *contiguity* (in time or place), and *cause and effect*. The first of these principles, resemblance, refers to the mind's propensity to conjure up ideas from impressions that resemble the impression in some way. For example, when presented with a photograph of some location previously visited, ideas of that trip will readily come to mind. Secondly, the mind makes connections between ideas that are close in time or space due to the principle of contiguity. Things that are commonly found together elicit thoughts of each other. A famous example would be the conjunction of the impressions of salt and pepper. Finally, the principle of cause and effect leads the mind to connect ideas that are resultant of one another. If one sees an object in motion, ideas come to mind of what set that object into motion. Thoughts and ideas "are often connected on the basis of perceived causal connections."[47] Unlike his empirical predecessors, Hume gave an account of the psychological processes occurring

[46] Carlin, *The Empiricists: A Guide*, 160.
[47] Ibid., 160-161.

inside a human mind with his relation of ideas, which leads into a sceptical claim that the philosopher is well known for.

The idea of causal connections were of a great concern to Hume. Particularly, the force, or in the philosophers own terms, *necessary connexion,* that seemed to entail an effect from a cause.[48] The concept that Hume was attempting to bring to light here was not the cause or the effect itself, but what happens in between the two. Necessary connection is the power that one supposes to always result in the same effect from the same cause. Hume took issue with this. If no knowledge is known *a priori* and all knowledge is derived from experience in the form of sensation and reflection, where does the idea of necessary connection come from? It is an invisible process. One can perceive effects following causes, but never that thing, that force, that compels them to. Therefore, according to Hume, there is no real justification to assume that the same cause will always produce the same effect.[49] That the future will always resemble the past is a flaw in human reasoning, this is Hume's *Problem of Induction*. Using induction to predict the future is faulty logic. There is no basis to believe that similar events will always, without fail, unfold. Appealing to past experience as a grounds for such reasoning is faulty as well. Witnessing the act of similar causes producing similar effects in other instances is also inductive. There is no guarantee that cause and effect, as a principle of nature, will operate the same in the future.[50] There is a strong basis for the belief in cause and effect, given its prevalence in past experience, but there is no justifiable grounds for this inductive form of reasoning.

[48] Hume, *Human Understanding*, 44.
[49] Carlin, *The Empiricists: A Guide*, 170.
[50] Ibid., 172-173.

So how does this overview of the British empiricists and their projects relate to the present research? It is certainly important to have an understanding of the content to be analyzed, but its purpose extends beyond just providing background to the texts that will be pulled apart and reassembled by the machine. This framework, this *mode* of thinking, can be applied to the analysis of text using a distant reading approach. These philosophers of the enlightenment were behind many of the practices that hard, empirical sciences follow today. Locke emphasized the importance of *observation* by first establishing the ways that sensory experience can come to furnish the mind through ideas, Berkeley highlighted the importance of the *observer* with his immaterial idealism, and Hume showed the strength of using *deductive* reasoning to arrive at proofs. These empiricists were among the first humanist scholars that attempted to weld intellectual, interpretive pursuits with a hard, quantifiable science and in many ways helped to lay the foundations for discovery through the scientific method.

At times, the goals of the empiricists and the goals of digital humanists sound strikingly familiar. Both groups argued for a more grounded, scientific examination of material in a field that was originally based upon conjecture and interpretation. For the empiricists it was a condition of the time. These enlightenment philosophers were reacting to a millennia long hold on natural philosophy—hylomorphism.[51] Thanks to the findings of early astronomers like Copernicus, Kepler, and Galileo, the geocentric model of the universe was gradually falling out of favour. In turn, this sparked the Scientific Revolution

[51] A philosophy of Aristotle that holds that "nature consists of individual things—*substances*—each of which is a complex of *matter* and *form*. Matter is the fundamental stuff of which all things are composed [...] Form is what gives matter its specific properties; it is what *actualizes* the potency of matter" (Carlin, *The Empiricists: A Guide*, 8).

by opening up possibilities that were free from theological and metaphysical theories. The empiricists saw a chance to overthrow Aristotelian hylomorphism and took full advantage of it. The inductive reasoning behind *substances* and *forms* could instead be replaced with a philosophical science that was informed by experimentation and mathematics.[52] Digital humanists find themselves in a similar epoch. The 'digital revolution' is changing everything humans know about how work and research can be done. Data mining and textual analysis techniques are opening up unexplored avenues, or, at the very least, streamlining traditional techniques through computation. This is a transitionary period. Distant reading techniques promise a shift from the purely interpretive analysis of literature to an ability to explain through data why an analysis is objectively correct.

Justification

Historical Precedent

There may be a noticeable lack of contemporary philosophical projects in DH, but in no way does this imply that the field is completely absent of such work. On the contrary, the DH field boasts a rich philosophical history. Philosophy and computation, the modern day process that lies at the heart of DH and text analysis, share an intricate relationship that can be traced all the way back to the work of the 19th century logician, George Boole. It is difficult to describe any type of coding method that does not reference Boolean logic in one form or another. Boole was influenced by, and ended up replacing, the Aristotelian logic paradigm that preceded him. He agreed with and incorporated the Greek philosopher's deductions of existential conclusions from universal premises. Boole justified Aristotle's logic by "showing that it had much more in common with mathematics than had

[52] Carlin, *The Empiricists: A Guide*, 7-10.

previously been thought," which in turn revealed to the world that logic was capable of achieving far greater tasks than supposed.[53] He refined the purely philosophical logic that had been established millennia before in ancient Greece. The logician replaced Aristotle's phonetic Greek representations of propositions with algebraic equations, resulting in an unlimited supply of propositions that could be called upon. Boole also introduced two purely logical elements that transcended formalized language: "1 for 'everything' or the universe of discourse and 0 for 'nothing' or the empty class."[54] Boole's transformation of Aristotelian logic to a more mathematical representation was a pivotal point in the history of computation. By abstracting to a symbolic representation of logic, Boole allowed for formalized language propositions to be reduced to variables that could then interact through simple statements. The work of this 19th century logician, who was heavily influenced by ancient Greek philosophy, is behind the computer processes that make DH projects and text analysis possible.

More recently, and more closely associated with distant reading, is the work that was done by Father Roberto Busa starting in the 1940s. Considered by many to be a pioneering figure of DH, Father Busa's best known project is the *Index Thomisticus*. In the words of DH scholar, Steven Jones, the *Index* is "a massive (56 volumes in print), lemmatized concordance containing every word in the complete works of the thirteenth-century philosopher and theologian, St. Thomas Aquinas."[55] Concordances are a way to determine the context of a word and are a staple of text analysis that can trace roots back

[53] Corcoran, J., "George Boole," (*Encyclopedia of Philosophy*, 2nd edition, Detroit: Macmillan Reference USA, 2006), 3.
[54] Ibid., 4.
[55] Steven E. Jones, *Roberto Busa, S. J. , and the Emergence of Humanities Computing: The Priest and the Punched Cards,* (London: Routledge, 2016), 7.

to the middle ages. Every occurrence of the word is plucked from a text with a set number of characters preceding and following it. The present research does make use of concordances in later sections as additional evidence to some findings. Lemmatization was also used in the preprocessing stages of some of the methods found in this research.[56] This is the process of tracing derivative words back to their root in order to facilitate analysis. When he started his work, Father Busa did not have access to the powerful digital tools that are now commonly used in similar projects. The idea for the *Index* began when Busa became intrigued after his search through tables and subject indexes for the Latin derivatives of *presence* (*praesens* and *praesetia*) revealed that Thomas Aquinas' "doctrine of presence is linked with the preposition *in*."[57] Indeed, Busa was so intrigued that he proceeded to write out around 10,000 sentences from the work of Aquinas on cards that contained or were connected to the word *in*, which then became the basis for his doctoral thesis.[58] This tedious work led Busa to seriously consider the value of studying an author's vocabulary and of having a repository of such information to draw upon. The Jesuit priest and academic took it upon himself to create "a concordance of all the words of Thomas Aquinas, including conjunctions, prepositions and pronouns, to serve other scholars for analogous studies."[59] This was a lofty idea for the time and one that echoes present day ideals of shared and open data. The theologian quickly realized that his vision of the *Index* was well beyond the processing power of a single human mind. The priest traveled to IBM's

[56] A process that Father Busa would surely be in favour of. Upon reflection of his 30 years of research, Busa is still convinced that "pre-editing and lemmatizing are necessary in processing large texts." (Roberto Busa, "The Annals of Humanities Computing: The Index Thomisticus," *Computers and the Humanities* 14, (1980): 87.)

[57] Busa, "The Index Thomisticus," 83.

[58] Ibid., 83.

[59] Ibid., 83.

American headquarters where he struck a deal to complete the *Index* using their state of the art (circa 1949) punch-card terminals.[60] The project was a massive undertaking and, despite the dated technology (or perhaps because of it), the *Index* is still impressive to this day. Punch-cards require an incredible amount of labour, and even after 3 separate checks Busa's team still found "1600 punching errors and one full line lost because of an *homoteleuton*."[61] The hours put in by human hands and eyes is staggering. Busa states that the work of the computer totalled around 10,000 hours, "while man hours were much more than one million."[62] Looking back from an age where the computer is prized for its ability to simplify the work for human beings, it is clear that the *Index* was being carried out for very different reasons—to test the capabilities of machines on human literature. Father Busa's work with IBM on the *Index* not only exemplifies the use of computation to analyze classic philosophical literature, it was one of the first times such techniques were attempted on a text and the project continues to this day. As noted by Jones, "Busa's work continued into the beginning of the twenty-first century, involving the use of powerful computers like the IBM 705 and the IBM 7090, for example, as well as personal computers, CD-ROMs, hypertext links, and the Internet."[63] The *Index* has found a home on the web decades after its humble beginnings as punched holes in stacks of cards.

Despite this historical relationship between DH and philosophy, a number of academics have become aware of the conspicuous lack of contemporary philosophers and philosophy projects in this interdisciplinary humanities field. DH scholar, Lisa Spiro, asks

[60] Jones, *The Priest and the Punched Cards,* 9.
[61] Busa, "The Index Thomisticus," 87.
[62] Ibid., 87.
[63] Jones, *The Priest and the Punched Cards,* 9.

"why philosophy seems to be less visibly engaged in digital humanities."[64] Assistant professor of philosophy, Laura Kane, wonders where her fellow philosophers are, given that "the projects developed within the Digital Humanities are often built upon the same objects of inquiry that philosophical theories grapple with: knowledge, identity, communication, language, logic, etc." and that "scholarly research in the Digital Humanities employs the same rigorous dedication to methodology that philosophical research does."[65] McDaniel College's associate professor of philosophy, Peter Bradley, shares a similar realization after attending the DH conference, THATCamp Pedagogy. The professor remarks that there was one humanities discipline largely absent at the conference, "my own, philosophy," and that this was "not new, nor surprising. It is, however, deeply regrettable."[66] Bradley goes on to say that it is not that there are *no* philosophers making use of digital techniques, but that the discipline is highly underrepresented. He struggles with this, as the discipline of philosophy is "interested in the discovery, development, classification and analysis of human concepts and reasoning. We teach texts, concepts, arguments, and the historical and social development and influence of such texts, concepts and arguments," all tasks that Bradley finds completely "amenable to the digital humanities."[67] This question of where the philosophers of DH are is one that Bradley is never able to sufficiently answer in his article. However, Bradley is correct in his assertion that even though they are few and far between, there are still a number of exciting contemporary DH projects making use of philosophical material.

[64] Lisa Spiro, "Exploring the Significance of Digital Humanities for Philosophy," (Digital Scholarship in the Humanities, 2013), para. 2.
[65] Laura Kane, "No Room for Digital Humanities in Philosophy," (GC Digital Fellows, 2014), para. 1.
[66] Peter Bradley, "Where Are the Philosophers? Thoughts from THATCamp Pedagogy," (The Chronicle of Higher Education, 2011), para 2.
[67] Ibid., paras. 6-7.

Contemporary Projects

While philosophy may not be as well represented as other humanities fields, there are some DH projects that are deeply involved with the discipline. As has been mentioned, Father Roberto Busa's work lives on as an online digital artifact and serves as continuing inspiration for all DH researchers.[68] There are also interesting digital initiatives being undertaken by groups of DH philosophers and researchers. There is "Mapping the Republic of Letters", an expansive project sponsored by Stanford University that has created a variety of data visualizations to map the networks of correspondence between early philosophers, scientists, and other great thinkers.[69] Another digital project that focuses strictly on philosophical sources is "The Homer Multitext Project", a collaborative endeavour to digitize and house the various manuscript versions of Homer's *Iliad* and *Odyssey*.[70] Last but not least is "The InPho Project," an online resource that creates 'dynamic ontologies' of philosophical internet resources through data mining, machine learning, and visualization.[71] As impressive as these projects are they are conceptually quite different from the research presented herein. Therefore, it would be prudent to examine a contemporary distant reading of philosophical material that shares similarities with the methods and resources.

Geoffrey Rockwell and Stéfan Sinclair present such an example of philosophical text analysis in their book, *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. The analysis is philosophical not only due to the corpus, *Dialogues Concerning Natural*

[68] See: http://www.corpusthomisticum.org/it/index.age.
[69] See: http://republicofletters.stanford.edu/index.html.
[70] See: http://www.homermultitext.org/.
[71] See: https://www.inphoproject.org/.

Religion by Hume, but also because the project ties the philosophical scepticism raised in the text to the method itself. As noted by Rockwell and Sinclair, “it struck us at a certain point that text analysis was a method (or performance) of questioning, a thinking through similar to Philo’s scepticism.”[72] This is an intriguing aspect of using philosophical literature as a source for distant reading, and one that is not lost upon the researcher. The philosophical mode of enquiry, raising doubt that in turn fuels questions and debate, resonates with what text analysis can provide. The method is constantly questioning not only the source material, but itself. Exploratory text analysis may not provide human interpretation, but it does provide an abundant supply of hypotheses. In the words of Rockwell and Sinclair, “new questions will come faster than answers.”[73] However, there are some practical implications that need to be addressed when using philosophical literature as a source for distant reading.

In their analysis, Rockwell and Sinclair explore one of Hume’s dialogues—a uniquely philosophical form of writing. Dialogues are constructed much like a script for film or theatre. Character names are followed by their speech. This structure calls for a different handling of data than other literature. For example, Rockwell and Sinclair note that it is difficult to determine what it is that Hume the philosopher thinks when presented with a text in dialogic form, because “nowhere does Hume write in the first person.”[74] This uncommon writing style calls for additional preprocessing of data. Rockwell and Sinclair sidestep the issue by isolating the names with the use of semantic tagging in XML and then

[72] Geoffrey Rockwell and Stéfan Sinclair, *Hermeneutica: Computer-Assisted Interpretation in the Humanities*, (Cambridge: MIT Press, 2016), 187.
[73] Ibid., 169.
[74] Ibid., 171.

using word counts for each characters speech to determine which of the voices is most likely Hume's. In this case, words uttered by the character Philo vastly outnumber other characters in the dialogue, and by "assuming that quantity represents authority" the authors determine through textual analysis techniques that Philo is the likely representative for Hume's voice.[75] Dialogues are also found in the corpus used for this research and the treatment of this philosophical writing style will likewise need to be processed correctly for analysis. While Rockwell and Sinclair opted to include character references in their analysis, it was a choice based upon the question being asked. The use of XML semantic tagging was necessary for their purposes, but in other instances may be excessive "where a text can be treated as a simple sequence of words."[76] Such is the case with the analysis presented in this project. Character names, along with other unnecessary words, have been stripped in order to facilitate the analysis of the large selection of texts. More information on the preprocess cleaning is provided in the Methodology section.

As shown by Rockwell and Sinclair, textual analysis of philosophical literature can require a bit more from the researcher than other sources. As is always the case with digital text analysis, the analyst should be intimately familiar with the literature in order to determine the best processing techniques and methods to use. Additionally, philosophy can engage by not only providing documents to analyze, but by referring back to the text analysis process itself. For Rockwell and Sinclair, Hume's scepticism provides a sense of similarity between text analysis and the philosophical endeavour. Likewise, the projects of the British empiricists can be seen to share commonalities with the methods proposed in

[75] Rockwell and Sinclair, *Hermeneutica*, 171.
[76] Ibid., 170.

this research. The scientific method, observation, and deduction all play important roles in decoding British empiricism when using the raw textual data in visualizations. Philosophy, while perhaps under utilized as a resource in text analysis, can begin to bridge the gap between the humanities and the use of digital tools.

METHODOLOGY

"Truly my opinion is, that all our opinions are alike vain and uncertain. What we approve today, we condemn tomorrow. We keep a stir about knowledge, and spend our lives in the pursuit of it, when, alas! we know nothing all the while: nor do I think it possible for us to ever know anything in this life. Our faculties are too narrow and too few. Nature certainly never intended us for speculation."

George Berkeley, *Three Dialogues Between Hylas and Philonous*, p. 61.

Method

Given the research questions, it should come as no surprise that the nature of this analysis is primarily exploratory. Although the project aims to add to foundational distant reading practices on philosophy source material, there was a vague notion that the analysis could also add to the discussion surrounding British empiricism. As much has been shown already by the examination of Rockwell's and Sinclair's exploration into Hume's *Dialogues on Natural Religion.* Breaking down the text and examining its constituent parts revealed the author's narrative identity and Hume's scepticism helped to explain the method itself. Computational text analysis has a gained advantage over other qualitative humanities research. Working with data allows for theories and refined research questions to bubble up to the surface during the course of an analysis. One need not start from a concrete, preconceived theory. This is at the heart of the thesis, and the approach can best be described through Anselm Strauss' and Barney Glaser's *Grounded Theory*. Despite the name, this is in fact a research method that has been in development since the late 1960s. The method, for these research purposes, can be seen as an attempt to weld analyses that are generally considered at odds. Namely, the *quantitative analysis of qualitative data*, which "may include the analysis of word frequencies and word combinations."[77] Analysing qualitative data quantitatively requires the researcher to exercise restraint. Grounded Theory started as "a more or less theory-less approach: everything the researcher theorized before the analysis ('preconceived theories') was thought to inhibit rather than promote his or her perceptions during the analysis process."[78] It is difficult for research to be completely free from any preconceived theory, but this was Strauss and Glaser's first definition of the method and it has since been updated to include more concepts from classical research. The research questions at the beginning of the thesis have been left intentionally vague so that the data may speak for itself through the results. Text analysis, as has been stated by Rockwell and Sinclair among others, typically leads to more questions than answers. It is best for the distant reading researcher to start with somewhat of a *tabula rasa* of their own as they begin the endeavour. This is not to say that there is no value in having a general direction of enquiry, as the original conception of Grounded Theory seems to deny. Indeed, Udo Kuckartz says as much himself in *Qualitative Text Analysis: A Guide to Methods, Practice & Using Software*: "Prior knowledge is always a factor, as the researcher's brain is never 'empty'."[79] Positing research questions that are open enough to be answered at some level by the research and findings is valuable, and it is what the researcher has strived to achieve.

That being said, text analysis is a technical pursuit and also requires a practical approach. Appealing to Grounded Theory in order to define the research questions is fine,

[77] Udo Kuckartz, *Qualitative Text Analysis: A Guide to Methods, Practice & Using Software,* (London: SAGE Publications Ltd, 2014), 7.
[78] Ibid., 40.
[79] Ibid., 22.

but to justify this rigorous methodological procedures need to be established. To these ends a 5 step process has been created and applied to all of the Empiricist's documents: corpus selection, data preparation, data coding, data visualization, and culminating with interpretation of the results.

Corpus

The first step after establishing the research questions was to pull together a collection of relevant texts for the analysis. The corpus is not an exhaustive selection of the literature written by the British empiricist's, but it does capture the best known works from each of the philosophers. These late 17th to 18th century philosophical writings are now well within the public domain and so copyright infringement is a non-issue. The documents were collected from freely available online PDF versions that have been scanned from the original documents and optimized using Optical Character Recognition (OCR) technology.[80] The text was then transferred to Sublime Text 3 in order to save plain text versions of the documents for analysis.[81] There are online resources, like Project Gutenberg, that do provide plain text versions, but the documents are not always accurate. The site makes use of volunteers to digitize texts and proofread the results.[82] Human error is unavoidable in this situation, especially given the length and language of some of these philosophical works. By directly transferring the original PDF text to the correct format, one can ensure that the corpus meets the quality required to carry out this type of research. Some texts have been intentionally left out of the analysis and the reasons behind this will be explained in further detail as each of the empiricist's are examined. Word counts were

[80] OCR technology converts scanned documents into an editable and searchable digital format.
[81] Sublime Text 3 is a freely available programming text editor. See: https://www.sublimetext.com/3.
[82] See: https://www.gutenberg.org/.

taken from the cleaned version of the texts in order to provide a more accurate picture of the tokens used in the analysis. Word counts for the corpus section were coded using basic Python 3 functionality.[83]

While Locke may have the least amount of sources in the corpus, his total word count is still comparable to that of the other empiricists. His two best known works have been included: *An Essay Concerning Human Understanding* and *Two Treatises of Government*. Locke did write some important texts on the philosophy of education (*Some Thoughts Concerning Education*, 1693) and theology (*The Reasonableness of Christianity*, 1695), but his major epistemological and philosophical projects are expressed in the two sources used in the corpus.[84] Word count total was also taken into consideration. Locke's texts are quite lengthy when compared to Berkeley and Hume. Even after restricting his contributions to the corpus, Locke does still have the largest amount of words.

Table 1. John Locke corpus word count.

John Locke Texts	Word Count (Cleaned Texts)
An Essay Concerning Human Understanding (1689)	285, 205
Two Treatises of Government (1689)	101, 621
TOTAL:	386, 826

The selection of Berkeley's texts reflects his 'golden period,' these are the works that defined the philosopher's epistemology and immaterialism. His empirical project started with an examination of the senses in *An Essay Towards a New Theory of Vision*, which he latter expanded upon in both *A Treatise Concerning the Principles of Human Knowledge* and

[83] View the code here: https://github.com/jasonrbradshaw/DBE/blob/master/TotalWordCounts.ipynb.
[84] Graham A. J. Rogers , "John Locke," *Encyclopædia Britannica*, 2018.

his *Three Dialogues between Hylas and Philonous*. The one text that stands apart in the Berkeley corpus is *Alciphron*. Also presented as a dialogue, *Alciphron* is an example of Berkeley's moral philosophy related to his theological background. God being the source of human ideas, Berkeley's theology finds a voice in his empiricism as well. It would be remiss to not include something from Berkeley that highlighted this facet of his philosophy. Theology played a major role in not only Berkeley's philosophy, but in his daily life. He was a deacon and priest for quite some time.

Table 2. George Berkeley corpus word count.

George Berkeley Texts	Word Count (Cleaned Texts)
An Essay Towards a New Theory of Vision (1709)	27, 795
A Treatise Concerning the Principles of Human Knowledge (1710)	35, 710
Three Dialogues between Hylas and Philonous (1713)	36, 301
Alciphron (1732)	83, 460
TOTAL:	183, 266

As with the other empiricists, the collection of Hume's writings reflect his best known works. Beyond the inclusion of his obvious contribution to empiricist ideals (*An Enquiry Concerning Human Understanding*), the other texts are definitive of Hume's philosophy. They can also be seen to complement the selections from both Berkeley and Locke. *Essays, Moral, Political, and Literary* can be contrasted to Locke's political philosophy. Likewise, *Dialogues Concerning Natural Religion* provides a dialogic source similar to Berkeley's *Alciphron* and *Three Dialogues*. Those that are familiar with the works of Hume may notice that one of his most famous pieces is missing. Carlin notes that Hume's "most popular work was not a philosophical one. Between 1754 and 1762, Hume wrote and published his six-volume *A History of England*."[85] This collection has been excluded from the analysis in order to avoid misrepresentation. This collection is massive and would most likely throw off the results with its historical, rather than philosophical, content.

Table 3. David Hume corpus word count.

David Hume Texts	Word Count (Cleaned Texts)
Essays, Moral, Political, and Literary (1758)	229, 525
An Enquiry Concerning Human Understanding (1777)	52, 310
An Enquiry Concerning the Principles of Morals (1777)	52, 021
Dialogues Concerning Natural Religion (1779)	38, 734
TOTAL:	372, 590

Data Preparation

After the corpus documents were selected, the .txt files that were created from the PDFs in Sublime Text 3 needed to undergo a cleaning process for the text analysis techniques.[86] Front matter and table of contents were removed easily enough by hand, but for some elements it was necessary to clean the texts with the use of regular expressions (regex) in Python. Regex can be implemented within most programming languages and uses ASCII characters to represent search patterns, making it an effective tool to remove or replace characters in a string of text. Each of the text documents required varying degrees of cleaning, depending upon their contents and because of the nature of philosophical writing from this time period. For example, some documents make use of roman numerals

[85] Carlin, *The Empiricists: A Guide*, 157.
[86] View the code here: https://github.com/jasonrbradshaw/DBE/blob/master/CleanTexts.ipynb.

to delimit sections, philosophical dialogues have speaker names prefacing dialogue, all of the documents contained page numbers, excessive whitespace, and there were some characters that were unique to the documents being analysed (bullet points and tildes). All of this content can be considered superfluous for the analysis and was removed making use of regex patterns.

Cleaning the unnecessary elements was only the first stage of the data preparation. Next, it was necessary to define stopword lists for the documents in order to remove the propositions, conjunctions, and other non-essential words from the text.[87] This was accomplished by using the Natural Language Tool Kit module (more on modules later) in Python. At each phase of the data preparation process, versions of the texts were saved. Additionally, the empiricists individual documents were concatenated and saved as "complete" versions. The corpus ended up containing all of the original sources as .txt files, cleaned versions of these .txt files, cleaned versions with stopwords removed from these .txt files, and a complete cleaned stopwords removed collection for each Locke, Berkeley, and Hume. This may seem a bit excessive, but it is a good practice to have these multiple versions for different types of analysis. A complete collection of an empiricist's works may yield vastly different results than comparing their individual texts. This difference is expressed most fully through topic modelling. Models change quite dramatically depending upon the input source.

Coding

All coding has been done in Jupyter Notebooks, a cell based Graphical User Interface (GUI) for Python 3 whose environment operates in the browser window. Jupyter is free to

[87] View the code here: https://github.com/jasonrbradshaw/DBE/blob/master/LexicalRichnessIndex.ipynb.

download as part of the Anaconda Navigator bundle, a suite of software packages that includes both Python and R coding tools. Other scripting languages do have text analysis capabilities, but coding in Jupyter with Python has several advantages. Scripting language largely comes down to user preference and familiarity, but there are reasons to choose Python for this specific type of research over another language like JavaScript, R, or Perl. Python has stripped out many of the idiosyncrasies that are particular to programming languages. For example, Python does not require variables to be explicitly stated, conditional statements do not need to be surrounded by ellipses (they are defined by indentation), and statements do not require a closing character (they are delimited by new lines). Though it might not seem like much, doing away with these common coding elements is beneficial from the reader's perspective. It makes for code that is clean and easier to understand for an audience that may be unfamiliar with coding. The Jupyter Notebook environment further enhances the accessibility of Python scripts. The cells break up the code nicely and allow for the insertion of comments and markup so that readers may follow along with the programmer's thought process.

Python also allows for the implementation of modules. These are libraries of code created by others that can quickly execute specific functions. An example has already been provided with NLTK. This module contains functions that can instantly tokenize strings of characters and perform word counts, along with a host of other text based procedures. Modules can be downloaded freely and fall under a number of software licences. Each of the text analysis techniques for this analysis make use of modules to some extent and these modules are explored further in the following sections. All of the coding for this research has been uploaded to a public GitHub repository.[88] This was done in the spirit of open and transparent data. Readers are encouraged to view the coding process for themselves and to adapt the code to their own projects.

Visualization

Some visualizations for analysis have been created by making use of modules within the Jupyter Notebook environment. For others, visual representations of the results have been crafted in Adobe Photoshop. Informative and interesting visualizations are an important aspect of this research, as they will strengthen ties to the British empirical framework that this thesis draws upon. Experience through the senses, like sight, is how ideas come to rest in the mind. Consideration of User Experience (UX) and interactivity have become an integral component of data visualization, and rightfully so. Transparency of data and results is important for any type of analysis that makes use of quantitative measures. However, it is highly impractical to present readers with copious amounts of data and no way to interpret the results. Visualizations are a means to impart research findings to audiences that may not be intimately familiar with the methods and techniques used. In this way, information is "*represented* to make it perceptually accessible and available for use and interaction."[89] Representation of information does make it accessible, but it can also obscure by creating an abstraction from raw data. Therefore, it is of the utmost importance that the researcher judiciously designs or makes use of existing visualization frameworks. Such an approach has been taken with the British empiricist's word frequency count charts that complement the LRI analysis. This approach to the

[88] View the code here: https://github.com/jasonrbradshaw/DBE.

[89] Kamran Sedig and Paul Parsons, *Design of Visualizations for Human-Information Interaction: A Pattern-Based Framework*, (San Rafael, California: Morgan & Claypool, 2016), 1.

visualization can roughly be considered equivalent to an 'infographic' and will be explained in further detail in the LRI section.

Interpretation

While textual analysis does have its quantitative components (raw text data), qualitative analysis still plays a large role in the process. The data and some of the visualizations produced by the research may come into existence free from interpretation, but to be of any value a human researcher needs to intervene at some point to make sense of the findings. This final stage of the methodology requires the researcher to make a qualitative assessment of the results provided to them. There are two important resources that will be drawn upon for interpreting the results: a working knowledge of the source material and the wealth of secondary sources on British empiricism, specifically, prior close readings by experts in the field. Knowledge of the original source material may seem an obvious requirement, and it is one that has been mentioned a number of times, but it is relevant to address the point as not all text analysis projects require this of the researcher. For example, classical content analysis is typically unconcerned with a prior knowledge of the texts and, depending upon the research question, not knowing the details of the content may prove to be beneficial and more revealing.[90] The modern day equivalent of such 'blind' textual analysis can be seen in the data mining of text from social media sites, like Twitter. These types of analysis scrape sites for user posts referencing certain hashtags or words to gather a corpus that then undergoes some type of coding procedure to identify thematic similarities. This is NOT that type of research. To get to the core of what text analysis has to offer for examining the works of the British empiricists and philosophy in general, prior

[90] Kuckartz, *Qualitative Text Analysis,* 29 -30*.*

knowledge is key. Secondary sources, like close readings, can enhance the understanding of these texts by providing further information into the projects of the empiricists, as well as becoming a resource to compare findings from the analysis to. In the spirit of the philosophic tradition, research involves not only a close examination of the source material, but also what others have had to say about it.

Techniques

Lexical Richness Index

Franco Moretti may have popularized text analysis with his "Howitzer" of distant reading, as Jockers so eloquently puts it, but by no means was he the first to examine texts by breaking them down to their constituent parts. In the mid-twentieth century, statistician George Udny Yule found himself wanting more than simple comparative analysis to determine authorship. Particularly in relation to the debate surrounding Thomas à Kempis and the *De Imitatione Christi.* Yule was searching for a "summary, some picture of the vocabulary as a *whole.*"[91] This led the statistician down a path that was novel for the time, though one that is now considered a foundational step in text analysis research. Yule began by counting and creating a word frequency chart of the nouns used in the document. The counts were telling, but what Yule found most interesting was how "the whole form of the distribution at once raised the question how far peculiarities of form might be characteristic of one author as compared to another."[92] Yule saw an opportunity to employ his statistical methods to answer a question that was formerly reserved to literary analysis. He hypothesized that authorship could be determined through empirical data by

[91] George Undy Yule, *The Statistical Study of Literary Vocabulary*, (Cambridge, England: Cambridge University Press, 1944), 2.
[92] Ibid., 3.

comparing vocabulary and diction across several samples. Moreover, the technique could be used as a function to compare the lexical indices of two or more authors. His research eventually led to the *K-Characteristic*, a mathematical property that returns a quantitative evaluation of an authors vocabulary by comparing set sample sizes of word types across a corpus.[93] Using manual methods, Yule was able to derive an equation to determine authorship with far greater accuracy than subjective comparative analysis. The modern day equivalent of Yule's seminal work with the K-Characteristic can be found in the form of the Lexical Richness Index (LRI). The technique, in the words of Matt Jockers, "is a measure of the ratio of unique words to total word tokens in a given text."[94] LRIs take a 'bag of words' and, using the K-Characteristic, count the number of unique tokens in a document to derive a Type Token Ratio (TTR) score. Given this quantifiable output of an individual's writing style, the technique naturally lends itself to an interpretive comparison of the works of two or more authors.

There are a few issues that arise when using a LRI to analyze literature. First, what should be considered a unique word? Statistically the term refers to the word counts that are unique to an author. Yule found that every author he examined had word frequency distributions with words that occurred only once in a given text, and that these one-off words constituted the largest portion of the distribution tables.[95] However, these authors have varying ranges of frequencies beyond the single word cases. This is where comparison can begin. Word frequency distributions are like fingerprints of authorship. They remain relatively constant for a single author across documents, but when compared to other

[93] Yule, *Literary Vocabulary*, 57.
[94] Jockers, *Macroanalysis*, 63.
[95] Yule, *Literary Vocabulary*, 22.

writers there is significant variation. Yule spotted another function of variation in word frequency counts for an author: "Everything—apparently—about the distribution will in fact alter as the size of the sample is increased."[96] This leads to a second issue with LRIs that has been commented on by a number of researchers, including Yule himself. Vocabulary, when examined by the K-Characteristic, is directly related to text length. Therefore, as noted by DH researchers Fiona Tweedie and R. Harald Baayen, "it is extremely hazardous to use lexical 'constants' to compare texts of different length."[97] Results of LRIs are dependant upon and change with text length, but there is a way to mitigate discrepancies. It is considered a best practice to take random samples of 10, 000 words from each text used in a corpus to produce accurate LRIs.[98]

LRIs have been determined for each of the empiricists using the LexicalRichness Python 3 module, available through PyPI.[99] The LRI module creates mean scores for the authors that can be used to compare unique word count, type token ratio (TTR), root type token ratio (RTTR), corrected type token ratio (CTTR), mean segmental type token ratio (MSTTR), moving average type token ratio (MATTR), measure of textual lexical diversity (MTLD), and hypergeometric distribution diversity (HD-D).[100] The analysis makes use of the full cleaned corpora from Locke, Berkeley, and Hume. It would have been redundant to break down each and every text for processing and so a collection of the respective authors works provides a more accurate representation of their vocabularies. Following the best

[96] Yule, *Literary Vocabulary*, 22.
[97] Fiona J. Tweedie and R. Harald Baayen, "How Variable May a Constant Be? Measures of Lexical Richness in Perspective," (*Computers and the Humanities*, 32, 1998), 324.
[98] Jockers, *Macroanalysis*, 103; Yule, *Literary Vocabulary*, 57.
[99] LexicalRichness created by Lucas Shen. Available under the MIT Licence: https://pypi.org/project/lexicalrichness/.
[100] For full documentation of LRI measures, see: https://github.com/LSYS/LexicalRichness/blob/master/lexicalrichness/lexicalrichness.py.

practices mentioned by Tweedie and others, random 10, 000 word samples were extracted from the tokenized word lists for the LRI. Analyzing these equal random text chunks from the empiricists certainly helps reduce inconsistencies of results, but even this may not be enough to deliver the accuracy that this research is trying to achieve. Therefore, an additional step has been added to the LRI process. A function has been defined to iterate through multiple random 10, 000 word 'chunks' and then determine their mean score. This additional step delivers results that are more accurate than a single pass through of the data, which tends to deliver varying results each time the code is run. Iterating the process and calculating the mean produces LRI measures that are a better reflection of the authors true scores and, by extension, their vocabularies. Word frequency counts of the empiricists' corpora were also conducted during this phase of the analysis using the NLTK module. Not only are word frequency counts valuable for comparing the author's lexical indexes, but they are an important resource for topic modelling as will be shown in the following section.[101]

As has been mentioned, the LexicalRichness module does not produce associated visualizations like the Topic Model or MCA modules. The code returns decimal numbers that reflect the empiricist's LRI scores. Visualizations of these numbers have been created in Python using the MatPlotLib module. The LRI results have been represented with bar graphs and the three authors have been plotted side by side for analysis. However, the word frequency counts from Locke, Berkeley, and Hume were visualized outside of the Jupyter Notebook shell. Heeding the advice of Sedig and Parsons, as well as taking note of trends in popular data analytics, new visualizations have been created using Adobe

[101] View the code here: https://github.com/jasonrbradshaw/DBE/blob/master/LexicalRichnessIndex.ipynb.

Photoshop. These new visualizations can be considered a form of infographic, an aesthetically pleasing way to represent data that is often employed in data journalism. This method works well for numerical data that is not visually appealing on its own. Furthermore, for a comparative analysis, like that of the empiricist's LRI measures, infographics can be well utilized to highlight the important discrepancies and similarities between the author's vocabularies. There are risks when working with these types of visualizations. They have the potential to be misrepresentative of the results if not enough information is supplied. Therefore, all numerical data from the analyses has also been included in the visual representations.

Topic Modelling

Topic modelling, in its most general sense, is "a way of describing what a text is about."[102] This is a unique text analysis method because of its emergent properties. The corpus topic lists are generated by code in the absence of any preconceived categorization. This hands-off approach exemplifies the objective, data driven nature of computational text analysis. However, topic modelling is not completely free from subjective interpretation. The technique operates under the 'bag of words' principle, which is also to say that "it is linguistically naïve and pays no attention to the grammatical or semantic structure between words."[103] Texts are reduced to only those words that are considered important for the analysis. Using a stopword list, pronouns, prepositions, and conjunctions are often removed for the sake of clarity and precision.[104] Topic modelling then takes this 'bag of

[102]Akira Murakami, Paul Thompson, Susan Hunston, and Dominik Vajn, "'What Is This Corpus about?': Using Topic Modelling to Explore a Specialised Corpus," (*Corpora* 12 no. 2, 2017): 244.

[103] Ibid., 246.

[104] Stopword lists are a commonly employed practice in text analysis. All of the methods explored in this thesis do make use of the reductive technique.

words' and creates lists by identifying words that have a high probability of co-occurrence within a corpus. It is here where the objectiveness of the procedure breaks down. As noted by Murakami et al., these lists can be considered to "characterise 'topics', and researchers may choose to refer to them using topic-like titles, but these are only convenient abstraction from lists of words."[105] The lists of topic modelling are self-generating, but the topic titles are not. It is left up to the researcher to determine just how it is that the list words are associated and, if they are so inclined, to label the group under a common topic heading. Such labelling, though not required, is often done in order to facilitate analysis. It provides a means to compare findings from the documents being examined. This brief overview of the process should provide an idea as to the results that topic models can produce, but what is happening in the backend to make these models a reality? Topic modelling is made possible through a probabilistic model known as Latent Dirichlet Allocation (LDA).

Considered to be the original authority on the topic, Blei and his colleagues define LDA as "a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words."[106] This can also be considered the current definition of topic models in general: the probabilistic generation of word lists that belong to subjectively determined topics. Without a solid background in computer science and mathematics the concepts and equations expressed in the article are somewhat difficult to grasp. Thankfully,

[105] Murakami et al., "What is this Corpus about?" 244.

[106] David M. Blei et al., "Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3, (2003): 996.

Blei offers up a more humanist-friendly explanation of LDA in "Topic Modelling and Digital Humanities." He begins with two underlying assumptions of LDA:

1. "There are a fixed number of patterns of word use, groups of terms that tend to occur in documents. Call them *topics.*"
2. "Each document in the corpus exhibits the topics to a varying degree."[107]

There is a clear distinction made that *topic* serves only as a placeholder term for the lists generated through LDA. It is noted that these lists only appear as topics "because terms that frequently occur together tend to be about the same subject."[108] One must not confuse LDA's statistical models with immutable subject matter to be found within a collection of documents. They provide a launching point for investigation and, as stressed by Blei, "it is still the scholar's job to do the actual interpreting and understanding."[109] There are a number of misconceptions in the field over just what it is that topic models can actually do for a researcher. Following are some of the most common issues that digital humanists may encounter when using the method on their corpora, as well as possible solutions.

Benjamin Schmidt, in "Words Alone: Dismantling Topic Models in the Humanities," highlights several shortcomings of topic models that researchers may face when using LDA. First, the simplification of topic models for humanists, who tend to ignore the underlying algorithms of the process, can lead to inaccurate insight. As noted by Schmidt, "the apparent ease and intuitiveness of topic models comes from a set of assumptions that are only partially true."[110] One assumption being that the topics contain two properties: they

[107] David M. Blei, "Topic Modeling and Digital Humanities," *Journal of Digital Humanities* 2, no. 1 (2013): 9.
[108] Ibid., 9.
[109] Ibid., 10.
[110] Benjamin M. Schmidt, "Words Alone: Dismantling Topic Models in the Humanities," *Journal of Digital Humanities* 2, no. 1 (2013): 49.

are *coherent* (word sets that appear together have things in common) and *stable* (similar topics appearing in different document sets mean the same thing in both).[111] These assumptions may or may not be true in any given analysis. When they do prove to be faulty, researchers have the potential to misconstrue results and infer meaning where there is none. However, knowledge of assumption beforehand can be empowering. As long as the researcher is aware of the "messy, ambiguous, and elusive" nature of topics produced through LDA, they can correct their inferences by engaging with the word counts that actually produce the model.[112] Schmidt offers a suggestion to avoid this issue, "visualization that uses the individual word assignments, not just the topic labels, can help dramatically change the readings that humanists give to topics."[113] Topic lists should never be studied in isolation, they require the researcher to have an intimate knowledge of the individual word counts. Moreover, researchers run a risk by isolating the topics themselves. They should instead be viewed as a whole, with analysis looking to "networks of interconnection through an entire model, because individual topics cannot stand in for discrete objects on their own."[114] A final issue worth examining that Schmidt raises is the *linguistic drift* that plagues human language. Literature that is over a century old may contain words that mean something very different today then when they were initially penned down. Therefore, a model studying classical literature from before the 20th century must account for this by either assuming that the vocabulary has not changed, or by "forcing all topics to occupy narrow bands of time."[115] Given the era of the British empiricists, this is something

[111] Schmidt, "Words Alone," 49.
[112] Ibid., 49.
[113] Ibid., 50.
[114] Ibid., 56.
[115] Ibid., 57.

that will have to be accounted for in the thesis and can be mitigated during the cleaning phase. Topic modelling has its fair share of sceptics, but when the technique is executed correctly it can provide invaluable information. The researcher must always be aware of underlying assumptions when using this, or any other, digital method.

There are modules to produce topic models within Python, such as Gensim, but none of them hold up to the power of MALLET, a popular Java-based language processing package.[116] MALLET is open source software that operates on the command line. There is a method to implement MALLET directly in Python, but the results do not lend themselves to the visualization module used in this analysis. Working on the command line is not as straightforward as coding with Python. There is an invaluable resource for topic modelling with MALLET written by Shawn Graham, Scott Weingart, and Ian Milligan that details the process quite nicely, for those who would try their hand at it.[117] Using MALLET and the British empiricist corpus two topic models were produced containing 20 topics each. The first model made use of every cleaned document from the full corpus. These are topics that can be seen to express commonalities between the philosopher's vocabulary, writing styles, and philosophical concepts. However, seeing as this is a project devoted to decoding British empiricism, a second model was created using only the documents that contain the author's major epistemic projects—Locke's *An Essay Concerning Human Understanding*, Berkeley's *A Treatise Concerning the Principles of Human Knowledge*, and Hume's *An*

[116] MALLET was created Andrew McCallum and an army of graduate students. Available under the Open Source Software Licence: http://mallet.cs.umass.edu/index.php.

[117] Shawn Graham, Scott Weingart, and Ian Milligan, "Getting Started with Topic Modeling and MALLET," (*The Programming Historian,* 2012), https://programminghistorian.org/en/lessons/topic-modeling-and-mallet.

Enquiry Concerning Human Understanding. This second model delivers a more accurate picture of the topics discussed by the philosopher's related to empiricism alone.

Visualizations of topic models have been created in Python using the pyLDAvis module.[118] The visualizations produced by pyLDAvis give the individual word counts that Schmidt recommends. The module also plots the topics on a Principal Component axis, representative of the distance between them. Principal Component Analysis (PCA) is a close relative to MCA and will be explained in further detail in the following section. The topic points are also numbered and weighted. Point size directly reflects the topic list's overall frequency within the corpus. To use the MALLET results, they needed to first be written out on the command line and then imported into Python. George Mason University PhD candidate, Jeri E. Wieringa, has done a fantastic job of showing just how this is done in "Using pyLDAvis with Mallet."[119] While static images will have to be used for the analysis, pyLDAvis does create wonderful interactive visualizations that allow for the selection and viewing of each topic, as well as a relevance metric slider.[120]

Multiple Correspondence Analysis

Multiple Correspondence Analysis (MCA) is an extension of Geometric Data Analysis (GDA), an analytical method with a rich history in social and empirical science. Le Roux and Rouanet, in *Multiple Correspondence Analysis,* define GDA as an overarching classification of geometric analysis encompassing MCA, its counterpart PCA, and the two method's precursor, Correspondence Analysis (CA). CA is an efficient analytical tool for two-way

[118] pyLDAvis created by Ben Mabey. Available under the MIT Licence: https://pypi.org/project/pyLDAvis/.
[119] Jeri Wieringa, "Using pyLDAvis with MALLET," (*From Data to Scholarship*, 2018), http://jeriwieringa.com/2018/07/17/pyLDAviz-and-Mallet/.
[120] View the code here: https://nbviewer.jupyter.org/github/jasonrbradshaw/DBE/blob/master/TopicModel_MALLET.ipynb.

frequency tables, while PCA excels at numerical values and MCA is most often used for categorical variables.[121] Le Roux and Rouanet trace the history of GDA back to when CA was first established as a method in 1963. In these formative years, the technique was most popular in France. It was here that MCA and PCA were used in the 1970s to eventually become the "standard analysis of questionnaires."[122] Since then, GDA methods have come to be favoured around the world for this type of analysis and are often employed in statistical software. However, MCA has not fared as well as the other members of GDA. As noted by Le Roux and Rouanet, "this method, which is so powerful for analyzing full-scale research studies, is still rarely discussed and therefore is underused."[123] While CA and PCA have found their place in statistical literature, largely due to their numerical propensity, MCA has not received nearly as much attention despite its usefulness as an analytical tool in the social sciences.

So just how does a geometric model deliver results in the analysis of literature? Le Roux and Rouanet provide three key ideas that are at play in every GDA method: *geometric modelling*, two sets of data entry points (variables) define two cloud points in a geometric representation; *formal approach*, the methods are based on the mathematical theory of linear algebra; and *description first*, geometric modelling defines the model first by following the data points.[124] In order to explore the concept fully, an illustration of a typical MCA model has been provided in Figure 1. For MCA, the geometric model is contained upon a 2d surface that is split into 4 planes and delineated by two axis. The two sets of data entry

[121] Brigitte Le Roux and Henry Rouanet, *Multiple Correspondence Analysis*, (California, USA: SAGE Publications Inc, 2010), 3.
[122] Ibid., 4.
[123] Ibid., 4.
[124] Ibid., 2.

points can be made up of a variety of textual elements, but for the purposes of the explanation they will refer to *documents* and the *categorical variables* shared amongst them. Processing the data with MCA results in clouds of information, documents cluster around the categorical variables that they express. This is the driving force behind GDAs *description first* paradigm. The clouds define the model on the representative 2d surface. The boundaries are defined in the 4 planes by 4 *active variables* that are at the extremes of the raw data, with the rest of the categorical variables and documents placed within those bounds.[125] The researcher can begin making inferences by examining the distance between points. Documents that are relatively close in the geometric representation and share clusters of categorical variables can be considered similar, while those that lie on the outskirts of the cloud can be assumed to differ from the rest of the data. It is in this way

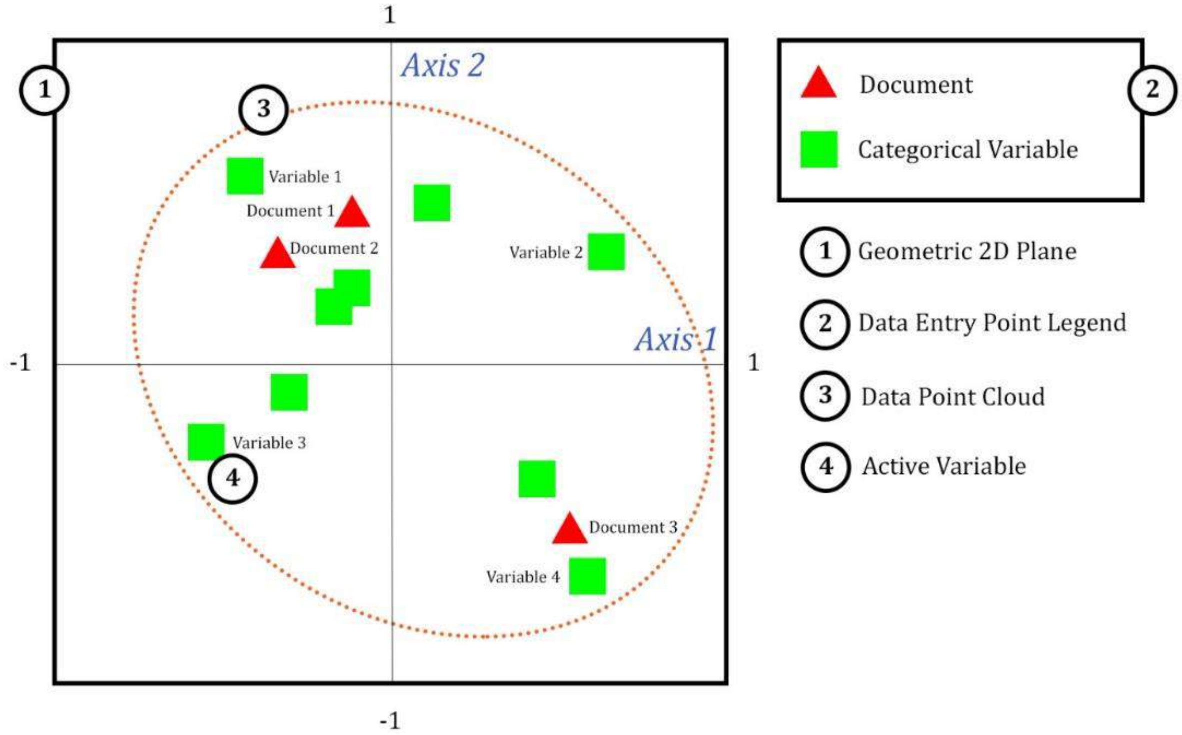


Figure 1. An example of Multiple Correspondence Analysis. Created in Photoshop CC.

[125] Le Roux and Rouanet, *Multiple Correspondence Analysis*, 10.

that MCA can give insight into a corpus. The distance between documents and the variable clusters, dependent upon which categorical variables are used, yields information on the similarity or dissimilarity of the texts being analyzed. This overview of MCA modelling and operation is by no means exhaustive, but it does highlight the core concepts of the method.

CA is certainly a well established method in the empirical sciences, but some researchers have raised concern over using the multivariate version of the model (MCA) to represent points in geometric space. In particular, the 2d planar surface can be somewhat misleading when it comes to interpretation. The distance between data points in MCA, like CA, is what fuels interpretation, but CA makes use of only 2-way frequency tables. MCA, on the other hand, contains as many categorical variables and associated data points as the researcher needs. Despite this, CA interpretations are often applied to MCA analysis, which can lead to faulty inferences. CA makes good use of the 2d geometric space provided, whereas MCA data points do not. In essence, MCA is a 3d model mapped onto a 2d surface, introducing "artificial dimensions into the full space."[126] Displaying MCA results in such a limited way can cause distances between some plotted data points to appear closer than they actually are. There are a number of ways to combat this inaccurate representation of MCA data on a planar surface. For example, the addition of relative size to weighted points facilitates MCA interpretation on a 2d surface.[127] Knowledge of software functionality can reduce inaccurate interpretation of MCA as well. As noted by Michael Greenacre, most software packages present CA and MCA "in much the same way, giving the same style of graphics, inertia decompositions and point contributions," resulting in a "misleading

[126] Michael J. Greenacre, "Interpreting Multiple Correspondence Analysis," *Applied Stochastic Models and Data Analysis* 7, no.2 (1991): 198.

[127] Le Roux and Rouanet, *Multiple Correspondence Analysis*, 9.

impression that the rules of interpretation for the two techniques are the same."[128] While the reduced dimensionality of MCA can certainly hinder interpretation, as long as the researcher is aware of the limitations of the model and software being used the method can provide valuable insight into the data. Moreover, modern visualization software can now enhance MCA research by plotting the points within a 3-dimensional space for interpretation.

MCA has been prepared for the analysis by following the freely available recipe and code from TAPoR, a site devoted to assisting DH scholars with text analysis.[129] TAPoR was created by Geoffrey Rockwell and the original code for the notebook was written by a University of Alberta colleague, Kynan Ly. The notebook makes use of several modules, including NLTK and Emre Safak's *MCA* to derive the geometric data points.[130] The entire British empiricist corpus was used for the MCA and the categorical variables were established through the Harvard general inquirer categories.[131] This is a downloadable CSV comprised of 182 columns of categories (affiliated with certain word groups such as 'power', 'positive', 'negative', etc.) and 11, 788 rows (the words that belong to each of the category dictionaries). The tokens from each of the British empiricist documents are compared to the Inquirer categories and the counts are then translated into a *pandas* dataframe.[132] These values are then run through the MCA module to determine their geospatial coordinates for plotting.

[128] Greenacre, "Interpreting Multiple Correspondence Analysis," 209.
[129] Original code by Kynan Ly for TAPoR.ca: http://tapor.ca/tools/672.
[130] MCA created by Emre Safak and is available under the BSD Licence: https://pypi.org/project/mca/1.0.3/.
[131] See: http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm.
[132] Pandas is an open source, BSD-licence library: https://pandas.pydata.org/.

Visualizations for the 2d and 3d MCA projections were coded using the Python matplotlib module, a library for creating static and interactive visualizations in the Python environment.[133] For the MCA, a traditional static 2d model was created along with an interactive 3d model. While the MCA data does not provide information for the z axis, the 3d model does offer some value when being manipulated to view the data from a different angle. The distance quite literally takes on new dimensions when represented in 3d space. By comparing and contrasting the 2d and 3d models, greater understanding of the document's connections to other documents and to the inquirer categories can be gained. Naturally, the images in the analysis will be static, but the interactive 3d visualizations can be accessed and manipulated on the *Decoding British Empiricism* GitHub repository.[134] The basic MCA that makes use of the author's ten documents and the first ten Harvard general inquirer categories has been supplemented with an extended example. The second visualization breaks the documents up into their chapters and plots all 182 categories. The result is a richer, more complex look at what MCA has to offer on a philosophical corpus. The results and potential findings from this additional analysis are massive, and so only select data points have been chosen for discussion. Readers are invited to view the full visualization of the extended MCA on the GitHub repository.[135]

[133] Hunter, John, et al. "Matplotlib: A 2d Graphics Environment." *Computing in Science & Engineering,* 9, no. 3 (2007): 90 – 95. https://matplotlib.org/.
[134] View the code here: https://github.com/jasonrbradshaw/DBE/blob/master/MultipleCorrespondenceAnalysis.ipynb.
[135] View the code here: https://github.com/jasonrbradshaw/DBE/blob/master/MCAadditional.ipynb.

FINDINGS & RECOMMENDATIONS

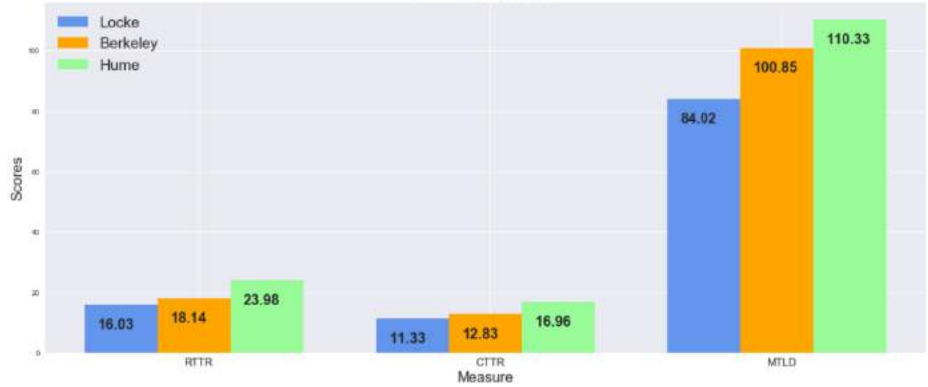
"The improvement of the understanding is for two ends; first, for our own increase of knowledge; secondly, to enable us to deliver and make out that knowledge to others."
John Locke, *The Works, vol. 2 An Essay concerning Human Understanding Part 2 and Other Writings*, p. 405.

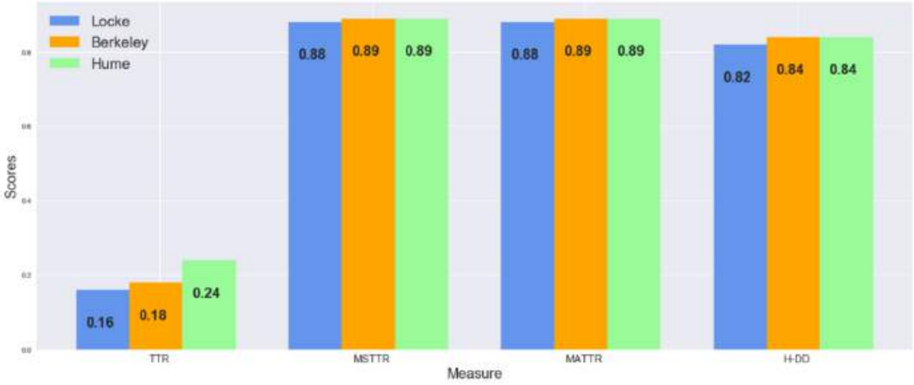
Findings

LRI

In the following visualization, the British empiricist's LRI measures have been expressed in 2 bar graphs. Some of the measure results produced by the LRI module tend towards lower decimal values (TTR, MSTTR, MATTR, H-DD), while others produce higher whole number values (RTTR, CTTR, MTLD). By separating these measures into 2 graphs the results are easier to interpret. The bars have been colour coded for each empiricist and the values are annotated upon them. Beneath the LRI measures is some information regarding the word frequency counts of Locke, Berkeley, and Hume. The top 10 frequency words have been given for each empiricist's full corpus and the documents making up that corpus, as well as a percentage of the total word count that the term comprises. Beneath the empiricist's names an alias has been given by the researcher. These aliases are meant to provide some insight into the lexical vocabularies of each empiricist and their meanings will become clear throughout this LRI analysis.

Lexical Richness Index





John Locke
The "Idea Man"

Full Corpus
Total Word Count: 386, 826
idea - 29, 846 ***(7.71%)***
make - 13, 978 ***(3.61%)***
mind - 10, 899 ***(2.81%)***
man - 10, 655 ***(2.75%)***
power - 10, 603 ***(2.74%)***
men - 9, 655 ***(2.49%)***
think - 9, 013 ***(2.32%)***
name - 8, 324 ***(2.15%)***
part - 7, 690 ***(1.98%)***
word - 7, 059 ***(1.82%)***

An Essay Concering Human Understanding
Total Word Count: 285, 205
idea - 29, 839 ***(10.46%)***
mind - 10, 788 ***(3.78%)***
make - 10, 728 ***(3.76%)***
name - 8, 037 ***(2.81%)***
think - 7, 849 ***(2.75%)***
man - 7, 728 ***(2.70%)***
men - 6, 968 ***(2.44%)***
knowledge - 6, 952 ***(2.43%)***
body - 6, 432 ***(2.25%)***
part - 6, 322 ***(2.21%)***

Two Treatises of Government
Total Word Count: 101, 621
power - 6, 344 ***(6.24%)***
right - 4, 696 ***(4.62%)***
make - 3, 250 ***(3.19%)***
man - 2, 936 ***(2.88%)***
government - 2, 871 ***(2.82%)***
father - 2, 760 ***(2.71%)***
men - 2, 703 ***(2.65%)***
adam - 2, 598 ***(2.55%)***
law - 2, 541 ***(2.50%)***
god - 2, 396 ***(2.35%)***

George Berkeley
The "Sensationalist"

Full Corpus
Total Word Count: 183, 266
idea - 3, 450 ***(1.88%)***
mind - 3, 064 ***(1.67%)***
thing - 2, 884 ***(1.57%)***
think - 2, 530 ***(1.38%)***
object - 2, 138 ***(1.16%)***
men - 2, 132 ***(1.16%)***
say - 2, 120 ***(1.15%)***
sense - 2, 022 ***(1.10%)***
see - 1, 946 ***(1.06%)***
make - 1, 830 ***(0.99%)***

Alciphron or: The Minute Philosopher
Total Word Count: 83, 460
men - 1, 616 ***(1.93%)***
man - 1, 280 ***(1.53%)***
think - 1, 226 ***(1.46%)***
god - 1, 010 ***(1.21%)***
say - 940 ***(1.12%)***
make - 892 ***(1.06%)***
thing - 888 ***(1.06%)***
religion - 860 ***(1.03%)***
good - 810 ***(0.97%)***
see - 786 ***(0.94%)***

A Treatise Concerning the Principles of Human Knowledge
Total Word Count: 35, 710
idea - 1, 640 ***(4.59%)***
mind - 898 ***(2.51%)***
thing - 706 ***(1.97%)***
sense - 508 ***(1.42%)***
without - 500 ***(1.40%)***
motion - 472 ***(1.32%)***
exist - 466 ***(1.30%)***
say - 426 ***(1.19%)***
body - 424 ***(1.18%)***
abstract - 398 ***(1.11%)***

An Essay Towards a New Theory of Vision
Total Word Count: 27, 795
object - 1, 160 ***(4.17%)***
distance - 840 ***(3.02%)***
visible - 696 ***(2.50%)***
sight - 636 ***(2.28%)***
idea - 616 ***(2.21%)***
eye - 592 ***(2.12%)***
tangible - 520 ***(1.87%)***
mind - 448 ***(1.61%)***
magnitude - 438 ***(1.57%)***
see - 390 ***(1.40%)***

Three Dialogues between Hylas and Philonous
Total Word Count: 36, 301
mind - 984 ***(2.71%)***
thing - 984 ***(2.71%)***
perceive - 722 ***(1.98%)***
idea - 702 ***(1.93%)***
exist - 700 ***(1.92%)***
think - 662 ***(1.82%)***
sense - 628 ***(1.72%)***
matter - 616 ***(1.69%)***
know - 582 ***(1.60%)***
existence - 544 ***(1.49%)***

David Hume
The "Humanist"

Full Corpus
Total Word Count: 372, 590
reason - 5, 597 ***(1.50%)***
great - 4, 935 ***(1.32%)***
nature - 4, 857 ***(1.30%)***
make - 4, 787 ***(1.28%)***
man - 4, 658 ***(1.25%)***
men - 4, 563 ***(1.22%)***
human - 4, 458 ***(1.19%)***
people - 3, 991 ***(1.07%)***
cause - 3, 914 ***(1.05%)***
time - 3, 864 ***(1.03%)***

An Enquiry Concerning Human Understanding
Total Word Count: 52, 310
reason- 1, 476 ***(2.82%)***
cause - 1, 410 ***(2.69%)***
object - 1, 375 ***(2.62%)***
nature - 1, 288 ***(2.46%)***
effect - 1, 247 ***(2.38%)***
experience - 1, 108 ***(2.11%)***
mind - 1, 036 ***(1.98%)***
idea - 928 ***(1.77%)***
human - 896 ***(1.71%)***
power - 847 ***(1.61%)***

An Enquiry Concerning the Principles of Morals
Total Word Count: 52, 021
man - 1, 001 ***(1.92%)***
sentiment - 951 ***(1.82%)***
society - 941 ***(1.80%)***
human - 842 ***(1.61%)***
us - 837 ***(1.60%)***
moral - 826 ***(1.58%)***
make - 826 ***(1.58%)***
reason - 817 ***(1.57%)***
justice - 812 ***(1.56%)***
think - 731 ***(1.40%)***

Dialogues Concerning Natural Religion
Total Word Count: 38, 734
cause - 1, 149 ***(2.96%)***
reason - 1, 020 ***(2.63%)***
human - 970 ***(2.50%)***
world - 831 ***(2.14%)***
nature - 806 ***(2.08%)***
god - 795 ***(2.05%)***
say - 695 ***(1.79%)***
think - 686 ***(1.77%)***
mind - 651 ***(1.68%)***
universe - 637 ***(1.64%)***

Essays, Moral, Political, and Literary
Total Word Count: 229, 525
great - 3, 941 ***(1.71%)***
men - 3, 184 ***(1.38%)***
government - 3, 173 ***(1.38%)***
people - 3, 004 ***(1.30%)***
make - 2, 877 ***(1.25%)***
time - 2, 870 ***(1.25%)***
man - 2, 677 ***(1.16%)***
state - 2, 315 ***(1.00%)***
reason - 2, 284 ***(0.99%)***
power - 2, 121 ***(0.92%)***

Figure 2. LRI measures and word frequency counts for Locke, Berkeley, and Hume.

One thing that is immediately apparent from the visualization is that LRI measures, even when visualized in a bar graph, are not particularly useful on their own. There is clearly a distinction between the three philosophers. Hume has the highest lexically 'rich' corpus, while Locke has the lowest and Berkeley's falls somewhere in between. But how does this discrepancy in LRI measure actually reflect the vocabularies of the British empiricists? To answer this question, one needs to turn to the word frequency counts associated with each respective philosopher. These counts play a pivotal role in not only the formation of LRI measures, but also in deciphering what these measures mean. For example, Locke, who has the lowest scoring Lexical Richness measures, has the highest percentages of words making up his word frequency counts. The 10th word in Locke's full corpus word frequency count ("word") makes up 1.82% of the entire corpus. Compare this to the empiricist with the highest LRI measure values, David Hume. Hume's first word in his full corpus frequency count ("reason") only comprises 1.50% of his entire corpus, a value lower than Locke's 10th word. Indeed, these types of percentages are consistent across each of the empiricist's documents and provide evidence for the accuracy of the LRI values. By no means does this suggest that Hume's writing is superior to Locke's in any way, rather it reflects a stylistic difference. This is an interesting trajectory of Lexical Richness scores from the three empiricists. Locke's contributions to the corpus were written towards the end of the 17th century, Berkeley's around the beginning of the 18th century, and Hume's at the tail end of the 18th century. This implies a correlation between time period and Lexical Richness value—it increases in the empiricist corpus over time.[136]

[136] It should be noted that *Decoding British Empiricism* only examines the works from three enlightenment empiricists. The case for the correlation between time period and LRI could be bolstered by introducing additional authors and sources into the analysis.

There are several possible explanations as to why such a trend has occurred. It could be argued that the empiricist's (excluding Locke) had the gained advantage of their predecessor's materials to draw upon and structure their own arguments against. Borrowing some of the vocabulary from past philosopher's while putting forth their own agenda could have led to increased Lexical Richness. Perhaps it is even a condition of the level of education each of the empiricist's received during their respective time frames. A century is not a huge gap of time in the history of literature, but it may have been enough for education to improve to where the empiricists were able to slightly surpass the last's Lexical Richness scores. It also became easier to access written material throughout this time period. Advances in printing and distribution of printed material may have given the later authors an advantage. Finally, though the empiricists may have been geographically close, they all received their formal educations from different nations. Locke and Berkeley both finished their secondary educations in their respective homelands (England and Ireland), while Hume split his time between the university of Edinburgh and the Jesuit college in La Flèche, France. Is there something about the education that these countries provided? Examining the discrepancy from a geographic perspective raises a couple of interesting points. Locke was educated in the capital region of the British Isles (England) and scored lower than the other empiricists. Hume continued his education on mainland Europe and received the highest measure of Lexical Richness. Was the cultural shift he experienced enough to increase his vocabulary to achieve these scores? Unfortunately, these are questions that the current research is unable to answer, but they do emphasize the ability of textual analysis techniques to "raise more questions than answers," questions that are deserving of further research.

There is one term shared across the philosopher's full corpora that stands out. It is not a uniquely philosophical word or one related to the empiricist projects in any obvious way. "Make" is featured in the top 10 word count frequency for Locke, Berkeley, and Hume. To discover why this common word is used so extensively by the British empiricists, it is best to turn to the definition of the term when used as a verb. Some of the definitions of "make" include: "to cause to happen to or be experienced by someone; to cause to exist, occur, or appear; to bring into being by forming, shaping, or altering material."[137] Given these definitions, the overrepresentation of "make" does make perfect sense in light of empiricist epistemology. Experience is pivotal to all three of these philosophers, and it is through sensory experience that ideas come to arise or appear in the mind. Furthermore, the ideas can be combined in many ways to make new and complex ideas. Creating (making) knowledge and understanding through sensory experience is the driving force behind these projects. This is a simple term that carries a lot of connotation when viewed through the lens of empiricism and it is reasonable to find it featured in the word frequency counts for each of the authors. Table 4 provides examples of the term in context through the use of concordances.[138]

[137] "make." *Merriam-Webster.com*. Merriam-Webster, 2011. Web. April 2019.

[138] View the code and full concordance lists here: https://github.com/jasonrbradshaw/DBE/blob/master/Concordances.ipynb.

Table 4. Examples of author concordances of the word 'make.'

Author	Concordance
John Locke	". . . all the acquaintance we can **make** with our own understandings . . ." ". . . the use of reason necessary to **make** our eyes discover visible objects . . ." ". . . be perceived by it so that to **make** reason discover those truths . . ."
George Berkeley	". . . what can there be that should **make** us believe or even suspect . . ." ". . . it follows that the judgment we **make** of the distance of an object viewed . . ." ". . . proceed unto the eye reason would **make** one think that object should appear . . ."
David Hume	". . . to give it grace or cause it to **make** any impression on the hearers . . ." ". . . nature has perhaps intended to **make** us sensible of her authority . . ." ". . . of superior powers because he has to **make** that inference from what he knows . . ."

Another general finding worth mentioning from the word frequency counts is the prevalence of masculine language and pronouns. Words like "man" and "men" feature extensively in each of the empiricist's word frequency counts. This is not surprising given the time period that these major philosophical works were written. The 17th and 18th centuries were not particularly observant of the roles and needs of women. Even in the 21st century, women and visible minorities often lack the respect that they deserve. It would be remiss to not acknowledge these word counts from the empiricists. It is interesting to note that while Hume does make excessive use of male pronouns, the philosopher also has word frequency counts reflecting non-gendered pronouns ("people" and "human"). This is a finding that will be returned to during the discussion on Hume's LRI measures and word frequencies.

John Locke has been dubbed as the 'Idea Man.' This playful alias refers to both his vast philosophical work on the constitution of ideas, as well as to the excessive use of the term itself within his corpus. "Man" alludes to the overrepresentation of male pronouns, which has just been discussed. The term "Idea" dominates Locke's discourse, at least in regards to his writing on empiricism. Indeed, "Idea" is used more than twice as much as the second highest term frequency in both Locke's full corpus and *An Essay Concerning Human Understanding*.[139] To emphasize the point, here is a sample paragraph from the text:

> By determinate, when applied to a simple *idea*, I mean that simple appearance which the mind has in its view, or perceives in itself, when that *idea* is said to be in it: by determined, when applied to a complex *idea*, I mean such an one as consists of a determinate number of certain simple or less complex *ideas*, joined in such a proportion and situation as the mind has before its view, and sees in itself, when that *idea* is present in it, or should be present in it, when a man gives a name to it [emphasis mine].[140]

In this single, albeit rather long, sentence, "Idea" is used by Locke five times. For a sentence that is 98 words longs the term comprises over 5% of the entire count, and this is before any stopwords have been removed. Philosophy is a discipline that values concise and accurate stylistic writing choices. The best term to describe something will be employed, regardless of its frequency within a document. Ideas play a central role in Locke's epistemology through sensory experience. As was discussed in the section on Locke, ideas can be simple, complex, and originate from sensory experience and self-reflection. In the words of Carlin, "sensation provides ideas of qualities of bodies, things external to the mind, while reflection furnishes the mind with ideas of things internal, namely, the operations of the mind itself."[141] Locke's word frequency count reflects his empiricist agenda. Idea, the central component of understanding through the senses, dominates the word frequency chart. "Mind," where ideas come to rest, closely follows. "Think" and

[139] It should be noted that all of the terms in the word frequency counts and LRI have undergone lemmatization or have been manually collapsed to their singular form for the analysis. This means that both plural and singular forms of a word are being counted. For example, "Idea" and "Ideas" both add to the count for the top frequency term.

[140] Locke, *An Essay Concerning Human Understanding*, 20.

[141] Carlin, *The Empiricists: A Guide*, 87.

“Knowledge” also make appearances in the top 10 word counts. The counts detail the most important aspects of Locke’s concept of the understanding.

There are a couple of noteworthy points that can also be taken from Locke’s *Two Treatises of Government*. The word use is overly patriarchal (“man”, “men”, “father”, “adam”), again not totally surprising given the timeframe and the systems of patriarchalism that formed government during this period. What does stand out is the mixture of language used to describe governance (“power”, “right”, “government”, “law”) and religious terminology (“father”, “adam”, “god”, “law”). Many of the concepts put forth by Locke in *Two Treatises of Government* revolve around the relationship between human beings and God. The philosopher views this relationship as a way to justify governance, as well as the liberties that all persons should be afforded. In a paraphrase of Locke by Peter Laslett in his forward to the *Two Treatises*: “When men think of themselves as organized with each other they must remember who they are. They do not make themselves, they do not own themselves, they do not dispose of themselves, they are the workmanship of God. They are his servants, sent into the world on his business, they are even his property.”[142] God plays an important role in both Locke’s philosophical and political agendas. The word frequency count of the *Two Treatises* provides further evidence as to how deep this connection truly is.

Recalling the discussion on George Berkeley and his empiricist philosophy, it would be fair to say that many considered him sensational. Berkeley’s immaterial idealism has raised some eyebrows and his denial of anything truly existing other than ideas in minds

[142] John Locke, *Two Treatises of Government*, ed. Peter Laslett, (New York: Cambridge University Press, 2003), 93.

has set him apart from Locke and Hume. This label can be applied to his word frequency count as well. In the figure, George Berkeley's alias is 'the Sensationalist.' However, it is used here in a very literal sense. Berkeley's full corpus contains terms that are typical of British empiricism ("idea", "mind", "think"), as well as a collection of words that describe the senses and their operations ("say", "sense", "see", "thing", "object"). It is true that the senses and sensory experience play an important role in the projects of each of these philosophers, but Berkeley's counts carry a strong representation of these types of words. One possible explanation is that the philosopher spent a great deal of time appealing to the senses in defense of his immaterial idealism. Berkeley denied the existence of an external world with external objects and, quite often, justified this position by referring to the human senses. For example, take the following passage from *A Treatise Concerning the Principles of Human Knowledge*: "As for our senses, by them we have the knowledge only of our sensations, ideas, or those things that are immediately perceived by sense, call them what you will: but they do not inform us that things exist without the mind, or unperceived, like to those which are perceived."[143] Sense and sensation are important components of Berkeley's argument against external bodies and the word frequency counts provide further evidence of this. To be fair, *An Essay Towards a New Theory of Vision* is a text all about sensory organs and so this may have impacted the full corpus results. However, this document does have the lowest word count in the Berkeley corpus at 27, 795 words and so its influence would be slight. Even though the document may not be as epistemically involved as the *Principles of Human Knowledge*, it does define who Berkeley was as a philosopher and enlightenment thinker.

[143] Berkeley, *Principles*, 39.

There is one term from Berkeley's *Principles of Human Knowledge* in the top 10 count that stands apart. "Motion" is the 6th most used term in the document, comprising 1.32% of the total word count. This term does not have any obvious relation to empiricism, but it does highlight an important facet of Berkeley's philosophy that has made its way into several of his writings. During the enlightenment the distinction between philosopher and natural scientist was not as well defined as it is now. Many enlightenment thinkers that are associated with hard science tried their hand at philosophy. Likewise, many names from this era that are now found in philosophy texts made important contributions to the natural sciences. Berkeley directed numerous criticisms towards his contemporary, Isaac Newton. Particularly in regards to Newton's formulation of absolute space and motion.[144] Berkeley's *Principles of Human Knowledge* devotes several paragraphs to exploring and denouncing the ideas of absolute space and motion put forth by Newton. Rejecting an absolute space that exists independently of perception also furthered the philosopher's empiricist agenda. Absolute space contradicts the existence of an omnipotent and omnipresent God, who, for Berkeley, is the source of all human ideas. Berkeley rejects Newton's view because it would imply that God just is absolute space, or that there are two independent and external things in existence that are "eternal, uncreated, infinite, indivisible, immutable," which the philosopher finds to be absurd.[145] Berkeley's word frequency counts are revealing. They highlight the major focus of this empiricist's work, the features that define Berkeley from the other philosophers, as well as his involvement and critique of the natural sciences being put forth by other enlightenment thinkers of the time.

[144] G. J. Whitrow, "Berkeley's Philosophy of Motion," *The British Journal for the Philosophy of Science*, IV, no. 13 (1953): 37.
[145] Ibid., 41.

Finally there is David Hume, 'the Humanist.' Not only is this a cleverly disguised pun, but the alias also accurately reflects the top 10 terms from the philosopher's word frequency count. Hume's corpus is guilty of patriarchal language, like the empiricists thus far examined, but it is interesting that non-gendered pronouns and terms also feature in the full corpus top 10 count ("human", "people"). Hume as humanist likewise extends to some of the other language found in the chart ("reason", "nature"). The word "reason" is the most commonly used term in the philosopher's full corpus, making up 1.50% of the total wordage. Reason is a uniquely human trait and, following Hume's religious scepticism, it is a means to derive truth without appealing to a God. As has previously been stated, Hume stands apart from the other two British empiricists because of this staunch religious scepticism. Berkeley was a philosopher and clergyman who appealed to God as the source of human ideas and Locke gave an account of how even simple ideas can be combined to arrive at the existence of an omnipotent God. Hume has been read by some researchers as one of the earliest modern humanists. In a lecture given by Amyas Merivale, University of Oxford philosophy and computer science professor, the academic recounts the story of Hume's change from a pessimist of human nature to a true believer in the benevolence of human beings after being persuaded by an argument made by Joseph Butler. However, Merivale finds it interesting that this change in no way persuaded Hume to believe in God, but instead to incorporate it into his anti-theist rhetoric and become "the first example of a recognisably modern humanist".[146] Hume believed the benevolence and empathy of human nature to simply be a *principle* of human nature, it need not be a trait inherited from being

[146] Amyas Merivale, "How David Hume became the First Modern Humanist, " (Lecture, Conway Hall Ethical Society, London, EN, June 5, 2016), para 2.

created in the likeness of a benevolent God. It is for this reason that religious terminology is absent in Hume's word frequency counts, and perhaps why patriarchal language, while dominant, is coupled with terms that are more human-centric. Merivale's reading of Hume as a modern humanist is certainly reflected in the analysis.

David Hume's word frequency counts also verify some of the philosopher's core empirical concepts. In the count for *An Enquiry Concerning Human Understanding*, "cause", "effect", "experience", "idea", and "mind" all make appearances. Cause and effect are the underlying forces behind Hume's idea of Necessary Connection. Sensory experience causes impressions, which in turn form less vivid ideas in the mind. Likewise, the counts for *An Enquiry Concerning the Principles of Morals, Dialogues Concerning Natural Religion,* and *Essays, Moral, Political, and Literary* all contain terms that reveal something about the philosophical agendas in the documents.[147] These word frequency counts give surprisingly accurate insight into the content. This trend is not reserved to Hume's corpus alone. Indeed, both Locke and Berkeley have similar word frequency counts for all of the texts from their corpora. Once superfluous textual elements like pronouns and quantifiers have been removed from the corpora, the word frequency counts produce accurate reflections of the empiricist's philosophical endeavours. It could even be argued that with little prior knowledge of British empiricism one could begin to form a basic understanding of its main philosophical tenets and motivations. Such discovery would not be at odds with empiricism. Rather, it would reinforce the deduction, observation, and scientific mindset

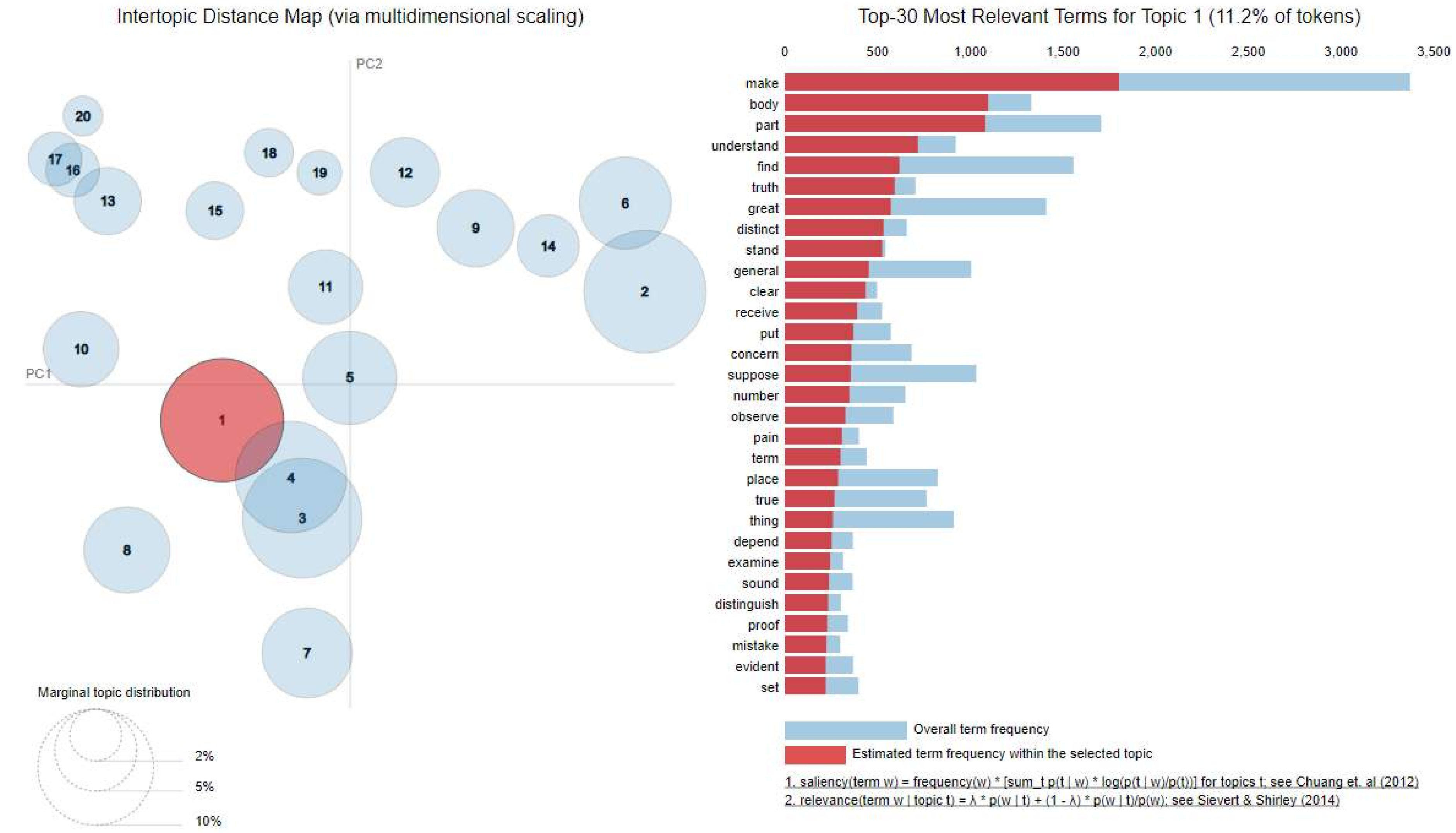
[147] Revealing terms from these word frequency counts are, respectively: "sentiment", "society", "moral", "reason", "justice"; "human", "world", "nature", "god", "universe"; "great", "government", "state", "power".

promoted by this school of philosophy—forming ideas through simple experiential impressions.

Topic Models

As was discussed in the methodology section, two topic models were created for the British empiricist documents using MALLET. The first topic model contains all of the documents that make up the corpus, while the second model contains only the main texts related to the philosopher's empiricism. An image of the pyLDAvis visualization has been included, followed by a breakdown of the 20 MALLET topic terms. The lists include a Topic ID number that corresponds to the numbered points on the pyLDAvis visualization, a Topic Name created by the researcher to facilitate analysis, the Total Token percentage of the topics for the given corpus, and the 20 Terms that make up the topic. There are a couple of things that should be noted about the assignment of topic names to the lists produced by MALLET. First, there is no real science behind this process. The topic names, as other scholars have discussed, are somewhat subjective. The names that have been assigned may be quite different than what the reader may have chosen to represent the term lists. Second, the connection between terms in a list is not always clear, but an approximation can be made by identifying a topic that the majority of terms in a list speak to. The task becomes increasingly difficult as the total percentage of tokens in the topic model decreases. For example, topic number 17 in the Full Corpus model ("**Indulgence**") contains several terms that are indulgent in nature ("riches", "increase", "art", "foreign", "delicacy", "emperor", "empire", "scarcely") along with several terms that do not really conform to the topic name ("history", "present", "found", "judgement", "preserve", "proceed", "farther"). This particular topic could have also been named something along the lines of **Imperialist**

Expansion, but given the close mixture of words it really comes down to determining a connection between the majority of the terms. The topic naming convention also proved difficult for the second topic model that made use of the author's main empirical projects. As is to be expected, many of these topics shared terms and were similar in nature. Regardless, these topic names should not be taken as objectively representative of the data. They are simply a convenient way to differentiate the term lists and provide an initial point for discussion. By nature, topic modeling is an exploratory type of analysis, much like LRI. The topic term lists do change each time the code is run, sometimes slightly and sometimes significantly. Therefore, in order to strengthen the points brought up in the findings, close reading examples of the texts have been scattered throughout this section.



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

Figure 3. pyLDAvis full corpus topic model.

Table 5. Full corpus MALLET term lists and topic names.

Topic ID #	Topic Name	Total Token %	Terms
1	**Physical Bodies**	11.2%	make body part understand find truth great distinct stand general clear receive put concern suppose number observe pain term place
2	**Ideas**	11.1%	ideas idea mind knowledge man men word things simple complex proposition power reason call substances action sort motion species real
3	**Epistemology**	10.6%	reason find natural nature present principles suppose matter common kind question produce man subject force give order form life mankind
4	**Life**	9.2%	make time give good bring sense leave nature true accord state man act live establish require judge long continue lose
5	**Knowledge**	6.5%	men power place learn great number age part mind rule author manner mention follow passions rise concern effect bear easy
6	**Sensation**	6.3%	mind perceive ideas sense exist things existence matter idea thing spirit motion real substance abstract qualities conceive colour god word
7	**Rule**	6%	power father government man men god adam author children law people authority dominion society nature laws give property make title
8	**Ancient Politics**	5.5%	great government essay lib state party hume liberty edition page money slave commerce industry roman observe war ancient arts rome
9	**Dialogues**	4.4%	men things god man religion don alciphron free sense world faith minute euphranor good philosophers isn virtue crito doesn mind
10	**Geopolitics**	4.2%	people public time life country write house support law political spirit modern government pleasure manners nations civil court authority land
11	**Human Nature**	4.1%	man human interest character laws sentiments society passion view case call arise beauty moral advantage virtue influence situation regard real
12	**Operations of the Mind**	3.6%	effect object experience human action philosophy nature idea power event mind instance fact concern operations reason regard infer derive sense
13	**Power Dynamics**	3.3%	world hand thing private enjoy mankind govern peace master death public live belong great submit lay force remain possession put
14	**Visual Sensation**	2.9%	object distance sight visible eye figure ideas perceive tangible mind hath suggest idea magnitude greater extension point touch things doth
15	**Morality**	2.5%	justice sentiment rule society feel general virtues moral happiness social humanity property praise source approval person virtue hume utility qualities
16	**Establishments**	2.2%	possess form constitution difference account afford commonly senate succession labour apt silver acquire magistrates arm occasion justly treatise family establishment
17	**Indulgence**	2.1%	history present general encrease art foreign empire employ found riches till scarcely judgment preserve pretend sentiment emperor delicacy proceed farther
18	**Metaphysics**	1.8%	world human god universe cleanthes religion argument philo mind system animal work order reply experience religious attribute gods existence design
19	**Contractions and Proper Names**	1.5%	dont isnt things doesnt admit immediately arent understand point people pain object exist philonous claim hylas shape experience show obvious
20	**Governmental Structures**	1.2%	sec temper circumstance tory commons lose provinces peculiar miles fund perfection literary levy reference extraordinary emulation encreased conclude mortgage recourse

What makes an immediate impression from the full corpus topic model is how accurately the first three topics ("**Physical Bodies**", "**Ideas**", "**Epistemology**") reflect the shared empiricist goals of Locke, Berkeley, and Hume. These term lists emphasize the attainment of knowledge through bodily senses that then inform and create different sorts of ideas in the mind, the main function at work in empiricism. There is something to be said about the space represented by the pyLDAvis intertopic distance map as well. **Physical Bodies** and **Ideas** are separated significantly. This is appropriate given the distinction made by the empiricists between the mind and the body.[148] While Locke may have rejected rationalist innatism due to the fact "that *all* ideas are acquired from experience," he did have to concede to the mind-body dualism proposed by Descartes and others.[149] Dualism was a result of the mechanistic view of the universe that was gaining in popularity during the 18th century, for both the empiricists and the rationalists. Looking again to the intertopic distance map, **Physical Bodies** and **Epistemology** are relatively close. This makes perfect sense as empiricism appeals to sensory experience as its primary source of knowledge.[150] Unlike mind-body dualism, this facet of empiricism is in stark contrast to the epistemology of the rationalists, who instead express "the view that *reason* (as opposed to sensory experience) is the ultimate justification of our knowledge".[151] It is interesting that

[148] For example, "If identity (to instance that alone) be a native impression, and consequently so clear and obvious to us that we must needs know it even from our cradles, I would gladly be resolved by any one of seven, or seventy years old, whether a man, being a creature consisting of **soul** and **body**, be the same man when his body is changed?" [emphasis added] (Locke, *Human Understanding*, 68.)

[149] Carlin, *The Empiricists*, 83.

[150] For example, "Thus greater Confusion having been constantly attended with nearer Distance, no sooner is the former **Idea perceiv'd**, but it suggests the latter to our **Thoughts**. And if it had been the ordinary Course of Nature, that the farther an *Object* were placed, the more Confused it shou'd appear, it is certain, the very same **Perception** that now makes us think an Object approaches, would then have made us to imagine it went farther. That **Perception**, abstracting from *Custom* and *Experience*, being equally fitted to produce the **Idea** of great Distance, or small Distance, or no Distance at all," [emphasis added] (George Berkeley, *An Essay Towards a New Theory of Vision*, ed. David R. Wilkins, (Dublin, 2002), 4.)

[151] Carlin, *The Empiricists*, 1.

these three topics take precedence as the full corpus contains many different types of documents from the philosophers. Including texts like dialogues, political essays, and moral philosophies. It suggests that empiricism, at least for these three enlightenment thinkers, played a large role in all of their philosophical endeavours. The terms that define the forces behind empiricist thought must creep into the philosopher's written work on other subjects. An injection of empiricist ideology can certainly be found in Berkeley's, *Three Dialogues between Hylas and Philonous*. The dialogues are presented in a conversational tone that often reflects on theology, but Berkeley uses the medium as a voice for his empiricism. For example, in the first dialogue Hylas questions Philonous on his "most extravagant opinion" that there is no such thing as material substance, stating that there is nothing "more fantastical, more repugnant to common sense, or a more manifest piece of scepticism, than to believe there is no such thing as matter".[152] This is clearly alluding to Berkeley's own immaterial idealism. This empiricist concept, along with many others, are expressed throughout the dialogues. This is likewise seen in Locke's *Two Treatises of Government* and Hume's *Essays*. Empiricism is not just a past time for these philosophers, it is a core component of their world view that informs their other writings, resulting in empiricist topics dominating the top of the term lists.

There are several other topics related to empiricism that are not as prominent as the three so far examined. **Knowledge, Sensation**, **Operations of the Mind**, and **Metaphysics** are more particular examples of the philosopher's individual expressions of empiricism. The **Sensation** term list readily brings Berkeley to mind. There are a number of terms that were originally expressed by Locke ("ideas", "substance", "qualities"), but

[152] Berkeley, *Three Dialogues*, 8.

Berkeley often called upon these concepts in criticisms of his predecessor. Coupled with the terms that conform to Berkeley's empiricist agenda, it is reasonable to assume that the term list reflects the work of this unique, theologically inclined philosopher ("matter", "spirit", "motion", "real", "colour", "god"). **Operations of the Mind** has a distinctly Humean voice. Several of the terms directly reference his empiricist philosophy or his language of deduction ("effect", "object", "nature", "power", "event", "instance", "fact", "operations", "reason", "regard", "infer", "derive"). **Knowledge** and **Metaphysics** are a little more difficult to pin down to only one of the philosophers. **Knowledge** is comprised of terms that are objective and quantifiable, while the language in the **Metaphysics** topic is primarily religious in nature. It could be argued that **Metaphysics** is, more than not, representative of Hume's *Dialogues Concerning Natural Religion*. Religious terminology was used by all of the empiricists in their major epistemic projects, but the presence of the proper name "philo" in the term list suggests that these terms could be from Hume's *Dialogues Concerning Natural Religion.*

Another category of topics that are prominent in the full corpus analysis are those relating to politics. **Rule**, **Ancient Politics**, **Geopolitics**, **Power Dynamics**, **Establishments**, and **Governmental Structures** all contain term lists that are highly politicized in nature. These results do make sense given the scope of the full corpus. At least half of the British empiricist's documents contain philosophical material relating to government and politics of the time. Locke's *Two Treatises of Government* make up about a third of the philosopher's total word count contribution to the corpus. In the *Treatises*, Locke's main tenet of governance stems from the expression of God's will made manifest through the Law of Nature: "Conceived of as a law (the law of nature), or almost as a power, it is sovereign over all human action. It can dictate to a man as conscience does and to more than one man in the social situation, since it is given by God to be the rule betwixt man and man."[153] Locke's doctrine of the law of nature is reflected by the term lists associated with **Rule** ("power", "father", "man", "men", "god", "children", "law", "authority", "dominion", "society", "nature", "give") and **Power Dynamics** ("govern", "master", "death", "public", "live", "belong", "submit", "lay", "force", "possession"). The law of nature places the will of God above all else, it informs humankind on how society should be run and creates space for hierarchical, patriarchal power dynamics. **Geopolitics** can be accounted to Locke as well. The philosopher spends a deal of time on ownership and property in the second half of the *Treatises*. In particular on how property can function as a means to restrict freedoms from certain groups. In the words of Laslett, "a slave lacks all political rights because he is incapable of property: despotical power, not properly political at all, can only be exercised over the propertyless."[154] Property, for Locke, is not only granted via the will of God through the law of nature, but it imbues the owners of property with similar powers. Hume's *Essays, Moral, Political, and Literary* make up the majority of the philosopher's contribution to the full corpus. Although not strictly political in nature, there are many references to Hume's own ideas regarding governance. The term list for **Ancient Politics**, which also happens to contain the philosopher's name, have words that are present in several of the political essays that are part of the anthology. For instance, Essay III ("That Politics may be Reduced to a Science") contains a number of references to the governance of medieval Europe and the absolute rule and conquest by ancient Rome. As for the last 2

[153] Locke, *Two Treatises of Government*, 95.
[154] Ibid., 102-103.

political topics, **Establishments** and **Governmental Structures**, the lists contains terms that are used often by both authors.[155] The terms are fairly standard for any political discussion. These topics are clustered relatively closely on the intertopic distance map in the top left corner. **Rule** and **Ancient Politics** are somewhat outliers, but still fall within the left hand side of PC2. Why such a distribution has taken place is not entirely clear. **Rule** does contain many terms that are theological, making the separation from the other topics reasonable. However, **Ancient Politics** shares terms with **Geopolitics** and **Power Dynamics**. There may just be some deeper function of the distribution matrix at play that is not readily obvious.

Hume's moral philosophy finds a voice in two of the topics from the full corpus: **Morality** and **Human Nature**. The enlightenment revolutionized many perceptions and attitudes towards the external world. Copernicus revolutionized the way people saw their place in the world after discovering that the earth moves around the sun, Newtons laws of motion opened the doors for a mechanistic world view, and some have argued that Hume marked the beginning of modern moral philosophy. In *Hume's Place in Moral Philosophy*, Nicholas Capaldi notes the shift from the belief in an external source of morality (God) to an inward source during this time, and that "the *first issue* faced by modern moral philosophers is to determine whether there is an *inwardly discoverable moral domain*."[156] Modern moral philosophers debated the faculty to which this source of moral insight

[155] For example, "Most writers, that have treated of the British government, have supposed, that, as the lower house represents all the **commons** of Great Britain, its weight in the scale is proportioned to the property and power of all whom it represents," (David Hume, *Essays Moral, Political, Literary*, ed. Eugene F. Miller, (Indianapolis 1987), 39) ; "So that, in our author's sense, all that was said here to Noah and his sons, gave them no dominion, no property, but only enlarged the **commons**; their **commons**," [emphasis added] (Locke, *Two Treatises of Government*, 28.)

[156] Nicholas Capaldi, *Hume's Place in Moral Philosophy*, (New York: Peter Lang Publishing Inc., 1989), 2.

belonged. One camp believed that morality was derived from *reason*, while the other believed that it originated from *sentiment*. Hume straddled the line between these two camps, with the likes of Joseph Butler and Francis Hutcheson, all of whom "opted for some combination but one in which sentiment was the dominant element."[157] We see this view reflected in the Hume's moral topics. "Sentiment" makes its way into the lists for both **Morality** and **Human Nature**, along with a number of associated terms like "passion", "moral", "feel", and "source".[158] Capaldi notes that Hume was opposed to the utilitarianism favoured by some of his predecessors, like Thomas Hobbes. Hume voiced a "persistent rejection of the move to reduce morality to self-interest."[159] There are echoes of Hume's opinion in the topics, through terms like "utility" and "happiness". Hume rejected self-interest as the source of morals for a view of morality as a developmental process that was both natural and artificial. Morals are artificial because they are "partly conventional, being social products over time" and they are natural due to our "capacity for sympathetic identification which can reconcile conflicts among self-interest, moral motives, and the social good."[160] This emphasis placed by Hume on the artificial and natural origins of morals is represented in the topics with terms like "interest", "character", "laws", "society", "situation", and "social".

[157] Capaldi, *Hume's Place in Moral Philosophy*, 3.

[158] For example, "The ancient philosophers often *assert* that virtue is nothing but conformity to **reason**; but their writings generally suggest that they think that morals derive their existence from taste and **sentiment**. And on the other side, our modern enquirers talk a great deal about the 'beauty' of virtue and 'ugliness' of vice, seeming to imply that their basis is **sentiment** or **feeling**; but they have commonly tried to account for the virtue/vice distinction by metaphysical **reasonings** and by deductions from the most abstract principles of the understanding," [emphasis added] (David Hume, *An Enquiry into the Sources of Morals*, (1751), 2.)

[159] Capaldi, *Hume's Place in Moral Philosophy*, 6.

[160] Ibid., 6.

The final set of topics are directly related to the empiricist's individual projects and stylistic differences between the documents. **Visual Sensation** is made up of terms related to Berkeley's *An Essay Towards a New Theory of Vision* ("object", "distance", "sight", "visible", "eye", "figure", "perceive", "tangible", "magnitude", "extension", "point"). The **Dialogues** topic contains some of the names and stylistic structures that are commonly found in philosophical dialogues, in this case Hume's *Dialogues Concerning Natural Religion* along with Berkeley's *Alciphron* and *Three Dialogues between Hylas and Philonous* ("god", "religion", "alciphron", "faith", "minute", "euphranor", "philosophers", "virtue", "crito"). MALLET also managed to successfully sort through the tokens to produce **Contractions and Proper Names**, superfluous textual elements ("don't", "isn't", "doesn't", "aren't", "philonous", "hylas"). There is not much to be said about this selection of topics other than the term lists accurately reflect specific materials from the empiricists. The topics are also scattered across the intertopic distance map, as they do not share much in common with one other. If anything, these topics showcase the power of MALLET. The software was able to successfully analyze these tokens and create term lists that are representative of very particular philosophical concepts and styles found within the corpus, as well as the messy contractions that did not belong anywhere else.

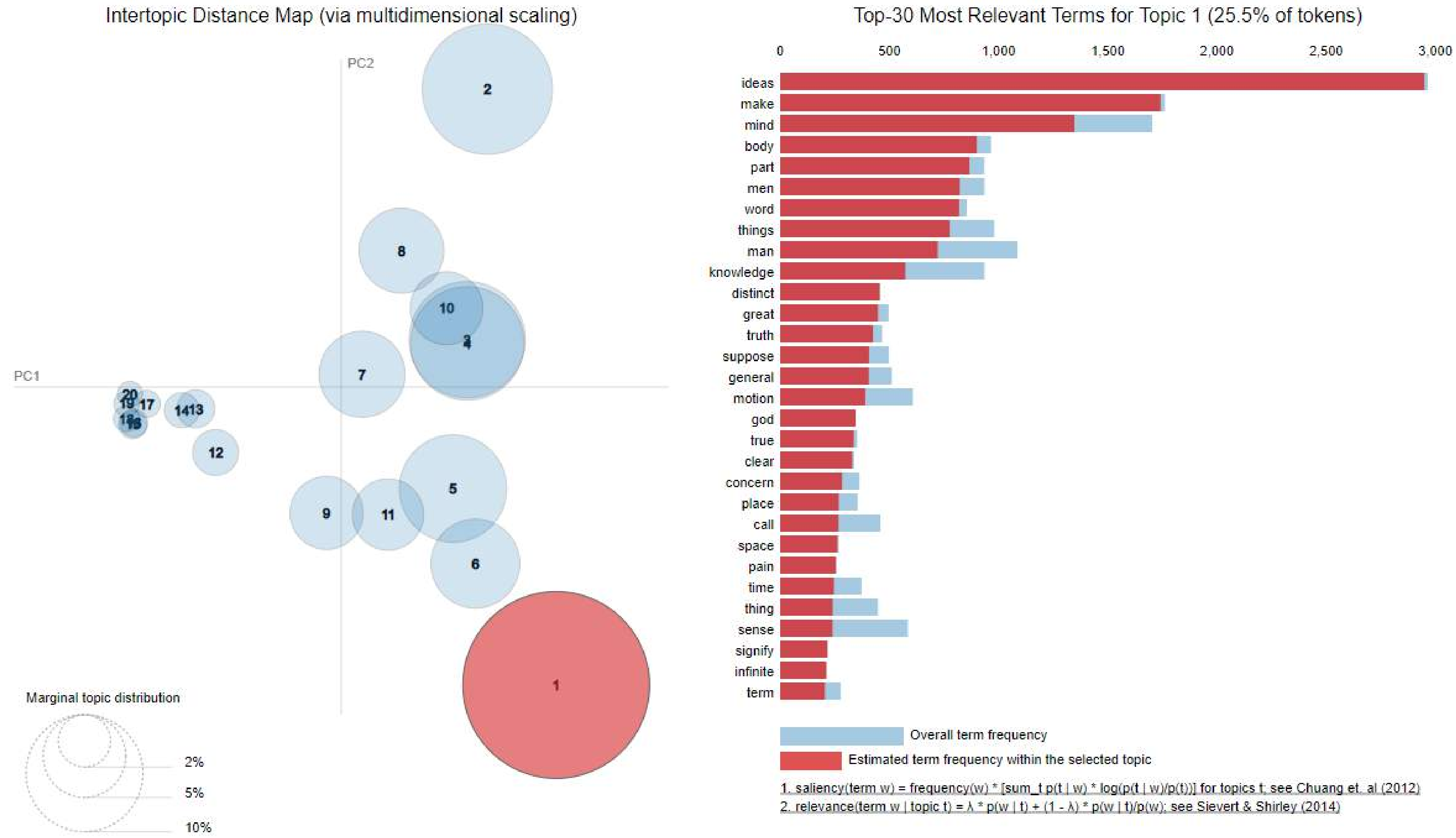

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
2
8
10
3
4
7
20
19
17
18
15
14
13
12
5
9
11
6
1
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (25.5% of tokens)
0
500
1,000
1,500
2,000
2,500
3,000
ideas
make
mind
body
part
men
word
things
man
knowledge
distinct
great
truth
suppose
general
motion
god
true
clear
concern
place
call
space
pain
time
thing
sense
signify
infinite
term
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 4. pyLDAvis empirical projects topic model.

Topic ID #	Topic Name	Total Token %	Terms
1	**Empiricism**	25.5%	ideas make mind body part men word things man knowledge distinct great truth suppose general motion god true clear concern
2	**External World**	12.4%	idea reason matter qualities real number understand thoughts find world common connexion soul move depend evident end imagine change rest
3	**Human Understanding**	9.9%	power find man doubt understand present case examine species learn liberty carry certainty greater instance experience easily state serve life
4	**Ideas**	9.4%	simple proposition knowledge action innate substances produce show assent good determine call species reason mens happiness complex mix gold truths
5	**Formulation of Ideas**	8.4%	nature object mind principles force discover follow mankind draw form principle proceed age account long attribute pretend dispute proper immediately
6	**Observing External Bodies**	5.8%	exist perceive abstract spirit sense thing things figure hath substance existence motion effect extension answer matter mind frame observe finite
7	**Measurement**	5.4%	sort complex stand duration modes agreement essences receive measure disagreement thoughts identity signification nominal rational specific yellow observation obscure frame
8	**Human Judgement**	5.3%	idea put essence pleasure whilst judge natural capable moral bring wrong greatest judgment eternal desire children relations belong maxims discourse
9	**Objectivity**	3.9%	give observe proof reflection view commonly voluntary universe country step probable endeavour water stop organs precise design circle minute copy
10	**Passing Judgement**	3.9%	reason effect experience human concern sense fact action suppose event infer events testimony philosophy argument instance derive similar necessity result
11	**Creation**	3.7%	produce natural manner general motion place order existence philosophers sect doctrine acknowledge regard external laws conceive impossible opinion ascribe remain
12	**Observation**	1.5%	appearance employ experiment ignorant direct acquire guide single report opposite secret chain hope chiefly blame gods trust return arrive hundred
13	**Deduction**	1%	inference conduct found conclusion common part command public conjecture pronounce diminish behaviour bring government security requisite attempt superstition solution operation
14	**Enlightenment Language**	0.9%	shew line humane unthinking absolute support intirely wit inert abstraction plain phænomena unperceived reflexion accidents innumerable premise scepticism forego separate
15	**Hedonism**	0.6%	obvious confess public partake privilege tune nestor annual edge diminish folly labour peripatetick scanty minutes drunkenness usefulness meat boast vouch
16	**Ancient Empiricism**	0.6%	extraordinary process priori criminal detect actuate facts occur epicurus stronger extremely spirit happen crime attendant superiority brute merit philosopher medium
17	**Earthly Desires**	0.5%	sober debate advantage excuse chessboard flatter testimony deal crave woman aright piety quicksighted luck imputation implicit assign finite personality backwards
18	**Made by God**	0.5%	play err thisthat atoms propagate sing star periodical decompounded grave chimerical dim antecedent conceptions artificial post million readiness condemn component
19	**Biblical**	0.5%	busy desire adam lay familiarly malleableness yards clock infants allowance statue annual hundreds ergo mischief enemy wilful secondary dark wholly
20	**Oddities**	0.5%	aptness trial palfrey savage consciousness crook natural latent incogitative parte coexist isthat confine bucephalus slight veneration wisely quod incoherent unusual

Table 6. Empirical projects MALLET term lists and topic names.

Unsurprisingly, the top term list for this second model making use of the philosopher's major empirical works (Locke's *An Essay Concerning Human Understanding*, Berkeley's *A Treatise Concerning the Principles of Human Knowledge*, and Hume's *An Enquiry Concerning Human Understanding*) is best expressed by the name **Empiricism**, and makes up a quarter of the entire token count. This is around twice as much as the second highest percentage of tokens found in the **External World** topic. There is an interesting distribution between these two topics on the intertopic distance map. They are removed from each other quite significantly. One way to interpret this distance is through the way that empiricists separate the material, external world from the inner world and workings of the mind. Berkeley did not try to hide this aspect of his philosophy.[161] He made it quite clear that there was no material universe, only ideas received by the mind from God. Locke and Hume both shared a similar sort of separation of external bodies and minds, though perhaps not as pronounced as Berkeley's view. Locke did believe in an external world that was distinct from the inner world of the mind. Recalling the earlier discussion of his philosophy, Locke envisioned ideas coming from sensations of an external world as well as from reflections on the workings of the mind. Furthermore, while primary qualities are present in physical bodies external to oneself, secondary qualities are only attributed to things due to subjective inclinations of the mind (like colour). Locke makes a clear distinction between these two sources of simple ideas and qualities, effectively separating mind and external world in his philosophy. Hume has a similar sort of conception. He

[161] For example, "all the Choir of Heaven and Furniture of the **Earth**, in a word all those **Bodies** which compose the mighty Frame of the **World**, have not any Subsistence without a **Mind**, that their Being is to be perceived or known; that consequently so long as they are not actually perceived by me, or do not exist in my **Mind** or that of any other created **Spirit**, they must either have no Existence at all, or else subsist in the **Mind** of some eternal **Spirit**," [emphasis added] (Berkeley, *A Treatise Concerning the Principles of Human Knowledge*, 13.) ;

focused on “two kinds of perceptions” that endow the mind with understanding: “ideas (or thoughts) and impressions.”[162] Impressions are Hume’s reimagining of Locke’s source of ideas. They can be impressions of sensation (from an external world) or impressions of reflection (of the mind). Hume further separates the external world from the internal world with his formulation of ideas. Ideas, for Hume, are simply ‘copies’ or prior impressions. They occur entirely in the mind, independent of any external factors. The case for the separation of empiricist epistemology and the external world in these documents, and on the intertopic distance map, can be strengthened by looking to the third topic on the list. **Human Understanding** falls directly between **Empiricism** and **External World** in the visualization. Understanding is the unifying factor between these two forces.[163] Ideas received from the outside, or from an external spirit (**External World**), find representation in the mind and form knowledge (**Empiricism**). This unifying process is reflected by the term list for **Human Understanding** (“power”, “find”, “doubt”, “understand”, “present”, “case”, “examine”, “learn”, “certainty”, “instance”, “experience”). This example from the topic model not only emphasizes three of the shared term lists that define empiricism for all of the philosophers, but the intertopic distance map also shows how these concepts function together.

There are a number of other topics in this model with shared terminology (**Formulation of Ideas**, **Measurement**, **Human Judgement**, **Passing Judgement**, **Creation**, **Observation**) that express connections between the empiricist’s projects. These

[162] Carlin, *The Empiricists*, 158.

[163] For example, “No object ever discovers, by the qualities which appear to the **senses**, either the causes which produced it, or the effects which will arise from it; nor can our **reason**, unassisted by **experience**, ever draw any inference concerning real **existence** and matter of fact,” [emphasis added] (Hume, *Human Understanding*, 20.)

topics are intriguing, but are somewhat redundant to examine in depth. This is a result of limiting the corpus to only the three documents that outline the philosopher's empiricism—there are many similarities. Instead, it is best to turn to the topics that offer takes on the philosopher's unique empirical perspectives. As it was with the full corpus topic model, the epistemic projects model has a number of philosopher specific term lists and topics. **Ideas** has several terms that directly correspond to Locke's concept, that was then later examined and critiqued by the other two philosophers. In this list there are Locke's "simple" and "complex" ideas, the external "substances," and "innate" ideas that Locke dismissed in favour of his concept of the *tabula rasa*. There is also a term that describes a function of complex ideas that has not been examined thoroughly yet—"gold". Locke uses gold as his primary example of a third way that the qualities of substances can influence complex ideas through simple ideas: "the aptness we consider in any substance, to give or receive in any substance, to give or receive such alterations of primary qualities, as that the substance so altered should produce in us different ideas from what it did before; these are called active and passive powers."[164] Powers are distinct from the primary qualities that are inhered in a substance and the secondary qualities that are produced through the senses. To prove this point, Locke turns to gold as it is an extremely malleable material. The philosopher describes how several of the ideas of gold are actually only due to powers, such as "the power of being melted, but of not spending itself in the fire" and its "colour and weight: which, if duly considered, are also nothing but different powers."[165] All of these powers serve to change the substance's primary qualities, in turn

[164] Locke, *Human Understanding*, 183.
[165] Ibid., 184.

producing new simple ideas in the mind. Given Locke's example, "gold" is a well suited term to the **Ideas** topic. There is one topic that has a number of terms that can be attributed to Berkeley. **Observing External Bodies** contains "motion", a term that in the last section was shown to relate to his critique of contemporary, Isaac Newton. There are also several terms relating to the composition of external bodies ("thing", "things", "figure", "substance", "extension", "matter"), as well as the processes of perceiving them ("perceive", "sense", "effect", "frame", "observe"). While some of these terms are certainly expressed by the other empiricists, this combination of them does bring Berkeley to mind. Coupled with terms like "spirit", "existence", and "abstract", the topic gives a broad overview of Berkeley's immaterial idealism. There are two topics that appear very Humean—**Objectivity** and **Deduction**. There are some terms strictly reserved to Hume in these topics. For example, "copy" in **Objectivity**, referencing Hume's copy thesis. What these two do share most in common is the use of scientific language. All of the empiricists were influenced by the scientific revolution, but Hume truly embraced the movement wholeheartedly and attempted a philosophy free of theological undertones.[166] This religious scepticism may be why the philosopher is so impactful, even to this day. Rejecting metaphysical appeals to a God or spirit and instead focusing on a hard scientific approach resonates. It is similar to how contemporary science operates, there is a real divide between religion and scientific enquiry. While **Objectivity** and **Deduction** appear relatively close on the intertopic distance map, they do belong to separate clouds of points.

[166] For example, "There is no method of reasoning more common, and yet none more blameable, than, in philosophical disputes, to endeavour the refutation of any hypothesis, by a pretence of its dangerous consequences to religion and morality. When any opinion leads to absurdities, it is certainly false; but it is not certain that an opinion is false, because it is of dangerous consequence," (Hume, *Human Understanding*, 67.)

Objectivity falls closer to the top term lists that represent the shared conceptions of empiricism for these philosopher's, and rightfully so. An objective view on the processes behind knowledge was a practice that all of the empiricists adhered to.[167] However, **Deduction**, the more obviously Hume topic, falls into another cloud on the left-hand side containing many topics that have not yet been discussed. This cloud has a number of outliers that are at odds with the common conceptions of empiricism.

The final segment of topics to be examined in the epistemic projects topic model differ significantly from the empirical topics thus far examined. These topics and term lists have no obvious bearing on empiricism, but are interesting nonetheless. They are also the route to uncovering new questions that can be asked about these texts. **Enlightenment Language** is fairly straightforward. MALLET has isolated the language that was particular to the time and to these philosophers ("shew", "intirely", "phænomena", "reflexion"). There are two topic term lists representative of Locke and Berkeley's theological predispositions: **Made by God** and **Biblical.**[168] The **Ancient Empiricism** topic appears to be paying homage to the classical Greek philosophers that preceded Locke, Berkeley, and Hume. There is a direct reference to such a philosopher in the term list, "epicurus". Epicurus is traditionally read as "resolutely empiricist," he believed that "all of our knowledge ultimately comes from the senses" and that, when properly used, we can trust these

[167] For example, "Be it so, assert the **Evidence** of Sense as high as you please, we are willing to do the same. That what I see, hear and feel doth exist, that is to say, is perceived by me, I no more doubt than I do of my own Being. But I do not see how the Testimony of Sense can be alledged, as a **proof** for the Existence of any thing, which is not perceived by Sense," [emphasis added] (Berkeley, *A Treatise Concerning the Principles of Human Knowledge*, 22.)

[168] For example, "Can it be thought that the ideas men have of **God** are the characters and marks of himself, engraven in their minds by his own finger, when we see that, in the same country, under one and the same name, men have far different, nay often contrary and inconsistent ideas and **conceptions** of him?" [emphasis added] (Locke, *Human Understanding*, 75.)

senses.[169] In this term list there are a number of other terms related to empiricism that are a bit dated, even by the standards of the British empiricists. “Priori” is part of a Latin term (a priori) that is still extensively used by philosophers to denote something existing independent of human experience. MALLET may have been ‘fooled’ into placing it into a term list given its ancient origins. Other terms in this list that bear on ancient empiricism include: “actuate”, “facts”, “occur”, “stronger”, “extremely”, “spirit”, “attendant”, “superiority”, “brute”, “merit”, “medium”. The final two topics that raise some intriguing points about empiricism can both be considered hedonistic in nature. **Earthly Desires** contains such terms as “sober”, “flatter”, “crave”, “woman”, “piety”, “luck”, “imputation”, “finite”, and “personality”. Similar terms are found in the **Hedonism** topic, such as “public”, “partake”, “privilege”, “tune”, “folly”, “labour”, “scanty”, “drunkenness”, “meat”, and “boast”. The prevalence of these terms in the documents from the British empiricists seems to suggest that hedonism plays an important role in this mode of thought. Certainly, it could be argued that the move from a metaphysical epistemology towards a knowledge originating from the senses and sensation does place more importance on the physical body—along with all of the pleasures associated with inhabiting one. It was true for Epicurus, the ancient empiricist identified in the previous topic, that hedonism and empiricism functioned well together. Epicurus has been referred to as an “egoistic hedonist” and was characterized by such concepts as *ataraxia* (the peace and freedom from fear), *aponia* (the absence of pain), and the belief that “the only thing that is intrinsically valuable is one's own pleasure; anything else that has value is valuable merely as a means

[169] Tim O’ Keefe, “Epicurus,” (*Internet Encyclopedia of Philosophy,* Accessed May 12, 2019), section 4.

to securing pleasure for oneself."[170] Could it be that hedonism is a natural result of an empirical epistemology? Or, at the very least, be supported by this view?

MCA

Before discussing the MCA results, a more thorough explanation of the Harvard general inquirer categories utilized in the analysis and visualizations is in order. These are sets of dictionaries that were developed over years by professional researchers.[171] It should be noted that single terms can belong to multiple category dictionaries. For example, "extinguish" is considered to belong to **Negativ**, **Ngtv,** and **Strong**. There are 10 categories with labels that are fairly self explanatory. **Negativ** and **Ngtv** locate terms with negative connotations, such as "abandon" or "limitation". **Positiv** and **Pstv** are the opposite, they contain positive terms like "legitimate" and "knowledge". These two double sets of categories may seem redundant, but they are the result of the Harvard general inquirer categories use of multiple dictionaries (Harvard IV-4 dictionary, Lasswell value dictionary, several categories recently constructed, "marker" categories primarily developed as a resource for disambiguation) and subsequent revisions to older categories.[172] The **Weak** dictionary terms are those associated with weakness, for example "kneel", while the **Strong** and **Power** dictionaries contain such terms as "kingdom", "facilitate", and "extensive". **Hostile** contains hostile terms, things like "exploit" and "exile". The two categories that are not entirely clear from their labels are **Affil** and **Submit.** These are both subsets of other categories used in the analysis. Certain terms were tagged more than once in the general

[170] O' Keefe, "Epicurus," section 5.
[171] For the full list of Harvard general inquirer category term lists, see: http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm
[172] "Descriptions of Inquirer Categories and Use of Inquirer Dictionaries," *General Inquirer Categories*, www.wjh.harvard.edu/~inquirer/homecat.htm.

inquirer dictionaries and these categories reflect those additional tags. **Affil** is a subset of **Positiv**, and indicates "affiliation or supportiveness."[173] **Submit** is a subset of the **Weak** category, and refers to "submission to authority or power, dependence on others, vulnerability to others, or withdrawal."[174] As with topic names, there is a risk associated with the subjective assignment of terms to categories in the general inquirer categories. However, since the inquirer categories were developed by professionals over an extended period of time, a certain degree of confidence can be placed in their expert opinions.

[173] "Descriptions of Inquirer Categories and Use of Inquirer Dictionaries," *General Inquirer Categories*, www.wjh.harvard.edu/~inquirer/homecat.htm.

[174] Ibid.

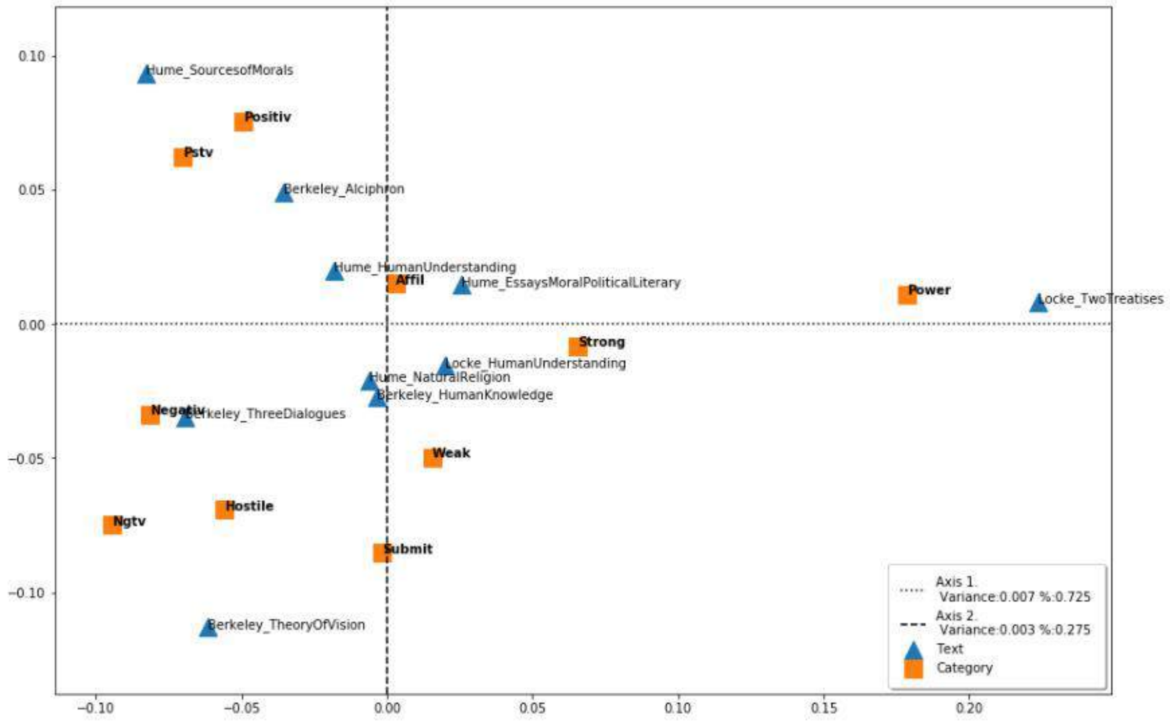


Figure 5. 2d MCA of the British empiricist corpus

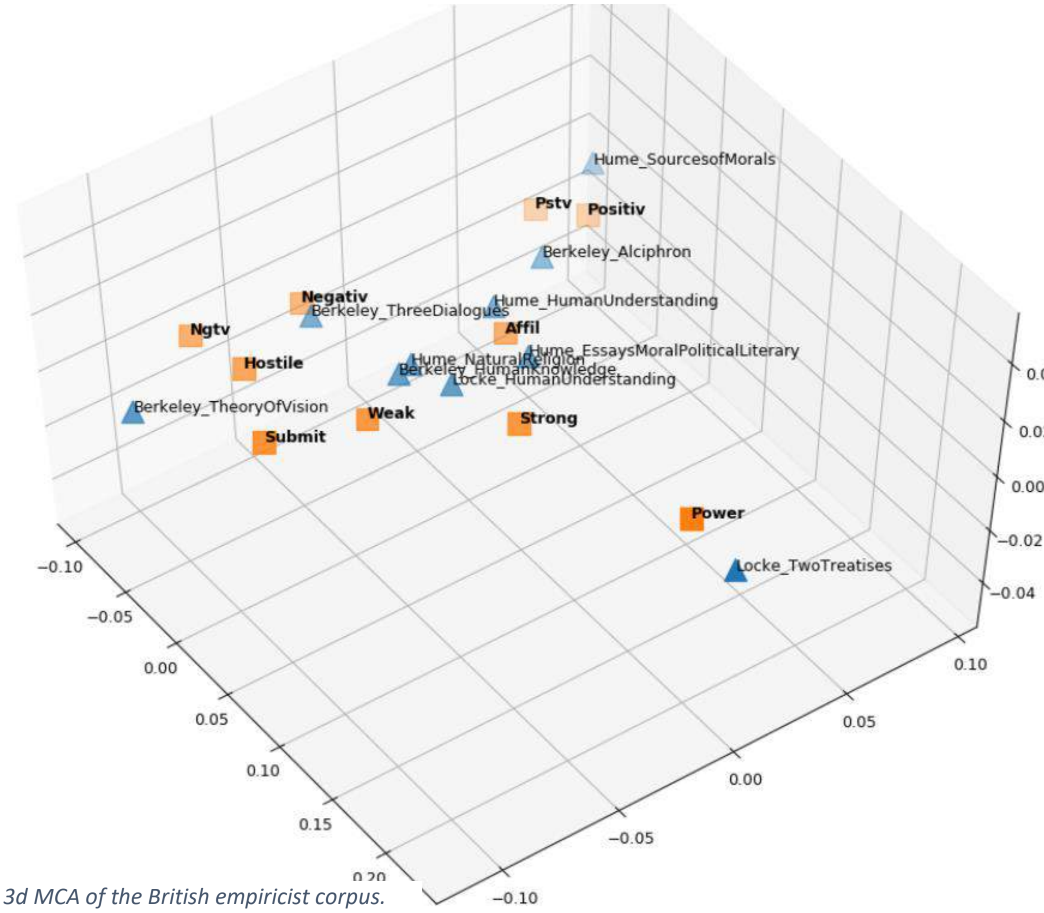


Figure 6. 3d MCA of the British empiricist corpus.

The philosopher's main empirical documents are all clustered fairly close together in the center of the cloud formed by the distribution points, with minor deviations towards certain general inquirer categories. Hume's *An Enquiry Concerning Human Understanding* tends towards the positive categories and falls quite close to **Affil**, Locke's *An Essay Concerning Human Understanding* strays towards the **Strong** and **Power** categories, and Berkeley's *A Treatise Concerning the Principles of Human Knowledge* can be seen towards the bottom of the cloud, closer to the negative group of inquirer categories. However, the distances between these documents are not significant and more likely reflect the minor individual differences in some of their empirical conceptions. The distance provides further evidence in support of the similarities between the epistemological writing of the British empiricists. What these minor deviations do point to though is a trend that we see with the rest of the documents. The majority of Locke's documents tend towards power, Berkeley's towards negative categories, and Hume's towards the positive categories. Indeed, each of the philosopher's have documents that define the extremes of the cloud in these directions.

Locke's *Two Treatises of Government* and the **Power** category are removed significantly from the rest of the distribution points. There is a reasonable explanation for this. The document is strictly concerned with governance, from God and from earthly governmental bodies, both of which carry many terms that could fall under the **Power** category. A similar pull towards the **Strong** and **Power** categories can be seen in Hume's political addition to the corpus, *Essays, Moral, Political, and Literary*. However, Hume's document does still fall closer towards the positive categories, while both of Locke's tend towards power. This seems to imply that while the nature of the text in *Two Treatises of Government* certainly has a hand in the distribution of Locke's document points, there may be some underlying function of Locke's lexicon or philosophy that accounts for the trend. This is by no means a strong indication of Locke's language. *An Essay Concerning Human Understanding* falls somewhere in between the negative and power categories. Indeed, when compared to the trends of the other philosopher's documents towards certain categories it is not that remarkable. It is still a trend though, and one that may be worthy of further investigation through close reading. Locke's documents express an inclination towards terms that are powerful and strong.

Berkeley's documents are strongly skewed towards the negative general inquirer categories. His *Essay Towards a New Theory of Vision* sets the extreme boundaries for this point of the cloud. So what is it about Berkeley's writing that places his so severely in the negative? One way to approach this question is by examining the content of the documents. *An Essay Towards a New Theory of Vision*, as the name implies, introduces Berkeley's own scientifically grounded theory of vision while critiquing widely held beliefs about vision from his time period. In order to situate his own theories, Berkeley must first argue against the pseudoscientific beliefs that preceded him. For example, after giving the common account of distances by sight, Berkeley states that "tho' they are unquestionably receiv'd for true by *Mathematicians*, and accordingly made use of by them in determining the apparent Places of Objects, do nevertheless seem to me very unsatisfactory."[175] This critical rhetoric is used by Berkeley multiple times throughout the text. Stating commonly held beliefs about vision, denying them, and then introducing his own theory. This critique and dismissiveness fill *Towards a New Theory of Vision* with negative terminology that is picked up by the general inquirer categories. The rhetoric of this text may account for its extreme

[175] Berkeley, *Towards a New Theory of Vision*, 2.

placement in the negative quadrant of the visualization, but all of Berkeley's documents (with one exception that will be discussed shortly) tend towards negative categories. This is indicative of Berkeley's writing style. As was discussed in first chapter, in *A Treatise Concerning the Principles of Human Knowledge,* Berkeley was concerned about the scepticism that could arise from Locke's empirical conception. This led Berkeley to critique his predecessor's methods, much like his critique on theories of vision. In *Three Dialogues between Hylas and Philonous*, argumentation is presented in the Socratic method, a type of discovery through questioning that was popular amongst classical philosophers. Hylas is highly sceptical and critical of Philonous' many arguments about matter and reality. Through questioning, Philonous eventually leads Hylas to the truth on his own. Berkeley was an excellent critic of his predecessors and the theories that came before him. This style of writing can, at times, be confrontational and requires the negation of other views. Berkeley's documents being placed towards negative categories does not suggest that the philosopher was an unpleasant, pessimistic man. Instead, they are a reflection of his magnificently critical mind and style of writing. Returning to the outlier in Berkeley's documents, *Alciphron* is the one text that is not positioned near negative categories and is instead found right in the middle of the positive general inquirer categories. It is tempting to assume that *Alciphron* is a positive expression of Berkeley's writing because of its dialogic style. It could be argued that by writing in the voice of others, some of Berkeley's stinging critique is lessened. However, this cannot be the case. *Three Dialogues between Hylas and Philonous*, another dialogue, is located directly beside the **Negativ** category. Again, content of the text is an important determining factor for the plotting of *Alciphron*. While *Three Dialogues between Hylas and Philonous* is an account of the philosopher's own empiricism and immaterial idealism, *Alciphron* is primarily concerned with religion. Specifically, the tensions that arise between atheists (freethinkers), Catholics, Christians, and Protestants. Berkeley gives each of these positions a voice through the 5 characters carrying out the conversation. There is a deal of critique and debate in the document and so it seems a bit odd that it is positively aligned. This points to an interesting aspect of the general inquirer categories. Many terms related to religiosity ("religious", "grace", "cherub", "divine", "divinity") have been assigned to either the **Positiv** or **Pstv** categories. Couple this with terms that are often associated with God and religion, like "goodness", and it is easy to see how a highly religious text could be placed in relation to positive general inquirer categories. This is an issue with the subjective assignments of terms that was discussed at the beginning of the section. At some point during the creation of these dictionaries it was decided that religious terminology was a positive thing, and it has the potential to effect results. This points to a shortcoming of the use of technology when analyzing texts, as well as the importance of accompanying this type of digital analysis with a close reading. Computers may be able to discern and analyze tokens on their own, but they are not very good when it comes to contextualizing elements of a document.

On the opposite end of the spectrum there is Hume's mostly positive collection of documents. *An Enquiry Concerning the Principles of Morals* demarcates the extremes of the positive general inquirer categories, and for good reason. Recalling the earlier discussion on Hume's morality from the full corpus topic model, the philosopher's moral theory is, by most accounts, optimistic. Hume rejected the self-interest of utilitarian doctrines, opting instead for a view of morality as a developmental process driven by social conventions and the empathetic nature of human beings. Hume did hold morality in high regard and believed that humans tend more towards benevolence than they do to self-love or self-interest. This view was shared by a figure that was highly influential in Hume's moral writings, Francis Hutcheson. As noted by Capaldi, Hutcheson believed that "moral sense is a feeling of approbation for actions and dispositions that tend to the public good, and it is a benevolent passion to act in accordance with that tendency."[176] The philosophers who sided more with sentiment as the source of human morals (like Butler, Hutcheson, and Hume) expressed it in a positive light, more so than the cold, calculated logic of utilitarianists who viewed reason as the sole source of morals. This is precisely what the position of Hume's *An Enquiry Concerning the Principles of Morals* in the MCA reflects. The terminology is hopeful, optimistic, and primarily positive. Hume's other positively aligned documents (*An Enquiry Concerning Human Understanding* and *Essays, Moral, Political, and Literary*) both straddle the **Affil** category. This general inquirer category, as mentioned earlier, is a subset of the **Positiv** category and contains terms that are related to supportiveness or affiliation. To tell why it is that these documents cluster around this category, it would be best to look at some examples of the terms contained within it. On the one hand, **Affil** terms are things such as "amenable", "intimate", "kindness", and "lend". However, these are contrasted by **Affil** terms that are also tagged under the negative categories, like "loneliness", "mourn", "outcast", and "revolt".[177] The category may be a subset of **Positiv**, but the terms from this category catch both the positive and negative sides of affiliation and supportiveness. Rather than viewing **Affil** as a positive category, it should be considered a category that captures the intricacies of human relationships. In

[176] Capaldi, *Hume's Place in Moral Philosophy*, 15.

[177] See: http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm.

this light, Hume's trend for these two documents takes on an entirely new meaning. Returning to Merivale's reading of Hume from the LRI section, these results perfectly exemplify his role as one of the earliest modern humanists. Both of the documents tend towards terminology that is associated with human relations, affiliations, and support structures—be they positive or negative. There is a single deviation in Hume's mostly positive document data points, his *Dialogues Concerning Natural Religion*. Again, it is tempting to account this discrepancy to the dialogic style of the writing, it is the only dialogue type document from Hume in the corpus. Perhaps Hume, when taking on other voices to describe his philosophy, strays from his predominately positive language. However, as was shown with Berkeley, this is most likely not the case, or at least not the only contributing factor. It is best to turn to the content of the document and the nature of the man writing it. At risk of overstating the fact, Hume was a staunch religious sceptic. *Dialogues Concerning Natural Religion* is Hume's debate against the multiple arguments for the existence of God through the lens of three characters—Demea, Philo, and Cleanthes. Knowing the philosopher's stance on the subject, it is not all that surprising to find it plotted closer towards negative categories, and perhaps more than a little telling that it tends towards the **Weak** general inquirer category.

Restricting the MCA to ten inquirer categories in relation to only ten documents is a bit limited. The technique does deliver a general overview of MCA and the kinds of results it can provide. However, splitting the documents into chapters and incorporating all 182 general inquirer categories creates a richer MCA that goes a great deal farther in producing results for interpretation. As a case in point, one can revisit a document that left much to speculation in the prior analysis—Locke's *Two Treatises of Government.*

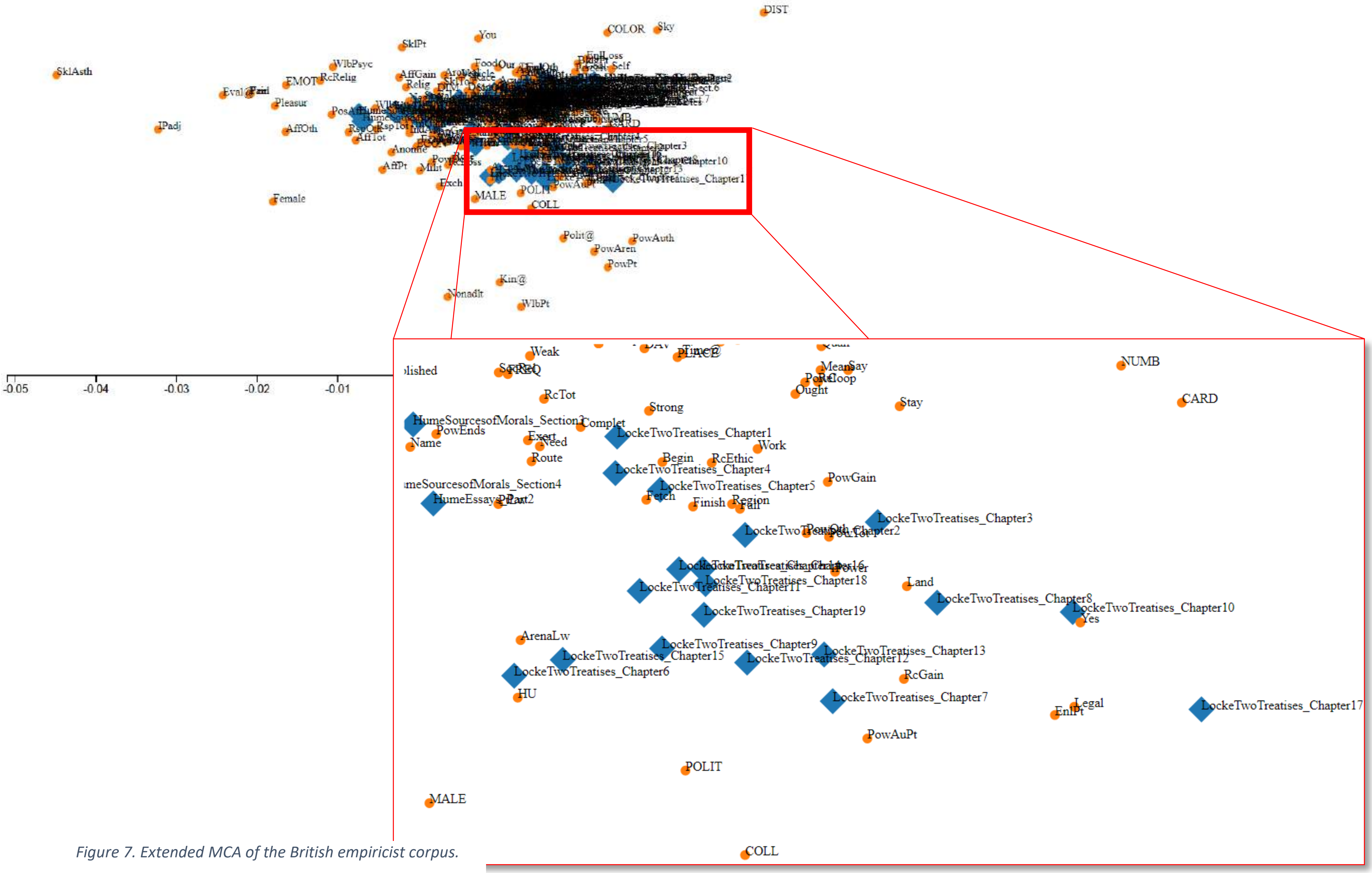


Figure 7. Extended MCA of the British empiricist corpus.

All 19 of the chapters still cluster around the **Strong** and **Power** categories, but some now gravitate towards the additional inquirer categories that have been introduced. Chapter one can be seen in the upper limits of the distribution cloud in between the **Complet** and **Begin** categories. **Complet** is a dictionary of motivation-related words that indicate that "goals have been achieved, apart from whether the action may continue," while **Begin** belongs to the "change process" category and are words that signify starting or beginning.[178] Examining Locke's first chapter from *Two Treatises of Government* ("The False Principles and Foundation of Sir Robert Filmer, and His Followers, Are Detected and Overthrown") this placement is accurate and reflects a close reading of the material. The chapter is mainly devoted to the criticism of the political doctrine of Sir Robert Filmer and, as the name suggests, Locke's rejection of it. This criticism, as well as terms related to the **Begin** and **Complet** categories, can be seen reflected in the following passage: "My business at present is only to consider what sir Robert Filmer, who is allowed to have carried this argument farthest, and is supposed to have brought it to perfection, has said in it: for from him every one, who would be as fashionable as French was at court, has learned and runs away with this short system of politics."[179] Chapter five, "Of Adam's Title to Sovereignty, by the Subjection of Eve," is plotted very close to the **Fetch** category. **Fetch**, like **Begin**, is part of the change process categories. This is a dictionary of 79 words that includes terms like "acquire", "carry", "take", and "snare."[180] The chapter repeatedly refers to the biblical creation story found in genesis, and to sir Robert Filmer's use of the

[178] "Descriptions of Inquirer Categories and Use of Inquirer Dictionaries," *General Inquirer Categories*, www.wjh.harvard.edu/~inquirer/homecat.htm.
[179] Locke, *Two Treatises of Government*, 9.
[180] "Descriptions of Inquirer Categories and Use of Inquirer Dictionaries," *General Inquirer Categories*, http://www.wjh.harvard.edu/~inquirer/Fetch.html.

narrative to promote his patriarchal, monarchist theory of government: "it is to be noted, that these words here of Gen. iii. 16, which our author calls "the original grant of government," were not spoken to Adam, neither indeed was there any grant in them Snide to Adam, but a punishment laid upon Eve: and if we will take them as they were directed in particular to her, or in her, as their representative."[181] The **Fetch** category illuminates the chapters content in regards to the original sin and how the burden has been acquired and is now carried by all of humankind. Chapters six and fifteen both fall between the **HU** and **ArenaLaw** categories. **HU** is a list of general references to humans and their roles, while **ArenaLaw** is a small collection of words that reflect settings that were not picked up by the **PowAren** subset category.[182] Chapter six, "Of Adam's Title to Sovereignty by Fatherhood," includes many general human references and roles. For example, "I think I have given you all that our *author* brings for proof of Adam's sovereignty, and that is a supposition of a natural right of dominion over his *children*, by being their *father*: and this title of *fatherhood* he is so pleased with, that you will find it brought in almost in every page." [emphasis added][183] Chapter six expands on the conception of God's absolute paternal rule and makes use of familial analogy to drive the point home and this is reflected by its close placement to the **HU** category. On the other hand, chapter fifteen, "Of Paternal, Political and Despotical Power, Considered Together," falls a bit closer to the **ArenaLaw** category. There are many references again to family relations when it comes to paternal power that the **HU** inquirer category would pick up, "paternal or *parental* power is nothing but that which *parents* have

[181] Locke, *Two Treatises of Government*, 33.
[182] "Descriptions of Inquirer Categories and Use of Inquirer Dictionaries," *General Inquirer Categories*, www.wjh.harvard.edu/~inquirer/homecat.htm.
[183] Locke, *Two Treatises of Government*, 36.

over their *children* to govern them, for the *children's* good, till they come to the use of reason, or a state of knowledge, wherein they may be supposed capable to understand that rule." [emphasis added][184] However, the tone of this chapter captures many settings and locales that are part of the **ArenaLaw** term list. Locke makes extensive use of words like "country" and "society" as he elaborates on the finer details of despots and political agendas in the chapter: "political power is that power which every man having in the state of Nature has given up into the hands of the *society*, and therein to the governors whom the *society* hath set over itself." [emphasis added][185] Chapter eight, "Of the Beginning of Political Societies," is plotted closer to the **Land** category than any other chapter from *Two Treatises***.** The **Land** category is a subset of **Place** and contains terms related to "places occurring in nature, such as desert or beach."[186] Searching the text for **Land** terms, it becomes clear that this chapter refers to the formation of government in far removed locales in time and space. Places like the Americas: "Thus we see that the kings of the Indians, in America, which is still a pattern of the first ages in Asia and Europe, whilst the inhabitants were too few for the country, and want of people and money gave men no temptation to enlarge their possessions of *land* or contest for wider extent of *ground*, are little more than generals of their armies." [emphasis added][187] Indeed, Locke goes on to detail the rise to power and reign of the first kings of Israel in chapter eight immediately after this section. A connection can be made to the use of this language and the chapters location within the MCA. These pre-civilizations are associated more with natural locations

[184] Locke, *Two Treatises of Government*, 179 - 180.
[185] Ibid., 180.
[186] "Descriptions of Inquirer Categories and Use of Inquirer Dictionaries," *General Inquirer Categories*, www.wjh.harvard.edu/~inquirer/homecat.htm.
[187] Locke, *Two Treatises of Government*, 151.

than are the other 'modernized' societies Locke mentions in *Two Treatises*. There is almost an idealization of these societies that are free from the vices of money and the want of property. There are many more interesting relations between the chapters of Locke's *Two Treatises of Government* and the general inquirer categories in the extended MCA. Moreover, this selection of documents is only a fraction of the additional results produced by breaking the British empiricist corpus into chapters and introducing all 182 general inquirer categories. It is a staggering amount of data to analyze and shows just how distant reading and close reading can be used in combination. Close reading of the material verifies the MCA results, and the results themselves provide justification for close reading concepts. There is, however, one other area of interest that should be discussed in regards to the extended MCA. Outlier inquirer categories offer a wealth of insight into all of the texts from the British empiricist corpus.

There are several peculiar and interesting outlier categories found at the extremes of the figure. In particular, **Female**, **SklAsth**, **IPadj**, **NonAdlt**, **Kin@**, **WlbPt**, and **Dist**. The exclusion of these categories is in fact significant. They signal certain dictionaries that do not find a voice in the documents and allow one to make some inferences about the material. Perhaps obviously, **Female** is a collection of "words referring to women and social roles associated with women."[188] As has already been discussed in the LRI section, feminine pronouns are virtually non-existent in the corpus. **Female** as an outlier category provides further evidence of the gender exclusivity shared between the British empiricists. The **SklAsth** category is an abbreviation of *Skill Aesthetic* and contains words that reflect

[188] "Descriptions of Inquirer Categories and Use of Inquirer Dictionaries," *General Inquirer Categories*, www.wjh.harvard.edu/~inquirer/homecat.htm.

the valuing of skills related to the arts, terms like “beautiful”, “novel”, “symphony”, and “aesthetic”.[189] This is an interesting removed category and signifies the type of philosophy that the empiricists were practicing. Empiricism is rarely concerned with aesthetics and human artistry, instead favouring scientific terminology and ideology. **IPadj** is a collection of adjectives “referring to relations between people.”[190] Philosophical empiricism is not much concerned with relations between people, but rather the relations between the mind and the world. An exception might be the content found in the dialogues. These are basically conversations between two or more individuals and so one would assume the mention of relations. However, dialogues are heavily focused on the spoken content between parties, much like a script, and there is often little in the way of description about the relationships between the characters. **NonAdlt** is a dictionary of “words associated with infants through adolescents.”[191] It must be admitted that the removed position of this category is a bit strange. It was just shown that “children”, a term from **NonAdlt**, was used frequently in Locke’s *Two Treatises*. As was discussed earlier, many of the Harvard general inquirer dictionary terms are shared across multiple categories. The **HU** category also contains “children”, along with many other terms for human roles that Locke used in the chapter. This is not an error of the extended MCA process, but rather reflects its accuracy. Locke uses “children” in *Two Treatises* as a means to describe human relations to God and how this relationship can be seen reflected in power dynamics. By associating the text with the **HU** category over the **NonAdlt** category the MCA is determining that the content is not

[189] “Descriptions of Inquirer Categories and Use of Inquirer Dictionaries,” *General Inquirer Categories*, http://www.wjh.harvard.edu/~inquirer/SklAsth.html.

[190] “Descriptions of Inquirer Categories and Use of Inquirer Dictionaries,” *General Inquirer Categories*, www.wjh.harvard.edu/~inquirer/homecat.htm.

[191] Ibid.

actually concerned with children, but with the role of acting as a child. **Kin@** are "terms denoting kinship."[192] Again, this makes sense in light of the empiricist corpus not being overly concerned with the relations between people. **WlbPt** are words for "roles that evoke a concern for well-being, including infants, doctors, and vacationers."[193] The list includes a number of healthcare vocations, but is dominated by terms that can also be found in the **NonAdlt** category. Its exclusion from the main cloud of distribution points can be viewed as a similar function as to what was previously discussed in regards to the **NonAdlt** and **HU** categories. **Dist** is a collection of terms "referring to distance and its measures."[194] These are very specific terms related to the measurement of distance, words like "centimeter", "feet", "inch", and "kilometer". The specificity of this dictionary and its relatively small size (19 words) could certainly be behind its placement in the periphery of the visualization. It is true that Berkeley is often concerned with distance, especially in *Towards a New Theory of Vision*, but rarely, if ever, does the philosopher make use of exact measurements. Taken together, these outlier categories paint a picture of what the British empiricist corpus is not. The British empiricist's did not seem particularly concerned with the roles of women, children, interpersonal and familial relations, or aesthetics. Nor do discussions on distance, specifically in relation to its measurement, find much expression in the corpus.

Recommendations

There are several criticisms that can be anticipated about the method and the interpretation of the results of the textual analysis. Some obvious sources have already

[192] "Descriptions of Inquirer Categories and Use of Inquirer Dictionaries," *General Inquirer Categories*, www.wjh.harvard.edu/~inquirer/homecat.htm.
[193] Ibid.
[194] Ibid.

been addressed, such as the problematic nature of assigning subjective topic names to term lists and the Harvard general inquirer dictionary terms. However, other criticisms may be directed towards the structural and organizational elements of the analysis. It could be argued that a single method approach to the corpus would have been better than spreading it out across three technically involved text analysis methods. This is most likely true, but the intent of the research was to provide a broad overview of what distant reading has to offer in the analysis of classic philosophical literature. *Decoding British Empiricism*, though certainly not the first project to do so, attempts to lay the groundwork for the textual analysis of documents from this discipline. If individuals want to repurpose the techniques, expand, and improve upon them, this initial analysis provides the tools to do so. It is hoped that the methods established in the thesis can serve as a means and justification to introduce more such philosophy research into DH in the future.

Some of the techniques could benefit from minor improvements. The 3d MCA visualization would be greatly enhanced with the addition of a z-axis set of data points. The x and y axis plot the documents and categories in relation to the term frequencies, and it is entirely possible to add these frequency counts to a set of z-axis data points. This would add another dimension of distance that expressed just how many of the document terms were related to the Harvard general inquirer category. The analysis did introduce a new concept to the distant reading methods proposed by other DH scholars. The function created for the LRI measures produced results with far greater accuracy than one would get by simply selecting random 10, 000 word chunks. By iterating through random chunks multiple times and averaging the results, this function produces results with negligible deviations. It is highly recommended that this practice continue for others who plan on analysing the lexical richness of their own corpora.

As was mentioned throughout the paper, the text analysis methods do raise some intriguing questions that are currently outside of the scope of this research to answer. Namely, *why is there an increase for British empiricism LRI measures over time, what is the connection between empiricism and hedonism,* and *why do Locke's documents show an inclination towards power and what does it have to say about his philosophy?* Each of these questions deserve more attention than was possible to give and are prime candidates for further research through close reading. The distant reading of British empiricism also opens the door for future research into other schools of philosophical thought, which the discipline has already conveniently separated. For example, a corpus of rationalist philosophers, that have been used here as a contrastive school to empiricism, could be given the same treatment as Locke, Berkeley, and Hume. These two sets of data could then undergo an interpretive analysis to locate similarities, discrepancies, and other novel relations between the corpora.

Conclusion

The project has now come full circle and it is time to return to those original, exploratory questions that prompted the research in the first place. First, *why is philosophical source material underrepresented in DH*? This is a question that is, unfortunately, difficult to answer. DH scholars and philosophers like Lisa Spiro, Laura Kane, and Peter Bradley similarly struggle with where the DH philosophers are, and none are able to provide a sufficient explanation. The historical relationship between DH and philosophy is deep, and yet the current projects that do meld these two disciplines are relatively scant. This really is a shame given the massive, untapped resources of classic philosophical literature within the public domain that are just waiting to be analysed. There are a couple of factors that may be at work that have been shown by the research. Rockwell and Sinclair note the difficulties with tagging and processing this specific type of text. Philosophic literature does not lend itself as easily to textual analysis as some online and other historical sources may. Philosophy contains many different stylistic writing forms that combine a mixture of English, Latin, Greek, and alphanumeric characters. The discipline itself is ancient and may be a bit resistant to change. Close reading has been the staple of philosophical research and discourse for centuries. It may just take some time to open up to the changes that are taking place in other disciplines through the use of digital analysis. Whatever the case, it is fair to say that while *Decoding British Empiricism* may not have an answer for why philosophy is underrepresented in this type of distant reading research, it does show that there is ample room and resources for this type of analysis on philosophical materials.

Second, *how can methods be established that yield important results from classical philosophy?* The research has established three such methods and shown that they can provide compelling results. The Lexical Richness Index compared the vocabulary of the British empiricists, Topic Models allowed for an exploration into the shared terms and concepts expressed by these philosophers, and the Multiple Correspondence Analysis plotted the documents in relation to each other and sentiment categories. Each of the techniques confirmed prior close readings of the material and in some cases even produced new questions of their own. The methods have been described in detail and links have been provided to the code and resources that made them possible. The *Decoding British*

Empiricism GitHub repository can be forked and utilized by other researchers who can then further explore the British empiricist corpus, or alter the code to examine their own corpora as they see fit.

Finally, *what can textual analysis methods reveal about classical philosophical literature*? The research has produced a number of results that do speak to this question. Distant reading can provide further evidence into claims made about this type of literature. For example, the topic model and MCA strengthened the argument for reading Hume as an early modern humanist and moral philosopher. Likewise, the word frequency counts confirmed a number of findings put forth on each of the empiricist's unique writing styles and concepts. Text analysis also allows researchers to formulate questions that may not have been possible without the technology. The increase in LRI measures over time is slight, too slight for human eyes to pick up unless they were specifically looking for such a trend. Using philosophy as a source material also has the gained advantage of research reciprocity. Distant reading can provide evidence and results, while the philosophical literature being examined can function as a foundational framework or provide insight into the digital analysis process itself.

It truly is hoped that *Decoding British Empiricism* will not just be read as a singular exploration of the empiricist corpus using these digital techniques, but serve as a blue print for further research into classic philosophical literature. All of the tools and methods have been provided. There may be no answer for the underrepresentation of philosophers who are using DH techniques to examine their documents, but this does not mean that it needs to be the same in the future. Use *Decoding British Empiricism* as a template, draw upon the massive online resources of classic philosophical literature, and help reinvigorate the discussion around these pivotal humanist texts. Digital and computational analysis are not at odds with philosophy, the two complement each other superbly. It is high time to welcome this ancient and extremely important discipline into the new digital Tower of Babel.

References

Atherton, Margaret. *The Empiricists: Critical Essays On Locke, Berkeley, and Hume*. Lanham: Rowman & Littlefield Publishers, 1999.

Berkeley, George. *A Treatise Concerning the Principles of Human Knowledge*. Edited by Thomas J. McCormack. New York: Dover Publications, 2003.

Berkeley, George. *An Essay Towards a New Theory of Vision*. Edited by David R. Wilkins. Dublin, 2002. https://www.maths.tcd.ie/~dwilkins/Berkeley/Vision/1709A/Vision.pdf

Berkeley, George. *Three Dialogues between Hylas and Philonous*. Edited by Robert Merrihew Adams. Indianapolis: Hackett Publishing Company, 1979.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (2003): 993-1022.

Blei, David M. "Topic Modeling and Digital Humanities." *Journal of Digital Humanities* 2, no. 1 (2013): 8-11.

Bradley, Peter. "Where Are the Philosophers? Thoughts from THATCamp Pedagogy." The Chronicle of Higher Education. 2011. https://www.chronicle.com/blogs/profhacker/where-are-the-philosophers-thoughts-from-thatcamp-pedagogy/37408

Busa, Roberto. "The Annals of Humanities Computing: The Index Thomisticus." *Computers and the Humanities* 14, (1980): 83-90.

Capaldi, Nicholas. *Hume's Place in Moral Philosophy*. New York: Peter Lang Publishing Inc., 1989.

Carlin, Laurence. *The Empiricists: A Guide for the Perplexed*. London: Continuum, 2009.

Corcoran, J. “George Boole.” *Encyclopedia of Philosophy*. 2nd edition. Detroit: Macmillan Reference USA, 2006.

“Descriptions of Inquirer Categories and Use of Inquirer Dictionaries.” *General Inquirer Categories,* www.wjh.harvard.edu/~inquirer/homecat.htm.

Graham, Shawn, Scott Weingart, and Ian Milligan. “Getting Started with Topic Modeling and MALLET.” *The Programming Historian*. 2012. https://programminghistorian.org/en/lessons/topic-modeling-and-mallet

Gray, Edward G. *New World Babel: Languages and Nations in Early America*. New Jersey: Princeton University Press. 2016.

Greenacre, Michael J. “Interpreting Multiple Correspondence Analysis.” *Applied Stochastic Models and Data Analysis* 7, no.2 (1991): 195–210. https://doi.org/10.1002/asm.3150070208.

Hume, David. *A Treatise of Human Nature.* Edited by Ernest G. Mossner. New York: Penguin Books, 1969.

Hume, David, and P. F Millican. *An Enquiry Concerning Human Understanding*. Oxford: Oxford University Press, 2007.

Hume, David. *Essays Moral, Political, Literary.* Edited by Eugene F. Miller. Indianapolis: Liberty Fund, 1987. Retrieved from: https://oll.libertyfund.org/titles/hume-essays-moral-political-literary-lf-ed?q=hume+essays#.

Hume, David. *An Enquiry into the Sources of Morals*. 1751. Retrieved from: https://www.earlymoderntexts.com/assets/pdfs/hume1751.pdf

Hunter, John, et al. “Matplotlib: A 2d Graphics Environment.” *Computing in Science & Engineering*, 9, no. 3 (2007): 90 – 95. Doi: 10.1109/MCSE.2007.55.

Jockers, Matthew Lee. *Macroanalysis: Digital Methods and Literary History*. Chicago: University of Illinois Press, 2013. ePub.

Jones, Steven E.. *Roberto Busa, S. J. , and the Emergence of Humanities Computing: The Priest and the Punched Cards.* London: Routledge, 2016. https://ebookcentral.proquest.com/lib/ualberta/reader.action?docID=4470572

Kane, Laura. "No Room for Digital Humanities in Philosophy." GC Digital Fellows. 2014. https://digitalfellows.commons.gc.cuny.edu/2014/10/08/no-room-for-digital-humanities-in-philosophy/.

Kuckartz, Udo. *Qualitative Text Analysis: A Guide to Methods, Practice & Using Software*. London : SAGE Publications Ltd, 2014. doi: 10.4135/9781446288719.

Le Roux, Brigitte, and Henry Rouanet. *Multiple Correspondence Analysis*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE Publications, Inc, 2010. https://doi.org/10.4135/9781412993906

Locke, John. *An Essay Concerning Human Understanding*. Edited by Pauline Phemister. New York: Oxford University Press, 2008.

Locke, John. *The Conduct of the Understanding*. Edited by Jonathan Bennet, 2017. https://www.earlymoderntexts.com/assets/pdfs/locke1706.pdf

Locke, John. *Two Treatises of Government*. Edited by Peter Laslett. New York: Cambridge University Press, 2003.

Merivale, Amyas. "How David Hume became the First Modern Humanist." Lecture, Conway Hall Ethical Society, London, EN, June 5, 2016. https://conwayhall.org.uk/ethicalrecord/david-hume-became-first-modern-humanist/

Moretti, Franco. *Graphs, Maps, and Trees: Abstract Models for a Literary History*. New York: Verso, 2005.

Moretti, Franco. *Distant Reading.* New York: Verso, 2013.

Murakami, Akira, Paul Thompson, Susan Hunston, and Dominik Vajn. "'What Is This Corpus about?': Using Topic Modelling to Explore a Specialised Corpus." *Corpora* 12, no. 2 (2017): 243–77. https://doi.org/10.3366/cor.2017.0118

O' Keefe, Tim. "Epicurus." *Internet Encyclopedia of Philosophy*. Accessed May 12, 2019. https://www.iep.utm.edu/epicur/

Rockwell, Geoffrey and Stéfan Sinclair. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. Cambridge: MIT Press (2016). http://hermeneuti.ca/artifice-dialogue

Rogers, Graham A. J. "John Locke." *Encyclopædia Britannica*, 2018. Accessed February 13, 2019. https://www.britannica.com/biography/John-Locke

Schmidt, Benjamin M.. "Words Alone: Dismantling Topic Models in the Humanities." *Journal of Digital Humanities* 2, no. 1 (2013): 49-65.

Sedig, Kamran and Paul Parsons. *Design of Visualizations for Human-Information Interaction: A Pattern-Based Framework*. San Rafael, California: Morgan & Claypool, 2016.

Spiro, Lisa. "Exploring the Significance of Digital Humanities for Philosophy." Digital Scholarship in the Humanities. 2013. https://digitalscholarship.wordpress.com/2013/02/26/exploring-the-significance-of-digital-humanities-for-philosophy/

Tweedie, Fiona J, and R Harald Baayen. “How Variable May a Constant Be? Measures of Lexical Richness in Perspective.” *Computers and the Humanities*, 32 (1998): 323-352. https://doi-org.login.ezproxy.library.ualberta.ca/10.1023/A:1001749303137

Wieringa, Jeri. “Using pyLDAvis with MALLET.” *From Data to Scholarship*. 2018. http://jeriwieringa.com/2018/07/17/pyLDAviz-and-Mallet/

Whitrow, G. J. (1953). “Berkeley’s Philosophy of Motion.” *The British Journal for the Philosophy of Science,* IV (13): 37–45. https://doi.org/10.1093/bjps/IV.13.37

Yule, G. Undy. *The Statistical Study of Literary Vocabulary.* Cambridge, England: Cambridge University Press, 1944.