Rose M Combs. Searching for Unknown Allusions: A Need to be Filled. A Master's Paper for the M.S. in I.S. degree. November, 2015. 44 pages. Advisor: Jaime Arguello

In the growing world of technology, there are tools that have been developed to explore texts for those in the social sciences or humanities. These tools allow searching and analysis to occur that previously had to be done manually. While the tools that are available meet many needs, there is one need that is not being met. The ability to locate unknown allusions has not yet been addressed. This paper explains the benefits of having the ability to locate unknown allusions. In addition, it examines some of the tools that are available and what they are capable to producing. In conclusion, a description of the needed ability of a future tool is provided.

SEARCHING FOR UNKNOWN ALLUSIONS: A NEED TO BE FILLED

by
Rose M Combs

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

November 2015

Approved by

_____

Jaime Arguello

Table of Contents

1 Introduction

For those who are familiar with Romeo and Juliet, in the beginning, Act 1 Scene 1, men from the Montague's and the Capulet's begin to sword fight. Shortly after, Benvolio comes in and exclaims, "you know not what you do" after admonishing them to stop. If that phrase, "you know not what you do," seems vaguely familiar, then you may be aware of Luke 23:34 of the Bible where Jesus is on the cross. He is being crucified and begs forgiveness from God for those who put him on the cross because "they know not what they do" (1599 Geneva Bible).

This awareness of what a few words can refer to comes from what can be defined as a literary allusion, the referencing of one word, phrase or story to another work. These allusions are found all around us and are an important part of the development of a culture. People are unable to distance themselves from speech and because of this, meaning is created through discourse. People rely on language to understand and interpret the world around them (Glucksberg, 1991).

Julia Kristeva, basing her work on that of Mikhail Bakhtin, who "was one of the first to replace the static hewing out of texts with a model where literary structure does not simply *exist* but is generated in relation to *another* structure" (Kristeva 1986, 35-36), examines the importance of realizing that text, or conversation, in and of itself is simply a collection of words. These words only develop meaning when they have something to reference. The understanding of allusions allows for a better understanding of history and

is necessary to comprehend the interactions that are encountered. Allusions are a literary device that provides deeper meaning to the text that is using it. Allusions allow for a writer to expound on their intent with either a few succinct words or through the parallel of a whole story. Allusions come from personal awareness of other words or can be gained through documented work that is then shared so that others may find greater meaning in their own reading or research. For example, a book by Naseeb Shaheen, offers extensive research on allusions found in the works of Shakespeare to the Bible. Other references such as this can be found through a simple library search.

Allusions are found everywhere and are used to better describe, interpret and relate to previous instances, whether they are found in modern movies (Carroll, 1982), the understanding of how citizenship works (Joskowicz, 2011), or for interpreting poetry (Junhui, 2014). The studies that surround literary allusions are numerous with each new article or book suggesting new connections or ways to interpret previously labeled allusions. Research regarding aspects of allusions will continue to be produced. If a system can be created that will improve the ability to locate literary allusions, then more allusions will be able to be interpreted and give greater meaning to the texts in which they are found.

Locating a literary allusion is a difficult task for a human who does not have prior knowledge of what they are looking for. In order to locate an allusion, there needs to be an initial text reference that a user would use to compare a second text to. As an example, imagine that there are two texts. Within the first text exists a single phrase that is a reference to a phrase in the second text. The user now is searching through the first text but has no knowledge of the second text. Because of this deficiency in knowledge, the

user would not become aware of the reference, or allusion, because there is no reference to bring that allusion to the users awareness. The question then, is "is there a system or tool available that can help users in their research process to search for unknown allusions?"

The challenge surrounding literary allusions is that while references to a single word may be easier to detect automatically because there is a term that can be searched for, attempts to find a phrase or a whole story may not be as easily resolved. It is necessary for the user to have prior knowledge or have a reference work that allows for them to search for allusions elsewhere. The process of determining a literary allusion is a nuanced one. Using the knowledge gained from a previous text "requires an act of interpretation to complete its sense… but it is not always easy to account for the connection between the text and its reading" (Hays 1989, 25). In order to make that connection, there is a human element that is required. This has been described as "precisely nothing more than a relational mass. It maintains relationships to other texts and to the one 'general text,' which Kristeva designates as culture… Every text is written and read in relation to that which is already written and read" (Hays et al., 2008, 4). The work of social scientists provides research on those connections that can be made between current events or discourses and throughout history. This in turn affects how cultures come to new understanding.

Even though locating an allusion requires some form of interpretation, there are some guidelines that have been established. Hays presents "seven tests" (Hays 1989, 29) that can be used to determine if a text is "an echo" of a prior text. An "echo" refers to the allusion made to the original text. These tests will be reviewed and further explained so

that they can be used as a guideline for evaluating some of the text analysis tools that are available with respect to their ability to locate allusions.

The repetition of words in a text, how closely those words are in relation to each other, and how the surrounding text may correspond with the queried text will all contribute to defining the search for an allusion. In order for an allusion to be found, as Kirsteva points out, there must be previous awareness gained form encountering the original text. If there were a connection between two texts, however, and the person reading one text was unaware of the second text, then the person would not have the ability to make that connection. There may even be times where a person reads two related texts but overlooks allusions that are present. Both of these concerns could be resolved if a system can be used to make those connections – even without the text being tagged with allusions.

This paper is meant to explore some of the tools that are currently available for those who study in the humanities or social sciences and how those tools have yet to provide an easy method that could be used to automatically identify literary allusions. This research is important to those who conduct research within the humanities and social sciences because they recognize that " '…we can no longer speak of individualized discrete texts without recourse to the universe of texts. Intertextuality thus designates the totality of connections between texts in the universe of texts' " (Hays et al., 2009, 5). Allusions are used to strengthen an argument when presenting new information or to find new connections or interpretations of what is being researched. This is especially significant as the humanities and social sciences continue to move into the digital environment and systems are being created to help in that research, such as: WordSeer, Voyant-Tools,

Monk, TAPoR, Consilience, Tessarae and Databridge. However, these systems search in differing ways – some require known words or phrases to be searched for while others have the ability to compare two texts.

Using the work of Shaheen, four Shakespeare plays with the most biblical allusions were chosen to be compared against the 1599 version of the Geneva Bible. The tools were used to examine how well each tool was able to recognize allusions. Using these tools that are mentioned, searches were conducted based on their requirements to explore the extent to which they were able to reproduce known literary allusions. Though hopeful that at least one of these would produce something close to the desired effect of locating known allusions, none of the tools were able to present anything that would represent searching for allusions. Most of them require known information to be searched for, in other words, if I were to enter a specific allusion, the tool would be able to locate something to return. Searching for a single allusion was not performed because the intent was to search for allusions based on a full text.

Concillience and DataBridge were considered; however, these tools do not offer the ability to create searches in the manner that was being explored here. These allow for links between texts to be created with the ability to search through those connections that already exist. This can be offered as a solution to tracking allusions, but would not provide any base to create a tool that could allow automated searches to be conducted.

This paper will describe the results of searching for known allusions based on the work of previously mentioned Naseeb Shaheen using each of these tools (if they are available), concluding with suggestions of where future designs could go.

2 Literature Review

2.1 Approaches to Text-Mining

There are various methods that are currently used and have been used to conduct text analysis. The research surrounding text categorization, or text classification, has significantly increased over the last decade or so. This is due to the growing availability of documents, both modern and historical, that are being converted into digital texts. The increase of digital texts allows for new ways to access and manipulate texts (Sebastiani 2002; Bijalwan et al. 2014; Kahn 2010). The exploration of these texts using text-mining techniques allows users to "extract explicit and implicit concepts and semantic relations between concepts" (Gupta & Lehal, 2009, 73).

Text classification is the organization of documents into predefined categories based on features within the document (Joachims, 1998). This provides a tool used in Web searches, spam filters, recommender systems, ad placement, credit scoring, fraud detection, stock trading, drug design, Wikipedia, as well as other applications (Domingos, 2012; Wang et al, 2009). These approaches are used in text mining to be able to analyze and extract data from texts for a variety of reasons and there are several text classification methods.

The Naïve Bayes model is based on the idea that a term, or instance, being searched for can be sorted into specific categories based on the probability of that term occurring. The probabilities that are assigned to a term are representative of previous observations, which are then used to determine the probability of a combination of terms occurring at the same time. This is used in text-mining to predict the likeliness of a term falling into a particular category (Baxter, 1997; Dumais et al., 1991; Yang & Liu 1999). This algorithm

is what the code for Tessare is built on (http://tesserae.caset.buffalo.edu/blog/latin-greek-search-competing-methods/). I think that this could be a useful tool in locating allusions based on terms because the text that was input into a system could then have its text parsed and then use those as queries looking for text with similar terms against the corpus.

The challenge of attempting to use text mining to locate unknown allusions is that it is not only a simple search and retrieval function, but also an attempt to create the terms or phrases that will be used in the search and retrieval.

2.2 Text Retrieval Software

The methods listed above all contribute to text retrieval or data mining. They are designed to assist in the searching and analyses of a document or documents based on user supplied criteria. Each method described above provides a way to make connections to other relevant documents at a faster rate than a user would be able to manually (Knoth et al., 2010). The variations that have been created around some of these methods usually require a single phrase or word to be searched or compared against the remaining texts. The progression that has moved forward with regards to searching capabilities has been diverting away from the single term search. Programs have been developed that now allow a user to compare two texts, search out similar phrases within a single text or multiple texts, or use some variation/combination of the given methods to produce searches that better facilitate searching for specific connections between texts.

The question "What allusion detection software is publicly available?" posted by "tedunderwood" in The Association for Computers and the Humanities website (http://digitalhumanities.org/answers/topic/what-allusion-detection-software-is-publicly-

available) represents the desire for such software to be available. This question only generated two responses; one of the responses seems to come from a teacher or professor who offers personal help to accomplish such a task. In addition, the company Brepolis will perform customized searches manually for professors and researchers who are searching for specific allusions. This demonstrates that the ability to locate allusions is possible, only with individualized work-arounds to produce the desired results. In response to similar requests, software that detects various rhetorical similarities is in the process of being researched and designed. Further evidence that such tools are needed is presented through Text Analysis Portal for Research or TAPoR, located at (http://www.tapor.ca), which shows that collections are being created to help researchers find the means that they need for interpreting and analyzing texts. Below are descriptions of some of the software that has been developed for use with text analysis. These tools will be explored to see if they possess the capability of locating known allusions by using the four chosen Shakespeare plays with their associated, known, allusions, and comparing them with the Geneva Bible.

2.2.1 WordSeer

WordSeer is in the creation and testing stages at this time. It is being designed to be a web-based text analysis tool to be used by those in the humanities or social sciences (http://wordseer.berkeley.edu). In 2012, a group of undergraduate students in a Shakespeare class used and analyzed several text-mining tools that were available. While learning these tools, they were required to blog about what they liked and did not like. These blogs were later used as a guide for what abilities a user would like to have when researching in the humanities. WordSeer offers a visual environment for text analysis that

will allow a user to compare texts and find similarities that will be of use to the researcher (Hearst & Muralidharan, 2014). The program was developed from the dissertation work of Aditi Shrikumar (Shrikumar, 2013) as a way to address the lack of software available to those in the humanities.

2.2.2 Voyant-Tools

Voyant-Tools is a website designed as part of a collaborative work to compare and analyze texts (http://voyant-tools.org). The idea behind this project is to create the ability to compare texts across different formats and locations. The texts can either be uploaded as one of several files or referenced from an online location. After the two, or more, texts are inserted into the search function, an analysis of the text(s) is returned that presents comparisons as well as term frequencies and distribution. This is a program to support text analysis for those in any field and is still being developed further (Sinclair & Rockwell, 2014).

2.2.3 Monk

Monk was researched between 2007 and 2009 through the University of Illinois at Urbana-Champaign. The Andrew Mellon Foundation supported the research and it created a digital system that would help scholars in the humanities to make connections between the texts that they were analyzing. The project ended in 2010 and they have provided some of the algorithms they developed so others can have access to them through the HathiTrust Research Center. Documents following the research conducted during this period can be found at the MonkWiki page: https://apps.lis.illinois.edu/wiki/display/MONK/The+MONK+Project+Wiki. In addition, the wiki page contains information pertaining to the tools and suggested reading

surrounding the project (Unsworth, 2010). The final version of MONK can be downloaded at http://quest.library.illinois.edu/monk/project/downloads/.

2.2.4 TAPoR

TAPoR provides a central location for a user to go in order to find various tools to analyze text with. The intent was to create a space where tools can be evaluated and commented on by users. From the main page, a user can determine what they want to discover and use what will best suit their needs. The tools are sorted into groups to allow a user to browse through was is available.

2.2.5 Tesserae

The Tesserae project is another creation of a literary detection software (http://tesserae.caset.buffalo.edu). The Tesserae project began with a focus on Latin poetry and has since developed into searches in Greek and English. When conducting a search, the user chooses the two texts to compare and then submits the query. This returns a list of the comparisons based on the score assigned and the list can then be sorted. In addition to this capability, they are expanding into other development projects, which can be found under the “Other Tools” section. Within this section is the ability to compare a text to a corpus to find parallels, matching of texts between Latin and Greek sources, searching for thematic comparisons, as well as four addition search functions. The sources used for this project are texts from the Latin Library and The Perseus Project. They have a corpus of just over 300 texts that the project conducts searches on and the markup of those texts have been adjusted to meet the necessary requirements for the search functions.

2.2.6 DataBridge

Databridge is a "sociometric system for long-tail science data collection" (http://databridge.web.unc.edu). It is currently under development through the University of North Carolina at Chapel Hill. This system has been created for the use of scientists who are managing a significant amount of data, or what is now known as big data. The idea is to create a collaborative environment where connections between research efforts are found automatically. There is so much data to now search that attempting to make connections manually would prove to be a struggle, especially for outlying or less well-known, research. However, by using the sociometric network analysis algorithms, it would allow connections to be made between "data, human interactions, and usage methods and practices" to create "rich models of social networks inter-connecting massive long tail science" (Rajasekar, King & Zhan, 2013).

While this tool is directed towards the sciences, this could easily be adapted to be used in the humanities. The tools that are available have made significant progress, but perhaps using a tool that utilizes the social network available to humanities researchers, would offer the ability to share knowledge as well as open up a forum for discussion surrounding allusions. This tool will not be evaluated, but is placed here as a possible solution to locating allusions.

2.2.7 Consilience

Consilience is a project currently underway at Harvard University with the goal of creating system that allows for connections to be made between unstructured texts. With the increase of digital material in blogs, websites, news, publications and others, they are designing a way to organize the data with the help of clusters (https://consilience.com/text/). This is appears to be similar to DataBridge.

3 Methodology

The scope of this project could entail anything that falls into the realm of an allusion, which could include anything from ancient texts to modern television shows. In order to determine whether or not a tool can successfully find literary allusions, tests of available tools using text with known allusions will be considered. In order to facilitate a more manageable project, the works of Shakespeare will be used. There has been significant research on understanding Shakespeare and the works that he produced. This offers a selection to choose from to compare. Within his works, there are several types of allusions. For this project, the focus will be on allusions to the Bible.

Using the work of Naseeb Shaheen in his book *Biblical References to Shakespeare's Plays*, four plays have been chosen, those with the most biblical references recorded. Each tool presented here offers different searching and analyzing abilities, which may not allow use of each of these works. With this in mind, each of these plays will be evaluated as far as each tool will allow them to be.

3.1 Defining Allusions

In a book written by Hays, he describes tests that can be used to determine whether or not terms or phrases can be defined as an allusion. The given outline refers specifically to biblical allusions; however, they can be expanded for use in the defining of literary allusions as well. For this research, we will be exploring the presence of biblical allusions found in the works of Shakespeare.

1. Availability: When comparing two texts, it is necessary to understand whether or not the original text was available for the author when the text was created. This

will require determining which text came first. In this case, we know that Shakespeare was familiar with the Geneva Bible. A clear example of this is in Richard II where it reads “Lions make leopards tame” which is a reference to “Can the blacke More change his skin? Or the leopard his spottes?” Based on the work of Shaheen, he states that this is “[a] clear reference to the Geneva Bible” because “[a]ll other versions have ‘catte of the mountaine’ instead of ‘Leopard’.” (Shaheen, 1999, 363).

2. Volume: This is determined based on the repetition of the term and how distinctive it is with regards to the text.
3. Recurrence: This refers to how often a particular passage is cited. This may not be apparent within the small design of this project, but for a larger design, the more that an original passage is referenced, the more likely one will be to find it as an allusion.
4. Thematic Coherence: Understanding whether or not a similar theme can be found between the two texts. The argument for a text to be an allusion is that there is some way that it is able to flow easily with the surrounding story compared to the original.
5. Historical Plausibility: This test is based on the requirement that the queried text would be understood in a way that agrees with the historical setting of the original text. This looks at how readers of the time would have understood a passage.
6. History of Interpretation: This relates to how others have previously interpreted and compared texts. It looks at whether or not others have detected the “echo” and

how it was understood. For this project, the allusions are based on a single interpretation, though based on a significant amount of work.

7. Satisfaction: Questions to ask here are based on the reader. It is necessary to review the text and decide if it makes sense within the surrounding text. Following that, deciding whether or not the reader is happy with the resulting interpretation and the connections that were made.

The tests described, again, are specific for recognizing biblical intertextuality, however, the tests that will be useful for this paper are (2) Volume, (3) Recurrence, and (4) Thematic (Hays, 1989). These tests have been chosen because based on the texts that will be used, Shakespeare and the 1599 Geneva Bible, it is has already been determined that the Bible predates the work of Shakespeare and other comparisons are beyond the scope of this project; this removes the first test. The second, third and fourth all will be relevant in analyzing the results of the comparisons being made. The fifth, sixth and seventh tests are not relevant for this project because the allusions that will be used have already been defined and the work that will be reviewed surrounds the ability of a system to locate these known allusions. The purpose of this is to determine if a system can be used to locate known allusions in an attempt to find a tool to locate unknown allusions.

3.2 Preparing the Texts

Using the work of Naseeb Shaheen in his book *Biblical References to Shakespeare's Plays*, he lists all of Shakespeare's plays and the biblical allusions found within. In Shaeheen's book, he suggests that there are 1604 instances of biblical allusions from Shakespeare's work and frequency table, compiled by David Crystal, displays each play and the number of possible biblical allusions. Based on that frequency table, the four

plays with the most allusions will be used in evaluating the software listed above, with the exception of DataBridge. DataBridge will not be evaluated because it is not meant as a stand-alone. This will offer 288 biblical allusions that have already been identified to use as a reference. In addition, the Geneva Bible will be used to compare the four Shakespeare texts. According to Shaheen, Shakespeare used the Geneva Bible more frequently at the time. The intent is to record which of these allusions can be recognized and note the difference between those that are and are not recognized.

The four Shakespeare plays that are going to be used are *Richard II* with 77 references, *Hamlet* with 76 references, *Richard III* with 68 references and *Henry VI part 2* with 67 references. These will be downloaded from http://shakespeare.mit.edu and entered into an excel spreadsheet so that each line corresponds to a row and the location of allusions can be marked. The first column in the spreadsheet will contain an id, the second column will be the Shakespeare line and the third column will be the class – or yes/no – in response to whether or not the line is an allusion. The defining of an allusion is based on Shaheen's work that explores Shakespeare text and explains which passages are allusions as well as what they are allusions to. Having this information is necessary in reviewing what is found from each system.

In addition, xml files of the Shakespeare plays were downloaded from http://www.ibiblio.org/xml/examples/shakespeare/ so that they could be used with some of the tools. A copy of the Geneva Bible was downloaded from https://drive.google.com/folderview?id=0B7vzRsRM2aOQOVRIeGJ0bFhydTA&usp=sharing&tid=0B7vzRsRM2aOQMW96LW1KOUxrSVk as an xml file.

3.3 Discussion of Tools

The systems that will be looked at have already been described but each needs to be reviewed to determine what aspects would be useful in locating literary allusions. In DataBridge, they use techniques to "measure semantic similarity" as well as "Probabilistic Latent Semantic Analysis (PLSA)… used to identify both synonyms (words that refer to the same topic) and polysemy (words with multiple meanings)" (Rajasekar et al., 2013, 7-8). Consilience is developing a clustering tool that uses computer-assisted methods: "1. calculate the words count matrix…2. apply clustering methods… 3. calculated a similarity distance… 4. Project this matrix of distance across all clusters… 5. calculate new clustering solution" (Rajasekar et al., 2013, 11). Other systems that will be reviewed are Voyant-Tools, Monk, Tessarare and WordSeer to compare.

Using the texts downloaded compared to the Bible, each of these systems should return some results. As mentioned previously, the total number of known allusions that will be used is 288 and are defined as the standard. Those 288 allusions will be compared to the results returned by each system and if results outside of the known 288 allusions are returned, it will be suggestive of not returning a correct result. The accuracy of each system should be determined based on precision and recall where precision is defined by how many of the returned allusions were correctly assigned as such and recall is evaluated based on the percentage of the actual allusions that were located.

Unfortunately, after evaluating each of the tools, none of them produced any results that could be compared to the desired results. The ability to determine accuracy based on precision and recall is zero for each tool.

4 Results

The desired ability for each of these tools would be to take a given text and compare it to a second text in an attempt to locate allusions and return highlighted locations in both the given text and the second text that could be related. None of these tools were designed for this use. While the goal was to mimic this process with each tool, the desired functions could not be reached. This is due to the process attempted and not to the actual functionality of each of these tools.

The following are the results of running texts through the different systems. The main page for the system is shown in the appendix as well as some of the images taken from running the program to find the allusions.

One of the problems that came up was the design of the Geneva bible that was used. While there is a pdf of the Geneva Bible, there is not a good online version that could be used to compare those tools that used a website. There are some that allowed for specific books and chapters of the Bible to be used, but not for the entire Bible.

4.1 WordSeer

The initial interaction with WordSeer offers a premade Shakespeare collection (appendix 1.15.1). After searching through the collection located at http://wordseer.berkeley.edu/shakespeare/index.php, it appears that the abilities that are publicly available allows for searching to compare individual terms or sentences. It was necessary to reexamine the four plays that were chosen so that word frequency could be taken into account between the allusions found within the four plays and use those terms to search with on WordSeer. The spreadsheets of the four texts were sorted and those

lines that contained allusions were put into a separate file. That file was entered into CheckText and the terms with the highest frequency was returned (appendix 1.14). WordSeer allows for the visualization of the text and using the terms “god” and “heaven” as a result of high frequency terms, a heatmap is created that visually represents where in the corpus the terms occur. If the user hovers over a location, an explanation is displayed of what is occurring (appendix 1.15.2). A single term or multiple terms can be entered here.

The ability to compare files or texts seems to be only available to those who download the code and enter in xml files. This is not something that was done in this project though it could be done if xml files can be found for Shakespeare and the Bible.

This tool offers many benefits for the user who is aware of what they are searching for and can enter in a term or sentence that they would like to explore. In addition, the tool allows for a user to locate a term and explore other terms that are prevalent in relation to the given term (Muralidharan and Hearst, 2012). As of right now, there are only slave narratives collection and a Shakespeare collection.

While this tool is very successful in allowing researchers to search texts for new meanings, it does not allow a user to search for allusions. If a user is able to provide a term, then searches can be conducted, but WordSeer does not offer the ability to search for unknown allusions, nor was I able to manipulate the data so that a comparison could be made.

4.2 Voyant Tools

At first look, Voyant Tools may not seem like much, but once a text or list of websites is entered, it seems to offer many great tools for a user. The home page for Voyant Tools is

a simple interface with request that you enter in the text you want to explore or enter in a single or multiple websites (appendix 1.16.1). It does not seem to have the capability to evaluate a combination of entered text and websites, only one type can be explored at a time. The initial test was to enter in both a created text file of some of the chapters of the Geneva Bible and the text of *Richard II*. When the text was not recognized, the website for Matthew 27 of the Bible was used (http://goodbooksfree.com/scriptures/genevabible/index.html). However, because of the design of the web page, this could not be interpreted well by the system.

The ability to compare the two texts was lacking, so the evaluation of *Richard II* occurred without comparing it to the Bible. Instead, the terms that were high frequency were displayed and after removing stop words from the results and marking "king," "duke," and "Richard," Voyant Tools created a visualization that displays the analysis of the single text (appendix 1.16.2).

There is a possibility that this could have allowed for a user to locate an allusion, however, locating that allusion requires more work than simply using this tool. This conclusion is based on the results found when comparing *Richard II* to *Hamlet*. It appears that one of the features presents the trend in terms (appendix 1.16.3). This could allow for terms that are similar to be located which could lead to an allusion. Though, this was not something that could be confirmed due to inability to use the files that I have.

This tool is one that allows a user to explore how the text is associated with itself as well as with additional ones, so long as a website can be provided. There is a possibility that this could locate allusions based on comparing term similarity between the two documents independently, though even then, it would only provide the comparisons, not

a location of an allusion. Based on the information that was produced using Voyant, this tool was unable to produce anything resulting in the location of an allusion.

4.3 Monk

This tool is no longer supported but does offer a downloadable format. However, in attempting to download this for use, there were other files that were required to have the program run correctly. After attempting to access the correct files, I was still unable to get the program to run.

4.4 TAPoR

When first accessing this website, it presents an extensive list of tools that are available to the researcher who would like to perform text analysis or utilize retrieval tools (appendix 1.17.1). Of the tools that are offered by TAPoR, the three that were explored were the use of concordance, co-occurrence and comparator (http://taporware.ualberta.ca/~taporware/htmlTools/findtext.shtml?). In order to use these services, TAPoR requires the user to enter a web address or text file and choose how the system should evaluate it. When attempting to upload a file for use, the system returned an error stating that the incorrect file type was uploaded but the correct file type was not mentioned. For concordance, the website for *Richard II* was used because it had the most allusions. In order to conduct a search, it was necessary to enter in a term that would be evaluated. Based on the top 10 terms that had the highest frequency among the four texts, the term “god” was entered. The results displayed that 66 entries for “god” was found and provides a list of the terms surrounding “god” (appendix 1.17.2). Similar to concordance, the use of co-occurrence requires the user to enter two terms to be compared and the system presents the user with the locations of the terms within the single text. This is only

helpful in the search for allusions if an allusion is already known and the user is searching for a specific term.

To use the comparator tool, *Richard II* was used along with an online version of the Geneva bible. The difficulty with this was that the online version of the Bible (http://goodbooksfree.com/scriptures/genevabible/index.html) is set up so that the user must choose a specific book and chapter. This meant that comparison between the two full texts was not possible. To remedy this, I searched through the Shakespeare references and found that there were two locations of allusions that referenced Matthew 27. Using Matthew 27 as the second website to compare, the results summarized what was found within each text and then a list of terms and their frequency was created (appendix 1.17.3). Similar to concordance and co-occurrence, this search did not result in locating the allusions that are present.

4.5 Tesserae

Tesserae offers a comparison between two texts, however, they must be based on the collection that is on their website, http://tesserae.caset.buffalo.edu. This limitation hinders this project in that while it does offer Shakespeare, there are only four texts that have been added to their site. In addition, the bible that is offered is not a copy of the Geneva Bible, nor is the whole bible listed. The bible appears to be divided into four sections similar to how the Tanakh (Hebrew Bible) is divided into three sections. The sections that are present are: Pentateuch, Revelation, Prophets and Writings. While these do provide an exact match with what is being explored in this project, a comparison can still be made.

From the main web page, the English tab is chosen to display the works that are available there (appendix 1.18.1). The "First Source" chosen is "Shakespeare" with the play "Hamlet" specified. The "Target Source" is the "World English Bible." With the bible offering four sections to compare, four searches were conducted to include each section: Pentateuch, Prophets, Revelation and Writings. When the results were displayed, it shows that only four books of the bible were among the first 100 results of each search. These books were Genesis, Jeremiah, Revelation and Psalms. Using these as a reference, the Shaheen book was reviewed for *Hamlet* and each allusion was looked at to see where the bible reference was located. These were searched for within each search page and while the search results demonstrate some similarity, none of the allusions listed in Shaheen's book was presented among the first 100 results across the four searches (appendix 1.18.2).

Tesserae benefits researchers who desire to search for texts that are among the library that is offered. If a user wishes to view what is among the library, they can access the list through the "Sources" button along the top of the page. The tool offers a comparison of terms along with a ranking of how similar the terms are and how close together they are located. There appear to be additional features that should be able to be adjusted, such as stopwords or distance, but this was not accessed. With regards to this project, however, it did not provide a match for any of the allusions that could have been returned.

5 Discussion

There are several tools that are currently available to those in the humanities and social sciences to desire to explore text further. While these tools all provide improved benefits over what was previously available, there are still needs that are not being met. None of the tools that were examined were able to produce any of the allusions that were initially listed as part of the gold standard. Because of this result, there was no way to use precision and recall in expressing the accuracy of these tools.

These tools are best used if a user can approach a topic with a specific theme, phrase, or word in mind. Approaching the search with a previous knowledge, a user is able to create and alter searches in order to reach some conclusion. Another limitation was the availability of databases. For instance, WordSeer only has a Slave Narrative and Shakespeare database and Tesserae is limited to classical texts. There were some comparisons that could be made if previous knowledge was available, but the intent was to find a system that could locate allusions by comparing two texts and not having to provide specific terms to search for.

These tools meet some of the needs, but one that has yet to be reached is the ability to find literary allusions in a corpus. Ideally, a user would be able to provide a text and have that text compared to a corpus. The results would be locations in the corpus that are similar to the queried text. This is not something that comes up often because of the difficulty in locating allusions. When it does arise, such searches are sent to companies that already have software that can be altered to meet the needs of the researcher, such as Brepolis. However, the ability to find allusions is something that is relevant to the study of humanities.

This request would not be a simple one to achieve because it would require a tool that not only searches for a term or phrase, but also determines what those terms or phrases should be. If a system was designed to take in a full text as its input and then was able to compare that text to a corpus, it could then offer suggestions for locations that a user to explore to search for comparisons from the corpus that may previously have been unknown. In addition, comparisons could be conducted across disciplines, without having to have a previous knowledge of the existence of an allusion.

Parsing the queried text into several smaller queries, perhaps by sentences, and then using those smaller queries to search against the corpus could provide a way to achieve this for phrases. The search would not have to be for exact sentences, but rather, for words to be similar, both in meaning and spelling, and for the user to decide what the minimum number of terms would have to be for it the search to produce a result.

For comparing overall stories, rather than just terms or phrases, the tool would not have to parse out the sentences, but rather create a map of the terms from the queried text and then use that to compare against the other texts in a corpus. Similar to how WordSeer is able to produce a map of the terms in a document, those maps of terms would be compared to see if words (either term or meaning) seem to appear in a similar fashion. This would need to be fleshed out a little more to be able to know how to best compare the terms and their locations in a text to know whether or not they would produce a positive result.

6 Conclusion

The programs or tools that were reviewed above are all beneficial and worthwhile to researchers in the humanities or social sciences to make connections where they might not be able to otherwise. However, there is still room to build upon what is currently offered to expand to the ability to recognize allusions. The difference from text mining and searching for a value is that text mining allows the user to search for unknown values while in a search "the user is typically looking for something that is already known and has been written by someone else… In text mining, the goal is to discover unknown information" (Gupta & Lehal, 2009). The challenge is to combine these two to provide the ability to search for unknown allusions.

At this point, none of the tools above were able to produce search results that matched any of the allusions that were among the standard. With none of the expected allusions located, the tests that would have been used to determine whether or not an allusion was found, were not able to be used. These are tests that would need to be used when a system is developed in order to determine if an allusion is located.

For future development, a tool could be built upon the current designs to assist in the process of locating unknown allusions. The ideal tool would allow for a user to submit a text that would be queried against an entire corpus, similar to how a user can search for a phrase in Google Books. The submitted text would then be parsed into manageable phrases, perhaps by sentence or restricted to a number of words. These would then form the values that would be queried. Each of these phrases would then be searched for and those that have multiple terms, in the same order, within a set amount of terms would be

returned as a suggestion to be further explored as an allusion. The result would document the location of the similar phrases in all texts. If there were more than one text located from the corpus, then the results could be ranked based on the number of correct terms that exist in the phrase.

For example, using the phrase “they know not what they do” is six terms long. If this were one of the phrases parsed out of the original text and I wanted the phrases to be restricted to a ten term length, then the search of a “new” text would attempt to locate some of these six terms, in that order, within a span of ten terms. Upon locating this phrase, the result returned would be the location of the suggested phrase from the queried text and the “new” text so that the user would be able to analyze the presence of an allusion.

In addition, the texts would not have to be from the same genre or discipline to be compared. For some of the tools above, they were able to conduct searches within a corpus that was already provided rather than directing to another corpus (though Tesserae used external corpora, they only used a select few). Voyant Tools offers the ability to compare websites, but anything larger is still not available, nor could it compare text to a website.

Searches between disciplines do not happen often because a prior knowledge or awareness of topics needs to present in order for such a search to be initiated. For those disciplines that are closely related, there is a greater likeliness to find research that bridges them – for example, religious studies and classics are closely related. If drawing connections beyond a single discipline is possible to automate, then a user would be able to submit a single text and compare that text across multiple corpora, finding connections

where they would not have normally been found. This model would be similar to DataBridge in the attempt to connect information that would not normally or easily be discovered. The benefit would be to not have to look at linked data, which is information that has already been produced (DataBridge or Consilience), and have users expand their research into areas where new connections can be automatically suggested.

7 References

Apté, C., Damerau, F., & Weiss, S. M. (1994). Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems (TOIS)*, *12*(3), 233-251.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*(4), 673-721.

Baxter, J. (1997). A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, *28*(1), 7-39.

Bijalwan, V., Kumari, P., Pascual, J., & Semwal, V. B. (2014). Machine learning approach for text and document mining. *ArXiv Preprint ArXiv:1406.1580.*

Carroll, N. (1982). The future of allusion: Hollywood in the seventies (and beyond). *October*, 51-81.

Crystal, D. (2011). Citing scripture for my purpose. *Around the Globe*, 48, 20-21.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78-87.

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on information and knowledge management*, 148-155.

Frank, E., & Hall, M. A. (2011). *Data mining : Practical machine learning tools and techniques.* Burlington, MA: Morgan Kaufmann.

Fuhr, N., Hartmann, S., Lustig, G., Schwantner, M., Tzeras, K., & Knorz, G. (1991). *AIR, X: A rule based multistage indexing system for large subject fields.* Citeseer.

Glucksberg, S. (1991). Beyond literal meanings: The psychology of allusion. *Psychological Science*, *2*(3), 146-152.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60-76.

Hays, R. B. (1989). *Echoes of scripture in the letters of paul.* New Haven: Yale University Press.

Hays, R. B., Alkier, S., & Huizenga, L. A. (2009). *Reading the bible intertextually.* Waco, Tex.: Baylor University Press.

Hull, D. (1994). Improving text retrieval for the routing problem using latent semantic indexing. In *SIGIR'94*, 282-291.

Hearst, M. A., & Muralidharan, A. (2014). WordSeer: A text analysis environment for humanities scholars. [Web page] Retrieved from http://wordseer.berkeley.edu.

Iwayama, M., & Tokunaga, T. (1995). Cluster-based text categorization: A comparison of category search strategies. In *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval*, 273-280.

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features.* Springer.

Joskowicz, A. (2011). Jewish anticlericalism and the making of modern jewish politics in late enlightenment prussia and france. *Jewish Social Studies*, *17*(3), 40-77.

Junhui, W. U., & Qingshun, H. E. (2014). Poetry translation: An intertextuality approach. *Studies in Literature and Language*, *9*(1), 55-62.

Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology*, *1*(1), 4-20.

King, G. (2014). Consilience. [Web page] Retrieved from http://datascience.iq.harvard.edu/consilienc.

Knoth, P., Novotny, J., & Zdrahal, Z. (2010). Automatic generation of inter-passage links based on semantic similarity. In *Proceedings of the 23rd international conference on computational linguistics*, 590-598.

Kwon, O. -W., & Lee, J. -H. (2003). Text categorization based on *k*-nearest neighbor approach for web site classification. *Information Processing & Management*, *39*(1), 25-44.

Kwon, O. -W., Jung, S. -H., Lee, J. -H., & Lee, G. (1999). Evaluation of category features and text structural information on a text categorization using memory based reasoning. In *Proceedings of the 18th international conference on computer processing of oriental languages (ICCPOLÕ99)*, 153-158.

Lam, W., & Ho, C. Y. (1998). Using a generalized instance set for automatic text categorization. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*, 81-89.

Larkey, L. S., & Croft, W. B. (1996). Combining classifiers in text categorization. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval*, 289-297.

Masand, B., Linoff, G., & Waltz, D. (1992). Classifying news stories using memory based reasoning. In *Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval*, pp. 59-65.

Muralidharan, A. & Hearst, M. A. (2012). Supporting exploratory text analysis in literature study. In *Literary and Linguistic Computing* 28(2), 283-295.

Rajasekar, A., King, G., & Zhan, J. (2013). Big data: Databridge: A sociometric system for long-tail science data collection. [Web page] Retrieved from http://databridge.web.unc.edu.

Rajasekar, A., Kum, H. -C., Crosas, M., Crabtree, J., Sankaran, S., Lander, H., Carsey, T., King, G., Zhan, J. (2013). The databridge. *Science*, *2*(1), pp-1. doi:databridge.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, *34*(1), 1-47.

Shaheen, N. (1999). *Biblical references in shakespeare's plays.* Newark: University of Delaware Press.

Shrikumar, A. (2013). *Designing an exploratory text analysis tool for humanities and social sciences research.* Thesis.

Sinclair, S., & Rockwell, G. (2014). *Voyant-Tools (version 1.0 beta (4669))* [Computer Software]. Retrieved from http://voyant-tools.org: http://voyant-tools.org.

Unsworth, J. (2010). Humanities text mining in the digital humanities (monk). [Web page] Retrieved from http://www.lis.illinois.edu/research/projects/humanities-text-mining-digital-library-monk.

Vidhya, K. A., & Aghila, G. (2010). A survey of naïve bayes machine learning approach in text document classification. *International Journal of Computer Science and Information Security (IJCSIS)*, *7*(2), 206-211.

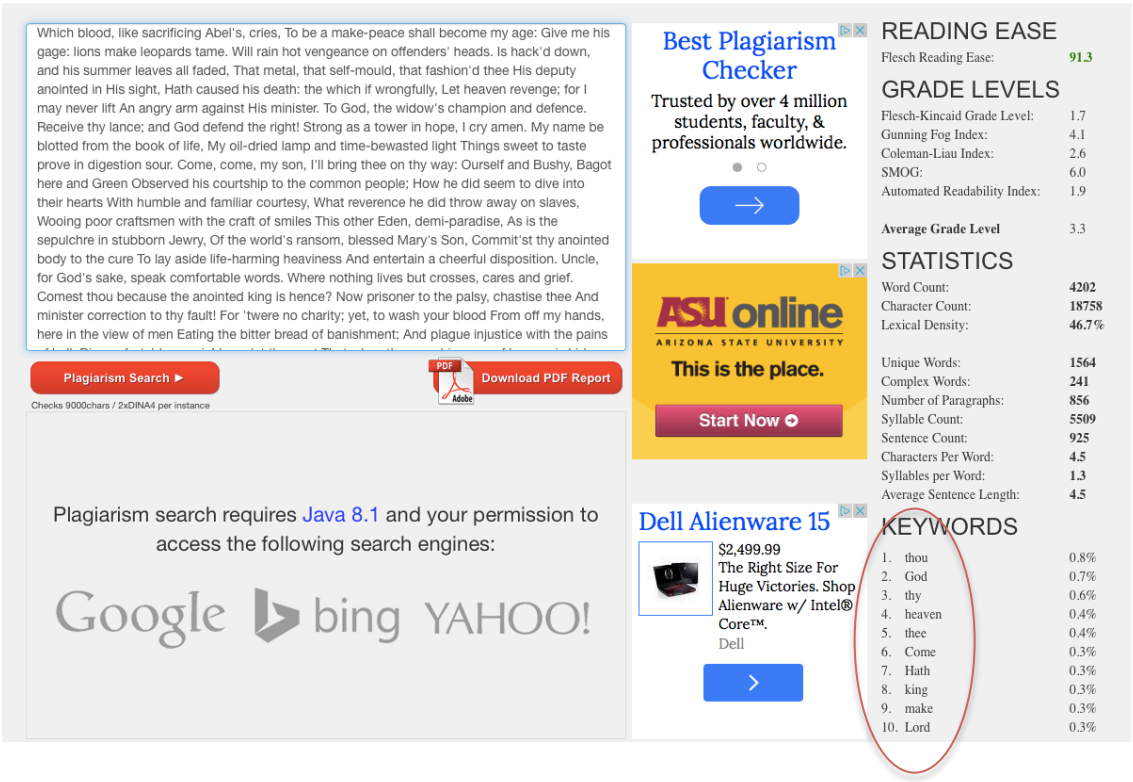
Wang, P., Hu, J., Zeng, H. -J., & Chen, Z. (2009). Using wikipedia knowledge to improve text classification. *Knowledge and Information Systems*, *19*(3), 265-281.

Xia, F., Zhang, W., Li, F., & Yang, Y. (2008). Ranking with decision tree. Knowledge and Information Systems, 17(3), 381-395.

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 42-49).

8 Appendix

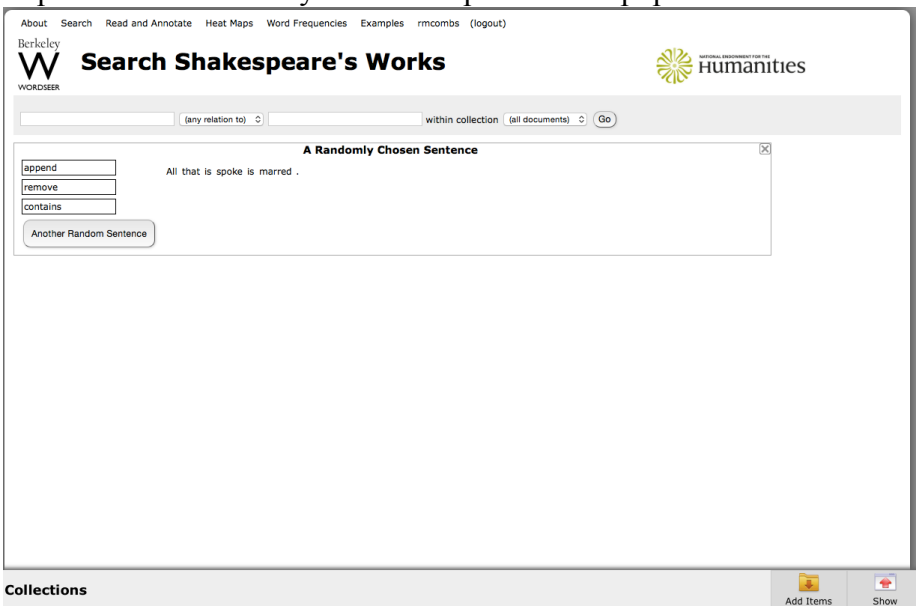
8.1 CheckText



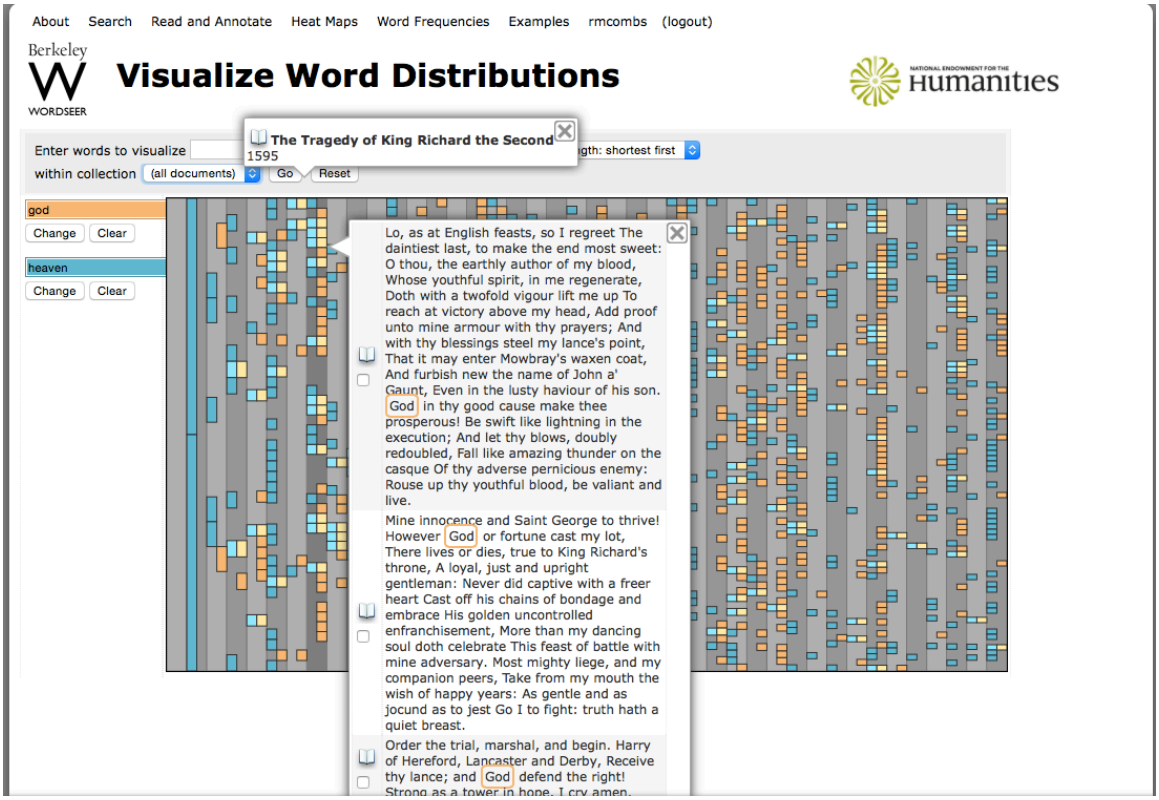
8.2 Images of WordSeer

8.2.1 Home page

http://wordseer.berkeley.edu/shakespeare/index.php

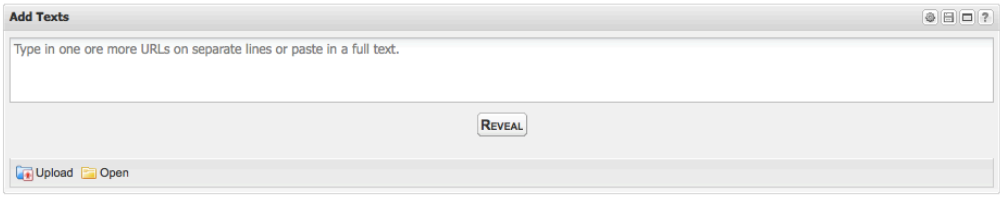


8.2.2 search terms “god” and “heaven”



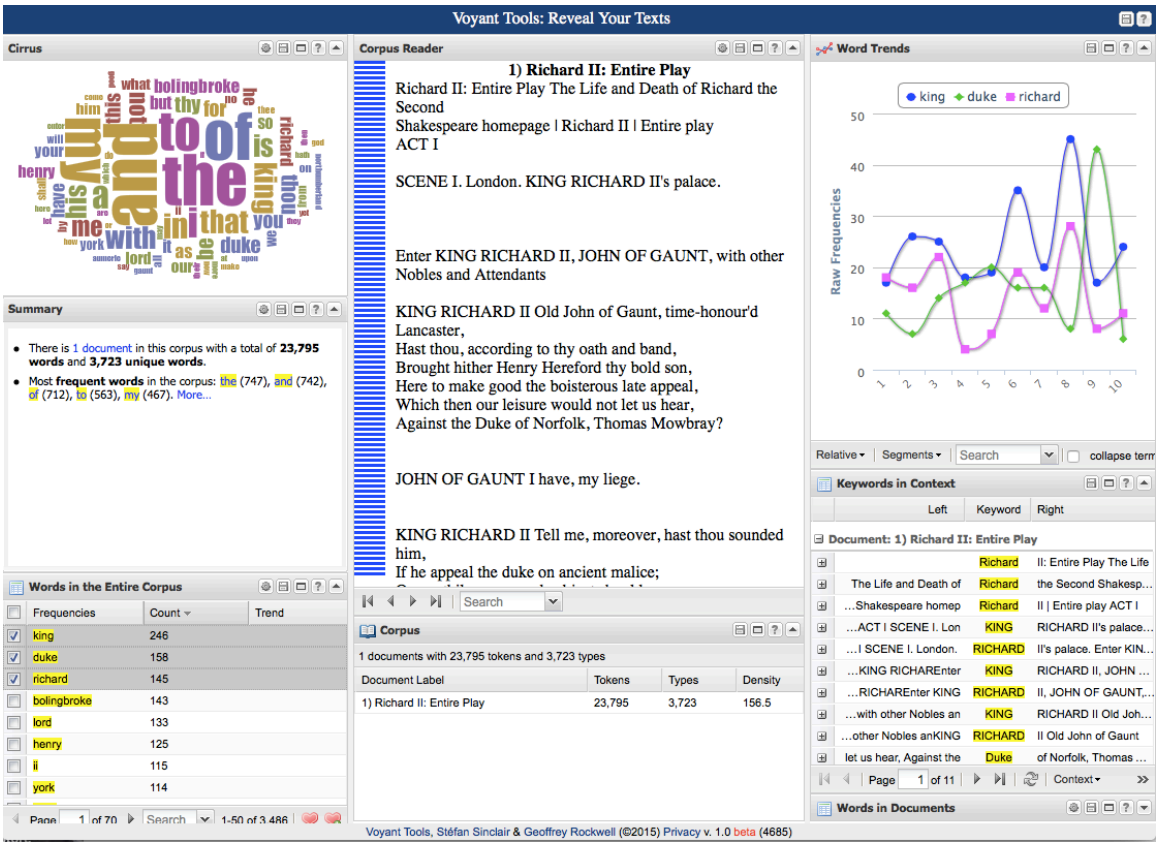
8.3 Images of Voyant Tools

8.3.1 Home Page

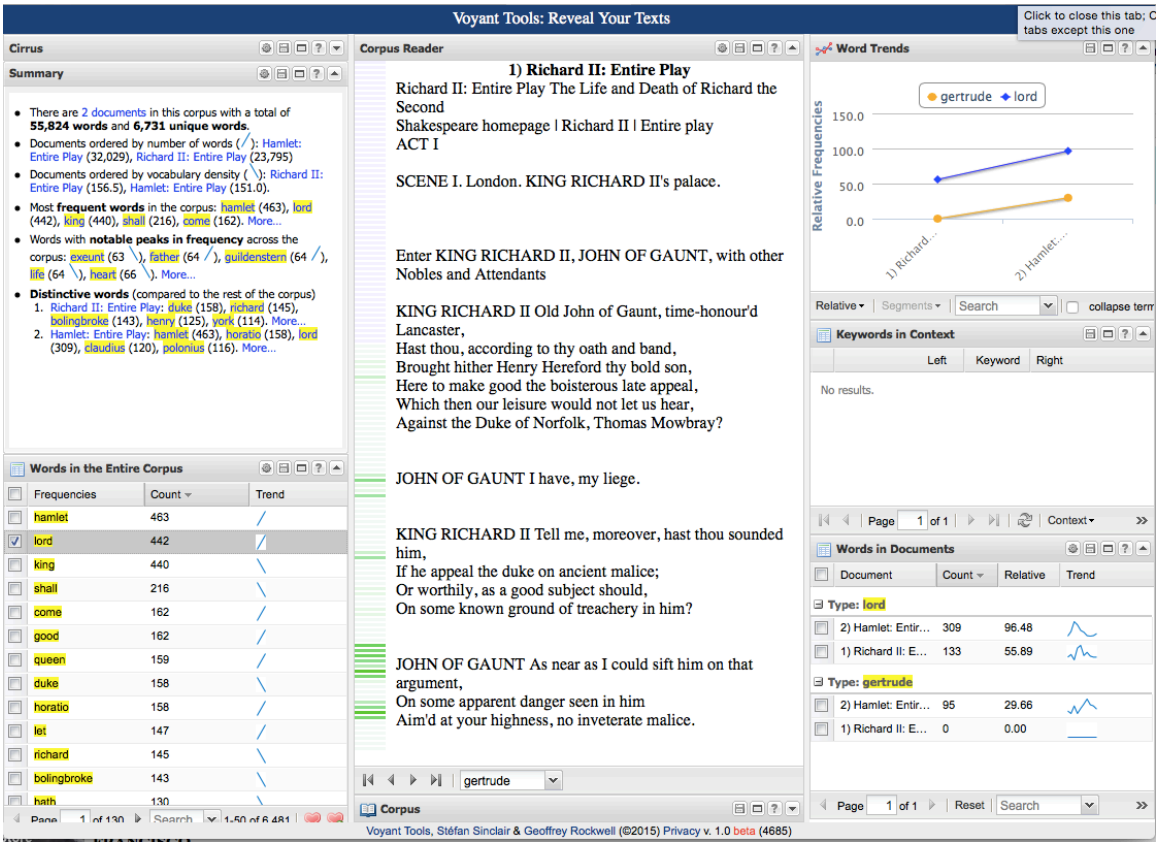


Voyant Tools is a web-based reading and analysis environment for digital texts. Find out more.

8.3.2 Richard II



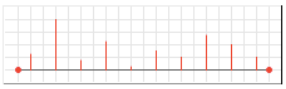
8.3.3 Richard II and Hamlet



8.4 Images of TAPoR

8.4.1 Home Page

8.4.2 TAPoR concordance



Summary: 66 entries found.

of his blood , How	**God**	and good men hate so
. HENRY BOLINGBROKE O ,	**God**	defend my soul from such
death . JOHN OF GAUNT	**God's**	is the quarrel ; for
is the quarrel ; for	**God's**	substitute , His deputy anointed
? JOHN OF GAUNT To	**God**	, the widow's champion and
cause . Lord Marshal In	**God's**	name and the king's ,
engaged by my oath-- Which	**God**	defend a knight should violate
my loyalty and truth To	**God**	, my king and my
, by the grace of	**God**	and this mine arm ,
, A traitor to my	**God**	, my king , and
, To prove , by	**God's**	grace and my body's valour
foul and dangerous , To	**God**	of heaven , King Richard
son . JOHN OF GAUNT	**God**	in thy good cause make
thrive ! THOMAS MOWBRAY However	**God**	or fortune cast my lot
Receive thy lance ; and	**God**	defend the right ! HENRY
Derby , Stands here for	**God**	, his sovereign and himself
, A traitor to his	**God**	, his king and him
, and Derby , To	**God**	, his sovereign and to
duty that you owe to	**God--**	Our part therein we banish
so help you truth and	**God**	! Embrace each other's love
But what thou art ,	**God**	, thou , and I
A brace of draymen bid	**God**	speed him well And had
II Now put it ,	**God**	, in the physician's mind
go visit him : Pray	**God**	we may make haste ,
succession ? Now , afore	**God--God**	forbid I say true !
. NORTHUMBERLAND Now , afore	**God**	, 'ti s shame such

8.4.3 TAPoR Co-Occurrence

Basic information for the two texts

	Text 1	Text 2
Text Source	http://shakespeare.mit.edu/richardii/full.html	http://goodbooksfree.com/scriptures/genevabible/40027.html
HTML page title	Richard II: Entire Play	Matthew 27 · Geneva Bible (1599) · Scriptures (Good Books Free · beta)
First HTML Heading/Bold text	ACT I	Geneva Bible (1599): Matthew 27
Opening words (50)	The Life and Death of Richard the Second Shakespeare homepage | Richard II | Entire play ACT I SCENE I. London. KING RICHARD II's palace. Enter KING RICHARD II, JOHN OF GAUNT, with other Nobles and Attendants KING RICHARD II Old John of Gaunt, time-honour'd Lancaster, Hast thou, according to	You can skip to local navigation , content or closing (global) navigation . JavaScript is disabled within your browser, and certain features are not available without it. goodbooksfree. com: good. books. free. Good Books Free (beta) Contextual Menu Chapter: ←Previous · All/ Index · Next→ Page: ↑Top · Passage
Total words	23913	2385
Unique words	3711	586
Words that occur once	2043	368
Words that occur twice	590	91
Highest word frequency	746	268
Average words frequency	6.4421	4.0630
For element <body>	Number of this element: 1 Maximum length in words: 23807 Minimum length in words: 10000 Average length in words: 23807.0000	Number of this element: 1 Maximum length in words: 2068 Minimum length in words: 2068 Average length in words: 2068.0000

Words comparison results of two texts

Common words
Words in text 1 only
Words in text 2 only

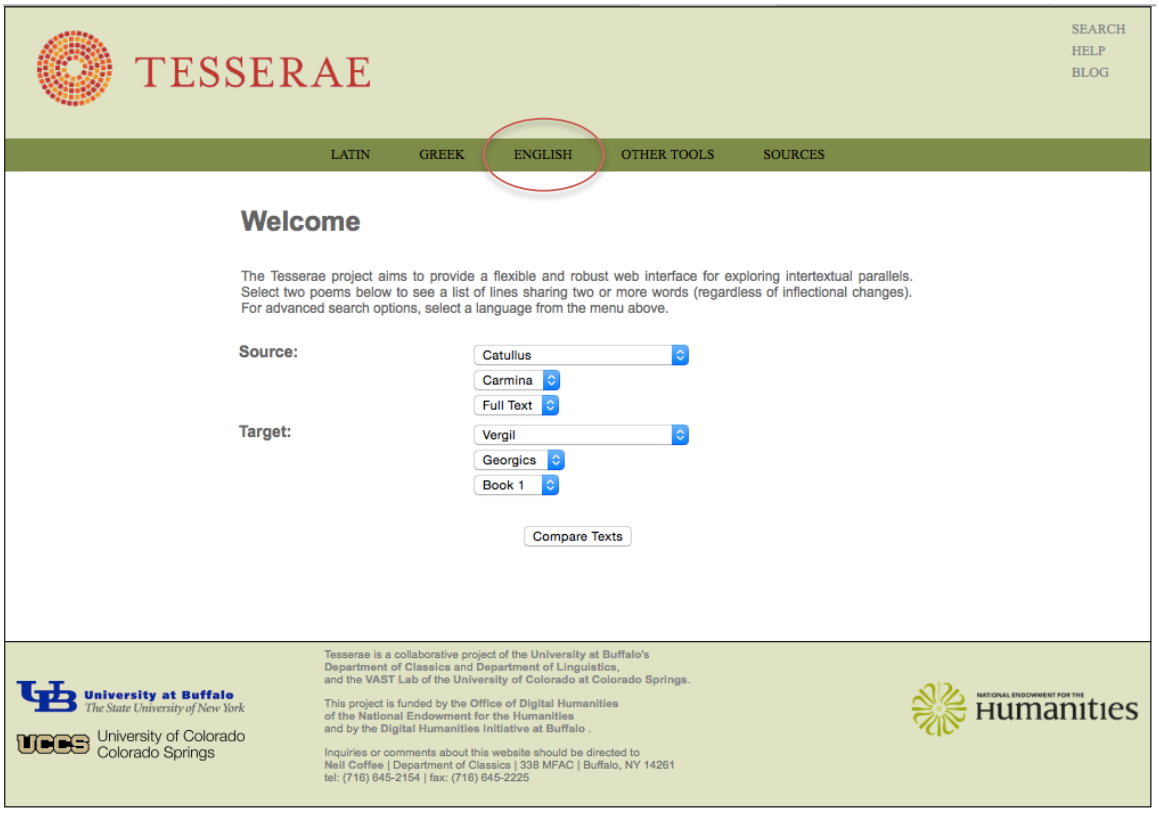
(Click any word to get its concordance and collocates.)

Common words

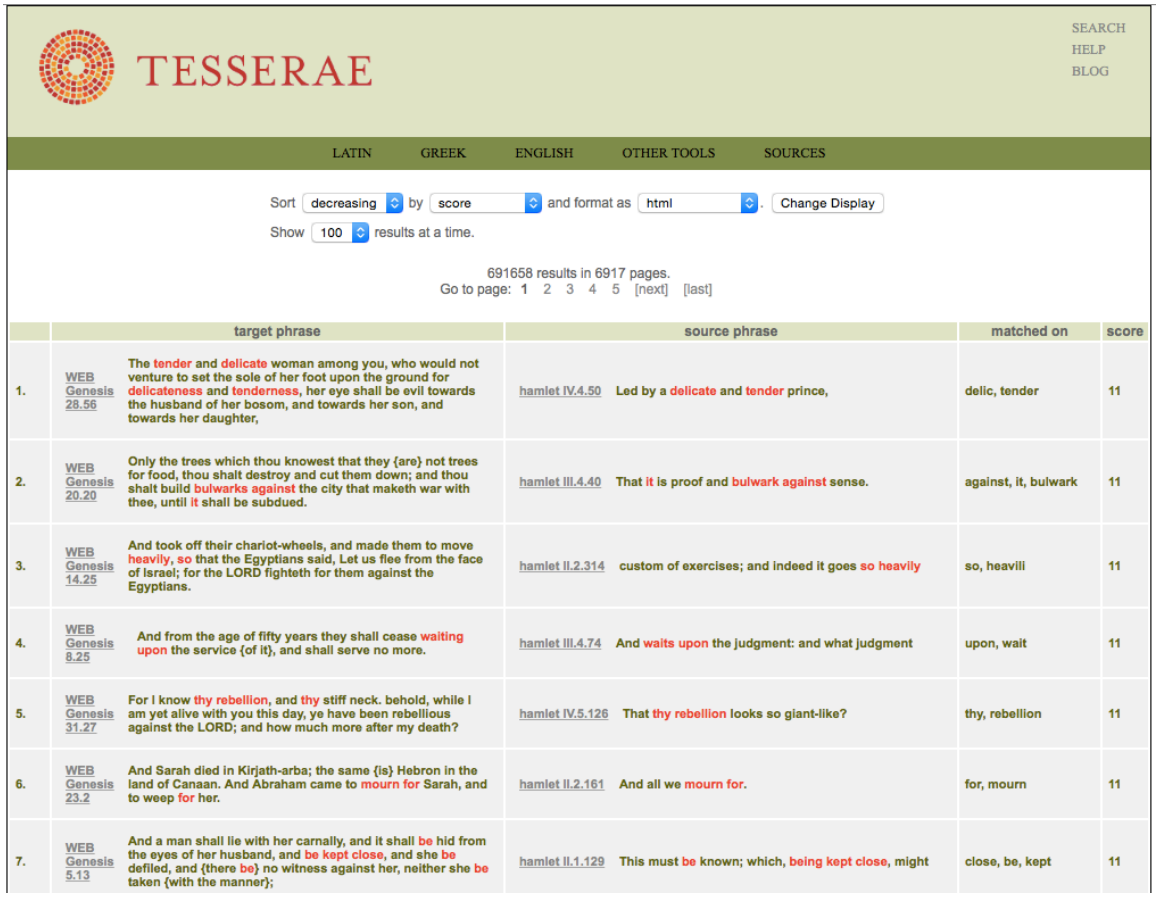
Words	Text 1 counts	Text 1 relative	Text 2 relative	Text 2 counts	Relative ratio (text1/text2)	Word distribution in text 1	Word distribution in text 1
king	245	0.0102	0.0025	6	4.0739		
thou	177	0.0074	0.0034	8	2.2074		
thy	175	0.0073	0.0004	1	17.4598		
lord	133	0.0056	0.0004	1	13.2694		
henry	125	0.0052	0.0008	2	6.2356		
shall	102	0.0043	0.0008	2	5.0883		
thee	88	0.0037	0.0008	2	4.3899		
hath	68	0.0028	0.0004	1	6.7844		
come	55	0.0023	0.0025	6	0.9146		
god	53	0.0022	0.0029	7	0.7554		
good	53	0.0022	0.0013	3	1.7626		
make	53	0.0022	0.0004	1	5.2878		
let	52	0.0022	0.0038	9	0.5764		
say	49	0.0020	0.0008	2	2.4444		
death	42	0.0018	0.0004	1	4.1903		
blood	41	0.0017	0.0004	1	4.0906		
john	38	0.0016	0.0008	2	1.8956		
hand	34	0.0014	0.0008	2	1.6961		
man	30	0.0013	0.0021	5	0.5986		
day	30	0.0013	0.0017	4	0.7483		
land	29	0.0012	0.0004	1	2.8933		
earth	29	0.0012	0.0004	1	2.8933		
away	26	0.0011	0.0013	3	0.8647		

8.5 Images of Tesserae

8.5.1 Home Page



8.5.2 Pentateuch Search



TESSERAE

SEARCH
HELP
BLOG

LATIN GREEK ENGLISH OTHER TOOLS SOURCES

Sort decreasing by score and format as html . Change Display
Show 100 results at a time.

691658 results in 6917 pages.
Go to page: 1 2 3 4 5 [next] [last]

	target phrase	source phrase	matched on	score
1.	WEB Genesis 28.56 The tender and delicate woman among you, who would not venture to set the sole of her foot upon the ground for delicateness and tenderness, her eye shall be evil towards the husband of her bosom, and towards her son, and towards her daughter,	hamlet IV.4.50 Led by a delicate and tender prince,	delic, tender	11
2.	WEB Genesis 20.20 Only the trees which thou knowest that they {are} not trees for food, thou shalt destroy and cut them down; and thou shalt build bulwarks against the city that maketh war with thee, until it shall be subdued.	hamlet III.4.40 That it is proof and bulwark against sense.	against, it, bulwark	11
3.	WEB Genesis 14.25 And took off their chariot-wheels, and made them to move heavily, so that the Egyptians said, Let us flee from the face of Israel; for the LORD fighteth for them against the Egyptians.	hamlet II.2.314 custom of exercises; and indeed it goes so heavily	so, heavili	11
4.	WEB Genesis 8.25 And from the age of fifty years they shall cease waiting upon the service {of it}, and shall serve no more.	hamlet III.4.74 And waits upon the judgment: and what judgment	upon, wait	11
5.	WEB Genesis 31.27 For I know thy rebellion, and thy stiff neck. behold, while I am yet alive with you this day, ye have been rebellious against the LORD; and how much more after my death?	hamlet IV.5.126 That thy rebellion looks so giant-like?	thy, rebellion	11
6.	WEB Genesis 23.2 And Sarah died in Kirjath-arba; the same {is} Hebron in the land of Canaan. And Abraham came to mourn for Sarah, and to weep for her.	hamlet II.2.161 And all we mourn for.	for, mourn	11
7.	WEB Genesis 5.13 And a man shall lie with her carnally, and it shall be hid from the eyes of her husband, and be kept close, and she be defiled, and {there be} no witness against her, neither she be taken {with the manner};	hamlet II.1.129 This must be known; which, being kept close, might	close, be, kept	11