EEBO-TCP as a tool for integrating teaching and research

Heather Froehlich, Richard J Whitt, and Jonathan Hope

I. Introduction.

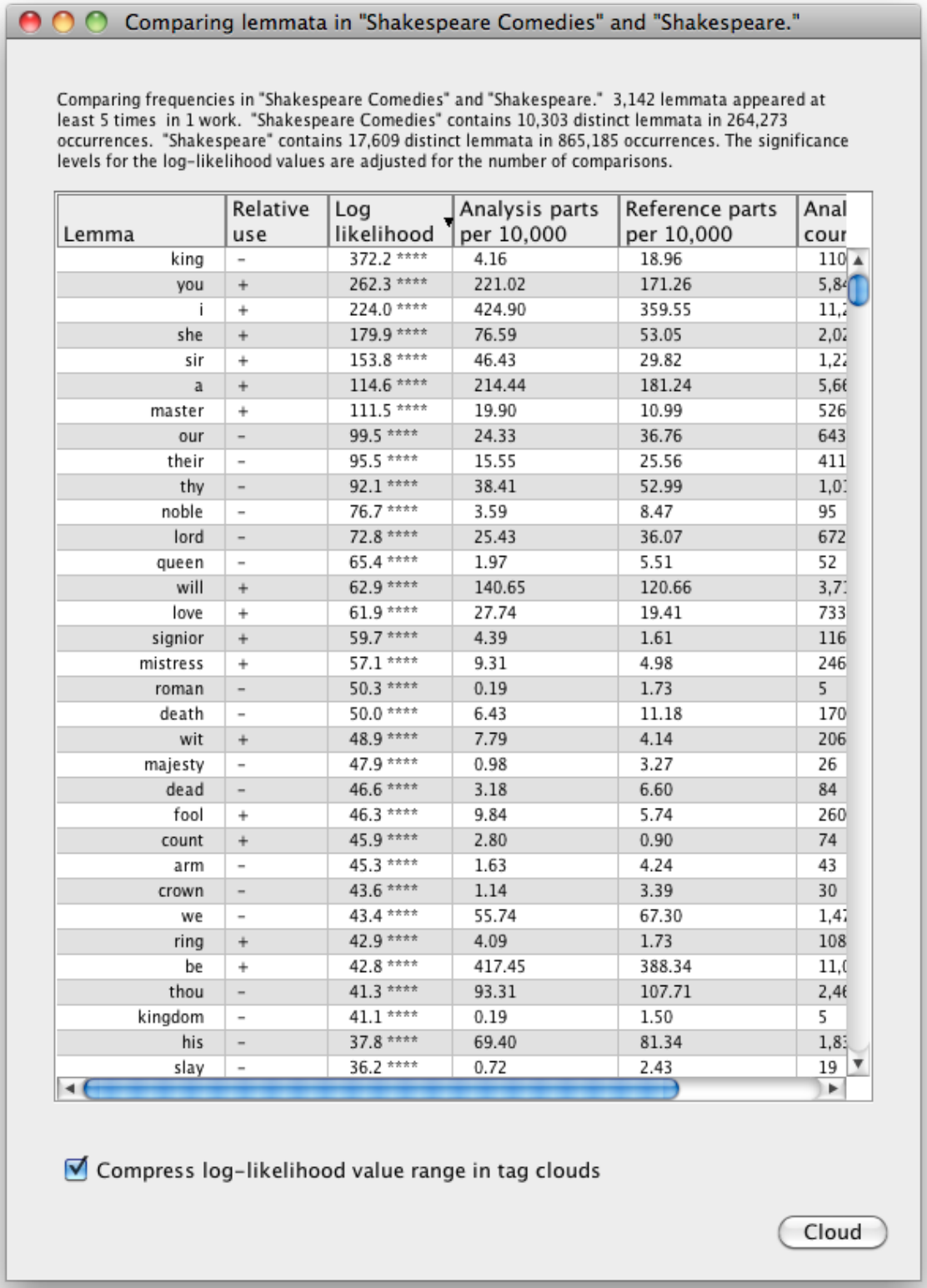
TextLab is a new course at the University of Strathclyde that introduces students to methods of digital text analysis. As part of the Vertically Integrated Projects initiative, students work in teams, testing newly developed software programmes to identify specific linguistic features of literary texts. The Vertically Integrated Projects initiative puts students from a variety of different backgrounds (from different faculties, from first year through to postgraduate study) and they work on a project together. This system, first developed at the Georgia Institute of Technology by Professor Ed Coyle, focuses on hands-on, interdisciplinary research-led teaching, wherein students contribute directly to ongoing academic research. Textlab serves as a Digital Humanities Vertically Integrated Project at Strathclyde, fostering interdisciplinary collaboration between the faculty of English Studies and the faculty of Computer and Information Science.

In collaboration with the Mellon-funded Visualizing English Print 1470-1800 project (VEP) between the University of Strathclyde, the University of Wisconsin-Madison, and the Folger Shakespeare Library, students enrolled in Textlab were introduced to the same tools that researchers at these institutions use. The Visualizing English Print project seeks to ask what makes some texts more or less similar to other texts, what do specific genres of written language have in common, and what are the functional properties of literary language from the beginnings of English print (at ca. 1470) to 1800. Using a combination of newly-developed software packages designed for literary-linguistic analysis, students and researchers are able to conduct computer-aided analyses of texts to better understand, classify and quantify large numbers of texts in ways which were previously much more difficult. Students are arranged into mixed groups of five. In the first iteration of Textlab, each group was assigned one Shakespeare play to focus on. Using and evaluating the tools developed for this process, students were able to make genuine findings about their assigned play. Through their research, they would log their progress individually (in a notebook) and as a group (on a wiki).

II. Software Programs.

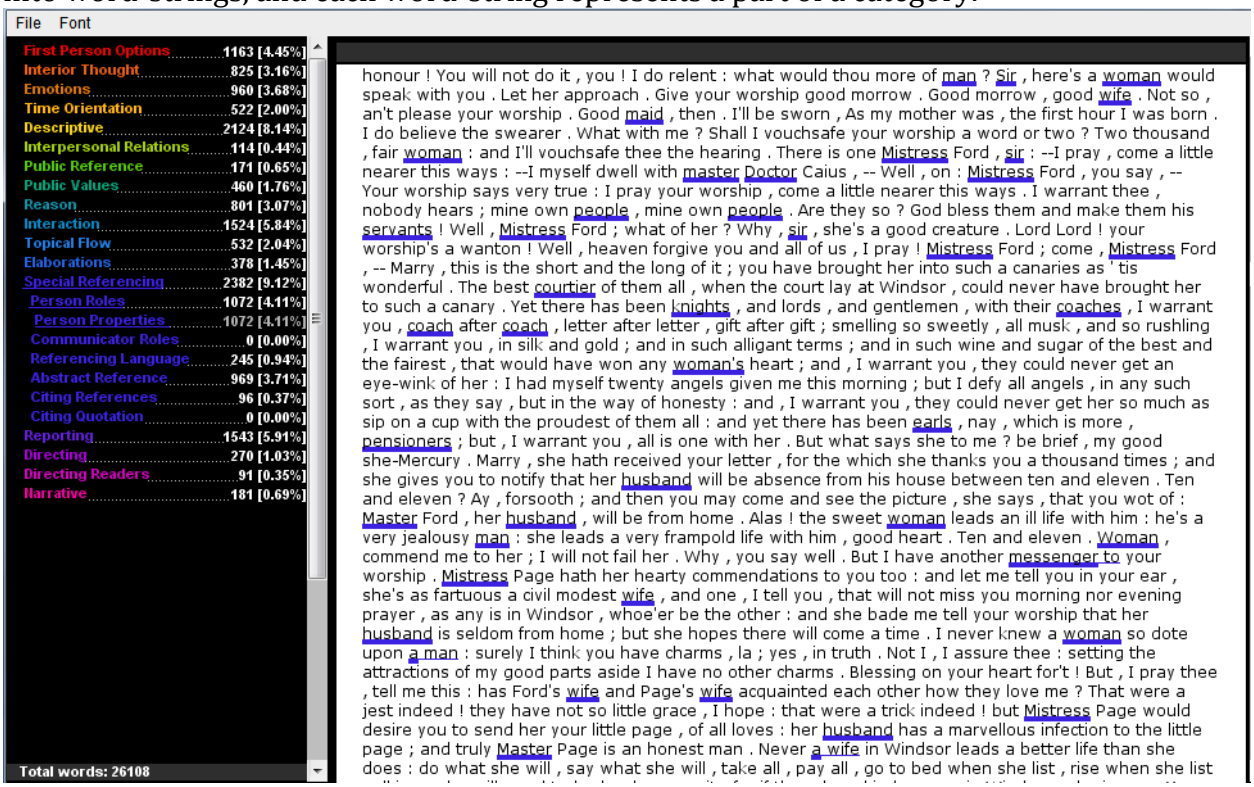
Students enrolled in Textlab used three software programs, some of which were developed as part of the Visualizing English Print project. WordHoard, Docuscope and LATtice all show different information about the same texts, allowing students to dig deeply into their data.

WordHoard is a corpus-analysis software programme developed at Northwestern University by Martin Mueller. We found WordHoard to be particularly useful for first-time users of large-scale corpora by presenting the question “what happens when we compare one (or more) text(s) to another group of one (or more) text(s)?” It produces a log-likelihood analysis of words and lemma, and allows the user to see these words in the context of Shakespeare’s corpus.



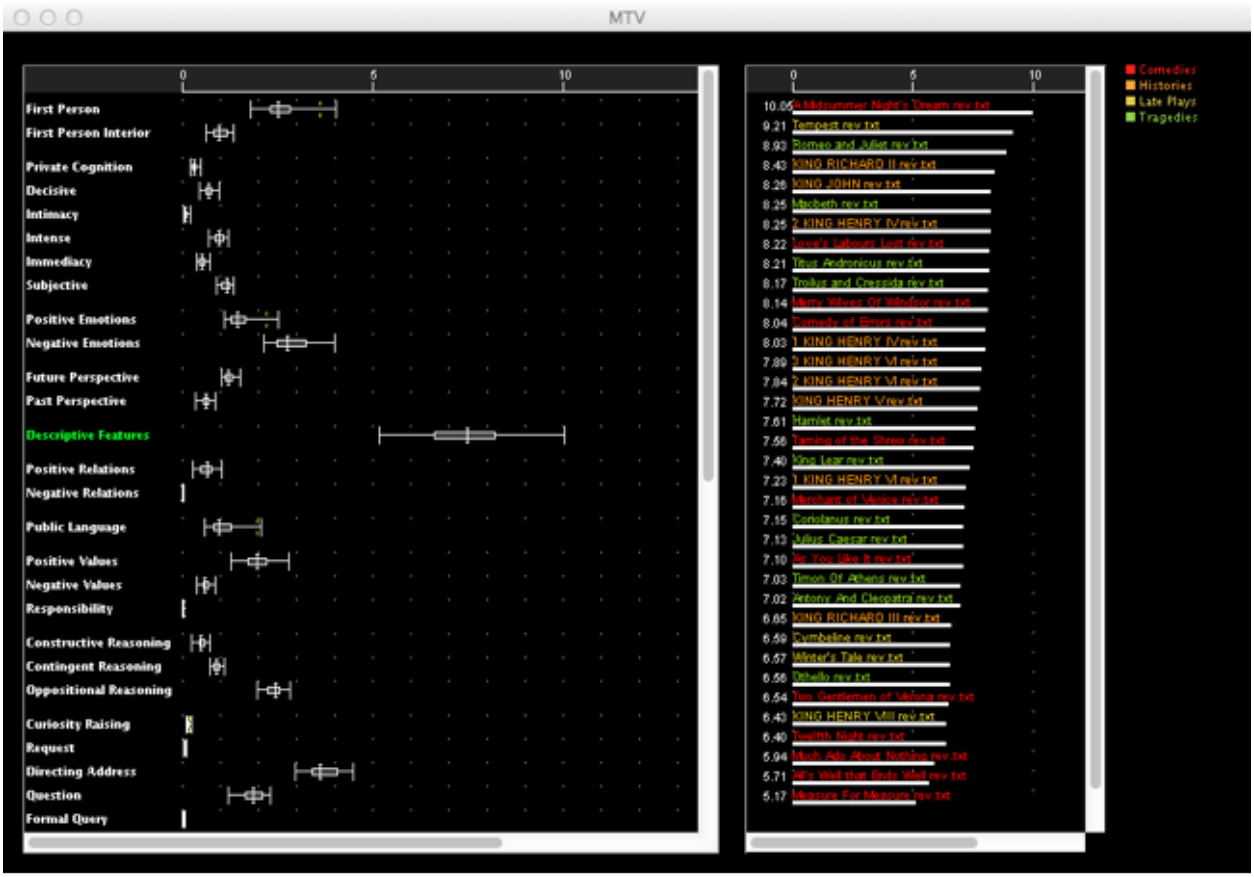
Docuscope is a a rhetorical analysis software developed at Carniege Mellon University by Suguru Ishizaki and David Kaufer (2004, 2011). When developing their program, Ishizaki and Kaufer encoded 175,000 of the most frequent words in English as well as the most frequent 2-4 word combinations into strings classified by rhetorical effect. Each category of rhetorical effect is broken down into specific functional features, called a “Language Action Type” or LAT. Each LAT is grouped

into a larger category ("Dimension"), which in turn are organised into larger, superordinate classes ("clusters") as a hierarchy. Each LAT is underlined and color-coded for analysis. They did this by breaking rhetorical patterns of English into word-strings, and each word-string represents a part of a category.[1]

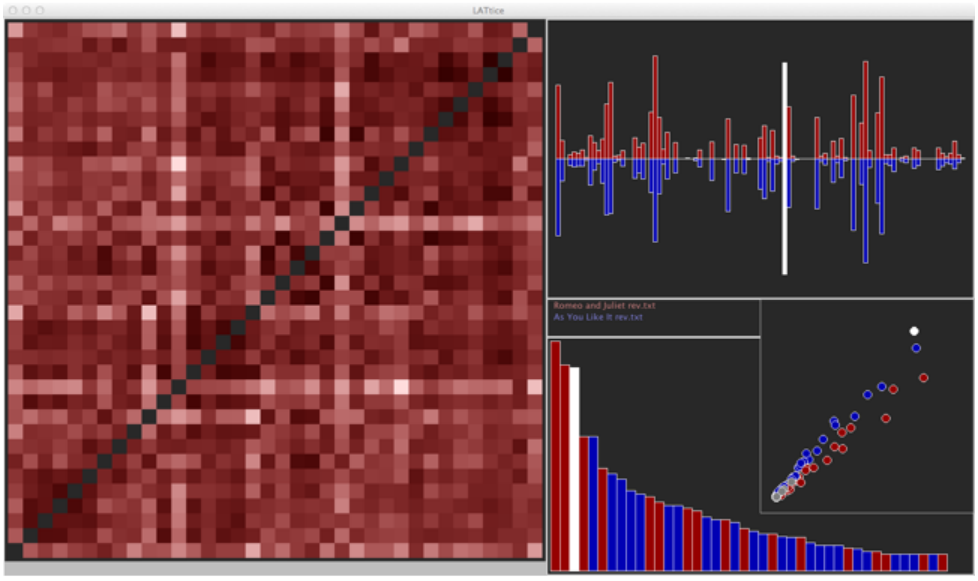


Docuscope can run a functional-linguistic analysis on many texts at once using its Multiple Text Viewer, which shows each dimension as box plots and a roster table.

[1] Ishizaki and Kaufer continued creating LATs until they felt confident that they could categorise most of English into their existing LAT structure. The creators tested their rhetorical classifications against "as broad a range of English prose text as we could find" covering newspapers, fiction, short stories, speeches, journals, student writings from a multi-genre course; essays on social criticism, reminiscence and narratives; information system documentation including meeting minutes, design plans, public announcements, and proposals; talk radio, song lyrics, fables, Inaugural speeches, and the New Yorker, among other sources. Docuscope also covers the 100,000 frequent words and 75,000 of the most common words in English, as well as the most frequent 2-4 word combinations (2011: 280; 2004: Chapter 1).



And its output can be exported to LATtice, which produces a heatmap visually showing similarity and difference using dark and light color shading based on LAT frequencies in the texts run through Docuscope.



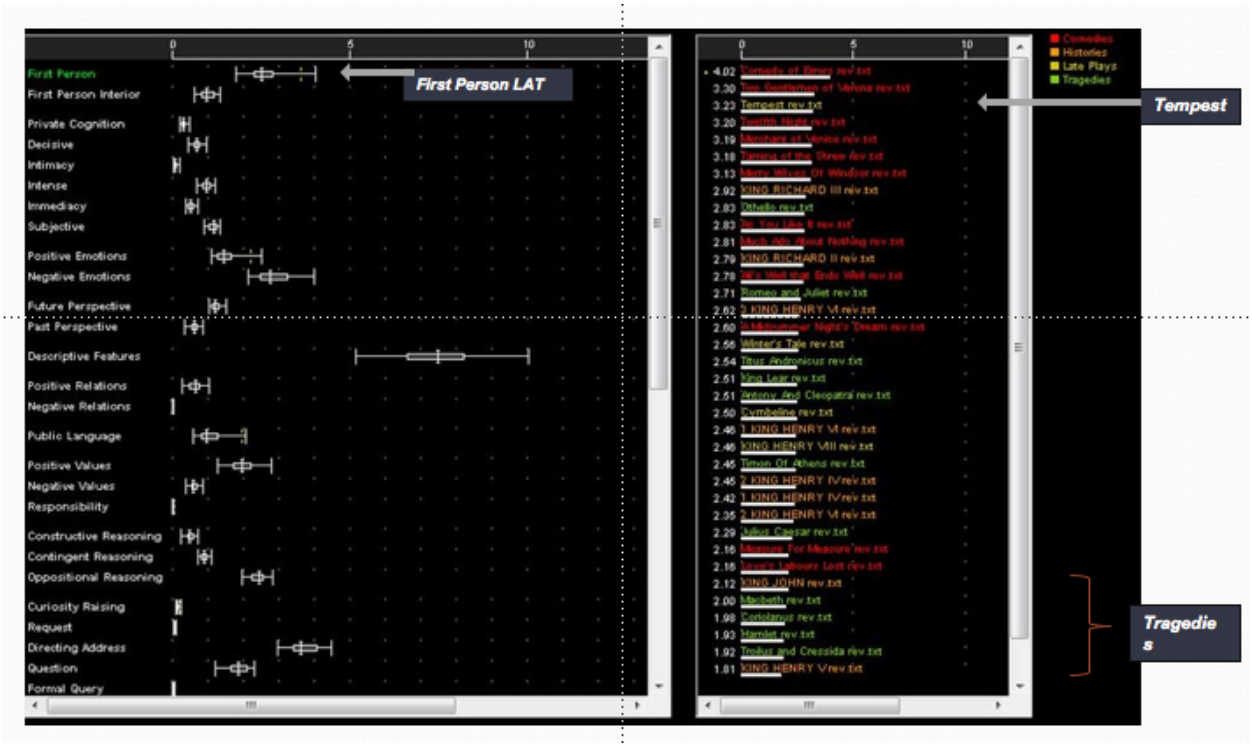
Anupam Basu's LATtice software takes the data values of each LAT from Docuscope and creates a heatmap of similarity or difference based on Euclidean distances between vectors. Using an N x N grid, LATtice illustrates difference or similarity through color coding: the darker a square is, the more similar the two texts are; the lighter, the less similar the two texts are. The black diagonal line

intersecting the graph represents a text compared against itself.[2] Relationships between any two texts in the corpus can be explored at the level of the LATs, as illustrated on the right-hand side of the software. This sort of mapping is ideal for digging into larger datasets, guiding the user to identify which texts should be looked at more closely.

III. Student Work from the First Iteration of Textlab

Students were made familiar with the software in a trickle-down approach: researchers presented the tools, and students enrolled in the course became comfortable with them by testing their assigned texts with them. This process encourages students to be active learners: they have a problem that they need to solve using the tools provided. What makes their assigned play different – or similar – in some way in comparison to other plays by the same author?

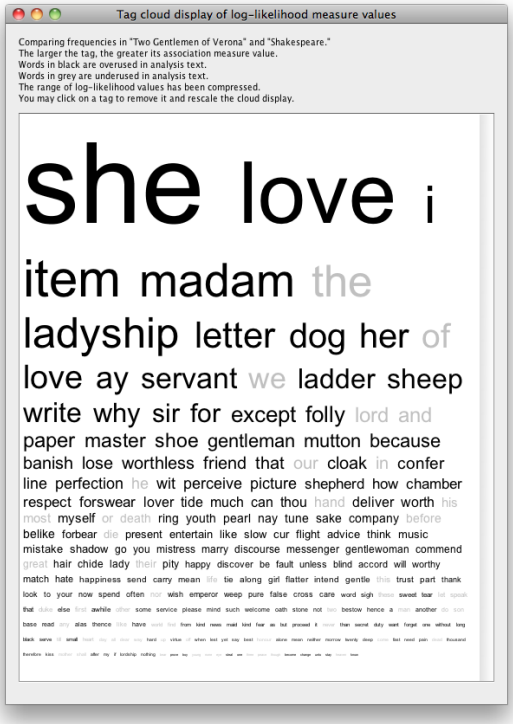
Jamie-Leigh Green, a fourth-year undergraduate student, discovers using WordHoard that "love" is less likely to appear in *The Tempest* compared to all other Shakespeare plays despite it being categorised as a "romance". This early discovery guided her and her group's work throughout the rest of the semester. Ultimately, they discovered that Shakespeare's late plays are only loosely connected by this generic division – which definitely does not correlate to our modern conception of "romance".



The group working on *The Two Gentlemen of Verona* discovered that *she* and *love* are especially important words in their play. A play about two men talking about a woman is not particularly surprising, but as Jessica Wagstaff points out, "language in *The Two Gentlemen of Verona* focuses heavily on the discourse of women as object[s] of attainment", adding that the feminine pronoun *she* is far more likely to appear in *The Two Gentlemen of Verona*, but only as the object of the

[2] http://winedarksea.org/?p=1285

male characters: women in this play are discussed but are not playing an active role. Close readings with Docuscope's Single Text Viewer and noting the role of relevant LATs in LATtice show that her claim is true.



Students enrolled in this first iteration of Textlab have come to terms with three separate textual analysis programs and have produced results in a ten-week course. And, not only have they produced results, but they can explain them in a variety of mediums: through oral presentation, through wiki writing, and through report writing. In addition to these 'hard skills', 'soft skills' were emphasised as well: teamwork and group management skills were a focus of the course as well. Naturally, the next step is to enter the larger world of early Anglophone print: the period itself is host to types of linguistic variation that span the rise of several major generic forms, Atlantic exploration and colonization, and orthographic standardization. EEBO is poised to become the leading depository of searchable Early Modern records. One of the driving forces behind Visualizing English Print is that through the digital analysis of non-canonical texts, larger patterns of linguistic practice can be identified and explored.

IV. EEBO Texts and the 1K Corpus

In order to investigate non-canonical texts, the selection and preparation of a 1000 text corpus has been the focus of much of the VEP this year. A focused analysis of a subset of EEBO and ECCO (Eighteenth Century Collection Online) texts representing the beginning of print culture to 1800, containing a random selection of 40 texts per decade from 1530 onward was made using Python's "random" number generator, picking out texts with no repetition. The process of text selection revealed that it was not always possible to get 40 texts per decade before 1530. Hence at this stage, texts that pre-date 1530 are not considered our analysis. The one criterion was that the text had to be longer than 500 words.[3] We

[3]Our text selection is admittedly quite varied in text size. The focus here was to get an equal number of texts representing each decade from 1530 to 1800. But as Docuscope

call this sampling of texts the "1K Corpus" (although there are slightly more than 1000 texts here; about 1080 to be exact).

These texts were then sent through Docuscope, which as discussed earlier, is one of the programmes used in Textlab. Initial output revealed that later texts were receiving significantly more tags than earlier texts, which pushed the need for some degree of modernisation. A variant detector software package (VARD 2, developed by Alistair Baron at The University of Lancaster[4]) was employed to modernise the spelling of the 1K Corpus for tagging purposes. After texts in the 1K Corpus were "Varded", tagging frequencies were much more uniform across the board.

Since one of the main goals of the VEP project is to trace the development of genre through the history of English print, genre labels were then assigned to the texts in the 1K Corpus. The labels themselves are based on those used in the Early Modern section of the Helsinki Corpus and ARCHER (A Representative Corpus of Historical English Registers)[5], with certain additional labels added to account for broader text selection. Labels were assigned based, in the first instance, on information made available in the metadata (e.g. keywords such as play, sermon, medicine, etc. easily give away a text's genre). Where metadata were unavailable or proved insufficient, the text was inspected manually.

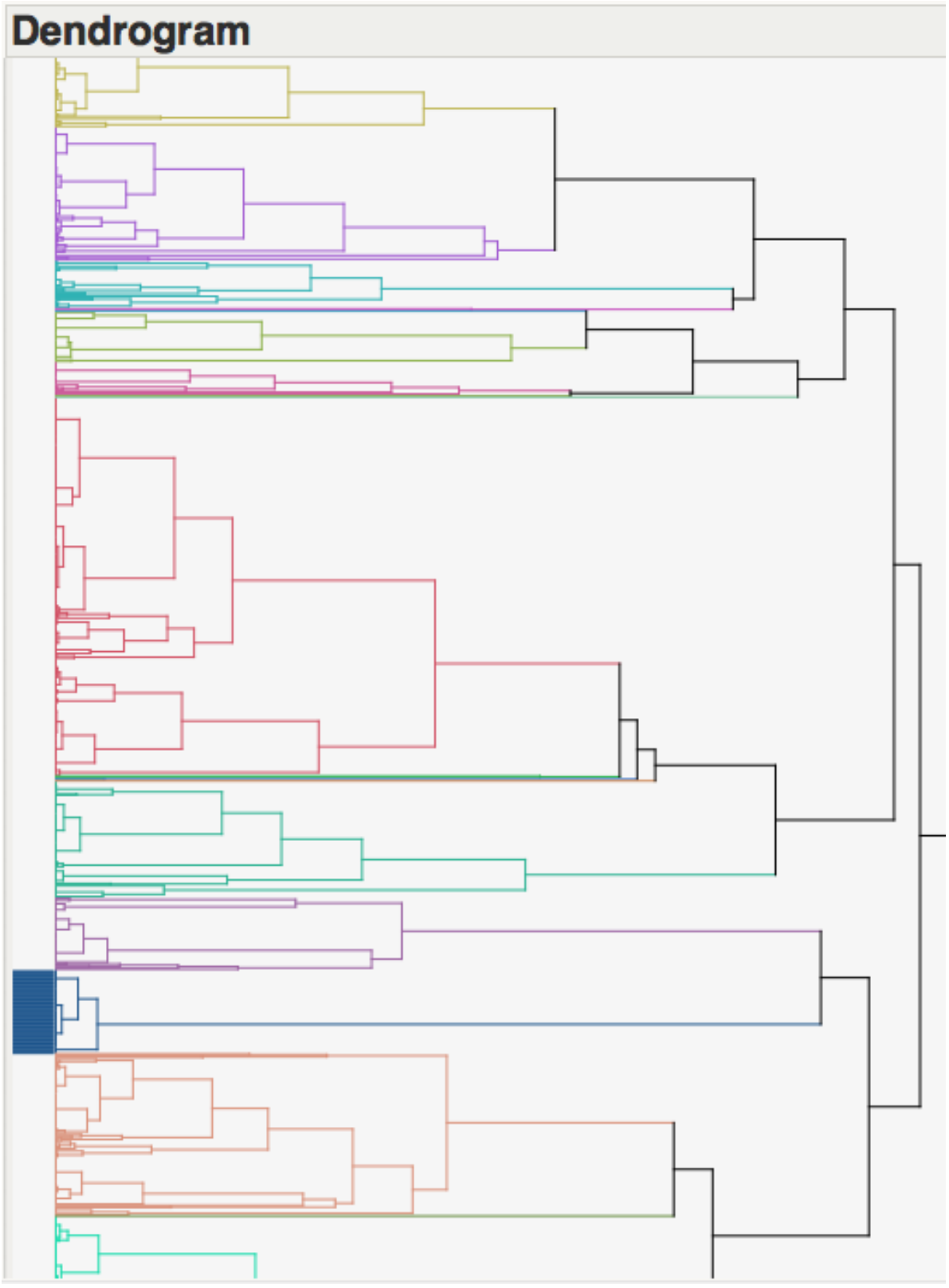
Using Docuscope output, a number of statistical tests—Ward's clustering method, PCA (Principal Component Analysis) and ANOVA (Analysis of Variance)—were run using the JMP statistical software package to see if any correlation could be found between the LATs assigned by Docuscope and the genre assignments. Indeed, all three tests revealed independently that Docuscope output could be used to see what features of language (vis-á-vis the LATs) distinguish one genre from another, and what features cause statistically significant separations among the genres: the dendrogram—the result of Ward's clustering—shows most dramas grouping together in the same cluster:

---

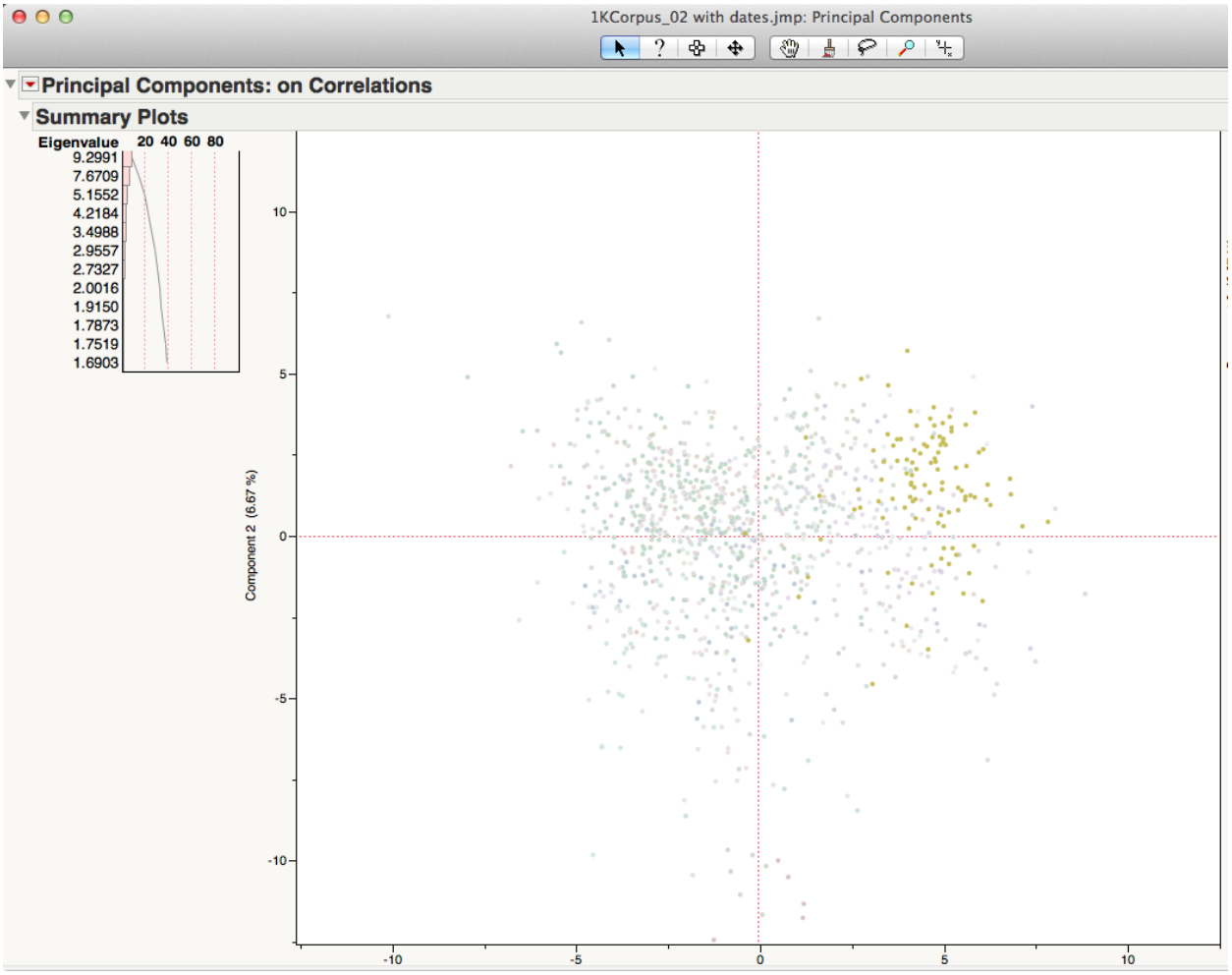
provides frequency results in percentages (relative to text size) rather than absolute numbers, we hope to have offset at least some disparities in text length.

[4]For more information about the VARD 2 programme, see http://www.comp.lancs.ac.uk/~barona/vard2/.

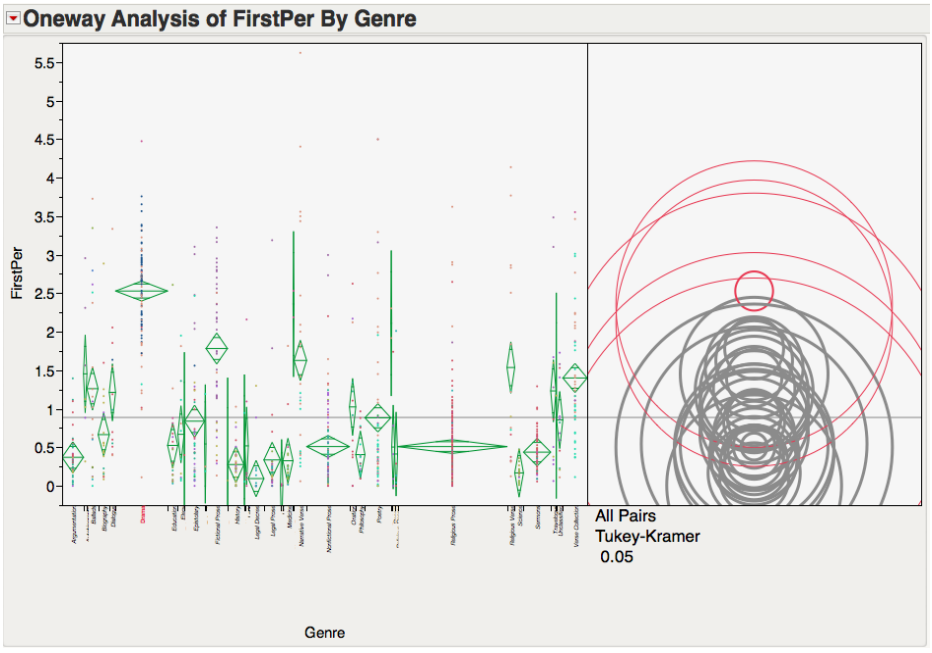
[5]For more information about these corpora, see http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/earlymodern2.html and http://www.llc.manchester.ac.uk/research/projects/archer/archer3_2/, respectively.



The dramas also appear, for the most part, to group together in the same quadrant of PCA space:



And ANOVA reveals that the FirstPer LAT is one of key factors causing dramas to evince statistically significant separation from most other genres (dramas are represented by the highlighted magenta circle in the visualisation of the Tukey test, with grey circles representating those genres with statistically significant lower frequencies in the use of the FirstPer LAT):



This combination of statistical and (digital) literary analysis promises many opportunities for interdisciplinary student collaboration. On the one hand, students of literature can become acquainted with the principles of statistics as they inform quantitative aspects of literary computing. On the other hand, students of statistics and computer science can see how the methods of their disciplines can be applied to the study of literary texts. As the point of the VIPs is to integrate students from a number of diverse disciplines, the intersection of digital literary analysis and statistical computation seems an ideal avenue to foster such cross-disciplinary interfaces, while stressing transferable skills amongst participants.

V. Digital Humanities at Strathclyde

As members of the VEP found, modernizing texts makes them much more consistently comparable to each other. In order to teach with the TCP texts, they must be modernised, but after that process is complete, a much wider variety of texts can be used by students. Whereas the first iteration of Textlab focused on the analysis of Shakespeare's plays, future iterations can focus on a broader range of texts currently begin used by the VEP project. Each group of students will be given a set of two texts of different genres and asked to use the software programmes to find the similarities and differences among both literary and non-literary genres. Possible questions include: how is the language of medicine different from the language of law? How are dramas and travelogues similar? What features differ academic prose from fictional prose? These results will obviously feed in to the larger VEP project.

Students enrolled in Textlab are primed to make genuine discoveries about texts in the early modern period in ways which were previously inaccessible. Students at Strathclyde are in a unique position to conduct this sort of work: Dr Anouk Lang teaches a complimentary course at Strathclyde in the first semester, Introduction to Digital Humanities.[6] Her course introduces students to digital tools for text analysis and digital mapping, emphasizing digital literacy and critical thinking about digital tools with a strong emphasis on transferable skills from the classroom to real-life projects. Textlab serves as a continuation of this line of research-led teaching with a hands-on workshop component, continuing to bridge the gap between computer and literary studies.

This first iteration of Textlab shows that interdisciplinary, research-led collaborative work allows for students to actively contribute to ongoing research while building an impressive skillset on their CVs. Looking to the future, having students create and maintain generic classifications in large text corpora allows students to continue practice both close and distant reading – that is, looking at individual texts closely and as part of a larger whole. This idea of distance reading allows us as scholars to ask more questions using new methods, and pedagogical approaches will be quick to follow. As students have been practicing close reading for years, the shift away from close reading and towards considering texts as part of an as-yet unknown whole will allow them to continue to make genuine contributions to the future academic landscape.

---

[6] For a full overview of Introduction to Digtial Humanities, please see Dr Lang's webpage for the course: http://aelang.net/wordpress/intro-to-dh/

Works Cited.

Ishizaki, Suguru and David Kaufer. "Computer-aided Rhetorical Analysis" in *Applied Natural Language Processing and content analysis: Identification, Investigation, and Resolution*. Ed. Philip McCarthy and Chutima Boonthum. 276-296. 2011.

Kaufer, D., Ishizaki, S., Butler, B., & Collins, J. *The Power of Words: Unveiling the Speaker and Writer's Hidden Craft*. Routledge. 2004.