Renaissance and Reformation
Renaissance et Réforme

Eder, Marciej, Jan Rybicki, and Mike Kestemont, project creators. Stylo(). Other

Thomas Merriam

Volume 44, numéro 4, automne 2021

URI : https://id.erudit.org/iderudit/1089358ar
DOI : https://doi.org/10.33137/rr.v44i4.38651

Aller au sommaire du numéro

Éditeur(s)

Iter Press

ISSN

0034-429X (imprimé)
2293-7374 (numérique)

Découvrir la revue

Citer ce compte rendu

Merriam, T. (2021). Compte rendu de [Eder, Marciej, Jan Rybicki, and Mike Kestemont, project creators. Stylo(). Other]. *Renaissance and Reformation / Renaissance et Réforme*, *44*(4), 223–228. https://doi.org/10.33137/rr.v44i4.38651

Works Cited

Cecconi, Elizabeth. 2020. "British Colonial Ideology in the Language of Pamphlets on Virginia (1584–1624)." In *The Language of Discovery, Exploration and Settlement*, edited by Nicholas Brownlees, 8–24. Newcastle Upon Tyne: Cambridge Scholars Publishing.

Ciambella, Fabio. 2019. "Specialised Tourist Discourse: Translating Schott's *Itinerarii* in Early Modern England." In "Translation and the Non-literary Text: From Early to Late Modern English." Special issue, *Status Quaestionis* 17.

Eder, Marciej, Jan Rybicki, and Mike Kestemont, project creators. Stylo(). Other.

Krakow: University of Kracow, Poland. Accessed 17 February 2022. github.com/computationalstylistics/stylo.

R Stylo is a suite of programs, the most widely used of which is stylo(). Stylo() enables stylistic exploration of lexical corpora. It serves as a testbed for wide-ranging experimentation with texts of varying size. It enables the use of a multiplicity of parameters that scrutinize the authorship of early modern texts. As evidence of its flexibility, five settings of three parameters allows for 3^5, or 243, possibilities, more than is practical with manual operation. And there are far more than five choices of three parameters in stylo().

One appreciates the advantages of stylo() as a tool, by comparing its simplicity with the multiple stages ordinarily required for statistical stylistic analysis. For an early modern play, one selects reliable electronic texts in original or standardized modern spelling, with their speech headings and stage directions commonly removed. This preliminary stage is the same with or without stylo().

The ensuing stages, however, are the ones in which R Stylo excels. Stylo() comes with a graphical user interface (GUI) which enables a researcher without programming skills to regulate its parameters. To interrogate, for instance, the authorship of a disputed early modern play, one selects a set of linguistic parameters. Stylo() automatically selects the most frequent variables determined by those pre-chosen parameters. These include the number of most frequent words (MFW): single, bigram, occasionally trigram, or the number of 2-character n-grams (MFC), 3-character n-grams (MFC), 4-character n-grams, and so forth. Dramatic texts may be input as entire plays, or parts of plays. Texts may be input in equal-sized sections of any chosen length.

The choice of parameters does not allow foreknowledge of the working variables selected. Hence, there is no cherry-picking of variables to favour a desired outcome. The variables used and their relative frequencies can be later referred to in the saved file "Table with Frequencies." Pronouns may be retained or removed from the analysis. With 100 percent "culling," variables are tallied if present in all the texts under investigation. Culling can be adjusted to any percentage down to 0 percent, at which point the variables need not be present in all the texts. The choice of parameters is sufficiently swift to allow repeated, altering experimentation. This is one of the finest features of stylo().

Stylo() calculates the relative frequencies of most frequent words or characters chosen from the texts, and normalizes them to ensure that their comparisons are fairly weighted in a correlation matrix. Without stylo(), normalization would involve a language-dependent tool such as Word to extract raw frequencies from the texts, and then would require inputting the results to a spreadsheet such as Excel to calculate standard deviations as *Z* values. Instead of porting results between separate software, stylo() automates the steps in creating a correlation matrix, and applies matrix algebra to compute the principal components from which visualized results are produced. Besides graphic results based on correlation or co-variance matrices, the results may also be shown by multi-dimensional scaling or cluster analysis charts. Differing results may be evaluated by examining their graphic presentations for comparability. Visualization is well suited to the provisional nature of exploratory analysis.

The primary aim of statistical exploration is the rejection of the so-called null hypothesis. Stylo() offers unlimited possibilities for rejecting such a hypothesis. Let us say, for the purposes of illustration, that a manuscript copy of *Antony and Cleopatra* in the confirmed hand of John Fletcher was retrieved from the dust of some National Trust property. As only printed editions of the play previously existed, a null hypothesis is then proposed that Fletcher, not Shakespeare, was the author of the play. Any hypothesis concerning the authorship of a disputed text can be subjected to a set of trials that monitor evidence for it while undergoing progressive controlled degradation. Evidence for any hypothesis can be gradually degraded by stylo() with its many controls. A hypothesis that degrades readily and consistently is provisionally false, while one that conducts an orderly retreat in the face of hostile bombardment may be recommended for further examination and, in the final analysis, literary judgment.

A Fletcher control is constructed of *Bonduca, Chances, Demetrius, Elder Brother, Island Princess, Loyal Subject, M. Thomas, Rule a Wife, The Captain, Mad Lover, Pilgrim, Valentinian, Wild Goose Chase*, and *Women Pleased*. A Shakespeare control is constructed of *1H4, 2H4, Ado, Cor, Cym, Ham, Lear, LLL, MND, MV, Oth, R2, Rom, Shr, TGV, Tmp*, and *WT*. The imagined manuscript of *Antony and Cleopatra* is then tested for its affinity to the Fletcher or to the Shakespeare controls. Figure 1 initiates testing with the Most Frequent Character single n-grams, a single alphabetical letter or sentence marker. Shakespeare plays are coded with +, while Fletcher plays are coded with Δ. *Anthony* is coded with a small circle found in the left-hand Shakespearean side of the chart in Figure 1.

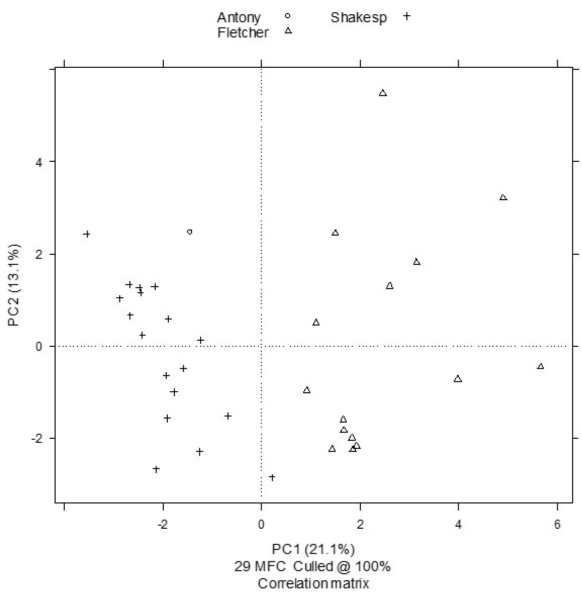

Figure 1. The diagram shows that *Anthony and Cleopatra* is correctly attributed to Shakespeare rather than to Fletcher by stylo()'s use of alphabet letters and spaces between words.

Further testing at 100 percent culling with two to ten Most Common Characters favours Shakespeare unequivocally as the author of *Anthony and Cleopatra*. However, at 100 percent culling with eleven Most Common Characters, discriminatory degradation begins. Of the eighteen Shakespeare plays, only eleven are placed together on the right, while of the fifteen Fletcher plays, only nine are placed together on the left. *Antony and Cleopatra* is placed with the Shakespeare majority on the right, but the evidence is no longer unequivocal. Note that the entire experiment is now based on only twenty-four character variables.

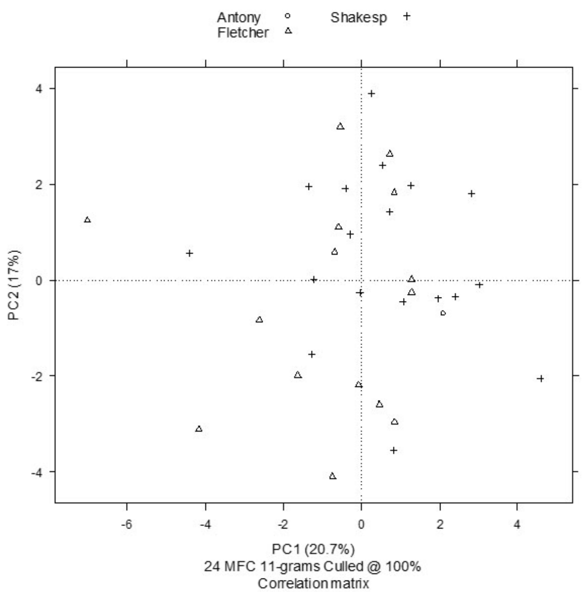

Figure 2. The discriminatory power is reduced from that in Figure 1. This is due to increasing the size of character n-grams from one to eleven. Fletcher and Shakespeare points are intermingled with each other.

However, if the parameter controlling the culling is changed from 100 to 80 percent, the discrimination between Shakespeare and Fletcher is restored and *Antony and Cleopatra* is firmly placed within the Shakespeare +'s on the left. The number of variables incorporated in Figure 3 is increased from twenty-four to 241.

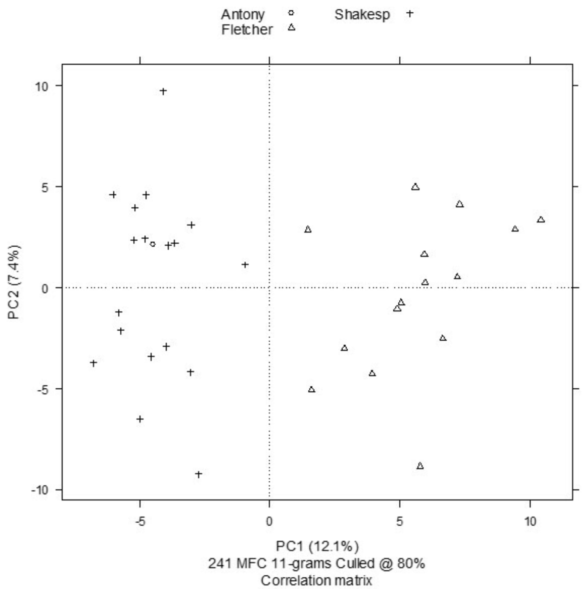

Figure 3. In this case, discriminatory power is restored by adjusting the culling parameter from 100 to 80 percent. This admits a far greater number of character variables: fourteen to 241. *Anthony and Cleopatra* are correctly assigned to Shakespeare.

By reducing the percentage of culling, one can extend the discriminating power of greater n-grams, with *Antony and Cleopatra* clearly within the Shakespeare bloc. For every correlation matrix chart, corresponding multidimensional and cluster analysis charts can be produced with stylo(). The number available is large, and for each chart the authorship of *Antony and Cleopatra* can be attributed.

The program in short permits "fast exploratory analysis of textual corpora."[1] The parameters encourage repeated experimentation to enable "testing to destruction" (my own expression, and not that of Eder) of any

1. Eder et al., "Stylometry with R," 108.

proposed authorship hypothesis. I therefore find stylo() one of the most useful and useable programs.

THOMAS MERRIAM
Basingstoke, England
https://doi.org/10.33137/rr.v44i4.38651

Works Cited

Eder, Maciej, Joanna Byszuk, and Jan Rybicki. "Stylometry with Stylo." Online tutorial. Accessed 17 February 2022. computationalstylistics.github.io/stylo_nutshell/#.

Eder, Maciej, Jan Rybicki, and Mike Kestemont. 2016. "Stylometry with R: A Package for Computational Text Analysis." *R Journal* 8 (1):107–21. journal.r-project.org/archive/2016/RJ-2016-007/index.html.

Craig, Hugh, director.
Craig's Zeta test. Other.
Newcastle, UK: Centre for Literary and Linguistic Computing (CLLC), University of Newcastle, 2009. Accessed 15 July 2021.
c21ch.newcastle.edu.au/ia.

The Zeta test has become a popular method, within computational stylistics, to determine the authorship of early modern texts. The Zeta test originated with John Burrows, but it is Hugh Craig's variant that has been most influential in early modern attribution studies.[1] This variant underpins several of the attributions in the New Oxford Shakespeare project.[2]

Zeta is a computational algorithm that is intended to detect a writer's style through determining which words they use more frequently compared with another writer or writers (or indeed themselves in a different period or working in a different genre). For example, an early successful application demonstrated clear shifts in vocabulary in Henry James's early and late styles.[3] The "style"

1. Burrows; Craig and Kinney.

2. Taylor and Egan.

3. Hoover.