Introducing English Studies
Chapter 10: Screencast Transcript

Slide 1 [Title slide]

Hello. I am Tamara O'Callaghan, co-author of the textbook *Introducing English Studies*. In this screencast, I will discuss "how to create and analyze a word cloud" using a literary work. Although you are likely familiar with the visualization of numbers in graphs, tables, and charts from math and science classes, you may not understand how the visualization of literature – specifically the word – can encourage you to "see" a literary work (or multiple works) in new ways. This screencast aims to help you with the process of applying data visualization tools to literature.

Slide 2 [Text]

Topics in this presentation include:

1. Visualizing Data: Anscombe's Quartet
2. Why Visualize a Literary Text?
3. Data Visualization with Text
4. *Pride and Prejudice* Word Cloud
5. Making Data Visualization Effective

Slide 3 [Image of a Table with Eight Paired Columns of Numbers]

You know the "A-Ha" moment we all experience…that moment when you suddenly understand a complex problem or idea? Often, you will indicate your sudden comprehension with the simple two-word phrase "I see." As humans, we typically use the concept of "seeing" as a metaphor for knowing. Consequently, it is not surprising that innovative visual representations can help us to "see" data in new ways, allowing us to better understand and interpret data and, in doing so, find new meaning. Visual elements can significantly aid in the comprehension of text-based content, whether that content is numbers or words.

The statistician, Francis Anscombe, created this table, known as Anscombe's Quartet, in 1973 to demonstrate that not all visualizations of data provide the same information. Don't worry! You do not have to be a statistician to understand the Quartet's significance. The four data sets presented in the table on this slide present numbers that do not differ significantly in value. As such, they do not appear to vary much in terms of statistical significance.

Slide 4 [Image of Four Scatter Plot Graphs]

However, if you graph the data in the table using scatter plots, the results are quite different and demonstrate that a table of numbers is not necessarily exact. As you can see here, the scatter plots show four very divergent patterns, suggesting significant differences not readily visible in the more traditional table visualization. Anscombe proved that re-presenting – or re-visualizing – the data allows us to see that data and interpret in new ways.

Slide 5 [Text]

Many academic disciplines, especially those in the social sciences and natural sciences, use such quantitative information as part of their research process. However, the humanities have typically resisted "crunching numbers" until recently. Developments in computer technology and the ready availability of texts in electronic form have encouraged the application of this approach, especially in English Studies. Visualizing textual and linguistic patterns in literature can yield some surprising results and allow scholars to work efficiently and productively with large and seemingly unwieldy texts and groups of texts as well as reshape our understanding and recognition of canonical texts, authors, and artifacts

Such a research process may seem daunting, but it is really not that difficult to do. Let's give it a try!

Slide 6 [Text]

Firstly, you need a literary work in electronic format. One of the best resources for such texts is Project Gutenberg. Project Gutenberg is the oldest digital library providing access to over 60,000 free ebooks. All you need to do is download or copy the text or texts you wish to analyze. You will need to strip the Project Gutenberg edition of any text, such as the front matter and public domain information, that is not part of the literary work in order to have an accurate edition.

Secondly, you need a data visualization tool that allows you to work with text rather than numbers. An excellent option is Voyant Tools, an open-source web-based application for performing text analysis. You simply paste text into the onscreen field, open an existing literary corpus, or upload files from your computer. Keep in mind that, although we are using a literary work in this screencast, you could choose to explore a non-literary text, such as a blog or website, with Voyant Tools.

Finally, once you have your text in Voyant Tools, click the "Reveal" button and select the data visualization(s), such as a Word Cloud, that you want to use.

Slide 7 [Text and Image of Word Cloud]

A Word Cloud is an electronic image of words from a text in which the size of each word indicates its frequency and importance. The Word Cloud in Voyant Tools visualizes to top frequency of words for the entire corpus. The most frequent words are centered and large, but it is important to remember that the color and position of those words are not significant

Here, on the right, is the Word Cloud generated using the Project Gutenberg edition of Jane Austen's *Pride and Prejudice* in Voyant Tools. You may not find it all that surprising that the most frequent words in the novel are "elizabeth," "mr," and "darcy." However, the Word Cloud does emphasize visually the fact that female characters are known by their first names while male characters are always addressed by their titles and last names. Such a realization may encourage us to think more carefully about the gendered roles of Austen's characters.

Slide 8 [Text]

For a data visualization to be effective, it must be informative, efficient, and beautiful. What exactly does that mean? Well, first, for your visualization to be informative, you need to plan out its purpose in the abstract before you try to execute it. What questions are you trying to answer? What story are you trying to tell? Be sure to distinguish between visuals designed to illustrate what we already know and visuals designed to aid in the exploration of the unknown.

Second, for your visualization to be efficient, you need to emphasize what matters by visualizing the relevant content. Keep in mind that changing the emphasis will change the message of the visualization. Don't overuse the options of the visualization tool, but instead apply standard representations and conventions in order to encourage learning and retention by your viewer

Finally, for your visualization to be beautiful, you need to remember that the aesthetics should reinforce the message. It is essential that the design, no matter how visually appealing, does not interfere with that message, especially if the message is complex.

By using such data visualization tools as Voyant Tools, we can explore favorite literary works in a new way, thereby gaining a better appreciation of their complexity and relevance across the literary canon.