Freebury-Jones, Darren. “Collocations and N-grams. Database.”

Citation details

Early Modern Digital Review, vol. 4, no. 4, 2021, https://doi.org/10.33137/rr.v44i4.38649.

Renaissance and Reformation / Renaissance et Réforme, vol. 44, no. 4, 2021, pp. 210-216, https://doi.org/10.33137/rr.v44i4.38649.

Peer review

This is a peer-reviewed article in *Early Modern Digital Review*, distributed in print by *Renaissance and Reformation / Renaissance et Réforme*.

Early Modern Digital Review

Early Modern Digital Review is an online, open-access, and refereed journal publishing high-quality reviews of digital projects related to early modern society and culture. It is committed to productive evaluation of both established digital resources and recent tools and projects. Its publications are distributed online by the journal and its partners, and in print by *Renaissance and Reformation / Renaissance et Réforme.*

In *The New Oxford Shakespeare: Authorship Companion*, edited by Gary Taylor and Gabriel Egan, 92–106. Oxford: Oxford University Press.

Taylor, Gary. 2014. "Empirical Middleton: Macbeth, Adaptation, and Micro-authorship." *Shakespeare Quarterly* 65 (3): 239–72. doi.org/10.1353/shq.2014.0030.

Taylor, Gary. 2019. "Finding 'Anonymous' in the Digital Archives: The Problem of *Arden of Faversham*." *Digital Scholarship in the Humanities* 34:855–73. doi.org/10.1093/llc/fqy075.

Taylor, Gary. 2002. "Middleton and Rowley – And Heywood: *The Old Law* and New Attribution Technologies." *The Papers of the Bibliographical Society of America* 96 (2):165–217. doi.org/10.1086/pbsa.96.2.24295710.

Taylor, Gary. 2020. "Shakespeare, *Arden of Faversham*, and Four Forgotten Playwrights." *The Review of English Studies* 71:867–95. doi.org/10.1093/res/hgaa005.

Taylor, Gary. 2019. "Shakespeare's Early Gothic *Hamlet*." *Critical Survey* 31 (1/2):4–25. doi.org/10.3167/cs.2019.31010202.

Weber, William. Personal Website. Accessed 2 Dec. 2021. williamweatherford-weber.wordpress.com.

Weber, William. 2014. "Shakespeare After All? The Authorship of *Titus Andronicus* 4.1 Reconsidered." *Shakespeare Survey* 67:69–84.

Rizvi, Pervez, creator.
Collocations and N-grams. Database.
London, 2017. Accessed 29 July 2021.
shakespearestext.com/can/index.htm.

Pervez Rizvi's electronic corpus of 527 plays dated between 1552 and 1657, titled Collocations and N-grams, is an invaluable aid for researchers aiming to ascertain the authorship and chronology of early modern texts. Results of automated searches enable scholars to check for phrasal repetitions between plays of the period. Rizvi's project, which is unfunded and not affiliated with any institution, is a gift to the scholarly community. Launched in 2017, it has already led to many fascinating discoveries concerning the dating of *Alphonsus, Emperor of Germany*;[19] the possibility of Cyril Tourneur's hand in *The Honest*

19. Jackson, "The Date of *Alphonsus, Emperor of Germany*."

Man's Fortune;[1] a new theory that John Ford was John Fletcher's posthumous collaborator on *The Noble Gentleman*;[2] as well as an "enlarged" Robert Greene canon of plays.[3] It seems fair to say that the database has raised the discipline to a new level by enabling researchers to examine objective, factual data linking texts according to such factors as common authorship, chronology, genre, and influence.

Collocations and N-grams consists of texts derived from Martin Mueller's corpus Shakespeare His Contemporaries[4] and the Folger Shakespeare Library Editions website. All source texts are downloadable on Rizvi's website with the exception of texts available on the Folger site. The fact that these texts derive from different sources is of little consequence given that early modern texts are in any case far from homogenous. In fact, the normalized and lemmatized texts allow a wider range of phrasal matches to be discovered than by searches using original spelling or unlemmatized forms of words. In corpus linguistics, the root form of each word (the lemma) is counted, so that "kind hearts" is matched with "kind-hearted," to offer one example. Although any researcher utilizing EEBO-TCP texts should be wary of transcription errors, such lemmatized searches reveal thousands of matches that no researcher conducting word-based searches has ever noticed before.

A phrase four words in length, like "of the brazen head," which co-occurs with Robert Greene's *Alphonsus, King of Aragon* and *Friar Bacon and Friar Bungay*, will contain different types of contiguous word sequences, known as "n-grams"—one tetragram (the four-word phrase as a whole); two trigrams (the three-word phrases, "of the brazen" and "the brazen head"); three bigrams (the two-word phrases, "of the," "the brazen," and "brazen head"); and four single words. These are what Rizvi calls "formal" n-grams. The four-word phrase itself would also constitute what Rizvi calls a "maximal" n-gram, in which case it would only be counted once.

Rizvi's database also enables users to search for discontinuous word sequences within ten-word windows: that is, matching phrases separated by intervening words, known as "collocations." For instance, the linguistic

1. Jackson, "Cyril Tourneur."

2. Freebury-Jones, "John Fletcher's Collaborator."

3. Freebury-Jones, "Determining Robert Greene's Dramatic Canon."

4. The latest incarnation of Shakespeare His Contemporaries, a collaboration between Northwestern University and Washington University in St. Louis, is EarlyPrint.

choices "distressful" and "wound" cluster in *The Spanish Tragedy* and *Arden of Faversham* but in no other play of the period. Unlike n-grams, these words are not adjacent; they are separated by non-matching words like "looks" and "to" in *Arden of Faversham* and "of" and "my" in *The Spanish Tragedy*. To reduce the matches listed in Excel files to a manageable number (i.e., tens of thousands rather than tens of millions), Rizvi has excluded matches involving very common words in his lists of "maximal" matches. These exclusions do not affect "formal" repetition counts. But in my experience, seemingly trivial groupings of words can provide excellent markers for a single authorial thought process when studied according to their contexts of use (i.e., dramatic situation, grammatical patterning, or prosodic placement), and it would have therefore been desirable to have the option of highlighting and conducting qualitative analysis of excluded "maximal" matches.

The phrasal matches between plays are "weighted," i.e., the raw figures are divided by composite word counts. Rizvi notes that, having conducted statistical tests on eighty-six plays of uncontested authorship in his corpus, dislegomena, or "unique n-grams" (i.e., occurring in just two of the 527 texts), "are better than all n-grams" for correctly identifying authors, despite the fact that n-grams unfiltered for rarity "provide a vastly greater amount of data." Rizvi has established that "unique 3-grams and 4-grams"[5] are the most reliable phrasal structures for authorship attribution purposes, with unique trigrams correctly classifying eighty-four and tetragrams correctly classifying eighty-three out of eighty-six of the attested texts in the corpus. Prior to the launch of Collocations and N-grams, I had placed emphasis on the applicability of trigrams and tetragrams for attribution purposes, while other scholars had claimed that "long word strings have been shown to be less effective in this regard than shorter ones."[6] Collocations and N-grams has enabled researchers to settle such questions through empirical, large-scale, automated means. Rizvi also points out that, in the case of researchers who are examining phraseology "for qualitative analysis, then maximal matches, and therefore counts of maximal matches, are appropriate. On the other hand, if we are looking at, say, 4-grams, in isolation to other n-grams, then of course we must use formal matches."[7]

5. Rizvi, "Which N-grams are the Best?"

6. Jackson, "Shakespeare, *Arden of Faversham*, and *A Lover's Complaint*," 130.

7. Rizvi, "The Counting of N-grams."

Importantly, the database is freely available for non-commercial purposes. The accessibility of the database, as well as the reproducibility of any tests a user chooses to conduct, renders Collocations and N-grams an excellent, scientifically-sound resource for scholars. The matching process is entirely neutral and takes no account of any scholar's authorship attributions. Some texts, however, like *Arden of Faversham* and Shakespeare's *Henry VI* plays, have been segmented according to prior attribution theories, even if these theories might seem untenable to some users. This obscures the data at times when a single scene from one of these plays features highly in the spreadsheet for a target text, but the database at least offers users the opportunity to search results for such aforementioned plays in their entirety. Moreover, Rizvi has provided instructions on his homepage so that researchers can access the database directly and filter the data according to their own choice of play divisions. This process is time-consuming and requires good skills in Excel, but I have found these instructions to be of considerable use for testing scholarly assumptions about the distribution of n-grams in sole-authored plays. The results are effective in determining the likely authors of whole plays but are less accurate in dealing with scenes or other small divisions, especially with unique matches, where the counts are small and therefore easily skewed.

Users can examine counts of n-grams in either "maximal" or "formal" variants. Beginning with the former, users should scroll to the bottom of the home page and click "Download the Files from my Microsoft Drive." They will then find a series of folders: to examine "maximal" data, click the folder titled "Lists_of_N-gram_and_Collocation_Matches." Users can select either the "Collocations" or the "N-grams" folder. Those clicking on the latter who are interested in cold, hard, numerical data should select the folder named "Summary." This in turn reveals folders for every play in the corpus, which can be sorted according to name, when the folders were last modified, or size. As the project is unfunded, there is no friendly web-based user interface, and finding the play one wishes to examine can be slightly trying. Say we are looking for results for the play *Soliman and Perseda*: users can scroll down until they reach play titles beginning with "S" and click the preview button to reveal the full play title names, as opposed to such examples as "Summary_N-grams_Sol." That folder enables users to download an Excel spreadsheet ranking Thomas Kyd's play against all other texts in the corpus according to all unique "maximal" n-gram matches:

1. *The Spanish Tragedy* [without Additions]
2. *Arden of Faversham* [excl. Act 3]
3. *1 Selimus*
4. *1 Tamburlaine*
5. *The Three Ladies of London*
6. *Edward the Second*
7. *The True Chronicle of King Leir*
8. *The Two Gentlemen of Verona*
9. *Henry V*
10. *The Battle of Alcazar*
11. *The True Tragedy of Richard the Third*
12. *Clyomon and Clamydes*
13. *A Larum for London, or The Siege of Antwerp*
14. *Alphonsus, Emperor of Germany*
15. *The Duchess of Suffolk*
16. *Troilus and Cressida*
17. *The Jew of Malta*
18. *Herod and Antipater*
19. *The Comedy of Errors*
20. *Cornelia*

It is notable that generations of scholars have attributed no less than one-fifth of the texts listed here to Kyd on entirely different evidentiary grounds. We should also note that Kyd's *Cornelia* is ranked twentieth, which reveals that the top twenty plays provide accurate indicators of common authorship.

Users wishing to examine phrases shared between *Soliman and Perseda* and other texts can go back and click on the "CSV" folder. They will once again need to navigate through the list of play folders and download the relevant Excel spreadsheet. Users can filter results in this spreadsheet according to lengths of repetitions by clicking the "No. of Matching Lemmata" tab, with the longest shared phrasal structure in the case of this play consisting of ten words; or they can rank the shared n-grams according to rarity by clicking the "No. of Plays Found in" tab. Users can also restrict results to just one other play or according to date: say, if we were only interested in plays of the period 1580–1600. I should note, however, that the dates for plays derive from Martin Mueller's

Shakespeare His Contemporaries, and one wishes that they corresponded to the more precise chronology afforded by Martin Wiggins.[8]

Users can repeat the above processes for collocations, bearing in mind the caveat that the numerical data are of considerably less value. I would, however, strongly recommend searching plays in the "CSV" folder for collocations if researchers are seeking to examine shared phraseology according to their contexts of utterance. Rizvi's database helpfully enables users to read these shared locutions according to the surrounding verbal texture by clicking the "Text" tabs.

Users seeking to examine "formal" n-gram data should click the folder titled "N-gram Counts." They will find a series of sub-folders, including those pertaining to function words, which are of little use for ascertaining authorship, chronology, or influence, as well as all n-gram matches, i.e., matching lemmata unfiltered for rarity. The most useful folder for attribution purposes is titled "Counts_Unique_N-grams." Users have the option of clicking either "Types" or "Tokens," but should disregard the "Types" folder and examine word-tokens. Selecting a play and downloading the Excel spreadsheet allows users to rank plays according to different unique n-gram lengths. For authorship attribution purposes, the weighted figures for "3-grams" and "4-grams" are most effective. The top dozen plays listed per spreadsheet for unique trigrams and tetragrams usually give a good indication of common authorship, but if researchers are seeking to examine, say, an Elizabethan text, the top twenty might be less restrictive given the vast date span covered by the 527 texts in the corpus.

Collocations and N-grams is a veritable treasure trove, consisting of a number of additional, helpful documents, including instructions on conducting your own tests. The database represents a watershed in corpus linguistic studies of early modern texts, and any researcher interested in authorship, chronology, influence, or source studies should consult it.

DARREN FREEBURY-JONES
The Shakespeare Birthplace Trust
https://doi.org/10.33137/rr.v44i4.38649

8. See Mueller, and Wiggins and Richardson.

Works Cited

Early Print. Evanston, IL: Northwestern University / St. Louis, MI: Washington University. Accessed 29 July 2021. earlyprint.org/.

Folger Shakespeare Library Editions. Washington, DC: Folger Shakespeare Library. Accessed 29 July 2021. folger.edu/folger-shakespeare-library-editions.

Freebury-Jones, Darren. 2020. "Determining Robert Greene's Dramatic Canon." *Style* 54 (4):377–98. doi.org/10.5325/style.54.4.0377.

Freebury-Jones, Darren. 2020. "John Fletcher's Collaborator on *The Noble Gentleman*." *Studia Metrica et Poetica* 7 (2):43–60. doi.org/10.12697/smp.2020.7.2.03.

Jackson, MacDonald P. 2019. "Cyril Tourneur and *The Honest Man's Fortune*." *Medieval and Renaissance Drama in England* 32:203–18.

Jackson, MacDonald P. 2019. "The Date of *Alphonsus, Emperor of Germany*: The Evidence of Unique N-Gram Matches." *Notes and Queries* 66 (4):512–14. doi.org/10.1093/notesj/gjz129.

Jackson, MacDonald P. 2017. "Shakespeare, *Arden of Faversham*, and *A Lover's Complaint*: A Review of Reviews." In *The New Oxford Shakespeare: Authorship Companion,* edited by Gary Taylor and Gabriel Egan, 123–38. Oxford: Oxford University Press.

Mueller, Martin. 2012. "Repeated N-grams in Shakespeare His Contemporaries (SHC)." Accessed 26 November 2021. scalablereading.northwestern.edu/?p=312.

Rizvi, Pervez. 2018. "The Counting of N-grams." Accessed 26 November 2021. shakespearestext.com/can/index.htm.

Rizvi, Pervez. 2018. "Which N-grams Are the Best?" Accessed 26 November 2021. shakespearestext.com/can/index.htm.

Wiggins, Martin, and Catherine Richardson. 2012–. *British Drama 1533–1642*. Oxford: Oxford University Press.