Renaissance and Reformation
Renaissance et Réforme

Kaufer, David, and Suguru Ishizaki, creators. DocuScope

Marcus Dahl

See table of contents

Explore this journal

Cite this review

Hoover, David L. 2007. "Corpus Stylistics, Stylometry, and the Styles of Henry James." *Style* 41 (2):174–203. jstor.org/stable/10.5325/style.41.2.174.

Rizvi, Pervez. 2019. "The Interpretation of Zeta Test Results." *Digital Scholarship in the Humanities* 34 (2):401–18. doi.org/10.1093/llc/fqy038.

Rudman, Joseph. 2016. "Non-Traditional Authorship Attribution Studies of William Shakespeare's Canon: Some Caveats." *Journal of Early Modern Studies* 5:307–28. doi.org/10.13128/JEMS-2279-7149-18094.

Rudman, Joseph. 1997. "The State of Authorship Attribution Studies: Some Problems and Solutions." *Language Resources and Evaluation* 31 (4): 351–65.

Taylor, Gary, and Gabriel Egan, eds. 2017. *The New Oxford Shakespeare Authorship Companion.* Oxford: Oxford University Press. doi.org/10.1093/actrade/9780199591169.001.0001.

Kaufer, David, and Suguru Ishizaki, creators.
DocuScope.
Pittsburgh, PA: Carnegie Mellon University. Accessed 25 November 2021. cmu.edu/dietrich/english/research/docuscope.html.

DocuScope is described by its creators as "a computer-aided text analysis tool that allows researchers to conduct both quantitative and qualitative analyses of how the designed reader experience is created by writers through the selection of micro linguistic composing patterns, i.e., words and phrases."[1] The program is able to generate a huge amount of potentially useful data—breaking down texts individually into tagged components, which can be compared quantitatively. One of the best aspects of the way in which the data is outputted is that a corpus can be viewed both at the level of a corpus and at the level of an individual text from that corpus. This allows the data to be interrogated at a much more detailed level than is perhaps usual: both in how the analysis for each word of a text can be viewed, and in comparison with output from a wider corpus, or group of texts.

From the point of view of early modern textual linguistics, however, the program uses a modern dictionary. My own use of DocuScope was limited by

---

1. *DocuScope*, 4.

the fact that many early modern spelling words are not recognized: e.g., the word "prodigall" (spelled with two *l*'s) provided the output "UNRECOGNISED." Much of the markup is recognizable to standard users: e.g., the word "out" provides the output "zSinglePreposition"; likewise, "with," "out," "upon," etc. However, other output is more obscure. For example, the phrase "the Earth that bred us for every trifle" provides the following output which is less obviously understandable for words such as "bred" and "us" (zOrphanedYetAnalyzed) and "For" and "Every" (ReasGeneralizeClassify):

the	The	s	DescriptSpaceRelation-General
Earth	Earth	s	DescriptSpaceRelation-General
that	That	s	zDeterminerDemon-strative
bred	Bred	s	zOrphanedYetAnalyzed
us	Us	n	zOrphanedYetAnalyzed
For	For	s	ReasGeneralizeClassify
every	Every	s	ReasGeneralizeClassify
trifle	trifle	c	NegValInsignificant

The details of DocuScope's markup originate from prior research in which analysis is made of a detailed patterning of words leading to semantic dynamics and theoretically stable rhetorical classifications.[2] The markup therefore is not strictly grammatical but involves an analysis of rhetorical features. Again, with the problems of early modern spelling, it was not possible to use our own corpus to assess the utility of these for our purposes; in principle, however, the analysis of rhetorical forms—which can be assessed for their quantitative and qualitative variance across different canons—seems intuitively useful for the purposes of authorship studies. One of the issues of early modern authorship studies is the availability of reliable modernized texts, and one of the issues for using modernized texts with a program such as DocuScope is presumably that small, consistent editorial changes to the text could result in errors in the

2. See Kaufer and Butler.

outcome. However, since the Excel output facilitates not only a generalized output of multiple documents but also (as above) a word-by-word account of the markup, painstaking analysis of key passages is available, meaning that large-scale statistical analysis could be checked at the level of the word/line once the text/texts had been agreed upon. Moreover, Suguru Ishizaki asserts that it is possible to edit the DocuScope dictionary app: "For example, if you search for a specific word or a phrase, you can find all the categories that include patterns with that word/phrase. This tool may allow you to locate a specific word in the dictionary and add variants of that word." However, "anyone who is interested in editing the dictionary needs to license it from [Carnegie Mellon] university. It is free for academic use, but it'll need to go through [the] university's licensing office."[3] This again potentially makes DocuScope a very useful tool for analytical work at both a corpus and close-up scale for all users.

I have found the DocuScope academic team to be keen to provide help in the tagging terminology, applications, and general use of the software. For modern spelling texts (and with some editing, early modern texts) it certainly seems that the idea and aim of the program are potentially of great use to authorship studies, particularly in combination with other data generating software. The possibility to analyze repeated rhetorical structures throughout wide ranges of texts across canons opens out huge opportunities in terms of thinking objectively about the construction of not just authorial canons but chronologies, "zeitgeists," and of course genres. I look forward to working with the program more over the coming years.

MARCUS DAHL
University of London

Works Cited

DocuScope. 2021. Version 1.9.7 User Guide. Pittsburgh: English Department, Carnegie Mellon University.

Kaufer, David S., and Brian S. Butler. 1996. *Rhetoric and the Arts of Design*. New York: Routledge.

3. Personal correspondence between Suguru Ishizaki and the author, 2–23 June 2021.