

CSDH 2020 | 3 JUNE 2020

WORD EMBEDDING FOR THE HISTORIAN:

EMPLOYING LSI TO
UNDERSTAND HOW WORDS
WERE HISTORICALLY USED

LISA BAER-TSARFATI | PHD CANDIDATE & THINC LAB FELLOW | UNIVERSITY OF GUELPH

[SLIDE 1 / TITLE SLIDE]

This paper explores the use of semantic text analysis and vector space modeling as a method for excavating an historically appropriate understanding of the ways in which word meanings were conceptualized in the past. It argues that word embedding can be used not only to understand the semantic features connecting the words within a text or multiple texts to one another, but also as a means for characterizing words based on their semantic distance within a corpus of primary source texts. That said, I would like to begin by first relating an historical episode.



ANGUS'S AMBITION WAS SO BLATANT THAT MARGARET WAS REMOVED FROM THE REGENCY IN SEPTEMBER [1514]...

BETWEEN 1525 AND 1528 ANGUS HAD ALMOST COMPLETE CONTROL OVER THE PERSON OF THE KING, REINFORCING HIS OWN POSITION WHEN HE TOOK THE POST OF CHANCELLOR IN AUGUST 1527.

MERRIMAN, MARCUS. "DOUGLAS, ARCHIBALD, SIXTH EARL OF ANGUS (c. 1489–1557), MAGNATE AND LORD CHANCELLOR OF SCOTLAND." *OXFORD DICTIONARY OF NATIONAL BIOGRAPHY*. 23 SEP. 2004; ACCESSED 27 MAY. 2020

[SLIDE 2 / EARL OF ANGUS]

In 1537, Janet Douglas, lady Glamis, was executed on fabricated charges of treason. Accused of conspiring and imagining the death of King James V by poison, she was convicted and burned at the stake on the esplanade of Edinburgh Castle.¹ In actuality, Janet was guilty only of providing relief and support to her brother, Archibald Douglas, sixth earl of Angus, who had been declared a rebel by the king and forced to flee Scotland. James V's hatred for the earl of Angus stemmed from the earl's political maneuverings and his treatment of James's mother, Margaret Tudor. Angus had married Margaret less than a year after the death of her first husband, James IV, and had spent the fourteen years of their marriage attempting to leverage his position as the king's stepfather into the office of regent and *de facto* ruler over Scotland's affairs. During that time, he was frequently separated from Margaret, who, due to the unpopularity of her marriage to Angus, had been forced to flee to the court of her brother, Henry VIII of England; however, this did not stop Angus from drawing upon Margaret's dower income to support himself and the mistress with whom he openly lived. By 1526, Angus had gained custody of James V, practically holding the king hostage for two years until James was able to escape and rejoin his mother in Stirling. At this point, James proscribed Angus and all of the members of the Douglas family (excluding his half-sister, Margaret) from coming within seven miles of his person.²

¹ Robert Pitcairn, *Ancient Criminal Trials in Scotland*, vol. 1 (Edinburgh: Bannatyne Club, 1833), 188, 189–90; C. A. McGladdery, "Douglas, Janet, Lady Glamis (c. 1504–1537), noblewoman," *Oxford Dictionary of National Biography*, 23 Sep. 2004; Accessed 27 May. 2020. <https://www-oxforddnb-com.subzero.lib.uoguelph.ca/view/10.1093/ref:odnb/9780198614128.001.0001/odnb-9780198614128-e-7906>

² Joseph Bain, *Calendar of the State Papers Relating to Scotland and Mary, Queen of Scots, 1547–1603*, vol. 1 (Edinburgh: H.M. General Register House, 1898), 81–2; Marcus Merriman, "Douglas, Archibald, sixth earl of Angus (c. 1489–1557), magnate and lord chancellor of Scotland," *Oxford Dictionary of National Biography*, 23 Sep. 2004; Accessed 27 May. 2020. <https://www-oxforddnb-com.subzero.lib.uoguelph.ca/view/10.1093/ref:odnb/9780198614128.001.0001/odnb-9780198614128-e-7866>



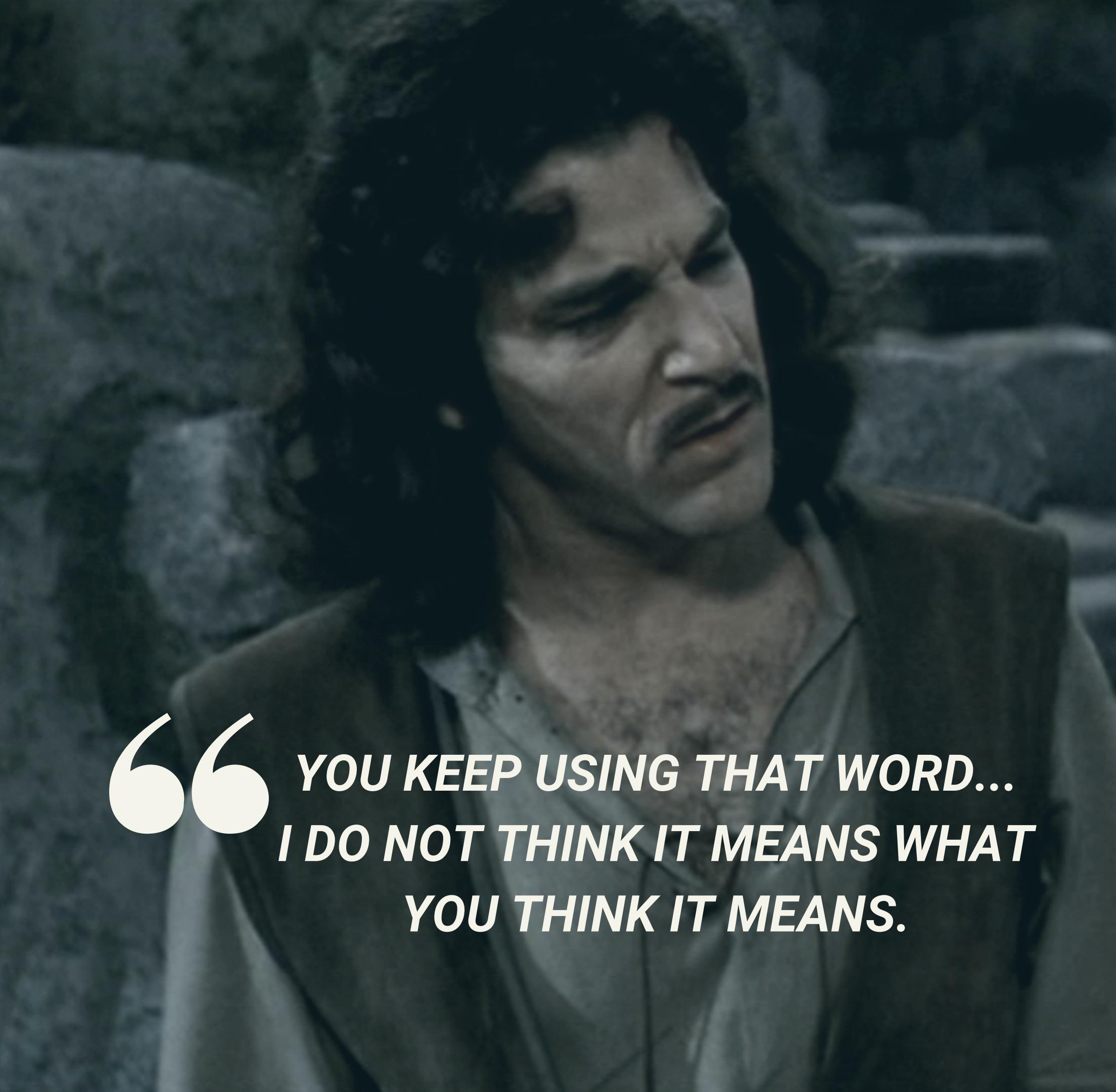
IT IS FUNDIN BE [TH]E SAID ASSISE, THAT JONET DOUGLAS LADY OF GLAMMYS HES COMMITIT ART AND PART OF [TH]E TRESSONABILL CONSPIRATIONE AND YMAGINATIOUNE OF [TH]E SLAUCHTER AND DESTRUCTIOUN OF OUR SOUERANE LORDIS MAIST NOBILL PERSONE BE POYSONE: AND FOR ART AND PART OF [TH]E TRESSONABLE ASSISTANCE, SUPPLÉ, RESSETT, INTERCOMMONYNG AND FORTIFYING OF ARCHIBALD, SUMTYME ERLL OF ANGUSE, AND GEORGE DOUGLAS, HIR BRETHER, TRAYTOURIS AND REBELLIS, IN TRESSONABILL MANER. -FOR [TH]E QUHILKIS TRESSONABLE CRIMES, [TH]E SAID JONET LADY OF GLAMMYS HES FOIRFALLIT TO OURE SOUERANE LORD, HIR LIFE, HIR LANDIS, GUDIS, MOVABLE AND VNMOVABLE: AND [TH]AT SCHO SALL BE HAD TO CASTELL HILL OF EDINBURGHE, AND [TH]AIR BRYNT IN ANE FYRE TO [TH]E DEID, AS ANE TRAYTOUR.

[SLIDE 3 / JANET DOUGLAS, LADY GLAMIS]

For the next nine years, James pursued Angus, Angus's brother, and Angus's uncle in an attempt to punish them for the outrage that he felt in relation to their ambition and misconduct. Yet, in the end, the only Douglas that James was able to punish was Angus's sister, Janet, whose only true crime had been her refusal to assist the crown in the capture of her kin. As the sixteenth century waned, historical accounts of Janet's trial and execution began to describe her offense as having been both treason and the crime of witchcraft.³ Although the legal reality surrounding Janet's conviction never included charges of witchcraft, accusations of witchcraft in this period were frequently employed to control the behaviour of women, particularly women who were perceived as behaving in too ambitious a manner.⁴ Moreover, as in the case of Janet Douglas, lady Glamis, attacking the kinswomen of particular men was an accepted tactic used by the political community to control the ambition of these particular individuals.

³ David Hume of Godscroft, *The History of the Houses of Douglas and Angus* (Edinburgh: Evan Taylor, 1644), 261.

⁴ Lisa Baer-Tsarfat, "Gender, Authority and Control: Male Invective and the Restriction of Female Ambition, 1583–161," *International Review of Scottish Studies* 44 (2019): 47–8.



THE PROBLEM

- Understanding language precisely within its original, historical context.
- Understanding what was meant by "ambition"; understanding how "ambition" was used.

“YOU KEEP USING THAT WORD...
I DO NOT THINK IT MEANS WHAT
YOU THINK IT MEANS.”

[SLIDE 4 / THE PROBLEM]

- Doctoral research examines the language of ambition and the ways in which it was employed to support and perpetuate misogyny and patriarchal social structures in early modern Scotland and England.
- ‘Ambition’ frequent topic of concern in early modern discourse.
- What is meant by ‘ambition’?
- Defining words with historical accuracy a common problem for scholars looking at texts created in the past.
- Historians are often confronted with the challenge of defining words or ideas in an historically appropriate manner. Language evolves; words lose some meanings and gain others over time, and it is important, when examining the past, for the historian to ensure that their analysis accurately reflects the language in use during the chosen period of study.
- Typically, historians look at etymological dictionaries.
 - Approach slow and cumbersome and unfeasible for sheer number of words in any given text.
- That said, some historians do not approach source content with linguistic caution, whether from no considering the way that language evolves or because there is no current robust and easy methodology to do so.
- I argue that it is not always possible to grasp the full and nuanced meaning of a text based on one’s knowledge of the way that language is used today: ‘ambition’ perfect example of this.

THE MODERN CONCEPT OF
AMBITION



AMBITION (n., v.)

An earnest desire for some type of achievement or distinction; the object, state, or result desired or sought after; the desire for work or activity; energy.

To seek earnestly or to aspire to something.

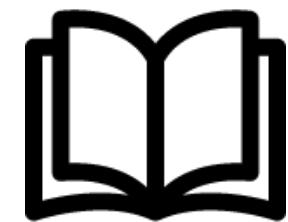
FROM DICTIONARY.COM

[SLIDE 5 / THE MODERN CONCEPT OF AMBITION]

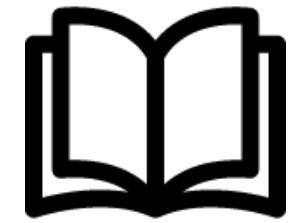
- 21st-century conceptions of ambition: see slide for more detail.
- Early modern conceptions of ambition = explicit desire for power causing or caused by traditional Christian vices like pride or greed.
- ‘Ambition’ today largely positive (except for female ambition which is still classified as illegitimate or undeserving).
- Early modern ambition universally negative and vilified regardless of gender.

SOURCES

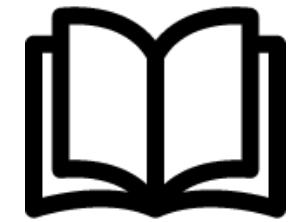
**TOTAL OF 280 PRINTED
DISCOURSE TEXTS**



Sermons, Religious Essays, &
Conduct Literature



Classical Texts, Neo-Classical
Humanist Texts, Histories &
Chronicles



Plays, Poetry, Songs, & Proverbs

[SLIDE 6 / SOURCES]

- Understanding derived from close reading of sources:
 - Sermons
 - Moral essays
 - Conduct literature
 - Mirror of princes literature
 - Classical and neo-classical humanist texts
 - Plays
 - Poetry
 - Songs
 - Proverbs
 - Histories and chronicles of ancient and contemporaneous kingdoms
- Results can be verified/reinforced by computational text analysis.
- Texts selected because they have instructional intent; that is, they were written to inform the public and mould the behaviour of prospective readers.
- In selected texts, discussion of ambition is both instructive and constructive.

NATURAL LANGUAGE PROCESSING (NLP)

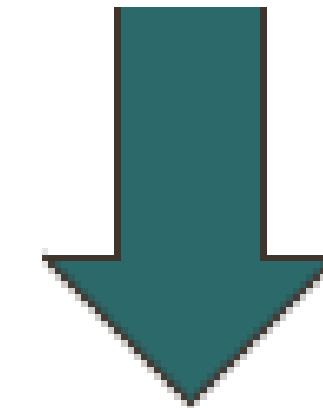
The process of teaching computers to interpret language in the way that humans naturally speak and write it.

MACHINE LEARNING

Statistical techniques used to identify parts of speech, entities, sentiment, and other parts of written text.

SEMANTIC TEXT ANALYSIS

Incorporates computer science and programming to locate relationships between words within texts and between texts within a larger corpus of works.



LATENT SEMANTIC ANALYSIS (WORD EMBEDDING MODELS)

Projects documents and terms into a low-dimensional space that represents the semantic concepts within the documents.

[SLIDE 7 / SEMANTIC TEXT ANALYSIS]

- Computational analysis drawn from research in the field of lexical statistical semantics, also known as vector space modelling, word space modelling, or word embedding.
- Uses statistics on word distributions to generate high-dimensional vector space.
- Words in space represented by vectors whose relative locations indicate semantic similarity to one another.

THEORY OF NEW APPROACH

DISTRIBUTIONAL HYPOTHESIS OF LINGUISTICS

- Lexemes with similar linguistic contexts have similar meanings." (Lenci, 2018: p. 152)
- Can be used to provide the meaning or characterization of a word.

Context words
(former)

Target word

Context words
(latter)

I read a *fascinating* book last week.



Looks like they have a similar meaning!

I read an *interesting* book last week.

Context words
(former)

Target word

Context words
(latter)

[SLIDE 8 / DISTRIBUTIONAL HYPOTHESIS]

- Word embedding models based on Zellig Harris's distributional hypothesis.
 - Words that occur in similar contexts also tend to have similar properties (i.e., meanings and functions).

The screenshot shows a Jupyter Notebook interface running on a MacBook. The notebook has two cells. The first cell, labeled 'In [10]', contains the command `tfidf = TfidfModel(corpus) # step 1 -- initialize a model`. The second cell, labeled 'In [11]', contains the command `corpus_tfidf = tfidf[corpus]` followed by a loop that prints each document's TF-IDF vector. The output for cell 11 shows the following data:

```
[(0, 0.47225121043470336), (1, 0.5461746229102505), (2, 0.5461746229102505), (3, 0.19140101671724688), (4, 0.37911872496653715)]
[(3, 0.6572080448736072), (5, 0.7537092183019993)]
[(3, 0.6572080448736072), (5, 0.7537092183019993)]
[(3, 0.129265879870477), (5, 0.14824664127328913), (6, 0.908432461546515), (7, 0.36886817219846935)]
[(3, 0.24052343645934593), (8, 0.6863484816453617), (9, 0.6863484816453617)]
[(5, 0.3729067629008573), (10, 0.9278688194905591)]
[(11, 1.0)]
[(3, 0.23404953520419178), (9, 0.3339374022796142), (12, 0.2887397836165715), (13, 0.8662193508497144)]
[(5, 0.19700972527583163), (7, 0.4902007670706823), (14, 0.4902007670706823), (15, 0.4902007670706823), (16, 0.4902007670706823)]
[(17, 0.5028931041726266), (18, 0.8643486135672457)]
[(3, 0.24052343645934593), (19, 0.6863484816453617), (20, 0.6863484816453617)]
[(5, 0.3729067629008573), (19, 0.9278688194905591)]
[(5, 1.0)]
[(3, 0.33071968076629255), (21, 0.9437290356632255)]
[(0, 0.5220791009700015), (5, 0.24266571713525517), (22, 0.35130269956021853), (23, 0.5220791009700015), (24, 0.5220791009700015)]
[(3, 0.22291815985524296), (22, 0.18505016283276532), (23, 0.5500146897079108), (25, 0.7832929032921402)]
[(20, 0.5940438871516261), (22, 0.3456250981782976), (24, 0.5136414856916389), (26, 0.5136414856916389)]
[(14, 0.6539284632122719), (17, 0.38046698937184087), (27, 0.6539284632122719)]
[(2, 0.6539284632122719), (17, 0.38046698937184087), (27, 0.6539284632122719)]
[(15, 0.5472976192550933), (22, 0.3184273039064339), (28, 0.5472976192550933), (29, 0.5472976192550933)]
[(22, 0.4188783096520597), (30, 0.5533677206959973), (31, 0.7199480032577005)]
[(1, 0.35540051210217025), (27, 0.37660036211767582), (24, 0.41106323378042636), (21, 0.4754074740521624), (22, 0.47
```

TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY

[SLIDE 9 / TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY]

- Latent Semantic Analysis [LSA] model behind word embedding.
- Originally developed by Thomas Landauer and Susan Dumais in 1997 to assist in information retrieval.
- Process has been used in last 20 years to model wide variety of semantic phenomena.
- Method analyzes a large linguistic corpus/collection of texts through creation of word-to-document matrix that is used to build linguistic vector space model.
- Matrix is built by using term frequency-inverse document frequency (tf-idf) term-weighting algorithm.
- Tf represents the number of times a certain content word represented by its lemma occurs within a particular document. Like stop words, “domain specificity” adds unnecessary noise to analysis. Domain specificity occurs when corpus is centered around a single concept or group of concepts with all texts including these concepts. In order to counter this, tf-idf weighs term frequency by the number of corpus documents in which the term appears.
- Each position in vector matrix corresponds to different word.
- Documents are represented by counting the number of times each word appears in said document.
- Word counts normalized by using logarithmic function to calculate frequency with which each word appears in all documents contained within corpus in order to give less frequent terms more weight.
- As raw data, tf-idf matrices can be misleading.
- Words whose meanings are unrelated do sometimes co-occur.
- Matrices are also usually massive but sparsely populated because most words only physically co-occur with relatively few other words.

- This is what makes LSA so robust. LSA focuses on latent structures that underlie this data by using a weighting scheme that de-emphasizes semantically uninformative words.
- LSA uses singular-value decomposition to reduce dimensionality of the original matrix to 300 active dimensions.
- Dimensionality reduction removes as much irrelevant noise in the corpus as possible in order to reveal its latent structure and fills in original sparse matrix in order to relate the main meaning-bearing words to one another whether they physically co-occur with one another in the corpus or not.
- In LSA, each word and document are represented by a vector of 300 numbers.
- By themselves, numbers are meaningless but together, they can be used to define a highly dimensional semantic space – a theoretically physical map of meanings.
- Like a two-dimensional map that can be used to locate any two points with respect to each other and then measure the distance between them, in a word embedding model, one is able to locate word meanings and document meanings within the 300-dimensional vector space and determine the distance between them.
- Distance is calculated by determining the cosine between the two vectors of interest. Words that are unrelated have a very large angle of separation, while words that are more similar have a much smaller angle between them.
- Practical application of the method was undertaken by employing LSA python package published by Gensim, an open-source library like NLTK whose programs can be used for unsupervised topic modelling and natural language processing through the utilization of modern statistical machine learning.
- Gensim designed to handle large corpora by shifting onus of processing to internet servers.

- Once texts cleaned and preprocessed, computer instructed to read each directory as its own corpus.
- Corpus then treated as a “bag of words”, with each word corresponding to an assigned integer in preparation for the mathematical analysis of these terms.
- After words translated into numbers, the tf-idf matrix was built by the machine, which was then instructed to transform this matrix into a vector space utilizing the Gensim Latent Semantic Indexing libraries.
- Final output of all processing was a collection of matrices, examples of which have been provided on the slide.
- Note that though the computer performs mathematical functions on texts to calculate physical distance between terms, it is not capable of performing analysis.
- Machine output must be read and interpreted by a human researcher to determine its value and utility for answering the question posed.

TABLE 1

Presence of words in first
ten documents of corpus.

Word Concept	Documents									
	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Ambition	1	1	1	1	1	1	1	1	0	1
Avarice	1	0	1	1	0	1	0	0	0	0
Church	0	0	1	0	0	1	1	1	1	0
Covetousness	1	1	1	1	0	1	1	1	1	1
Corruption	1	0	1	0	1	1	0	1	1	0
Destruction	0	0	0	0	0	0	0	0	0	1
Glory	0	1	0	0	0	1	1	0	0	0
Greatness	0	1	1	0	0	0	0	0	0	0
Honour	1	1	0	0	1	0	1	1	0	1
Pope	0	0	1	0	0	0	0	0	0	1
Power	1	1	1	0	0	1	1	1	1	1
Pride	0	1	1	0	1	1	1	1	1	1
Religion	0	0	1	0	0	1	0	0	0	0
Rome	0	0	1	0	0	0	0	0	0	0
Treason	1	0	0	1	0	0	0	1	1	0
Vanity	1	0	0	1	1	0	1	1	1	1
Virtue	1	1	0	0	1	0	0	0	0	0

[SLIDE 10 / TABLE 1]

- LSA processing of corpus suggested that ‘ambition’, ‘pride’, ‘avarice’, ‘covetousness’, ‘corruption’, ‘vanity’, ‘destruction’, ‘power’, and ‘treason’ were all semantically related.
- Also suggested that ‘ambition’ was related to the words ‘honour’, ‘glory’, and ‘greatness’, as well as the words ‘church’, ‘religion’, ‘pope’, and ‘rome’.
- These results are demonstrated by Table 1, which presents a snippet view of the relationship between the first ten sources in the corpus and the words introduced just now.
- One of the first steps conducted by LSA is the determination of whether or not a particular word exists within each of the texts within the corpus.
- The result of this test is then recorded in binary fashion, with “1” indicating that the word is present in the searched document and “0” indicating that it is not present in the searched document.
- This table is a representation of the first matrix (words by document matrix) necessary for the word embedding model.

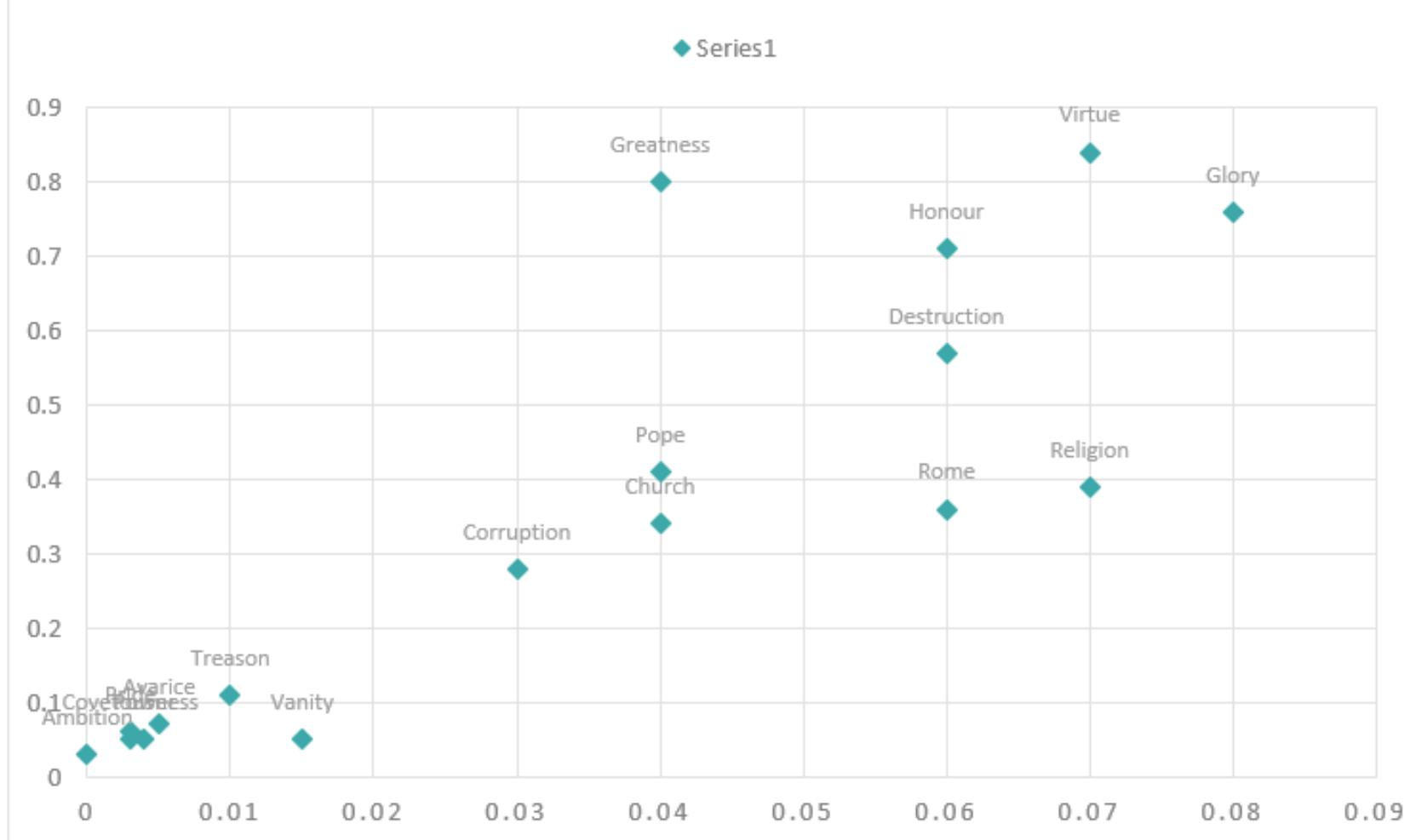
LSA MATRIX

Ambition	0	0.03
Avarice	0.005	0.07
Church	0.04	0.34
Covetousness	0.003	0.05
Corruption	0.03	0.28
Destruction	0.06	0.57
Glory	0.08	0.76
Greatness	0.04	0.8
Honour	0.06	0.71
Pope	0.04	0.41
Power	0.004	0.05
Pride	0.003	0.06
Religion	0.07	0.39
Rome	0.06	0.36
Treason	0.01	0.11
Vanity	0.015	0.05
Virtue	0.07	0.84

[SLIDE 11 / LSA MATRIX]

- LSA works by reducing the dimensionality of tf-idf analysis, resulting in an output of three values for each word concept and for each document in corpus.
- Values represent three dimensions that can be used to plot a physical, semantic position for each word within examined corpus.
- Values are calculated as cosine relationship between words and other words, between words and corpus documents, and between documents and other documents in same corpus.
- The frequency dimension, which calculates the number of times a word appears in a document or the length of each document is not useful when determining special relationships between words. It is therefore ignored in favour of the remaining two dimensions.
- For seventeen words determined to be most closely related to one another, LSA produced the matrix here.
- Vector values represent the vector relationships between these words to one another and to the corpus at large.
- The lower the value, the more closely related the terms are to one another.

LSA OF AMBITION IN DISCOURSE TEXTS



[SLIDE 12 / GRAPH]

- This is even more evident when looking at the graph.
- As we can see from the graph above, the terms ‘ambition’, ‘avarice’, ‘power’, ‘covetousness’, and ‘pride’ are very closely clustered together, with ‘vanity’, ‘treason’, ‘corruption’, ‘church’, ‘pope’, ‘Rome’, ‘religion’, and ‘destruction’ occurring further away from ‘ambition’ in the vector representation.
- When viewed this way, one can begin to construct a characterization of ‘ambition’ in 16th- and 17th-century Britain.
- If theoretical foundations of LSA and distributional hypothesis are correct, then ‘ambition’ in this period shared closest semantic features with words like ‘pride’, ‘avarice’, ‘vanity’, and ‘power’.
- Ambition also shares semantic features with religiously charged terms; however, a closer examination of the discourse texts used for this analysis provides some explanation: texts written at the height of the sixteenth-century Reformation are full of criticisms against the Catholic Church and against the various Protestant movements growing across Europe. These similarity features, then, can be mostly discarded and emphasis placed on those relationships that appear to exist independent of any historical political, social, or religious movement.
- The characterization produced is quite negative within discourse texts but discourse texts are likely to have both shaped and been shaped by the societies consuming the published works in this corpus.

DISCOURSE ANALYSIS

- People are both constructed, constrained, and shaped by the discourse to which they are exposed.
- Discourse, as an agent of power, can also be used to construct meanings and understandings of concepts like ambition, and that this construction is further enacted in the attitudes and beliefs that are formed about individuals who are described by such concepts.

[SLIDE 13 / DISCOURSE ANALYSIS]

- Discourse analysis most relevant theoretical framework for application of DH methods to study of early modern conceptions of ambition.
- Most studies in field of discourse analysis focus on speech but general theoretical approaches are still relevant to the study of written text.
- Of prime importance: language is both constructed and constructive.
- For instance, though “odd, discrepant, or deviant behaviour” has occurred throughout human history, the same behaviours have been described at various points of time in human history as “schizophrenia”, “witchcraft”, or “adolescent delinquency”.⁵ This behaviour is understood based on the words that are chosen to describe it. Thus, someone who is described as “schizophrenic” is not then also described as a “witch” because the use of the one term denies other possible meanings for the behaviour being described. In this way, the manner in which a person understands, reacts to, and judges events or behaviours is based on the words that they have available to them for description. For this reason, discourse analysts argue that discourse both produces and maintains the social reality in which we live.⁶
- An example of the ways in which this plays out is provided by Rosalind Gill. Take the use of “undocumented workers” and “illegal aliens” by American news services. The former suggests a liberal affiliation while the latter suggests a conservative affiliation. She argues that if the language we use is constructed, then in using it, we further affirm its

⁵ Margaret Wetherell, “Themes in Discourse Research: The Case of Diana,” in *Discourse Theory and Practice: A Reader*, eds. Margaret Wetherell, Stephanie Taylor, and Simeon J. Yates (London: Sage, 2001), 16.

⁶ Nelson Phillips and Cynthia Hardy, *Discourse Analysis: Investigating Processes of Social Construction: Volume 50* (London: Sage, 2002), 2–17.

construction. Yet, because no two experiences between individuals can ever be wholly identical, the use of language modifies, to some degree, its construction. Every conversation and discursive text is therefore an interaction between the constructed and the constructive elements of language and the world.⁷

- Foucault: individuals, like activities and events, cannot be thought of as immune to the constructive effects of language.
- People are both constructed, constrained, and shaped by the discourse to which they are exposed.
- Foucault and Fairclough: critical discourse analysis suggests that language and power are closely linked.
- Language users cannot stand outside the discourse they create and thus we are all caught up and shaped by the discourses that represent the social world.
- In order to understand this world, one must be a part of it and to be a part of it is to also be represented through its discourses.
- It is important to acknowledge that language does not exist as an objective material entity in the world. Instead, it only represents selected aspects of it. Because these representations are selective, they can be constructed through discourse to support some arguments and to refute others.
- Contested representations introduce a political dimension to discourse = language and power.

⁷ Rosalind Gill, "Discourse Analysis," *Qualitative Researching with Text, Image, and Sound* 1 (2000): 172–90.

- Social groups do not exist until they are constructed in discourse.
- Likewise, through discourse, identities are affirmed or denied and culture is transmitted and adapted.
- Discourse, as an agent of power, can also be used to construct meanings and understandings of concepts like ‘ambition’, and this construction is further enacted in the attitudes and beliefs that are formed about individuals who are described by such concepts.
- If ‘ambition’ shares semantic features with ‘pride’, ‘avarice’, and ‘vanity’, then it follows that using the word ‘ambitious’ to describe an individual in early modern Britain likely carried the weight of concepts associated with it, labeling target not only as ‘ambitious’ but also as ‘prideful’, ‘greedy’, and vain.



ADVANTAGES

Creates unbiased historically appropriate characterization of particular word concepts.

Allows historian to make connections between ideas that would otherwise remain hidden.

Provides insight into historical implications of language.

Expands number of useful source materials that can be consulted in the examination of research questions.

[SLIDE 14 / ADVANTAGES]

- Greatest implication of word embedding models for historical text analysis: by excavating semantic characterization of a particular word, it becomes possible to extrapolate across other texts that one is reading.
- Few manuscript texts explicitly include ‘ambition’ or its cognates; however, individuals are called avaricious, covetous, proud, vain, or treasonous.
- By demonstrating the semantic relationship between these words and the concept of ‘ambition’, a case can be made for arguing that in describing individuals with these terms, the authors of these texts may also have been implying that their targets were also ambitious.
- This opens a greater number of sources to analysis than might otherwise have been evident.
- Approaching the study of ambition in an historically sensitive way presents a number of challenges to the interested researcher. Language is not static; it is constructed by the world around it just as it shapes that world itself. It evolves, and therefore, it is necessary to first determine what early modern Scots meant and understood when they used the word “ambition” before any significant research can be undertaken in exploring the relationships that existed between ambition and gender, ambition and class, and the control of ambition as a method of exercising power.
- This paper introduced an original methodology for attacking this problem utilizing the tools and resources curated and developed by computational humanists and digital humanists for the past seventy years. These methods employed a corpus of discursive

instructional literature printed between the years of 1500 and 1625, and circulated within Scotland, to create an historically appropriate characterization of the concept of ambition in early modern Scotland and England based on the latent semantic analysis of the source material.

- Latent semantic analysis uses vector space modelling and dimensionality reduction, based on the mathematical principles of singular-value decomposition, to excavate the latent semantic relationships within text corpora and between words. Drawing from the distributional hypothesis, which states that words that tend to appear in close proximity to one another also tend to share semantic features of either meaning or function, latent semantic analysis calculates this underlying physical proximity in order to reveal those features. That said, the process of conducting this research is complex and involved. It requires a significant amount of text standardization and preprocessing, as well as a working knowledge of the Python programming language to be able to write the necessary instructional code utilized by the computer to perform LSA.
- But, the analysis of the text corpus curated for this task revealed many interesting features about the word ‘ambition’ and the ways in which it was characterized in early modern Britain. Words like ‘pride’, ‘avarice’, ‘covetousness’, ‘power’, and ‘vanity’ appeared to be most closely related according to their semantic features. The analysis also demonstrated semantic relationships between the word ‘ambition’ and words like ‘honour’, ‘glory’, ‘church’, and ‘pope’. Though this paper concluded with some discussion of the implications of these results, its primary purpose was to introduce and explain the

computational methodology that was employed in an effort to approach the study of early modern ambition with both historical sensitivity and deliberateness.

[SLIDE 15 / THANK YOU]

[SLIDE 16 / CONTACT]



THANK YOU!

CONTACT INFORMATION

EMAIL

baerl@uoguelph.ca

TWITTER

@baersafari

ACADEMIA.EDU

<https://lisabaer.academia.edu/>