Clustering and Classifying the Sonnets' Dateless Night

James Fisk, B.A.

Submitted in partial requirement for
Special Honors in the Department of English
The University of Texas at Austin

May 2015

Professor Douglas Bruster
Department of English
Supervising Faculty

Professor Pradeep Ravikumar
Department of Computer Science
Second Reader

Contents

Abbreviations

Alphabetical, dates from Riverside chronology

1H4	*1 Henry IV* (1596-7)
1H6	*1 Henry VI* (1589-90)
2H4	*2 Henry IV* (1598)
2H6	*2 Henry VI* (1590-1)
3H6	*3 Henry VI* (1590-1)
ADO	*Much Ado About Nothing* (1598-9)
ANT	*Antony and Cleopatra* (1606-7)
ARD	*Arden of Faversham* (1590)
AWW	*All's Well That Ends Well* (1602-3)
AYL	*Winter's Tale* (1610-11)
COR	*Coriolanus* (1607-8)
CYM	*Cymbeline* (1609-10)
E3	*Edward III* (1592-5)
ERR	*Comedy of Errors* (1592-4)
H5	*Henry V* (1599)
H8	*Henry VIII* (1612-3)
HAM	*Hamlet* (1600-1)
JC	*Julius Caesar* (1599)
JN	*King John* (1594-6)
LC	*Lover's Complaint*
LLL	*Love's Labor's Lost* (1594-5)
LR	*King Lear* (1605)
LUC	*The Rape of Lucrece* (1594)
MAC	*Macbeth* (1606)
MFM	*Measure for Measure* (1602)
MND	*Midsummer Night's Dream* (1595-6)
MV	*Merchant of Venice* (1596-7)
OTH	*Othello* (1604)
PER	*Pericles* (1607-8)
PHT	*Phoenix and Turtle* (1601)
PP	*The Passionate Pilgrim* (1599)
R2	*Richard II* (1595)
R3	*Richard III* (1592-3)
ROM	*Romeo and Juliet* (1595-6)
SHR	*Taming of the Shrew* (1593-4)
STM	*Sir Thomas More* (1594-5)
TGV	*Two Gentlemen of Verona* (1594)
TIM	*Timon of Athens* (1607-8)
TIT	*Titus Andronicus* (1593-4)
TMP	*Tempest* (1611)
TN	*Twelfth Night* (1601-2)
TNK	*Two Noble Kinsmen* (1613)
TRO	*Troilus and Cressida* (1601-2)
VEN	*Venus and Adonis* (1593)

Introduction

Writing on *Hamlet,* T.S. Eliot notes that in comparison to *Othello, Antony and Cleopatra* or *Coriolanus,* "the subject might conceivably have expanded into a tragedy like these, intelligible, self-complete, and in the sunlight." Instead, Eliot posits, *Hamlet* is "like the sonnets . . . full of some stuff that the writer could not drag into light, contemplate or manipulate into art."[1] This supreme intractability, which Eliot's reading considers the "artistic failure" of *Hamlet,* has been also the enchantment that has spellbound readers and critics of the Sonnets since their first publication in 1609. The collection of 154 of these 14-line poems can be foremost characterized by what Colin Burrow deems the "delighted mystification" it ceaselessly inspires in its readers.[2] The Sonnets are slippery, ambiguous and challenging, and have been a siren song for critics who have tried to make sense of them. They are neither intelligible, self-complete, nor in the sunlight. They exist in shadows.

Their enigma extends to the details of their inspiration, composition and publication. Dedicated to a mysterious "Mr. W.H.", printed by a mysterious publisher, and devoted to a subject that shifts from a fair young man to a sensual mistress, the 1609 Sonnet Quarto has entertained endless speculation, but has elucidated very little substantive information pertaining to its genesis. The first historical evidence of Shakespeare's Sonnets appears in 1598, indicating that there is a period of over a decade (at the very least) when Shakespeare could have been composing and revising his Sonnets before their 1609 publication. When Shakespeare wrote his sonnets and in what order during this period, a period that covers at least 17 plays and two other poetic works,

---

1   T. S. Eliot, *"Hamlet* and his Problems," *The Sacred Wood and Major Early Essays* (New York: Dover Publications 1998), 58.
2   Colin Burrow, ed. *William Shakespeare: The Complete Sonnets and Poems.* (Oxford: Oxford University Press, 2002), 92.

is anyone's guess.

This gap in our knowledge of when Shakespeare wrote one of his hallmark works would be alarming in its own right, as it would be for any of the plays or narrative poems, but the omission is particularly troublesome for the Sonnets. They are, perhaps more so than any other Shakespearean work, authorially aware and self-inflected. While reading the Sonnets absolutely autobiographically would be myopic and fruitless (the Sonnets intrinsically discourage an absolute reading of any kind), the self-conscious and self-referencing voice of the poet, frequent play on the word and name "Will," and the unique insertion of Shakespeare's name into the title of the work, *Shakespeare's Sonnets: Never Before Imprinted,* evinces an indisputable bind between Shakespeare and his Sonnets, as well as his presence within them. An understanding of the Sonnets contributes to an understanding of Shakespeare both as a man and a poet. Placing the Sonnets into a specific temporal context also places Shakespeare, his poetic subjects, patrons and peers. The question of when helps to answer these questions of who, where, how and why. Yet now, more than four centuries after their publication, our understanding of the Sonnets' chronology is still woefully scant.

Recently, however, scholars have begun to date the Sonnets using computational methods. Kent Hieatt, Charles Hieatt and Anne Prescott's 2001 study looked at the frequencies of rarely used words across Shakespeare's plays and in the Sonnets, searching for links between groups of sonnets and periods of playwriting.[3] Their work, later supplemented by MacDonald P. Jackson, indicated that the later sonnet subsequence, often called the Dark Lady sequence, was likely written earlier in Shakespeare's career, while certain subsections of the Fair Youth sequence aligned particularly with works

---

3   Kent A. Hieatt, Charles W. Hieatt and Anne L. Prescott, "When Did Shakespeare Write Sonnets 1609," *Studies in Philology* 88 no.1 (2001).

written in the seventeenth century.[4] Their work, while promising, is neither exhaustive nor conclusive. This study intends to take up this thread of investigation of using modern, computational techniques to answer the centuries-old question of when Shakespeare wrote his elusive Sonnets.

External Evidence and Publication History

Arguing a Shakespearean hypothesis in 1923 for a *Sir Thomas More* passage based upon internal evidence, R.W. Chambers compared the search to a situation in which, "You have to meet in a crowd a Mr. Harris, hitherto unknown to you, but who, you are informed, has red hair, wears a monocle, and walks with a limp," a conceit to which Jackson riposted in 2003,

> We might even reckon probabilities—one in ten men is left-handed, one in eight wears a moustache, and so on—and enlarge the list to the point where multiplying the separate odds would produce a billion-to-one coincidence. Harris must have met his *doppelgänger*! No, the passer-by is a stout Caucasian and Harris is a slim West Indian. The total absence of constraints on our search for resemblances renders the calculation meaningless.[5]

The sentiment communicates the crux of one problem with internal evidence: it can tend towards self-confirmation and become quickly detached from any realistic basis. It is paramount in any investigation into internal, stylistic evidence not to neglect the external evidence, which provides a crucial framework of constraints and indicators upon which to build a broader inquiry. Any method dealing with the probable and the possible must be grounded and constrained by the certain and the concrete. The internal evidence

---

4   MacDonald P. Jackson, "Vocabulary and Chronology: The Case of Shakespeare's Sonnets," *The Review of English Studies* 52 no.205 (2001).
5   From note 37 in Ward E. Y. Elliott and Robert J. Valenza, "Oxford by the Numbers: What Are The Odds The The Earl of Oxford Could Have Written Shakespeare's Poems and Plays?" *Tennessee Law Review* 72 (Fall 2004) 339, quoting R.W. Chambers "The Expression of Ideas—Particularly Political Ideas—in the Three Pages, and in Shakespeare," *Shakespeare's Hand in The Play of Sir Thomas More* (1923) 142; MacDonald P. Jackson, *Defining Shakespeare: Pericles as Test Case* (2003) 192.

must always yield to the external.

With this sentiment in mind, the external evidence surrounding the composition and publication of the Sonnets is slight. The history of Shakespeare's Sonnets begins in late 1598, with Francis Meres' publication of *Palladis Tamia*, entered in the Stationers Register on September 7, 1598. This smorgasbord of critical discourse includes a short and ambrosial account of "hony-tongued *Shakespeare*" and his "sugred Sonnets," circulated "among his private friends, &c" (Figure 1).[6]
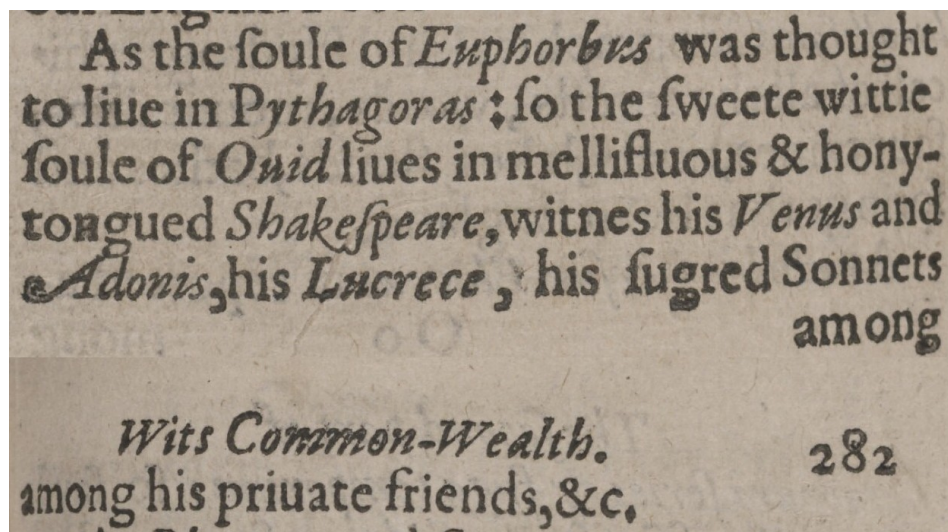

Figure 1 Excerpt from *Palladis Tamia*, the first record of Shakespearean sonnets.

While there is nothing explicitly linking the "sugred Sonnets" to any of the sonnets from the 1609 Quarto, *The Passionate Pilgrim* (1599), which does include specific sonnets from the 1609 Quarto, follows closely after, suggesting they both come from same crop. *The Passionate Pilgrim* includes Sonnets 138 and 144, as well as three sonnets from the 1598 Quarto of *Love's Labor's Lost*, and fifteen other poems that cannot be confidently attributed to Shakespeare. The miscellany was published by William Jaggard under Shakespeare's name and sold well, going through three editions with the

---

6 Francis Meres, *Palladis Tamia. Wits treasury. Being the Second part of Wits Common wealth. By Francis Meres, Maister of Artes of both Vniuersities.* (P. Short, 1598) 281-2. Accessed through Beinecke Digital Collections, Beinecke Rare Book and Manuscript Library, Yale University Library.

first surviving only as a fragment, implying much-thumbed pages among its buyers. Shakespeare apparently took umbrage with the publisher, and likely with the widespread circulation of his name attached to a volume Katherine Duncan-Jones describes as "quirky, occasionally sententious and frequently obscene."[7] Thomas Heywood reported in 1612 that Shakespeare was "much offended with M. *Iaggard* (that altogether vnknowne to him) presumed to make so bold with his name."[8]

It is clear that at least two sonnets, 138 and 144, existed by 1599, and it seems likely that there were more circulating in private. Duncan-Jones cites a Stationers Register entry for "A booke called *Amours* by J.D. with certain *other sonnetes* by W.S." entered on January 3, 1600, which she suggests could have contained a collection prepared by Shakespeare to mitigate the damage of the "piratical and mediocre" *Passionate Pilgrim*. This evidence, however, cannot amount to more than speculation. The identity of J.D. and W.S. is indeterminate, as well as the knowledge of whether the volume ever went to print.

The paper trail here dies until May 20, 1609—the date of entry of Thorpe's *Shakespeare's Sonnets* in the Stationers Register. The date of actual publication after this registration is unknown. Duncan-Jones notes that the 1609 Quarto did not appear to enjoy nearly the height of success as *Pilgrim's Progress.* It was never reprinted (*Venus and Adonis*, in contrast, went through at least sixteen editions in Shakespeare's lifetime), and there are thirteen copies which survive today in good condition.

There is one more important, though not immediately obvious, piece of external evidence to consider when evaluating the composition of the Sonnets: plague outbreaks. Duncan-Jones points out that both *Venus and Adonis* (1593) and *Lucrece* (1594) were

---

7   Katherine Duncan-Jones, ed. *Shakespeare's Sonnets* (London: Arden Shakespeare 1997) 2.
8   Thomas Heywood, *An Apology for Actors* (1612) . Accessed through Early English Books Online, University of Michigan Libraries.

published at a time when plague swells had caused civic orders for the closure of public theaters. These periods of hiatus for London theater naturally required dramatists to turn to other sources of income, one of which, for Shakespeare, included printed poetic works. Duncan-Jones suggests this may have accounted for the long gap between the "sugred Sonnets" of 1598 and the 1609 Quarto, as well as Shakespeare's anger at Jaggard's *Passionate Pilgrim*, suggesting that Shakespeare had a store of sonnets "saved up against a plague-ridden day," from which Jaggard's volume detracted value.

One exceptionally long period of plague outbreak closed theaters for thirteen months from 1603 to 1604. This particular period of plague was heralded by the death of Queen Elizabeth and the ascension of James I, a subject, as discussed below, of allusion in Sonnet 107 and possibly others. Although not entirely provable, it is easy to imagine a playwright out of work for over a year refashioning himself as a sonneteer, and using his time in unemployment to compose new sonnets or revise old ones.

At the time of the Sonnets' entry in the Stationers Register on May 20, 1609, plague deaths in London had reached nearly one hundred per week. Duncan-Jones notes Shakespeare "may well have been anxious to complete the transaction as quickly as possible before retreating from plague-ridden London," as there is very little evidence for authorial press-correction in the publication, the dedication is signed by Thorpe (something he was known only to do when an author was dead or out of the country), and there is a record of Shakespeare conducting legal business soon after in Statford on June 7, 1609.

Plague was clearly an impetus for Shakespeare to publish poetry, and, it seems likely, to compose it as well. This significant player in the working habits of Shakespeare and, consequently, the poetic habits of Shakespeare, should not be overlooked as

important external evidence of more probable periods of composition.

Authorial Consent

Scholars have historically debated over authorial consent and the 1609 publication of the Sonnets. Early editors of the Sonnets generally operated under a conception of Thomas Thorpe, publisher of the 1609 Quarto, as a rogue and scoundrel, surreptitiously gaining and publishing Shakespeare's manuscripts in poor form and not in their intended order. Duncan-Jones, in her comprehensive analysis of Thorpe and the Sonnets, traces part of this reputation to Sir Sydney Lee's 1905 *DNB* entry for Thorpe, which deems him both "predatory and irresponsible," who specialized in exploiting "neglected copy."[9] In this instance, it may be of value to note Lee's strong opinions of the Sonnets themselves. Lee's *DNB* entry on Shakespeare takes unusual pains to prohibit "a literal interpretation" of the sonnets, which, he advises, "credits the poet with a moral instability which is at variance with the tone of all the rest of his work," and "casts a slur on the dignity of the poet's name which scarcely bears discussion." Lee's protestations indicate not only a propensity for overstatement, but also, perhaps, an acute discomfort with the homosocial themes of the sonnet sequence and a desire to distance the poet from them as much as his powers permitted.

Duncan-Jones' own investigation reveals little to impart Thorpe with any fewer scruples than his contemporaries, and indicates that "Shakespeare must surely have known" and likely "viewed favorably" Thorpe and his practices. Duncan-Jones' conclusions have steered the scholarly consensus in recent years towards accepting authorial consent of the 1609 Quarto. As far as it affects this study, it is easiest to accept the Wells-Taylor fence-straddling position that "there is no evidence that Shakespeare had

---

9   Katherine Duncan-Jones, "Was the 1609 *Shake-speare's Sonnets* Really Unauthorized?" *The Review of English Studies* 34 no. 134 (May, 1983) 164.

anything to do with their publication," although "Thorpe was a reputable publisher, and there is nothing intrinsically irregular about his publication."[10]

A Lover's Complaint

The 1609 Quarto included the 154 sonnet sequence appended by *A Lover's Complaint*, which, while historically under disputed authorship, has been largely accepted as Shakespearean and is included with the Sonnets in many modern editions. The feminine complaint poem appended to complement the masculine-voiced sonnet sequence was a not uncommon feature in Elizabethan poetry.[11] Work dating *A Lover's Complaint* shows that it has a strong stylistic affiliation with works composed post-1600, and was likely penned while Shakespeare was collecting and sequencing his sonnets at some date near publication.[12] This study accepts that *A Lover's Complaint* was written by Shakespeare, and belongs to a period of composition between 1599 and 1609, although it is not of primary concern.

Other Editions

The next manifestation of the Sonnets after the 1609 Quarto came in John Benson's 1640 edition, published twenty-four years after Shakespeare's death. Benson conjoined, reordered, tossed out, titled the sonnets and intermixed them with miscellany from *The Passionate Pilgrim*, *The Phoenix and the Turtle* and others. He also altered pronouns to change the male fair youth into a female subject. The Benson edition, while not wildly successful commercially, remained the authoritative text for many years until the 1780 edition by Edward Malone, an edition that stayed largely true to the 1609

---

10 Stanley Wells and Gary Taylor with John Jowett and William Montgomery, *William Shakespeare: A Textual Companion* (Oxford: Clarendon Press, 1987) 444.
11 Examples include *Delia* (Daniel, 1592), *Phyllis* (Lodge, 1593), *Cynthia* (Barnfield, 1593), and *Amoretti* (Spenser, 1595).
12 See MacDonald P Jackson, *Shakespeare's 'A Lover's Complaint':Its Date and Authenticity*, University of Auckland Bulletin 72, English Series 13 (Auckland, 1965); Eliot Slater, "Shakespeare: Word Links between Poems and Plays," *NQ* 220 (1975) 157-63.

Quarto. It is Malone's legacy, along with the 1609 Thorpe Quarto itself, that precipitates modern editions of the Sonnets. Although many manuscripts exist of individual sonnets, none predates 1620, making *The Passionate Pilgrim* and the 1609 Quarto the only texts reliably relevant to the question of dating.[13]

Internal Evidence — Historical

Stonewalled by the external evidence, many scholars have turned to what the Sonnets have to say for themselves. Owing to the Sonnets' aesthetic and thematic intractability, however, the greater part of this exercise is left to speculation. Sonnet 2, for example, which begins "When forty winters shall besiege thy brow," seems to imply the age of the poet at composition, evoking a tone of experience, describing what has happened to himself in order to scare his fair youth into procreation. Shakespeare, born in 1564, would be forty in 1604, making Sonnet 2 a later sonnet if he wrote it at or around that time. This claim is, of course, indefensible. Indeed, the external evidence of Sonnet 2 of manuscripts suggests an early date of composition. There are eleven surviving manuscripts of Sonnet 2, four of them with the variant title of *Spes Altera*, some significantly altered from the 1609 text, suggesting the sonnet may have had a life of its own as an earlier text before its inclusion in the 1609 sequence. One manuscript even substitutes the word "forty" for "threescore," which, following the same internal logic, postdates the 1609 Quarto (and his own death) with a sixty-year old Shakespeare in 1624.

Only one sonnet, Sonnet 107, reproduced in Figure 2, offers compelling internal evidence for dating. In 1949, Leslie Hotson noted the association between the "mortal moon" of line 5 and the crescent-shaped Spanish Armada at their defeat by the English in

---

13   Wells and Taylor offer that, based on the consensus that the 1609 Quarto was set from transcript, at least two degrees removed from Shakespeare's hand, the individual manuscripts deserve greater attention for being potentially closer to the poet's hand. This hypothesis, while compelling, remains yet unproven, and the manuscripts do not offer additional insight into the dating question by my judgment.

1588.[14] Others have ascribed the mortal moon to Queen Elizabeth, who John Kerrigan notes was "frequently compared to Diana, the chaste *moon* goddess, though herself *mortal,*" who dies ("eclipse indur'de") in 1603.[15] Kerrigan follows the Elizabeth conceit through the subsequent lines to startling length, diving deep into the politics of the Queen's death, and identifies overwhelming evidence of a close coupling between lines 7 through 9 and the imagery and lexicon of the rhetoric surrounding the ascension of James VI in 1603. Kerrigan concludes that "on the basis of allusions, in short, 1603 seems the obvious date."
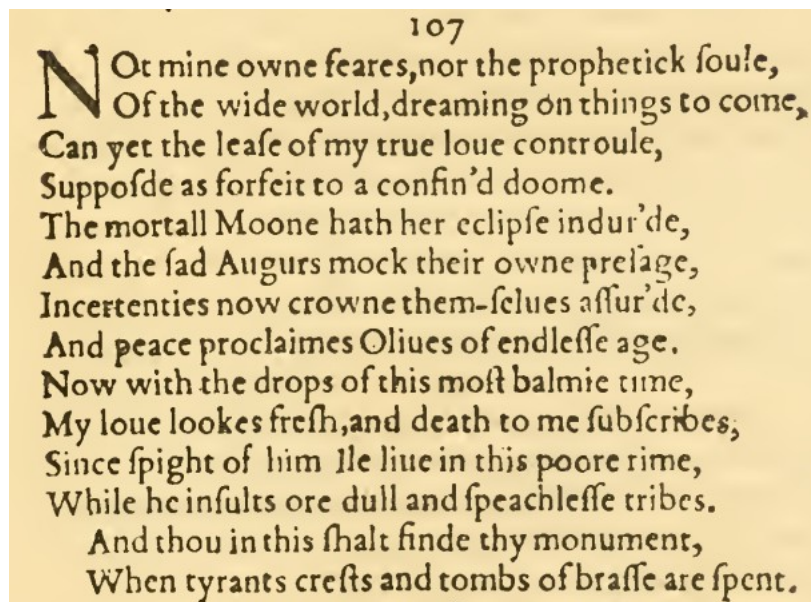
Figure 2 Sonnet 107, from 1609 Quarto

Other sonnets have arguable cases for ties to the same historical time. The pyramids of Sonnet 127 may correspond to the elaborate monuments erected to celebrate James's ascension, as likewise the canopy of Sonnet 125. Sonnet 124 invokes victimized political agents who "die for goodness, who have lived for crime," calling to mind any number of figures, but most prominently the conspirators of the Gunpowder Plot of

---

14 Leslie Hotson, *'Shakespeare's Sonnets Dated' and Other Essays* (New York: Oxford UP, 1949) 1-36.
15 John Kerrigan, ed. *The Sonnets and A Lover's Complaint* (London: Penguin, 1986) 314-20.

November 1605. These sonnets, however, do not offer the same level of confidence as the allusions in Sonnet 107.

Internal Evidence — Mr. W.H.

Burrow acutely describes the sonnet as "a form which was located at the intersection between private papers and printed record." This sentiment is most apparent in the frustrating and compelling dedication to the sonnets, which Burrow endears as "a dank pit in which speculation wallows and founders."[16] The dedication, reproduced in Figure 3, is the only place in the quarto where the reader is not reminded of the author's name (every page in the sequence has a header of either "SHAKE-SPEARES" on the verso and "SONNETS" on the recto), yet it has inspired the most investigation into Shakespearean biography and conspiracy than any other portion of the Sonnets.

The dedication inscribes the volume to a "Mr. W.H.", and is signed "T.T.", a convention likely indicating Shakespeare's absence from the publication, possibly, as discussed above, due to the summer plague. W.H.' s epithet as the "onlie begetter" strongly indicates him as the subject of the male-addressed sonnets (1-126). Kerrigan lists no fewer than seven possible candidates for W.H., including Shakespeare himself, as does Burrow. Both offer the same conclusion, albeit in different terms. Burrow cites the intentionality of the ambiguity and the trope of "targeted anonymity" in Elizabethan erotic verse, meant to "make their readers feel that they were both inside and outside a

---

16 See *supra* note 2. 98.

TO.THE.ONLIE.BEGETTER.OF.
THESE.INSVING.SONNETS.
M^r. W. H. ALL.HAPPINESSE.
AND.THAT.ETERNITIE.
PROMISED.

BY.

OVR.EVER-LIVING.POET.

WISHETH.

THE.WELL-WISHING.
ADVENTVRER.IN.
SETTING.
FORTH.

T. T.

Figure 3 Dedication, from 1609 Quarto

charmed circle of knowledge." Thorpe's dedication "is designed to play to those expectations," Burrow concludes. Kerrigan, in turn, offers the consolation that "none of this matters much."

The vexation of W.H. does matter, however, to the question of dating. Sonnet 104, for example, describes a three-year long period of friendship with the fair youth. In this case, as others, the sonnet dates can help to identify the youth, just as the identity of the youth sheds light on the period of composition. The two mysteries are codependent, and headway in one helps to result in headway in the other.

Significance

The fair youth as a poetic figure cannot correspond exactly to W.H., which, in his turn, may or may not correspond to a true person associated with Shakespeare. Likewise, the flattering admirer of the fair youth nor the "thy Will" of the dark lady can correspond exactly to Shakespeare biographically. These figures exist as aesthetic abstractions, self-contained and independent. All the same, however, the Sonnets come the closest out of Shakespeare's corpus to illuminating both biographical facts of his career as well as the inner workings of his imagination. The Sonnets are simultaneously intensely personal and intrinsically universal. They are, as Eliot described, "full of some stuff that the writer could not drag into the light," yet also, as Wordsworth calls them, the key with which "Shakespeare unlocked his heart."[17]

In the endeavor to understand Shakespeare and understand the Sonnets, dating matters. More accurate dating provides a sharper lens with which to view Shakespeare and his stylistic development, and provides a crucial framework for critical investigation. For example, academic Joel Fineman claimed in 1986 that the dark lady sequence is an

---

17 William Wordsworth, "Scorn not the Sonnet"

ironic commentary on the fair youth sequence.[18] This theory, however, loses substantial

credibility if the Hieatt study is to be believed and it is accepted that Shakespeare wrote

the dark lady sequence up to ten years earlier than the fair youth sequence. Dating the

Sonnets provides a common understanding of the context of their genesis, and allows the

critical community to work with coordinated efforts in the right direction.

      The Sonnets could be read in a vacuum, without any consideration for the cause

or the circumstances behind their being, but that is not likely how they were read in their

time, nor even how they were presumed, by their author, to be read. The 1609 Quarto is

the only work of Shakespeare's with both his name in the title and his name printed on

every single page (save the dedication). The numerous puns on "Will" or "Hugh" or

"Hathaway," and the mysterious dedication page all by design entice readers into

speculation and wonder. Shakespeare inserts himself into the Sonnets in no subtle way,

both physically on the page and in the contents of his verse. We are not meant to forget

the author, nor cease to probe his relationship to his work.

      In this investigation, I continue the centuries-long push to extract this contextual,

historical information from Shakespeare's enigmatic verse. I employ a suite of tests used

to quantify the Sonnets and Shakespeare's larger corpus, from the simplest of metric of

letter frequency, to measures of shared imagery, rhymes, phrases and more. I use this data

to cluster the sonnets into groups, and then to classify those groups temporally. The

sonnets may exist in shadows, but they are not beyond being dragged into light.

---

18   Joel Fineman, *Shakespeare's Purjured Eye: The Invention of Subjectivity in the Sonnets* (University of
California Press, 1986) 66.

Evidence

There exists no black-box computational machine that can take a text data input and return to the interested scholar any sort of meaningful analysis of what it is, who wrote it, nor, as this investigation concerns itself, when it was composed. There must be a human intermediary between the text and the machine who can inspect the text for relevant attributes, extract them, and groom them to be machine-friendly before a computer can be expected to be of any help. Into this mediating role steps the stylometrist. Stylometry is the discipline of finding and quantifying the variations of literary style through statistics, bridging the gap between the humanist intuition, inspection and understanding of literature and the computational number-crunching and data manipulation.[1] The field owes its origins to English logician Augustus DeMorgan who, in 1851, musing in a letter to his colleague, wondered whether a question of authorship might be answered by quantifying whether one text "does not deal in longer words" than another.[2] Although the field has evolved and technology has advanced since then, the fundamental question remains unaltered. How do we take a text and extract relevant, pure and meaningful data? Does this text deal with longer words than that one, and, if so, does it matter?

Consequently, the texts used, their preprocessing, the choice of relevant data and its method of extraction all occupy principal importance in this investigation. If the data is tainted, or if the stylometric attributes mined are not pertinent to the question of chronology, then any further effort in classification, clustering, dating or using any other

---

1   A good introduction may be found in Efsathios Stamatatos, "A Survey of Modern Authorship Attribution Methods," *Journal of the American Society for Information Science and Technology* 60, no. 3 (2009): 538; Javier Calle-Martín and Antonio Miranda-García, "Stylometry and Authorship Attribution: Introduction to the Special Issue," *English Studies* 93, no.3 (2012): 251-2.

2   F. J. Tweedie, S. Singh and D.I. Holmes, "Neural Network Applications in Stylometry: The Federalist Papers," *Computers and the Humanities* 30, no. 1 (1996): 1.

algorithm to engage with and understand the sonnets is an exercise in senselessness. Without thorough, consistent and accurate data, all conclusions are rendered moot.

Texts

This study uses the *Riverside Shakespeare* as its source text, the standard for most stylometric Shakespearean studies. It upholds a reputation as "the most widely respected edition of the Complete Works," notable for consistently being "reliably and conservatively edited."[3] The Riverside regularizes and cautiously modernizes spellings from the First Folio (1623), which, while potentially losing revealing characteristics of spelling, gains the advantages of a self-consistent and reliable corpus.

In light of recent publications[4] on the collaborative nature of Shakespeare's playwriting, certain disputably non-Shakespearean portions of collaborative plays were withheld from analysis.[5] Generous estimates place the figure of Shakespeare's career number of word tokens penned at over 900,000.[6] With copious amounts of data available, the increase in likely purity of the test corpus outweighs the potential loss of information by removing these passages.

Stage directions and other paratext (titles, speech headings and so forth) are removed from all plays.

Here it may be valuable to note that "word" generally refers to a word in the

3   John Jowett, Review of *The Riverside Shakespeare*, by G. Blakemore Evans, *The Modern Language Review* 94, no. 4 (1999): 1079.

4   See Brian Vickers, "Incomplete Shakespeare: Or, Denying Coauthorship in *1 Henry VI*," *Shakespeare Quarterly* 58, no.3 (2007), 311-352; Brian Vickers, *Shakespeare, Co-author: A Historical Study of Five Collaborative Plays* (Oxford: Oxford UP, 2002); *Shakespeare, Computers, and the Mystery of Authorship*, ed. D. H. Craig and Arthur F. Kinney (Cambridge: Cambridge UP, 2012); MacDonald P. Jackson, *Determining the Shakespeare Canon: Arden of Faversham and Lover's Complaint* (Oxford: Oxford UP, 2014) among others.

5   Redacted works consist of: *1 Henry VI* (2.4, 4.2-4.5 included), 2 *Henry VI* (1-4.2, 5 included), *3 Henry VI* (1.3-2.2, 2.4-3.1, 4.1, 5.1, 5.3-5.7 included), *Arden of Faversham* (2.47-52, 3-9 included), *Edward III* (1.2, 2.3, 2.4, 3.2.1-202, 5.1 included), *Pericles* (3-5 included), *Titus Andronicus* (2.3-3.2, 4.2-5.3 included), *Timon of Athens* (1.1, 2.1, 2.2, 4.1-4.3, 5.1-5.4 included), and *Two Noble Kinsmen* (1.1-2.1, 2.6-3.2, 4.3, 5.1, 5.4 included).

6   Marvin Spevack's *Complete and Systematic Concordance to Shakespeare* totals to 916,816 words with the inclusion of recently attributed works.

abstract, like a class, while "word token" refers to a word specifically, like an instance. One way to think of this type-token distinction is that the sum of all the word tokens used by Shakespeare is the number of times he put a pen to paper and lifted it back, while the sum of all the words used by Shakespeare would be a representation of the extent of his vocabulary. One further level of abstraction from the word is the "lemma," the uninflected lexical signifier of a concept, or, a dictionary definition. Each level of abstraction is a strict subset of the one before it. For one example of the interactions between these different levels, "disdaining," as seen in line 75 of Act III, Scene 5 of *Cymbeline* is a word token. "Disdaining," the word used three times by Shakespeare, is a word. "Disdain" is the lemma of the word "disdaining," as well as "disdains" (used six times by Shakespeare), "disdained" (used eight times by Shakespeare) as well as the word "disdain" itself (used 39 times by Shakespeare).

Chronology of the Corpus

Before attempting to date the composition of the sonnets, a credible chronology of Shakespeare's greater corpus must be established to test against. At this point in the academic conversation, the question of an exact chronology of plays continues to be discussed, although, fortunately, the broad strokes remain generally uncontested. This study mostly follows the model offered by Bruster and Smith's "A New Chronology for Shakespeare's Plays"[7] that modifies traditional dating paradigms and precedents to account for the internal data of changing verse line structure over Shakespeare's career. Dates for narrative and lyric poems come from the standardized Riverside chronology.

As the problem of assigning dates to the different sonnets is essentially a problem

---

7   Douglas Bruster and Geneviève Smith, "A New Chronology for Shakespeare's Plays," *Digital Scholarship in the Humanities* (Dec 2014) Table 1, 12.

of classification, constructing a workable chronology turns into a task of creating relevant

and representative classes for the different points of Shakespeare's career. The Hieatt,

Hieatt and Prescott effort at dating the sonnets[8] divides the Shakespeare corpus in half,

creating a class of "early" works of all texts before *Hamlet*, and a class of "late" works of

Hamlet and all texts that follow. This approach has the advantage that both classes

contain large amounts of data, which, by the law of large numbers, should mean that they

are closer to the expected ideal of that class, being more insulated from skewing by

anomalous data. These large classes encounter trouble, however, in their expansive

breadth. Each class represents over ten years of composition and hundreds of thousands

of words, but assumes a consistent and representative style throughout. The binary gives

little concession to continuity, and creating large arbitrary classes which may hold within

multiple distinct subclasses only serves to muddy the data and offer unrepresentative

classes. This weakness bears out in the results of that study, which, faced with many

sonnets that had attributes from *both* classes, concluded that those sonnets were

composed early and revised late.

      MacDonald P. Jackson, in his response to the Hieatt study, points out that many of

the sonnets have attributes "associated exclusively or mainly with works written neither

towards the beginning nor towards the end of Shakespeare's career, but in the middle."[9]

Taking the helm to further Hieatt's investigation, Jackson accommodates for the middle

sonnets by distinguishing the classes of plays further by a factor of two. His efforts result

in four chronological classes, which he terms "groups", arranged by Karl Wentersdorf's

---

8   Kent A. Hieatt, Charles W. Hieatt and Anne L. Prescott, "When Did Shakespeare Write Sonnets 1609," *Studies in Philology* 88 no.1 (2001): 69-109.
9   MacDonald P. Jackson, "Vocabulary and Chronology: The Case of Shakespeare's Sonnets," *The Review of English Studies* 52 no.205 (2001): 63.

modified version of E.K. Chambers's standard chronology[10] and by clusters of "dislegomena" and "trislegomena", twice- and thrice-used words as identified by scholar Gregor Sarrazin in a stylometric analysis from 1897.[11] Using four classes instead of two, Jackson tentatively classifies the middle-seeming sonnets with greater dexterity and specificity than Hieatt's model could. Jackson's efforts ultimately lead to a more confident and convincing proposal for the chronology of the sonnets' composition.

The greater success Jackson enjoyed over Hieatt through dividing the plays further begs the question: why stop there? Why not gain greater precision by having increasingly more classes, to the point of each play representing its own chronological class? Or, why bother with discrete classes at all, and instead discard them for a chronological continuum where one could trace Shakespeare's stylistic progression from year to year, month to month? At this point, the law of large numbers' reverse and any Shakespeare reader's common sense become relevant: as the sample size decreases, the sample mean increasingly diverges from the population mean. It is the same principle intuitive to any reader who recognizes that one scene from *Othello* does not imply that all plays past 1600 are about handkerchiefs and strawberries. Although Shakespeare's style did have a larger and possibly unconscious arc that spanned his career, it still remained well within his adept and conscious control to configure his style to whatever context or content he desired for a play, a scene or a character. A single play or any subsection of a play is far too sensitive to this contextual manipulation of style to be dependably representative of greater trends in Shakespeare's career. Additionally, the more temporally specific and small a sample for a class becomes, the more critical it becomes that those samples are dated at a level of precision and confidence that, for some plays, the

---

10 Karl Wentersdorf, "Shakespearian Chronology and the Metrical Tests," *Shakespeare-Studien. Festschrift für Heinrich Mutschmann* (Marburg, 1951).
11 Gregor Sarrazin, "Wortechos bei Shakespeare", *Shakespeare Jahrbuch*, 33 (1897).

academic conversation does not yet support.

The challenge therefore becomes creating a chronological model of classes that is broad enough to have large samples that can be confidently included, yet is also specific enough to represent distinct and meaningful periods of composition. To this end I use a modified version of Jackson's four groups, referred to in this study as "periods."

Period 1 (1589-1595)	Period 2 (1595-1599)	Period 3 (1599-1605)	Period 4 (1605-1613)
Taming of the Shrew (1593-4)	Richard II (1595)	As You Like It (1599)	King Lear (1605)
Two Gentlemen of Verona (1594)	King John (1594-6)	Hamlet (1600-1)	Macbeth (1606)
3 Henry VI (1590-1)	Sir Thomas More (1594-5)	Twelfth Night (1601-2)	Antony and Cleopatra (1606-7)
2 Henry VI (1590-1)	Merchant of Venice (1596-7)	Troilus and Cressida (1601-2)	Pericles (1607-8)
Arden of Faversham (1590)	1 Henry IV (1596-7)	Measure for Measure (1602)	Coriolanus (1607-8)
Titus Andronicus (1593-4)	2 Henry IV (1598)	Othello (1604)	Timon of Athens (1607-8)
1 Henry VI (1589-90)	Merry Wives of Windsor (1597)	All's Well That Ends Well (1602-3)	Winter's Tale (1610-11)
Richard III (1592-3)	Much Ado About Nothing (1598-9)	Phoenix and Turtle (1601)	Cymbeline (1609-10)
Edward III (1592-5)	Henry V (1599)	Lover's Complaint	Tempest (1611)
Comedy of Errors (1592-4)	Julius Caesar (1599)		Two Noble Kinsmen (1613)
Love's Labor's Lost (1594-5)			Henry VIII (1612-3)
Midsummer Night's Dream (1595-6)			
Romeo and Juliet (1595-6)			
Venus and Adonis (1593)			
The Rape of Lucrece (1594)			

Figure 1 Period breakdown. Dates from Riverside.

Each period is of relatively similar length,[12] and most have a mix of comedy, tragedy and history, as far as the chronology permits. One point of note is the timeline for Period 4 and its relationship to the publication of *Sonnets*. The Thorpe quarto, entered in the Stationers Register on May 20, 1609, predates most of the works in Period 4. The sonnets could not have been written contemporaneously to *Cymbeline* or *Tempest*, but those plays and others are still included in the data to flesh out a fuller arc of Shakespeare's style and provide a valuable trendline. A sonnet that matches stylometrically to Period 4 or to a work written years after the sonnet's publication does not imply an impossibly concurrent composition, but rather a trend emerging in the

---

12   Period 1 contains 232,226 words, Period 2 contains 200,697 words, Period 3 contains 168,368 and Period 4 contains 197,157 words.

sonnet and fully realized in later writing. The ineligibility of the later plays to date the sonnets within their own timeframe does not devalue their contribution to a fuller model of Shakespeare's evolving style in which to place the sonnets, even if it may mean extrapolating from those works backward.

Another pertinent detail to mention is the reasonable arbitrariness of these four periods. In delineating these four classes, I am not making a claim that one class marks a seachange from the class before it. I do not argue that *Hamlet* stands as a marked departure in style from *Julius Caesar* and the works before it, only to be usurped by another drastic overhaul in *King Lear*. The works are clustered only in so much as they have a common chronology and display a degree of commonality of Sarrazin link-words. They are a necessary and artificial imposition of discreteness on an attribute that is necessarily continuous and fluctuating: the literary style of an author. The attributes mined from each period should be viewed as evidence of trends, and not as something inherent and fundamental to that class or period specifically.

Attributes

With a pure and acute text corpus assembled, the next step is to abstract the text away from language and into numerical data based upon relevant features. The aim is to collect a wide array of numerical data that shrewdly captures the varying styles of each period and among the sonnets so that it may then be manipulated, analyzed and interacted with as any other kind of data, making the task of classifying a sonnet no different from the task of classifying a genome from genetic data, or customer habits from consumer data.

The attributes used for this investigation fall into three categories, all explored and

explained in detail in this section. The first category includes universal textual features inspired by Forsyth and Holmes's benchmark suite[13] of simplistic and general attributes involving letters, words and n-grams, the conjoining of two or more word tokens. The second category is comprised of Shakespeare-specific attributes identified by Elliott and Valenza in their comprehensive 2004 study,[14] Wells and Taylor in their Shakespearean reference[15] and others as variable attributes with chronological significance. The final category deals with rare and distinctive words, similar to the data used in the Jackson and Hieatt studies. These three categories of data combine to help illustrate a multidimensional and dynamic stylistic development across Shakespeare's career, numerically tracing his literary evolution.

Universal Text Attributes

The Forsyth and Holmes set of universal attributes derives from the perceived gap between the subjective and personal experience of literature and objective statistics inherent to the duality of stylometry. Almost every notable exercise in stylometry has been underpinned by a scholar with a preconceived set of assumptions about the text and predetermined attributes to test. Mosteller and Wallace, arguably the most noteworthy stylometrists, began their work on attributing *The Federalist Papers*[16] with a strong knowledge base of the specific features that would distinguish Madison from Hamilton, largely using the statistics to validate their previously-held suspicions. This property of stylometric study is troubling in that it introduces human bias and limits the

---

13 Richard S. Forsyth and David I. Holmes, "Feature-Finding for Text Classification," *Literary and Linguistic Computing* 11 no.4 (1996) 165-9.
14 Ward E. Y. Elliott and Robert J. Valenza, "Oxford by the Numbers: What Are The Odds The The Earl of Oxford Could Have Written Shakespeare's Poems and Plays?" *Tennessee Law Review* 72 (Fall 2004) 451-60.
15 Stanley Wells and Gary Taylor with John Jowett and William Montgomery, *William Shakespeare: A Textual Companion* (Oxford: Clarendon Press, 1987).
16 Federick Mosteller and David Lee Wallace, *Applied Bayesian and Classical Inference: The Case of the Federalist Papers.* (New York: Springer-Verlag, 1984).

reproducibility of the experiment. The subjective element comes close to susceptibility to confirmation bias, but can also lead to, as Forsyth and Holmes point out, "a 'tool-kit' of favorite marker types which encompasses only a small fraction of those that might be used" by a given scholar (163). The limitation of reproducibility slows the growth of the field by limiting the significance of discoveries and breakthroughs to the specific work to which they pertain. For example, a successful statistical investigation into the chronology of Plato's writing of dating certain works against the philosopher's larger corpus[17] would seem a likely candidate for emulation in answering the question of Shakespearean sonnet chronology, but the textual features which underpinned the analysis—the syllabic distribution of the clausula, a rhetorical feature unique to Roman oratory—are inept when applied outside of their domain.

The following combination of Forsyth and Holmes's benchmark suite of universal features and other tests are broad and general enough to be replicated across author, genre and language, yet robust enough to successfully distinguish and classify texts even in nuanced situations like chronology of a single author.[18] They have the advantage of the lack of inflection by subjective, human judgment, and, if effective, can be readily reproduced and applied to a similar problem.

To evaluate each feature, I provide information on the Euclidean distance and cosine similarity of each period to the others for that feature. The Euclidean distance is a metric that illustrates the distance between the two items if evaluated as points in Euclidean space. For data with only two dimensions, which can be graphically represented as x and y values, then the Euclidean distance is the Pythagorean theorem:

---

17 D. R. Cox and L. Brandwood, "On a Discriminatory Problem Connected with the Works of Plato," *Journal of the Royal Statistical Society* 21 (1959).

18 Forsyth and Holmes apply their set of features to distinguish early from late poems in the works of Emily Dickinson, John Pudney and W. B. Yeats. Their tests reported success across the board.

$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$, or, the length of the line drawn between the two points. This

principle of distance between two data points preserves in higher dimensions as well,

although it cannot be visually represented past the third dimension. For multidimensional

data, such as letter frequency, where each letter and its frequency represents a dimension,

the Euclidean distance is $\sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$. So, for the case of letter frequencies, *P* and *Q*

would both be periods, *n* would be 26, and q_i and p_i would be the frequencies found

in *P* and *Q*, respectively, for the *i*th letter of the alphabet. The higher the Euclidean

distance between two points, the greater the difference between the two data objects.

Cosine similarity provides another metric to compare two data objects. In contrast

to Euclidean distance, cosine similarity treats a dataset as a vector, not a point. The

difference measured is therefore not their distance from one another in space, but the

cosine of the angle between the two when represented as vectors. Cosine similarity takes

values from -1 to 1, where -1 would be the cosine of two diametrically opposed vectors, 0

would be the cosine of two vectors at 90°, and 1 would be the cosine of a vector with

itself. The closer to 1 the cosine similarity becomes, the more alike the two data objects.

I also provide test data on cross-validation classification for each period in an

attribute. I use a naive Bayes Classifier and test blocks of 4,378 words. I will expand

further on classification methodology in another section, but this metric essentially

demonstrates how well each period can classify its own members based on the given

attribute.

Letter Frequency

Measuring the style of a text by the frequency at which each individual letter

occurs seems simple to the point of being simplistic; however, Holmes and Forsyth cite

several previous studies[19] which employ letter frequency as a competent stylistic

indicator. At the very least, it provides a baseline for complexity. Being arguably the

simplest possible quantitative measurement of a piece of text, it supplies a threshold that

other, more sophisticated textual features must exceed in order to justify their additional

complexity.

In implementation, letters from 'a' to 'z' are counted as well as apostrophes and

dashes to accommodate words like "speak'st" and "life-blood". This test is case

insensitive–that is, it adds both the capitalized and lowercased instances of a letter to a

single count–and all words are evaluated, including proper nouns.

Euclidean Distance	Period 1	Period 2	Period 3	Period 4
Period 1	0.000	0.575	0.760	0.685
Period 2	0.575	0.000	0.709	0.899
Period 3	0.760	0.709	0.000	0.671
Period 4	0.685	0.899	0.671	0.000

Cosine Similarity	Period 1	Period 2	Period 3	Period 4
Period 1	1.00000	0.99974	0.99954	0.99963
Period 2	0.99974	1.00000	0.99961	0.99936
Period 3	0.99954	0.99961	1.00000	0.99964
Period 4	0.99963	0.99936	0.99964	1.00000

Bayes Classifier	Period 1	Period 2	Period 3	Period 4
Success %	48.6	70.4	72.9	66.7

Digrams

Increasing the complexity of letter frequency by one increment arrives at letter

digrams, or, the frequency of two concatenated characters in a given text, such as 'th',

---

19 See G. R. Ledger and T. V. N. Merriam, "Shakespeare, Fletcher, and the Two Noble Kinsmen,"
    Literary and Linguistic Computing 9 (1994): 235-248; L. Ule, "Recent Progress in Computer Methods
    of Authorship Determination," *ALLC Bulletin* 10 (1982): 73-89; G. R. Ledger, *Re-Counting Plato*
    (Oxford: Oxford UP, 1989).

'ea', 't ' and so on. This test evaluates the data in a manner similar to the test for letter frequency, but one significant departure is the inclusion of whitespace as a character. Only the 96 most frequent digrams are considered, which is a metric specific to each period. In order to evaluate Euclidean distance and cosine similarity, a common set of digrams is used, derived from the superset comprised of the 98 most frequent digrams from each period. This superset contains 105 digrams. Also displayed are the ten most common digrams in found in the corpus as a whole, with whitespace represented by an underscore.

Euclidean Distance	Period 1	Period 2	Period 3	Period 4
Period 1	0.000	0.004	0.006	0.006
Period 2	0.004	0.000	0.005	0.006
Period 3	0.006	0.005	0.000	0.005
Period 4	0.006	0.006	0.005	0.000

Cosine Similarity	Period 1	Period 2	Period 3	Period 4
Period 1	1.00000	0.99904	0.99813	0.99836
Period 2	0.99904	1.00000	0.99867	0.99815
Period 3	0.99813	0.99867	1.00000	0.99859
Period 4	0.99836	0.99815	0.99859	1.00000

Bayes Classifier	Period 1	Period 2	Period 3	Period 4
Success %	75.2	53.6	80.3	78.9

Overall Most Common Digrams
e_
_t
th
s_
t_
he
d_
_a
_s

Word Frequency

The next order of complexity, word frequency, is structured similarly to the digram test. Each word token is lightly lemmatized using the Python Natural Language Toolkit Wordnet Lemmatizer and Part of Speech Tagger.[20] In practice, that means genitives are converted to their noun root ("sister's" to "sister"), as are plurals ("children" to "child"), and verbs are lemmatized to their verb root ("told" to "tell"). Gerunds and participles are not lemmatized to a verb root. The test is case insensitive and includes

---

20 Steven Bird, Edward Loper and Ewan Klein, *Natural Language Processing with Python*. O'Reilly Media Inc. (2009).

proper nouns, although no proper nouns appear in the 96 most frequent words for any

period. The superset of the 96 most frequent words from each period includes 107 words.

Euclidean Distance	Period 1	Period 2	Period 3	Period 4		Overall Most Common Words
Period 1	0.000	0.008	0.011	0.011		the
Period 2	0.008	0.000	0.008	0.010		and
Period 3	0.011	0.008	0.000	0.009		i
Period 4	0.011	0.010	0.009	0.000		to
						of
Cosine Similarity	Period 1	Period 2	Period 3	Period 4		a
Period 1	1.00000	0.99494	0.99159	0.99164		you
Period 2	0.99494	1.00000	0.99465	0.99298		my
Period 3	0.99159	0.99465	1.00000	0.99427		that
Period 4	0.99164	0.99298	0.99427	1.00000		in
Bayes Classifier	Period 1	Period 2	Period 3	Period 4		
Success %	65.3	68.2	53.7	80		

Word Length

This test measures the frequency of words at different lengths two ways,

measured by letter count and by syllable count. Letter count tests for all words between

one and eighteen letters long. Word longer than eighteen letters, such as

"honorificabilitudinitatibus" (27 letters) from *Love's Labor's Lost* or, the longest,

"tragical-comical-historical-pastoral" (36 letters) from *Hamlet*, are discarded. Besides

being anomalous, both of these super-long words are employed for a specific comic

effect, so they hold no far-reaching stylistic implications.

Letter Length

Euclidean Distance	Period 1	Period 2	Period 3	Period 4
Period 1	0.000	0.810	1.125	0.816
Period 2	0.810	0.000	0.765	1.145
Period 3	1.125	0.765	0.000	1.490
Period 4	0.816	1.145	1.490	0.000

Cosine Similarity	Period 1	Period 2	Period 3	Period 4
Period 1	1.00000	0.99980	0.99961	0.99980
Period 2	0.99980	1.00000	0.99982	0.99960
Period 3	0.99961	0.99982	1.00000	0.99932
Period 4	0.99980	0.99960	0.99932	1.00000

27

Bayes Classifier	Period 1	Period 2	Period 3	Period 4
Success %	19.9	31.8	39.5	47.5

The test for syllable count accounts for words comprised of between one and six syllables. The only non-hypenated words in the corpus that exceed this upper bound are "honorificabilitudinitatibus" (13 syllables) and "anthropophaginian" (7 syllables) from *Merry Wives of Windsor*. Both of these words fall into the same criteria as the previous super-long words, so their omission is accepted.

Syllable Length

Euclidean Distance	Period 1	Period 2	Period 3	Period 4
Period 1	0.000	0.051	0.417	0.250
Period 2	0.051	0.000	0.450	0.262
Period 3	0.417	0.450	0.000	0.555
Period 4	0.250	0.262	0.555	0.000

Cosine Similarity	Period 1	Period 2	Period 3	Period 4
Period 1	1.00000000	0.99999980	0.99998677	0.99999779
Period 2	0.99999980	1.00000000	0.99998455	0.99999741
Period 3	0.99998677	0.99998455	1.00000000	0.99997935
Period 4	0.99999779	0.99999741	0.99997935	1.00000000

Bayes Classifier	Period 1	Period 2	Period 3	Period 4
Success %	23.3	15.7	29.5	40.3

N-Gram Word Chunks

Letters, digrams and words offer straightforward procedures to measure and quantify language from a computational perspective, but they diminish in relevance when viewed in a linguistic context. Certainly no author broods over the frequency of the letter 'e' while writing, and while that frequency may change as authorial style changes, it is almost unquestionably a symptom of a different stylistic development than a true indicator in itself. Even word use, Antonia, Craig and Elliott point out, plays a peripheral

role when considering style from a linguistic perspective. Their study affirms the "idiom principle" that, "language is characteristically produced in the brain in chunks, not in individual words."[21] One way to quantify these chunks is through word n-grams, where n represents the number of words in a chunk. The Antonia, Craig and Elliott study evaluates 1- to 5-grams–that is, word chunks of one to five consecutive words– concluding that "no one style of n-gram outshines the others,"[22] and that all can be valuable textual indicators in certain contexts.

Appendix A displays n-grams shared between the sonnets and themselves and between the sonnets and the corpus. Excluded is the 19-gram couplet "But do not so I love thee in such sort / As thou being mine mine is thy good report" repeated in sonnets 36 and 96. Also not tabulated are 3-grams and 2-grams, due to their quantity.[23] The n-grams are constructed in a hierarchy such that the n-grams with a greater n do not trickle down into the lesser n n-gram counts as well. For example, the 8-gram "lilies that fester smell far worse than weeds" (found in Sonnet 94 and *Edward III*) does not also count for two 7-grams of "lilies that fester smell far worse than" and "that fester smell far worse than weeds," or three 6-grams, four 5-grams and so on.

Appendix A also includes a visualization of shared n-grams across the sonnet sequence by plays in each period. The final visualization shows those shared n-grams in each period as relative to the size of those periods in words.

Shakespearean Attributes

Often, the most distinctive and successful stylometric features hinge upon

---

21 Alexis Antonia, Hugh Craig and Jack Elliott, "Language chunking, data sparseness, and the value of a long marker list: explorations with word n-grams and authorial attribution," *Literary and Linguistic Computing* 29 no.2 (2014) 147.
22 *Ibid*. 159.
23 The quantities of n-grams from sonnets are: 1 8-gram, 1 6-gram, 24 5-grams, 244 4-grams, 2,415 3-grams and 7,156 2-grams.

idiosyncrasies particular to the author in the case study. The most notable success of the field of stylometry, the attribution of the *Federalist Papers*, began with a tip to the statisticians Mosteller and Wallace that "the words *while* and *whilst* distinguished Hamilton from Madison quite well."[24] From there they investigated other prepositions, and stumbling on *upon*, a word used 18 times more frequently by Hamilton than Madison, the case unfolded quickly and plainly. These kinds of specific and particular authorial characteristics, while often hidden or peculiar, are also often the most successful and meaningful attributes in practice.

Fortunately, there is precedented scholarship in the evolving stylistic attributes of Shakespeare's writing. Elliott and Valenza in their expansive study stylistically discrediting the Oxfordian hypothesis of Shakespeare authorship collect and coalesce a wide array of internal chronological indicators for Shakespeare's plays.[25] They show beyond reasonable doubt that these metrical and stylistic indicators developed markedly over Shakespeare's career, making them an indispensable asset to any effort in stylometric studies in Shakespeare chronology.

All of these attributes will be illustrated and discussed here, and the corresponding data is found in Appendix B, including tabulations of these attributes in the corpus plays as well as in the individual sonnets.

Midline Speech Endings

Some of these chronological indicators, unfortunately, are less relevant to the question of sonnets than to the question of plays. For example, Midline Speech Endings– verse lines ended at or near five syllables–present "the best stylistic sequencer for the plays" according to Elliott and Valenza's studies, but provide little help for the sonnets,

---

24 Frederick Mosteller and David Wallace, "Inference in an Authorship Problem," *Journal of the American Statistical Association* 58 no. 302 (1963): 277.
25 *See supra* note 13 Appendix Ten: 451.

which include neither speech nor midline endings due to the sonnet form (451). The

nearest approximation of this metric for the sonnets is midline terminal punctuation, or,

the number of lines that contain a period, semicolon or other terminal punctuation mark

not at the end of the line. Take, for example, the first quatrain of Sonnet 104:

> *To me, fair friend, you never can be old,*
> *For as you were when first your eye I ey'd,*
> *Such seems your beauty still. Three winters cold*
> *Have from the forests shook three summers' pride,*

The third line of the quatrain, while not identical in semantics or metrics to a midline

speech ending, retains the sense of ending the line prematurely, then beginning another

after the period. This metric of Midline Terminal Punctuation, tabulated in the sonnets in

Appendix B, provides an imperfect but functional substitute for the best Shakespearean

indicator of stylistic chronology, the Midline Speech Ending.

Feminine Endings

Another important yet problematic attribute is the feminine ending. Halliday's

Shakespeare Companion (one touchstone and source of Elliott and Valenza's data) defines

feminine endings as "an extra or 'redundant' syllable at the end of a line," and mentions

that "for establishing chronology [it] is one of the most helpful of the verse-tests."[26]

However, any thought of feminine endings and the sonnets immediately recalls Sonnet

20:

> *A woman's face with Nature's own hand painted*
> *Hast thou, the master mistress of my passion;*
> *A woman's gentle heart but not acquainted*
> *With shifting change as is false women's fashion;*
> *An eye more bright than theirs, less false in rolling,*
> *Gilding the object whereupon it gazeth;*
> *A man in hue all hues in his controlling,*
> *Which steals men's eyes and women's souls amazeth.*
> *And for a woman wert thou first created,*

---

26   F. E. Halliday, *A Shakespeare Companion* (London: Gerald Duckworth & Co. 1968), 164.

> *Till Nature as she wrought thee fell a-doting,*
> *And by addition me of thee defeated,*
> *By adding one thing to my purpose nothing.*
> > *But since she prick'd thee out for women's pleasure,*
> > *Mine be thy love, and thy love's use their treasure.*

In this sonnet, every line contains one extra syllable and ends on a feminine rhyme. The form is mimetic of its aim, which Helen Vendler summarizes as "an explanatory myth to account for the young man's startling simultaneous possession of a man's penis and woman's face."[27] While this is a unique case in the sonnets,[28] it makes an important point. Feminine rhymes and their frequencies in Shakespeare are not an entirely subconscious development, nor merely a development of changing stylistic taste. Sonnet 20 demonstrates a proficiency of structural mimesis through feminine rhyme, and forebodes the potential context-sensitivity of such formalistic poetic elements.

Another important aspect of working with the quantification of feminine endings to acknowledge is the inescapable ambiguity. Metrical analysis relies heavily upon the reader's own voice and perception, and can require syllabically breaking down words that, by design, can have multiple valid pronunciations. "Spirit" and "hours," for example, can be monosyllabic or disyllabic depending on pronunciation and, in *Sonnets,* are both found as disyllables and monosyllables, depending upon the context. Appendix B shows that, while not extreme, there is pronounced variation between the counts of feminine endings in Shakespearean verse of different metrical scholars, due largely to this inherent ambiguity.

With these caveats in mind, feminine endings have nonetheless proven to be a valuable chronological indicator for Shakespeare's verse, evident in the corpus data in Appendix B. Early plays can be seen containing about 10-20 feminine endings, while

---

27 Helen Vendler, *The Art of Shakespeare's Sonnets* (Cambridge, MA: Harvard UP, 1997): 128.
28 With the possible exception of Sonnet 87 (12 feminine endings, 2 masculine endings).

later plays come closer to 20-30.

Open Lines

Some of the attributes identified by Elliott and Valenza and others do translate

well to the question of dating the sonnets. One of these is Open Lines, also called Run-on

Lines or Enjambment, which Halliday defines as "verse in which the sense runs on, or

flows over, from one line into the next." Open lines were seldom used in verse before

Shakespeare's time, but their prevalence steadily increased during his career, which

reflected in an increase in his own verse from about 10 percent use in the early plays up

to 40 percent in the later.[29]

Open lines, like feminine endings, are inherently subjective and ambiguous. The

presence or absence of punctuation at the end of a line can be one helpful indicator of

enjambment, but does not necessarily imply the sense of one line "flowing over" into the

next. We can see one example of this in the fourth quatrain of Sonnet 54 (ed. Kerrigan):

> *But for their virtue only is their show*
> *They live unwooed and unrespected fade,*
> *Die to themselves. Sweet roses do not so;*
> *Of their sweet deaths are sweetest odours made.*

The first line of this quatrain does not semantically flow into the second, and could

function grammatically (perhaps even function better) with a full stop after "show". The

second line, however, *does* flow semantically into the third, as "They" in line two is the

subject of the verb "die" in line three, yet there is punctuation at the end of line two,

before beginning line three. To make matters worse, punctuation of the sonnets varies

between editors, introducing another layer of variability to any scholar not using a

facsimile reference text. And for a scholar using the facsimile text, the burden of editorial

trust is only removed from modern editors to be placed upon the Elizabethan compositor,

---

29   *See supra* note 24: 427.

who Jackson has shown were in fact likely two different compositors, neither of them consistent nor often correct.[30]

In an attempt to minimize this variability and subjectivity, Appendix B tabulates two metrics for the sonnets: Open Lines and Unpunctuated Line Endings. Unpunctuated Line Endings counts all lines without any ending punctuation, excluding quotation marks and parentheses, and Open Lines counts only those lines that I judge to grammatically or semantically depend upon the line following to form a complete thought. Open Lines are a strict subset of the Unpunctuated Line Endings, thus, there are no lines with any ending punctuation mark that counts as an Open Line.

Light and Weak Endings

Other metrics derived from Halliday are light and weak endings. Light endings Halliday defines as lines ending with lightly stressed monosyllables, such as pronouns or auxiliaries, but excluding conjunctions and prepositions. Halliday terms endings with conjunctions and prepositions weak endings. Both of these metrics Halliday defines as subsets of the Open Lines. Light endings and weak endings both increase markedly over Shakespeare's career, although there are very few present in the Sonnets.

Colloquialism and Archaism

Wells and Taylor in their exhaustive Shakespeare companion identify what they call "colloquialisms," a set of 25 short words and inflectional endings that exhibit pronounced change in frequency over the corpus. To make this data more accessible, Elliott and Valenza split the Taylor colloquialisms into two categories: "colloquialisms," which constitute the 21 features that increase in frequency over the corpus, and

---

30 MacDonald P Jackson, "Punctuation of the Compositors of Shakespeare's *Sonnets* 1609," *The Library* 30 no. 1 (1975) 14-6.

"archaisms," the other 4 features with decreasing frequencies.[31]

The colloquialisms are largely more conversational and informal rhetorical particles. For example, the participial inflection "'d" is practically non existent in the first half of Shakespeare's career, with only four instances recorded by Taylor in verse from all works up to and including *Hamlet*. In the later works "'d" begins to be more used, peaking with 12 instances in *Winter's Tale*. Conversely, the archaism "-ed" declines throughout Shakespeare's career, from 89 recorded instances in verse from *1 Henry VI* to 10 in *Winter's Tale*. Of the 21 Wells-Taylor colloquialisms, only seven appear in *Sonnets*, displayed graphically in Appendix B. All archaisms are present in *Sonnets*.

Shared Rhyme

Appendix C catalogues all 367 rhyme pairs that the sonnets share with other sonnets, other texts from the corpus, or both. The triplicate rhyme ("chide," "pride," "dy'd") in Sonnet 99 from the additional line in the quatrain is treated instead as three rhyme pairs. Also appended are visualizations of the distribution of all rhyme pairs from a period by play, as well as the total and relative frequencies between periods.

External Citations

Internal, aesthetic evidence can often make the most compelling parallels to draw between plays and the sonnets, and have often been used before to make arguments for compositional chronology. It is difficult to deny, when confronted with the aforementioned 8-gram "lilies that fester smell far worse than weeds," found in both Sonnet 94 and *Edward III*, that the sonnet and the play have an important connection to

---

31 Colloquialisms include *'t, i'th', o'th', th', 'em, 'll, 'rt, 're, 'd/'ld, 'lt/'t, 'st/'ve, I'm, 'as, this', 'a'/ha', a', o', 's (us, his), 's (is), has, does*. Archaisms include *-eth, -ion, -ed* and "obsolescent pronunciations of unaccented syllables for metrical purposes." Wells and Taylor: 102.

one another and, possibly, that they may have been written not long after one another. The problem with this kind of imagistic evidence is that most of it, in the absence of verbatim quotation, relies upon evoking a scene, emotion or a phrase entirely and exclusively within the reader. It is subjective, unquantifiable and extremely prone to confirmation bias.

Appendix D offers one solution to this problem, prompted by MacDonald P. Jackson's successful look into repeated imagery in answering the dating question of *A Lover's Complaint*.[32] Adapting his methodology to the Sonnets, I consult three critically esteemed editions of *Sonnets,* and tabulate all of the citations in each sonnet to any of Shakespeare's plays or other poetic works.[33] The editors Booth, Burrow and Evans all have no stake in the compositional chronology of the *Sonnets,* bringing no bias to their readings and citations. They additionally have different editorial styles and instincts, and complement one another in their differing strengths. The Booth and Burrow citations derive from hand-counts, while the Evans derives from machine counts.

Rare and Distinctive Words

Vocabulary changes. New words may enter the lexicon of a writer as easily as old words may leave it, and finding these entrances and exits is the key to solving a question of chronological proximity. This hypothesis underpinned Gregor Sarrazin's work[34] with Shakespeare's texts in 1897 as he meticulously sought out words in the canon used two to three times (dislogmena and trislogmena). Sarrazin found that these words, which he deemed "link-words," clustered pronouncedly near one another in the plays. His work put

---

32 MacDonald P. Jacckson, *Determining the Shakespeare Canon: Arden of Faversham and A Lover's Complaint* (Oxford: Oxford UP, 2014) 65.
33 The three editions are: Ed. Stephen Booth, *Shakespeare's Sonnets* (New Haven; London: Yale University Press, 1977); Ed. Colin Burrow, *William Shakespeare: The Complete Sonnets and Poems* (Oxford: Oxford UP, 2002); Ed. G. Blakemore Evans, *The Sonnets* (Cambridge: Cambridge UP, 1996).
34 Sarrazin, *supra* note 10.

forth this observation, as well as the hope that they could be used for future chronological

studies. Eliot Slater was one scholar who notably upheld this hope, and expanded in

scope by searching for words from the canon used up to ten times in his systematic

attribution study of *Edward III*.[35] Since then, this thread of rare-word analysis has been

dominant in chronological studies of the sonnets. The Hieatt, Hieatt and Prescott study[36]

staked its claims on the grounds of "late rare words"— that is, those words which appear

in *Sonnets* and also appear in three or more works from *Hamlet* onward, but never in the

plays before it— and their proportion to "all rare words," or, those words in *Sonnets*

which are used in three to nine additional works exclusively, regardless of their position

relative to *Hamlet*. The study then compares the proportion of late rare words to all rare

words in *Sonnets* to that same proportion in four other, more confidently dated, control

works. Jackson's response[37] builds upon the Hieatt study by fleshing out a class of

"middle rare words" and complementing the data with an analysis of the sonnets' Sarrazin

"links," twice- and thrice-used words from the corpus.

      Because the Hieatt rare-word paradigm is particularly oriented for a bifurcated

corpus, it does not suit this study well. Sarrazin links are chosen instead for emulation.

His model is reproduced in Appendix E, which shows trislogmena and dislogmena from

the corpus that appear in *Sonnets*. The methodology is identical, but the data is not. This

can partially be attributed to editorial differences. For instance, Sarrazin counts "new-

risen" as a dislogmena, but its two instances in the Riverside appear as "new risen,"

neither of which are rare words. His base text also evaluates some scenes of disputed

authorship which this study excludes. One more divergence comes from errors

_____

35 Eliot Slater, *The Problem of the Reign of King Edward III: A Statistical Approach* (Cambridge: Cambridge UP, 1988).
36 Hieatt, Hieatt, Prescott, *supra* note 8.
37 Jackson, *supra* 9.

accumulated in his manual counts.

Finally, a difference in the level of lemmatization contributes to a divergence between the datasets. The algorithm used here lemmatizes genitives and plural nouns, but no verbs. This is largely due to the tools available. The Python Natural Language Toolkit's Part-of-Speech Tagger and WordNet Lemmatizer decrease steeply in accuracy when confronted with Elizabethan verb conjugations. Additionally, the Wells-Taylor data shows that Shakespeare's inflection of participles ("'d" or "-ed") as well as his conjugation of third-person singlular verbs ("-eth") do carry chronological importance. Lemmatizing verbs purges the data of these pieces (perhaps among others) of chronological information.

For reference, Figure 2 shows the result of running the lemmatizer on the section of *1 Henry VI* used. Out of the 2,699 total tokens, 109 were altered by lemmatization. Of these 109, three ("alas", "dies" and "whiles") are lemmatized incorrectly, producing an accuracy rate of 97.2%.

This dataset and Sarrazin's, however, are both self-consistent and extensive. The principle behind the data remains the same, even if the real words in the datasets diverge.

lords lord	god's god	wings wing	friends friend	limbs limb
gentlemen gentleman	yeomen yeoman	england's england	tours tour	stars star
means mean	place's place	curs cur	feeds feed	talbot's talbot
hawks hawk	words word	stags stag	commanders commander	flies fly
flies fly	king's king	hounds hound	whiles while	hopes hope
dogs dog	days day	heads head	lives life	mother's mother
blades blade	colors color	cowards coward	honors honor	mother's mother
bears bear	foes foe	friends friend	lands land	father's father
horses horse	friends friend	england's england	forces force	followers follower
girls girl	souls soul	colors color	cries cry	
blindman's blindman	gentlemen gentleman	scouts scout	legions legion	
thoughts thought	gates gate	espials espial	whiles while	
colors color	calls call	troops troop	drops drop	
lords lord	arms arm	horsemen horseman	limbs limb	
gentlemen gentleman	subjects subject	wars war	looks look	
roses rose	attendants attendant	england's england	england's england	
books book	towers tower	cornets cornet	succors succor	
cheeks cheek	nation's nation	talbot's talbot	lies lie	
roses rose	stands stand	traitors traitor	dies dy	
cheeks cheek	snares snare	smiles smile	horsemen horseman	
roses rose	squadrons squadron	talbot's talbot	hours hour	
whiles while	meets meet	hours hour	comes come	
friends friend	eyes eye	years year	lives life	
roses rose	dolphin's dolphin	lives life	stratagems stratagem	
scorns scorn	horsemen horseman	alas ala	talbot's talbot	

Figure 2: Lemmatized Words in *1 Henry VI*

Distinctive Words

Although the trends identified by both the Hieatt and Jackson studies seem

promising, it is difficult to escape the troublesome nature of the limited data, and harder

still to confidently offer conclusions. The Hieatt study identifies 23 late rare words in the

sonnets. Jackson's efforts cull 153 Sarrazin link-words. For the sonnets, totaling 17,520

words,[38] it is extremely problematic to stake a claim on such limited observed data. 153

words, approximately .8 percent of the whole of the sonnets, is a very thin thread from

which to hang conclusions. Being so few in number makes the words highly susceptible

to anomaly, and conclusions derived from them should draw skepticism from a critical

---

38 Spevack's count.

observer.

One other vexatious property of the rare-words analysis is the degree of scholarly arbitrariness involved in outlining the attributes. Sarrazin decided that words used twice and thrice were worth his attention, and Slater likewise limits his focus to words used ten times or fewer. To his credit, Slater does demonstrate that "there is a trend from the rarest words to the commonest" that the rare words show "the greatest tendency to cluster".[39] While this is easy to accept as the general case, it does not account for anomalous words that, while not rare, do demonstrate clustering behavior. "Vile" appears 54 times in the first half of Shakespeare's career, and 14 times in the later half. "Chaste" occurs 9 times before *Hamlet,* and 24 times after it. Even extremely common words can display remarkable clustering, like "something," a word with a frequency that more than doubles in the later half of Shakespeare's career.

To extract these non-rare words that nevertheless display a clustering behavior throughout the corpus, I applied a null hypothesis that assumed all words are evenly distributed throughout all plays and periods. From each period, I then extracted the words with observed frequencies in that period that exceeded a X^2 statistical significance of 0.05. The proportion of statistically significant words in each period proves substantial,[40] which does not come as a surprise. An even distribution of all of Shakespeare's words across his career is a fallacious hypothesis that disregards context, genre and plot. "Caliban" will of course cluster in Period 4 due to *The Tempest* just as "ides" will cluster in Period 2 due to *Julius Caesar*. While they do indicate things about his reading, topics and plots, these kinds of clusters, extremely context-sensitive and specific, elucidate little

---

39 *Supra* note 33. 88.
40 Period 1 has 1,236 significant words out of 10,120 words total, Period 2 has 802 significant words out of 12,108 words total, Period 3 has 931 significant words out of 10,513 words total, and Period 4 has 1,211 significant words out of 10,893 words total.

about Shakespeare's style broadly conceived.

　　To expunge these context-sensitive words, one must construct a new set of words. This set consists of the union of the four sets of period-significant words (4,097 words total) intersected with the set of all words in the sonnets. That is, the set of all words that are statistically significant to at least one period *and* that appear at least once in the sonnets. *Sonnets* contains no proper nouns other than a handful of figures from Roman antiquity, and more generally tends to avoid specificity in its language. Even the Fair Youth, Rival Poet and Dark Lady, the dominating figures of the sonnet sequence, have only been named so by readers. No name, title or epithet for them appears anywhere in the sonnets. The resulting set is therefore both pertinent to the question of the sonnets as well as largely insulated from these context-sensitive cluster words. This set of "distinctive words" evaluates to 800 words even.

　　The set of distinctive words contains most of the rare words identified by Hieatt, Hieatt and Prescott in *Sonnets,* but it is not a strict superset. This is more of a matter of methodology than theory. Their rare words are heavily and admirably lemmatized. For example, their model is robust enough to discover that "widow" as a verb is a late rare word, while "widow" as a noun is not. The corpus that provided the set of distinctive words is not lemmatized to a comparable level. Likewise, the words in this study are treated as a series of characters without considering any part-of-speech or definition data, making the model susceptible to noise from homographs. That is, the model has no way to distinguish the "bark" of a tree from the "bark" of a dog, and any engagement with those words will treat them identically.

　　Despite these drawbacks, a model of distinctive words has the important advantage of a broad base of relevant data upon which to draw. Each period has a

reasonably large base of distinctive words, which are fairly evenly distributed.[41]Appendix

F shows the relative use of each period's distinctive words compared to the other periods.

These sets are not mutually exclusive, so a single word may be distinctive to more than

one period if its frequencies in both of those periods are high enough to be statistically

significant.

---

41 Period 1 has 268 distinctive words, Period 2 has 162 words, Period 3 has 197 words, and Period 4 has
    228 words.

Clustering

We know that two of the Sonnets, 138 and 144, both existed at least a decade before the publication of the 1609 Quarto, and there is compelling historical evidence that Sonnet 107 was written at or soon after the ascension of James VI. We know, too, that analyses of Shakespearean vocabulary as far back as Gregor Sarrazin's[1] in 1898 have shown that many sonnets display a marked lexical similarity to plays written in the 1595-8 period. Trying to pursue this thread further, however, one soon runs into a vexatious problem. We cannot date the composition of the Sonnets as one complete work, as Sarrazin[2] or Eliot Slater tried to do, since it flies in the face of the internal and external evidence of long and irregular periods of composition. Conversely, however, we cannot date sonnets individually in the absence of external evidence, due to their short length. The Sonnets have an average length of 114 words each, making them far too small to withstand any individual stylometric interrogation. Figure 1 shows a table from Holmes and Forsyth's study[3] in textual feature-finding that lists the size of discrete text-blocks used in various prominent stylometric studies. The column labeled "Norm" shows the typical or recommended text size from that study, while the "Range" column shows the size in words of the smallest and largest blocks used. The average sonnet has fewer than half of the number of words used in the smallest text block recorded.

It becomes necessary, therefore, when working with the Sonnets stylometrically for the question of dating, to work with them in groups. The Hieatt, Hieatt and Prescott

---

1   Gregor Sarrazin, "Wortechos bei Shakespeare", *Shakespeare Jahrbuch*, 33 (1897) 119-69.
2   Eliot Slater, *The Problem of the Reign of King Edward III: A Statistical Approach* (Cambridge: Cambridge UP, 1988) 157-63.
3   Richard S. Forsyth and David I. Holmes, "Feature-Finding for Text Classification," *Literary and Linguistic Computing* 11 no.4 (1996) 165.

study[4] divides the Quarto into four Zones: Sonnets 1-60, 61-103, 104-126 and 127-154.

Table 1 Size of text blocks analysed by various stylometrists

Researchers	Subject	Norm	Range
Smith (1985)	Elizabethan Drama	20595	17965–24925
Ule (1982)	Marlowe's writings	13000	175–21106
Holmes (1992)	Mormon scriptures	10000	5715–12776
Butler (1979)	Sylvia Plath's poems	8000	6100–9340
Merriam (1989)	Federalist papers	6000	[unknown]
Burrows (1992)	Bronte sisters	4000	500–8000
Milic (1967)	Jonathan Swift	3500	3324–3777
Mosteller and Wallace (1984)	Federalist papers	2100	906–3506
Ledger (1989)	Platonic dialogues	1000	1000–1000
Matthews and Merriam (1993	Elizabethan drama	1000	[unknown]
Binongo (1994)	Nick Joaquim's short stories	1000	[unknown]
Elliott and Valenza (1991)	Elizabethan poetry	500	425–500
Thisted and Efron (1987)	Shakespeare and contemporaries	400	234–495

Figure 1 Text blocks, in words, used in various studies.
Holmes and Forsyth.

This system has merits as well as weaknesses. It accommodates discrete subsequences, such as the Dark Lady sequence (127-154) which, in its departure in tone and poetic subject from the Fair Youth sequence, suggests a different style, authorial mindset and, most importantly, period of composition.

This scheme of dividing the sonnets into to sequential groups, however, runs into many inherent problems. The Sonnets do have an internal structure thematically, and their sequencing does have consequence in the question of their interpretation, but there is little to suggest that the sequence has consequence in the question of their chronology. That is, there is nothing inherent to Sonnet 73 to suggest that it was written at the same time as Sonnet 72 and 74. Certain pairs of Sonnets are indisputably coupled, and some small groups riff on a singular theme, strongly indicating a shared period of composition,

---

4 Kent A. Hieatt, Charles W. Hieatt and Anne L. Prescott, "When Did Shakespeare Write Sonnets 1609," *Studies in Philology* 88 no.1 (1991) 90.

but, in the general case, there is nothing to chronologically link a given sonnet to its neighbors. On the contrary, there is strong evidence that Shakespeare rearranged the sonnets in the sequence. Sonnet 60, for instance, plays with the conceit of minutes, while Sonnet 12 plays with the conceit of hours between day and night. Both evoke imagery of hands on a clock. These examples among others[5] expose an interest in the numerology of the sonnet sequence that necessitates a conscious ordering and reordering of the Sonnets on Shakespeare's part. There are also, imaginably, other aesthetic motivations beyond numerology that may have inspired a reordering of the Sonnets from the original sequence of their composition, though those motivations may be less easy to expose.

This flaw in the sequential grouping of sonnet zones does not, in my judgement, entirely void the value of that system of grouping, but it does beg for an alternative. Since the 1640 Benson edition, there have been numerous attempts[6] to reorder the sequence. For many scholars it has served as a rabbit-hole of conspiracy and intrigue, indulging readings of secret narratives, hidden word links and poetic riddles. These reorderings, however, have been largely built upon ambiguous and subjective internal readings of the Sonnets, and none has earned widespread critical recognition. There has been no attempt to group the Sonnets chronologically.

In this section I offer one alternative method to group the Sonnets chronologically through clustering. By clustering, I mean that I evaluate each sonnet in terms of its chronological data as outlined in the last section (shared rhymes, metrical attributes, rare words, etc.) and group that sonnet into a cluster with the sonnets most similar to itself.

---

5   For a full numerological reading of the Sonnets, see Alastair Fowler, *Triumphal Forms: Structural Patters in Elizabethan Poetry* (Cambridge: Cambridge UP, 1970) 183-97.
6   See Denis Bray, *Original Order of Shakespeare's Sonnets* (M.S.G. House, 1925); Brent Stirling, *The Shakespeare Sonnet Order: Poems and Groups* (University of California Press, 1968) for a sample. Henry Rollins' New Variorum edition of the Sonnets (1944) tabulates twenty different reorderings.

Using a k-means clustering algorithm, this procedure produces four clusters of Sonnets for consideration and comparison to the four Zones of the Hieatt study. Clustering the Sonnets along these chronological attributes, we can see groups of which sonnets are most alike in terms of certain features of style, which we have shown to change over Shakespeare's career. Clustering offers a criteria other than proximity in the Quarto sequence to suggest contemporaneous composition between Sonnets, and one other lens with which to view the Sonnets and their chronology.

K-Means Clustering

The k-means clustering algorithm is one of the more straightforward and widely used clustering algorithms. In its simplest form, the procedure takes a set of data points and a predetermined number of clusters. It creates a "centroid" for each cluster, that is, the mean or average data point for that cluster, which may be a real data point from the sample or a manufactured ideal mean. Choosing these initial centroids can be a fraught prospect, which will be discussed shortly. After choosing initial centroids, the algorithm then loops between two steps. First, it assigns each data point in the sample to its nearest centroid. Each of those data points is now a member of that centroid's cluster. Second, it recomputes each centroid as an average of all the data points in that centroid's cluster. The algorithm continues this loop, alternating between recomputing clusters and recomputing centroids until the recomputed centroid is the same or significantly similar to its original value.

Figure 2 shows an example of the procedure with six iterations.[7] Centroids are denoted by crosses, data points are small circles and their clusters are delineated by color.

---

7   Image from Andrew Ng's Stanford CS221 slides, *K Means*.

(a) shows the original dataset in a two-dimensional space. In (b) two seemingly random centroids are chosen. (c) through (f) show the progressive averaging of the centroids into the two natural clusters.
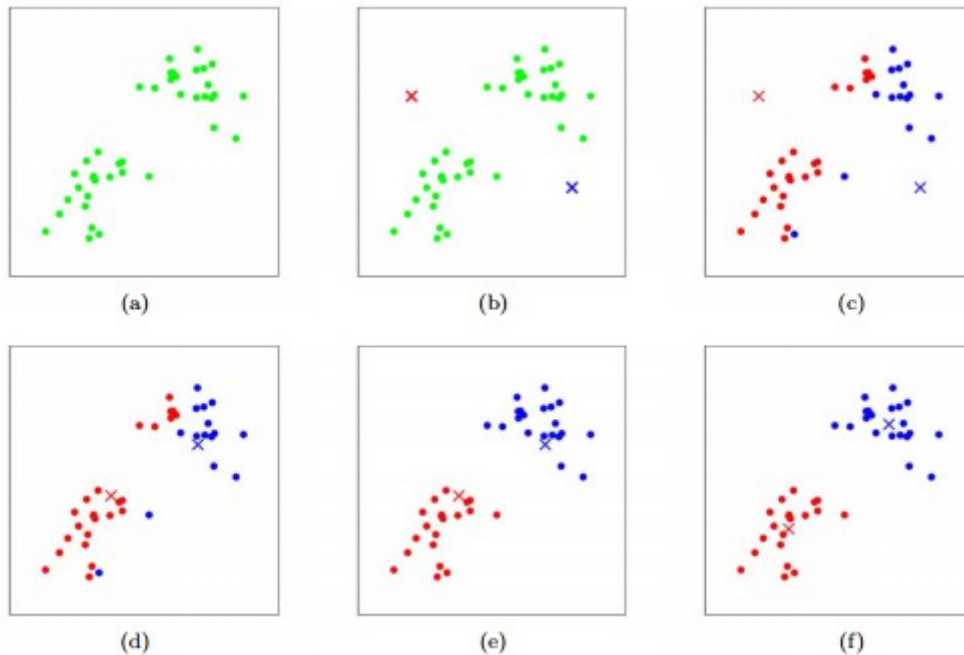
Figure 2 Six iterations of k-means algorithm.

Sonnets are represented as multidimensional data points as shown in Figure 3. A Sonnet data point includes three types of data. First, it has universal text attributes of letter frequency, frequencies of the 96 most common digrams from the overall corpus and frequencies of the 96 overall most common words. Second, it has the metrical data as shown in Appendix B. Finally, it has information about its relationship to the different play periods, including shared 4- and 5-grams, shared rhyme, citations to works to each period, number Sarrazin rare-word links (twice and thrice-used words) to works from each period, cataloged in Appendix E, and finally the number of distinctive words present from each period.

Because these data points are comprised of many different kinds of data, they

must be scaled. This preprocessing step prevents any given attribute from overpowering

the others out of proportion. For instance, if we were working with data points of people

and their proportions, we would not want the difference of one pound of weight between

two people to be proportional to the difference of one foot in height between two people.

```
Sonnet 43

Letter Frequency
a 5.97610 b 1.79283 c 1.19522 d 5.77689 e 14.14343
f 0.99602 g 2.78884 h 9.96016 i 6.37450 j 0.00000
k 1.19522 ...

96 Most Frequent Digrams
e_ 0.02652 _t 0.03276 th 0.03276 s_ 0.02340 t_ 0.01092
he 0.02652 d_ 0.00936 _a 0.00936 _s 0.02652 _i 0.01248
r_ 0.00624 ...

96 Most Frequent Words
the      0.02479 and      0.01653 i      0.02479
of       0.00000 a        0.00000 you    0.00000
in       0.03306 is       0.00000 ...

Metrics
MLE 0 FE 2 OL 2 LE 0 WE 0 COL 0 ARCH 1

Shared N-grams by Period
[1, 0, 0, 0]
Shared Rhyme by Period
[73, 17, 14, 16]
Shared Citation by Period
[4, 1, 1, 4]
Shared Sarrazin Links by Period
[1, 0, 0, 0]
Shared Distinctive Words by Period
[53, 13, 14, 11]
```

Figure 3 A Sonnet data point.

The scale function takes the data and converts it to a standardized scale, recomputing each data attribute as a function of its distance from the sample attribute mean. After scaling, the clustering algorithm can weigh the distance from a data point to its centroid equally for all attributes.

In another step of preprocessing, the data is subjected to a principal component analysis, or PCA, which transforms the data and

lowers its dimensionality. Dimensionality refers to the number of attributes evaluated, so,

for example, letter frequency in each sonnet has a dimensionality of 28, the frequencies

of the 26 letters and the frequencies of dashes and apostrophes. As dimensionality

increases, data becomes increasingly intractable. In highly dimensional spaces, the

Euclidean distance between data points becomes quickly inflated, and the difference in

distances between given pairs of data points becomes less pronounced. This problem, part

of the "curse of dimensionality", is particularly bothersome for k-means clustering, which

relies on a dependable distance function to operate properly.

The PCA used here transforms the given data into its two most principal significant single vector components. In addition to the other benefits of lower dimensional data, this two-dimensional data permits visualization. Something like letter frequency cannot be visualized in an untransformed state, as there is no way to conceptualize a 28-dimensional space. Reducing letter frequency to two dimensions, however, allows us to see a spacial representation of the data, as in Figure 4, where each point represents one sonnet, and the x and y values are the first and second principal components output by the PCA.

Figure 4 PCA of sonnet letter frequency

This study uses the clustering package from the Scikit-learn project.[8] Initial centroids are selected through the "kmeans++" routine, which chooses provably better centroids than random initialization.[9] Randomly initialized centroids are often susceptible

8   F. Pedregosa *et al.* "Scikit-learn: Machine Learning in Python" *Journal of Machine Learning Research* 12 (2011) 2825-2830.
9   David Arthur and Sergei Vassilvitskii, "kmeans++: The advantages of careful seeding" *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, Society for Industrial and Applied Mathematics (2007) 9-10.

to finding local minima of its distance function instead of the global minima of true clusters, which kmeans++ mitigates by choosing disparate starting centroids. Because it is still susceptible to finding local minima, each clustering routine is run 100 times, and the final clustering chosen is the one with minimal total inertia. Inertia is the distance function used to drive the k-means clustering, and is essentially a sum of squared Euclidean distance of each point in a cluster to its centroid. Total inertia is the average of the inertia of each cluster. The minimal total inertia, therefore, is the clustering in which the average distance from any given data point to its centroid is the smallest.

Evaluating Clusters

The sonnet clusterings, listed in Appendix G, are evaluated by two metrics: their total inertia and silhouette coefficient. Total inertia is not a normalized metric, but generally a lower score indicates more cohesive clusters. The lower the total inertia, the lower the distance between an average data point and its centroid.

The silhouette coefficient gives an indication of how well defined the clusters in a

k-means clustering are. It is computed as $s = \dfrac{b-a}{max(a,b)}$ where a is the mean distance between a data point and all other points its same cluster and b is the mean distance between a data point all other points in the nearest cluster that is not its own. A higher silhouette coefficient indicates a model with more well-defined clusters.

Results

With one exception, the Sonnets do not cluster with strong definition. Figure 5 shows visualizes the PCA of the three kinds of data used, and of the agglomeration of all data. None of them exhibit very strong visible clusters by inspection, an observation

validated by their silhouette scores.[10]
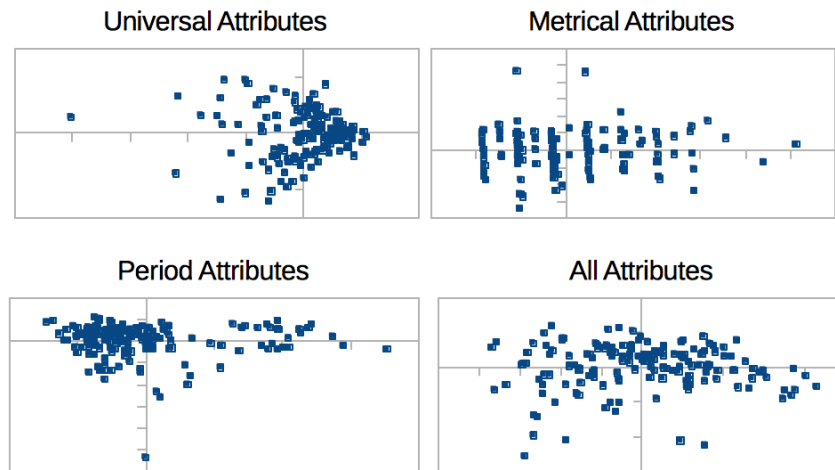
Figure 5 PCAs of Sonnet Attributes
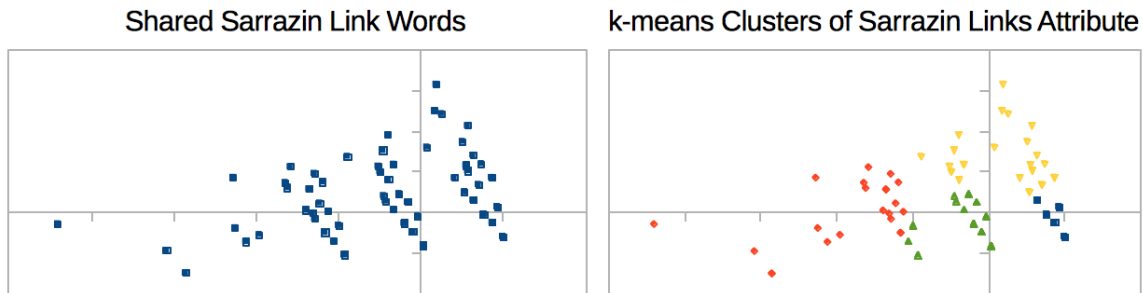
Figure 6 PCA of Sarrazin Link Attributes and k-means clustering on PCA

Only one attribute demonstrates a visible clustering. Sarrazin link-word data (represented in the sonnet data points as the number of links to each of the four periods, as, for example, Sonnet 43 in Figure 3 which has one link to Period 1 and zero links to Periods 2 through 4) shows what appear to be three to four elongated clusters in the visualization of its PCA in Figure 6. The result of the k-means clustering is also displayed, where each color indicates one cluster. The blue cluster, though it looks like the smallest, is actually largest, comprised of 54 Sonnets, some with extremely similar principal component values. K-means does not function well with irregularly shaped

_____

10   Silhouette scores, listed also in Appendix G, are: 0.3638 (Universal), 0.3546 (Metrical), 0.3803 (Period), 0.3570 (All).

clusters, which accounts for the divergence between the visible clusters and the computational clusters. Nevertheless, the k-means clustering of the Sonnets based on the Sarrazin link data scored the best of the clustering schema. It ended with a silhouette score of 0.4482 and a total inertia of 104.4367, the highest and lowest, respectively, of the various tested clustering routines.

Appendix G shows some of the results of the k-means clusterings with respect to the 1609 sequence, as well as the groupings of the Zones from the Hieatt study. The colors used have no relevance beyond distinguishing between clusters within a clustering. The green of one clustering has no significance for the green of another.

These clusterings offer one alternative to grouping the Sonnets sequentially when considering the question of chronology. A given Sonnet may have been composed contemporaneously to its neighbors, but there is evidence that the notion cannot be taken for granted. Considering the many factors aesthetic and otherwise which may have compelled Shakespeare to rearrange and reorder the Sonnets out of their chronological sequence, it makes sense to group them by inherent attributes instead.

Classification

The question of dating the Sonnets boils down to an essential question of classification. Reframing the dating question, we can view periods of composition as classes or labels, comprised of different representative samples like the various plays composed during those periods. Looking at a particular sonnet or group of sonnets, we can compare it to the different possible classes and decide which class or label suits it best. After establishing different criteria for grouping the Sonnets, we want to know their chronological significance. The groups and clusters only matters if we can find some way to date them.

There are many algorithms for classification, but one of the most prevalent in recent stylometric study has been the artificial neural network.[1] The artificial neural network models itself in a rudimentary way after the brain (a natural neural network). In the brain, learning takes place by observation and experience, which it then uses to extrapolate knowledge to unfamiliar circumstances. It is the principle that allows a person to see a duck, one never before seen, and instantly know it based on the patterns of webbed feet, feathers and quacking learned from previously seen ducks.

Artificial neural networks are likewise trained to recognize patterns in known, also called "supervised", data, and to use that information to make guesses about unknown, or "unsupervised", data. Like a natural neural network, the artificial neural network is arranged as an interconnected system of "neurons" or nodes. These nodes are then arranged into three layers: input nodes, hidden nodes and output nodes. The input

---

[1] Some examples include: Johan F. Hoorn, Stefan L. Frank, Wojtek Kowalczyk and Floor van der Ham, "Neural Network Identification of Poets using Letter Sequences" *Literary and Linguistic Computing* 14 no. 3 (1999) 311-338; F.J. Tweedie, S. Singh and D.I. Holmes "Neural Network Applications in Stylometry: The Federalist Papers" *Computers and the Humanities* 30 no. 1 (1996) 1-10; Matt Tearle, Kye Taylor and Howard Demuth "An Algorithm for Automates Authorship Attribution Using Neural Networks" *Literary and Linguistic Computing* 23 no.4 (2008) 425-42.

nodes each take one numerical input, and the output nodes each return one numerical

output. Between the input and the output is a series of interconnections between the

layers of nodes that transforms the data based on various weighted functions in each

node. Figure 1[2] provides one visualization of an artificial neural network which shows

three layers and the system interconnectivity of the nodes. If this model were used for

classification, it would classify a data point with three attributes (input layer) into two

classes (output layer).
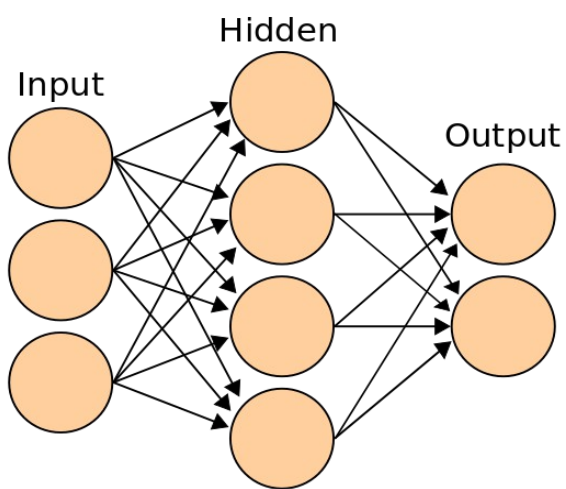
Figure 1 Artifical Neural Network

For classifying the Sonnets, I used the PyBrain Libary's[3] Feed-Forward Neural Network with backpropagation, one popular routine for weighting the transfer functions of each node during training. The neural network requires supervised training using data with known classes, as well as testing data with known classes in order to judge the effectiveness of the network. For this training and testing data I divided the

Shakespearean prose corpus (no Sonnets) into smaller text blocks, each block containing

4,377 words. This number anticipates the Sonnet groups used for classification, as 4,377

is about one fourth the word length of the entire sonnet sequence, making the test and

training blocks are similar in size to the sonnet data we will eventually classify. This

portioning out of the Shakespeare prose corpus results in 181 total text blocks: 53 from

Period 1, 45 from Period 2, 38 from Period 3, and 45 from Period 4.

The set of Shakespearean corpus blocks is then randomly partitioned 1:4 between

testing and training. Neural networks require a careful balance of testing and training.

2   "Artificial Neural Network" by Cburnett, lisenced under Creative Commons BY 3.0.
3   Tom Schaul *et al.* "PyBrain" *Journal of Machine Learning Research* 11 (2010) 743-746.

With too little training, the neural network does not learn enough about the patterns in the data and remains ineffective at classification. Overtraining of a network, in contrast, can lead to the network memorizing the training data instead of learning its general patterns, making it perform optimally with the training data but underperforming with anything else. Reserving a part of the supervised (known) data beforehand and not allowing the network to train on it provides an important opportunity to regularly judge the network's performance on data it has never before encountered.

Once the supervised data is divided between testing and training, we begin training the network on the data. The network is trained in iterations called epochs, where one epoch represents running each sample in the training set through the network once, adjusting the weight functions in the nodes, and then testing the network's accuracy on both the training and testing set. We want to choose the number of epochs that minimizes test error on the supervised testing set in order to avoid either overtraining or undertraining.

After the network is trained, we can apply it to unsupervised data, that is, data whose true class is unknown (Sonnets), and take the output as the best predicted class, with a level of confidence indicated by the test error.

Results

The first effort in classification for the Sonnets uses the Sarrazin rare word data. This is an important step in that, although it does not add greater insight into Sonnet chronology than was already offered by the Jackson or Hieatt studies, it vindicates both my process and theirs to arrive at the same conclusion by vastly different methodologies. It confirms that their studies are sound and reproducible, and affirms the effectiveness of neural-network classification for further tests. It increases confidence in the soundness of both my efforts and theirs.

The further tests use the Taylor chronological data from the Wells and Taylor Shakespeare Companion. These textual features are the speech particles and inflections that display marked change in frequency over the Shakespeare prose canon. They are summed and tabulated in Appendix B under "colloquialism" and "archaism," where the former indicates particles that increase in frequency and the later in decrease.

These two neural network classifications, one using rare-word data and the other using Taylor data, are supplemented with a look at some of the other data collected that does not well suit a neural network in this application, such as repeated imagery and shared rhyme. In this way, the problem of classifying the Sonnets is approached from many different and independent angles, and the results of any single test can be used to reinforce the results of the others.

Rare word tests

The neural network classifier using Sarrazin link word data takes four input nodes, the first being the number of links to Period 1, the second being the number of links to Period 2, and so on. A link to a period, in this sense, means a word found only in a given sonnet and in a play from that period. The output gives four probabilities of

membership in each respective period. The misclassification error of reserved testing

data minimized at 200 epochs.

Displayed below are the results of the neural network on Sarrazin data on both the

Hieatt groupings and on the k-means clusters based on Sarrazin data. These values should

be read as the probabilities of membership of a given group in a given period, where a

value of 1 is certainty of membership and a value of 0 is certainty of nonmembership.

Hieatt Zones	Period 1	Period 2	Period 3	Period 4
1-60	0.443	0.534	0.016	0.007
61-103	0.388	0.362	0.062	0.188
104-126	0.139	0.100	0.540	0.222
127-154	0.866	0.121	0.008	0.006

Sarrazin Clusters	Period 1	Period 2	Period 3	Period 4
Cluster 1	0.808	0.183	0.006	0.003
Cluster 2	0.746	0.245	0.005	0.004
Cluster 3	0.003	0.010	0.590	0.397
Cluster 4	0.068	0.068	0.321	0.544

Cluster 1
3 5 13 14 16 17 18 19 21 22 23 27 28 30 31 34 36 39 40 42 43 44 46
47 50 57 58 61 64 67 71 72 73 74 75 77 79 81 83 84 88 90 92 93 94 98
101 104 106 108 118 121 122 127 129 131 135 140 141 142 143 144
147 148 149 151 153 154

Cluster 2
7 10 12 15 20 25 29 32 33 35 37 54 66 76 78 80 82 85 87 95 96 99 109
114 119 128 138

Cluster 3
9 52 97 100 107 113 115 116 123 124 125 126

Cluster 4
1 2 4 6 8 11 24 26 38 41 45 48 49 51 53 55 56 59 60 62 63 65 68 69 70
86 89 91 102 103 105 110 111 112 117 120 130 132 133 134 136 137
139 145 146 150 152

The network ended with a training error rate of 25.7% and a testing error rate of

20.0%. While this is a wide improvement over a random classification error rate of 75%,

it still does not inspire great confidence. Looking at the behavior of the Hieatt Zones

would seem to confirm, however, some of the conclusions of their and Jackson's studies. The Dark Lady (Hieatt Zone 4) sonnets consistently scored the highest probability of being in Period 1 out of any zone or cluster in any period. Both the clusters and Hieatt Zones place Sonnet 138 (Zone 4, Cluster 2) and Sonnet 144 (Zone 4, Cluster 1) securely in the first half of Shakespeare's career—though not exactly with enough confidence to say between Periods 1 or 2—which accords with the external evidence of their existence by 1599. Likewise, Sonnet 107 (Zone 3, Cluster 3), with internal evidence tying it to the ascension of 1603, is dated solidly in the seventeenth century.

There are two complete sonnets in *Love's Labor's Lost* as well. When run through the network they return an output of [0.751, 0.126, 0.108, 0.014] placing it with high probability in Period 1, which accords with its 1594-5 composition date.

Taylor Data

After rare words, the next best chronological indicator for Shakespeare proves to be metrical data and the Taylor particle data. The first table in Appendix B, which metrically characterizes the Shakespeare corpus, shows how strong these features can be in combination. The last column of that table arrays the number of open lines added to the number of midline endings minus one-tenth of the number of Taylor archaisms. The result is a near-linear track from the beginning of Shakespeare's career to its close, with a 1 for *1 Henry 6* and 136 for *Two Noble Kinsmen.*

This metrical data, however, is less suitable to the Sonnets, which, unlike the prose or verse of the plays, are much more tightly constrained metrically. The counts of these metrical variations skew very low in the Sonnets, as tabulated and visualized in Appendix B. Curiously, the Taylor data also skews generally low compared to the prose corpus. This fact may be interpreted as an early date of composition, or, possibly, that the sonnet form and the voice of the sonneteer interferes. Many of the Taylor particles are

contractions and colloquialisms, such as *i'th* for "is the" or *'a'* for "have," which do not

entirely fit the style or tone of the Sonnets. Most of these sorts of colloquial contractions

do not even appear once in the Sonnets.

Whatever the case, a neural network trained on the Taylor data performed well

(31% error) in testing, but classified all Sonnets as members of Period 1. It performed

even better (4% error) when using two classes instead of four, that is, classifying

Shakespearean text as either either early or late. Figure 1 visualizes this phenomena

through the PCA of the Taylor data in each corpus text block. There is slight distinction

between Periods 1 and 2, but there is a comparatively defined distinction between Periods

1 and 2 combined and Periods 3 and 4 combined.

PCA of Taylor Data in Corpus Text Blocks
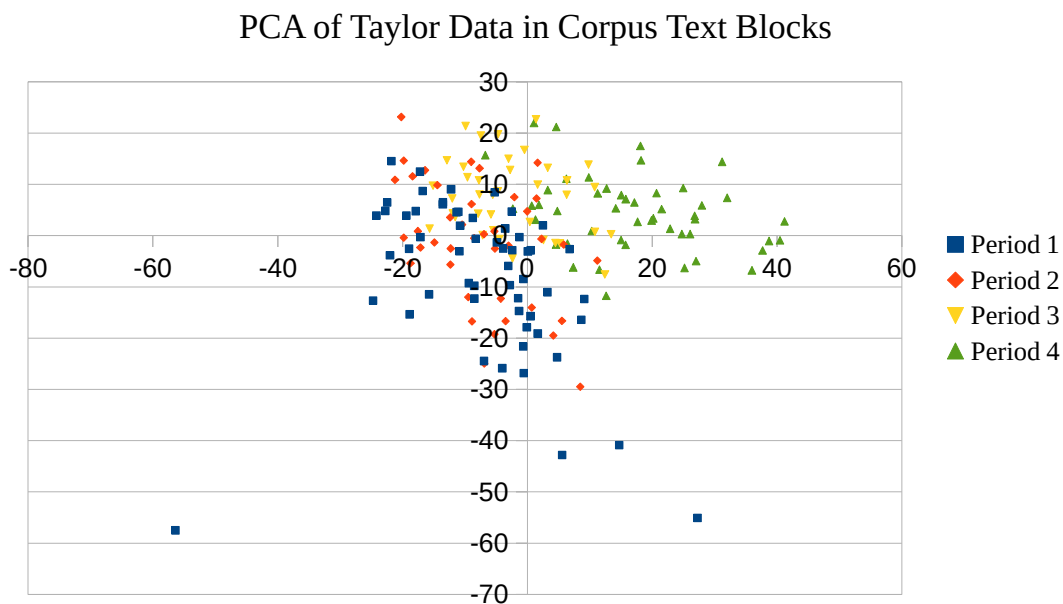
Figure 1

Using two classes, early and late, instead of four, while significantly increasing

the accuracy of the neural network in testing, still skews early for all Sonnets because

they contain so few colloquialisms.

To avoid this problem, we can narrow down the Taylor data only to that which is

relevant in the Sonnets. One intuitive way to do this is to narrow our focus only to the

particles "-ed" (archaism) and "-'d"/"-'ld" (colloquialism). This data is present in the

Sonnets, unlike many of the other colloquialisms, and in overall comparable amounts

(239 instances of "-'d"/"-'ld"; 235 instances of "-ed"). Additionally, since they are two

different morphological particles of the same grammatical inflection, one would expect

their frequencies to have an inverse relationship. The presence of one should mean the

absence of the other.

A neural network designed to only take these two pieces of the Taylor data as

input performed well enough. With only two outputs, early and late, the network achieved

34% test error rate. Displayed below are its results on the Hieatt zones of the Sonnets,

where Early indicates Periods 1 and 2 combined and Late indicates Periods 3 and 4

combined.

Hieatt Zones	Early	Late
1-60	0.7220	0.2780
61-103	0.7136	0.2864
104-126	0.3431	0.6569
127-154	0.7163	0.2837

This data largely corroborates the results of the neural network trained on rare

words. It tells us that based on the frequencies of these two pieces of the Taylor data, the

end of the fair youth subsequence displays a marked lateness compared to the rest of the

sequence. The graph in Appendix B that charts "-ed" and "-'d"/"-'ld" over the Sonnet

sequence visualizes this same conclusion to some extent. While this test is not entirely

independent of the rare words test— Appendix E shows that there are some Sarrazin links

that display either a "-ed" or "-'d"/"-'ld" inflection—it comes very close. A second,

unrelated metric independently comes to the same conclusion as the rare words test,

strengthening the conclusions of both.

Shared Rhyme and Repeated Imagery

The metrics of shared rhyme pairs between the Sonnets and the corpus (Appendix

C) and of editorial citations in the Sonnets to other Shakespeare works (Appendix D) are two datasets that can provide valuable chronological information, but cannot be used in a neural network, as there is no comparable supervised data for training. Still, by inspection they may help provide some indication of chronological affiliation of the various sonnet zones.

The results of summing and scaling the repeated imagery and the shared rhyme in the various Sonnet zones are tabulated below. Scaling is necessary because of the uneven size and composition of the periods. For example, Period 1 contains both the greatest number of works and the most lyric poetry, so it is unfairly weighted in both the shared rhyme and repeated imagery. Scaling transforms these raw values into standardized values based on their distances from the mean. As far as interpreting the data, a higher value indicates a higher affiliation with that given period.

Repeated Imagery

Hieatt Zones	Period 1	Period 2	Period 3	Period 4
1-60	-0.1459	-0.2823	-0.3861	-0.3807
61-103	-0.0451	0.0421	-0.2560	-0.1353
104-126	-0.2455	0.6897	1.0018	0.6322
127-154	0.5836	-0.0263	0.3975	0.5041

Shared Rhyme

Hieatt Zones	Period 1	Period 2	Period 3	Period 4
1-60	0.2811	0.2045	0.2527	0.3060
61-103	-0.2366	-0.1444	-0.2198	-0.2523
104-126	-0.4290	-0.4548	-0.3581	-0.3075
127-154	0.1133	0.1570	0.0902	-0.0156

Both of these tables reinforce the conclusions drawn by the rare words test, generally. Zone 3, Sonnets 104-126, reports the highest affiliation with Period 3 for repeated imagery, and with Period 4 for shared rhyme. The only major inconsistencies with the rare word results is in the repeated imagery of Zone 4, which shares citations almost evenly with Periods 1 and 4 and with the shared rhyme of Zone 1, which peaks in

Period 4. In most instances, however, this data backs up the claim of both the rare word and Taylor data tests, that Zone 3 was composed late, while the other three Zones tend earlier.

One other way to view this data that is perhaps more intuitive is in terms of the number of sonnets from each zone that have their best match with each respective period. The table below shows how many sonnets in each zone have their greatest affiliation with each period. By affiliation to a period, I mean that a sonnet has a highest value in that period out of the four after the raw data has been scaled.

Repeated Imagery

Hieatt Zones	Period 1	Period 2	Period 3	Period 4
1-60	20	20	10	10
61-103	13	11	11	8
104-126	0	6	10	7
127-154	9	5	7	7

Shared Rhyme

Hieatt Zones	Period 1	Period 2	Period 3	Period 4
1-60	14	12	19	15
61-103	7	17	9	10
104-126	4	4	6	9
127-154	9	8	7	4

Here we see again that Zone 3 tends toward lateness. In fact, for repeated imagery, there is not a single sonnet in that zone that is most affiliated with Period 1. There are some confusing aspects of the data, like Zone 1's rhyming affiliation with Period 3, but generally this data supports our hypothesis that Zone 1 and 4 were composed early, while Zone 3 was composed later.

Conclusions

Each of the different sonnet features analyzed for classification—rare words, Taylor data, shared rhyme and repeated imagery—arrive independently at similar conclusions of the chronological trends of the Sonnets. Zones 1 and 2, the opening sonnets of the Fair Youth subsequence, and Zone 4, the Dark Lady subsequence, both affiliate in style with the early parts of Shakespeare's career and seem very likely to have been composed before the turn of the century. Zone 3, the end of the Fair Youth subsequence, seems just as likely to have been written after the turn of the century, much closer to the date of publication.

Following Katherine Duncan-Jones's theory of plague as the impetus for Shakespeare's printed poetry, Figure 1 shows the points between 1590 and 1609 when London playhouses closed and the duration of those closures. The 1592 closure, trailed very closely by the next closure in 1593, represents nearly a year and a half straight where Shakespeare could not have a play performed. During this time he wrote and published both *Venus and Adonis* and *Rape of Lucrece*. It is possible that during this period he also wrote the earliest sonnets, like the beginning of the Fair Youth sequence and the Dark Lady sequence, as this is the only time of plague which coincides with Period 1. The black death killed 15,003 Londoners in 1593, more than forty people every day. It is easy to imagine a dramatist like Shakespeare looking to refashion himself as a poet and keep his financial options secure.

Date	Duration
1592.5	0.518
1593.1	1.115
1596.6	0.529
1597.6	0.205
1603.2	0.167
1603.4	0.918
1604.4	0.381
1605.1	0.123
1605.8	0.252
1606.2	0.085
1606.5	0.795
1607.3	0.962
1608.5	1.551

Figure 1

Another potential date for the composition of the early sonnets may be the periods

of plague around 1596 (6 months) and 1597 (2 months). While these plague outbreaks seem to be not as horrible as the black death in 1593, they are closer to the first record of the Sonnets. If Shakespeare were composing and circulating sonnets among his friends in 1596 and 1597, it would accord with their notice in *Palladis Tamia* in 1598 and their dubious pirating for the 1599 *Passionate Pilgrim*.

There are many periods of playhouse closure in the seventeenth century that may have been times of composition for the late sonnets. One intriguing possibility is the closure of 1608, which did not see the reopening of playhouses until 1610. This time would have been, for Shakespeare, the longest period of closure since the outbreak in 1593. Just as he chose to publish *Venus and Adonis* and *Lucrece* during that exceptional outbreak, he may have decided to publish a sonnet sequence in 1608. To pull together a proper sonnet sequence, he would have had to gather his old sonnets, find a way to arrange them, and write a feminine complaint to append it. In arranging the sequence, he may have encountered a tonal dissonance between his Fair Youth sequence and the Dark Lady sequence, and penned the Zone 3 sonnets to act as a bridge between them.

These theories are largely speculation. Hopefully, however, they are speculation in the right direction. We have the broad strokes of the chronology of Sonnet composition understood with a satisfying degree of confidence. Multiple indicators show that the beginning and the end of the sequence were composed early, while some in the middle were composed late. There are possibilities open for further research and refinement, with hopefully greater attention payed to sonnet clustering in the future. The Sonnets are enigmatic but not imperceptible. Examined using the right techniques they can reveal unknown facts about themselves, their genesis, and the man who made them.

Appendices

All data original unless otherwise attributed

N-Grams in Sonnets and Other Works

8-Gram	**Texts**
lilies that fester smell far worse than weeds	94 E3

6-Gram

for i have sworn thee fair | 147 152

5-Gram

or if it were it	127 LLL
which in thy breast doth	109 22
it is not so it	148 MFM
by heaven i think my	130 JN
or at the least so	122 OTH
it is but as a	17 2H4
thou hast the strength of	49 ROM
me that i am not	44 AWW
that which thou hast done	35 HAM
to set a form upon	89 JN
in me thou seest the	73 73
when i was wont to	102 LLL
the better part of me	39 74
in faith i do not	141 TRO
cannot choose but weep to	64 HAM
say it is not so	148 MAC
but do not so i	96 AYL
which i then did feel	120 H8
now i know thy mind	149 R2
to say it is not	148 MAC TNK
when thou art old and	2 MFM
give him leave to go	51 HAM
world is grown so bad	140 R3
and thou shalt find it	142 JN

4-Gram

is my love to	88 MND
in my head which	148 WIV
the treasure of his	63 LUC
why dost thou use	4 LR
the spirit of love	56 H5
the place where he	44 MFM
thee how to make	101 150
teach thee how to	101 R3
of the wide world	107 ERR
is this that thou	69 ANT
to have that which	64 TRO
this and this with	74 HAM
and all the rest	25 1H4 JC JN MFM R2 TMP TRO TRO
how can i then	22 28
to fear the worst	92 TRO
thou art all the	39 CYM
enough to torture me	133 JN

is the time that	3 52 WT
i must strive to	112 TMP
and all that is	133 LLL
say i love thee	149 MND
of good or evil	14 OTH
which cannot choose but	64 1H4 STM WT
nor no man ever	116 SHR
to hear her speak	130 MFM
me when thou wilt	90 1H4
love when it is	119 LR
me to my bed	27 R2
than thou art o	22 JN
is not so great	61 SHR
be it lawful i	142 LR
i shall see thee	49 ADO JC
bring him to his	52 HAM JC
me that i am	44 ADO AWW LLL TGV
no stronger than a	65 TMP
all the treasure of	2 E3
when i have seen	64 64 64
by that which i	72 MND
that i may not	140 ARD
the eye of heaven	18 E3 LUC R2
courses of the sun	59 H8 TRO
and therefore have i	83 ARD ROM
to come if it	17 HAM
in the eyes of	55 1H4 JN TRO TRO
by heaven i think	130 JN R3
for if i should	140 MND
say it is not	148 MAC MAC OTH SHR TNK
my love is as	147 21 TRO
and yet thou wilt	133 ROM
who is it that	84 JN LR MFM R3
i love thee in	96 TRO
thou dost seek to	142 R2
pray that thou mayst	143 JN JN
i may not be	140 R3
for my sake even	42 2H4
to the edge of	116 TRO TRO
when thou wilt if	90 1H4
dear friend and i	111 WIV
look on thee and	43 MND
thou canst not love	89 SHR
that tongue that tells	95 MAC
tir'd with all these	66 66
and by and by	75 1H4 JC MND OTH R2 ROM TGV TGV VEN
'tis better to be	121 H8 OTH

not the power to	131 H8	love for whose dear	151 TGV
i that i am	138 TGV TIM	so fair a house	13 TMP
i no more can	103 LUC	from the book of	25 R2
of my heart is	137 CYM JN	i think on thee	29 30 CYM
give him leave to	51 HAM MFM	thy beauty and thy	41 41
that the world may	75 ERR	who even but now	45 MND
i straight will halt	88 89	the cause of this	87 HAM HAM
that i do fawn	149 JC	in a kind of	95 ERR
me not to the	116 HAM	if for my love	40 LLL
so long to speak	100 ROM	the thing it was	49 TGV TRO
with her and she	138 H8 OTH	i never more will	110 MV
here and there and	110 TMP	not for that which	79 LLL
all his might to	80 OTH	the treasure of thy	136 2 2H6
of the time and	32 AWW COR HAM MV TN	to my content and	119 E3
there is no remedy	62 AWW MFM WIV	i love thee not	149 1H4 LR MND MND OTH WT
what should i do	57 ADO ANT COR JC OTH TN VEN	that all the world	112 LR ROM
better for my life	111 R3	the world well knows	129 LR
that thou mayst have	143 MV	not so much as	71 1H4 ARD CYM TGV TGV TN TNK
to the heart and	46 MAC	me when i am	71 2H4
that i am not	44 AWW AWW MAC SHR TGV TIM TIM TN WIV	my all the world	112 JN
him whom thou dost	89 TIM	but since i am	139 MV
to tell me so	140 2H4 R3 TRO	thou wilt for i	133 ROM
and yet it may	42 TRO	it is not so	148 ADO AWW AYL ERR H5 JN LLL LLL LLL LLL LLL MAC MFM MND MV
that which thou hast	35 HAM JN ROM	that you should love	72 AYL
closure of my breast	48 VEN	my love to thee	88 1H6 LLL VEN
that i am old	138 LR	by his side his	154 TIT
and yet i know	92 3H6 AWW AWW CYM CYM LR TN	i love to hear	130 LLL
the better part of	39 74 1H4 1H4 MV	with me for thou	47 ROM
and yet methinks i	14 TGV TMP	i cannot blame thee	40 SHR TMP
and more much more	103 LR	in my head to	27 SHR
to say it is	148 CYM MAC TNK WT	with you my most	113 OTH
not be so bold	131 VEN	mine eye and heart	46 47 3H6
mine own desert and	49 72	i will be true	123 TNK
for whose dear love	151 TN	so now i have	134 R2
i have look'd on	110 MV WT	but do not so	96 AWW AYL
i say 'tis so	85 TIT	in the eye of	88 H5
it is my love	61 MND ROM	but if thou live	3 ERR
wise as thou art	140 MND	with me and in	138 R3
more much more than	103 OTH R2	thine and all that	133 LR
of earth and water	44 H5	when i took my	48 TGV
me to go about	113 WT	all the world besides	112 JC
not to fear the	92 MFM	i grant sweet love	79 TGV
now is the time	3 MAC	for love of you	15 TGV
		i will not praise	21 TRO
		so as thou art	131 ADO

Appendix A

N-Grams in Sonnets and Other Works

N-Gram	Value	N-Gram	Value
my mistress' eyes are	127 130	what you will to	58 TN
my heart mine eye	46 LLL	the wide world and	19 ERR
give the lie to	150 AYL	day that i have	117 R2 ROM
when i shall see	49 WT	the likeness of a	141 MV ROM
in the spring when	102 98	in her cheeks and	130 TGV
the conquest of thy	46 TIM	because i would not	102 2H6 ARD AWW LR
leave me do not	90 MND		
when i am dead	71 2H6 CYM	i do forgive thy	40 TMP
to all the world	81 COR JC	the beauty of thy	69 95
thee for my love	40 3H6	the manner of my	140 MV
to whom i am	110 AWW MV MV TMP	and i myself am	134 TRO
		of you and you	76 R3
as fast as they	12 JC TRO	fair kind and true	105 105 105
as i am now	63 R3 TMP WT	where thou art not	48 2H6 MV
well i will not	21 SHR WIV	the world will be	9 ROM
how to make me	150 JN	the pleasure of the	97 3H6 MAC
if it do not	142 CYM	the world that i	71 ANT CYM LR
for my love thou	40 40		
beauty liv'd and died	68 VEN	night who like a	145 H5
i have seen the	64 COR LLL LR OTH OTH ROM TRO WIV	my soul doth tell	151 ADO
		in the rearward of	90 2H4
you shall hear the	71 CYM	mounted on the wind	51 AYL
or at the least	122 LUC OTH TGV TIT	i do not love	141 2H4 ADO TMP TNK
there is but one	36 AYL JC MV		
in me that you	72 H5	in thee it is	82 LLL
it may be said	42 AYL	see not what they	137 H5
it is that i	110 MAC	he will not be	134 AYL MAC R3 WIV
to mine eyes that	137 R3	that i might see	59 ANT H5
when thou shalt be	88 CYM	to the ground and	75 1H4 MV ROM
the world and i	112 MAC	and thou shalt find	142 77 1H4 ARD ERR JN LR
as fast as thou	11 1H4		
in time to come	17 1H4 2H6 TIT TRO	i hear and see	150 LR
and so should you	72 ANT	more than i have	119 R3 R3 TN
you for i am	72 ROM TMP	deeds must not be	121 MAC
so shall i live	93 2H4	own i may be	121 TNK
no marvel then though	148 LR	the level of your	117 WT
that love of thine	92 E3	how i do love	26 AYL
of my love to	26 1H6 AYL R3	to make me love	150 1H4
that i have look'd	110 CYM	i never saw a	130 2H6 SHR TMP WT
thou art all my	78 AWW		
brass nor stone nor	65 WT	as fair as any	21 MV
i pray that thou	143 ROM	may not be so	140 AWW
you and you and	76 JC	and for my sake	42 JC LUC LUC VEN
i do love thee	26 MND OTH R3		
for i have sworn	147 152 152	i love her and	42 H5
on the top of	16 2H6 HAM LLL MND	of a man thy	141 ROM
		i was wont to	102 JC LLL R3 WT
love is as a	147 TGV	is no stronger than	65 AYL
that i am and	121 ADO	all the world must	81 1H4
for thee and for	27 SHR	thou canst not then	133 HAM
since first i saw	104 TRO		

Appendix A

N-Grams in Sonnet Sequence

2- 3- 4- and 5-Grams of Period 1 Plays in Sonnets

Sonnet No.

Legend: 1H6, 2H6, 3H6, ARD, ERR, E3, LLL, R3, ROM, SHR, TGV, TIT, LUC, VEN, MND
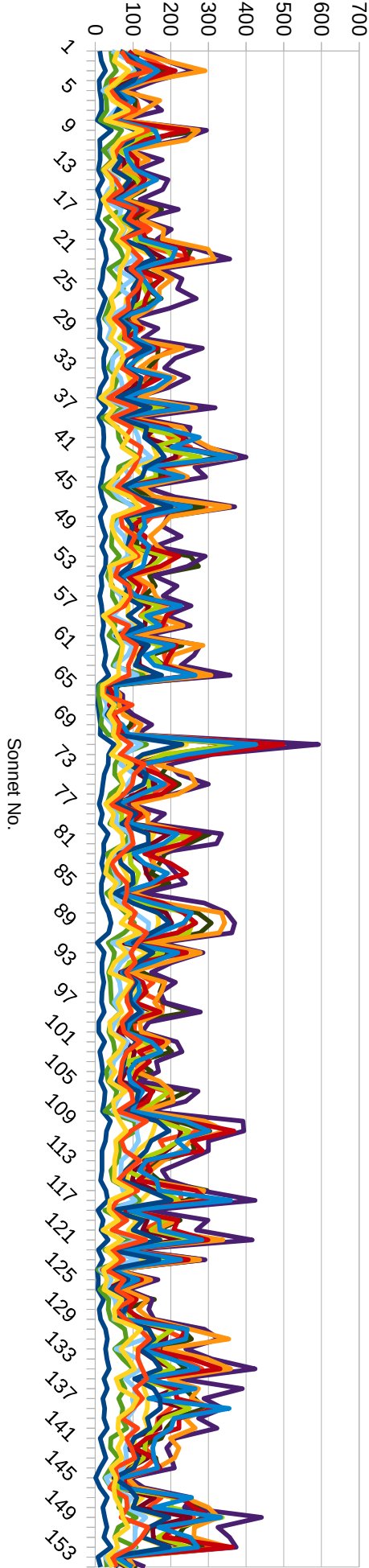
2- 3- 4- and 5-grams of Period 2 Plays in Sonnets

Sonnet No.

Legend: 1H4, 2H4, ADO, H5, JC, JN, STM, MV, R2, WIV
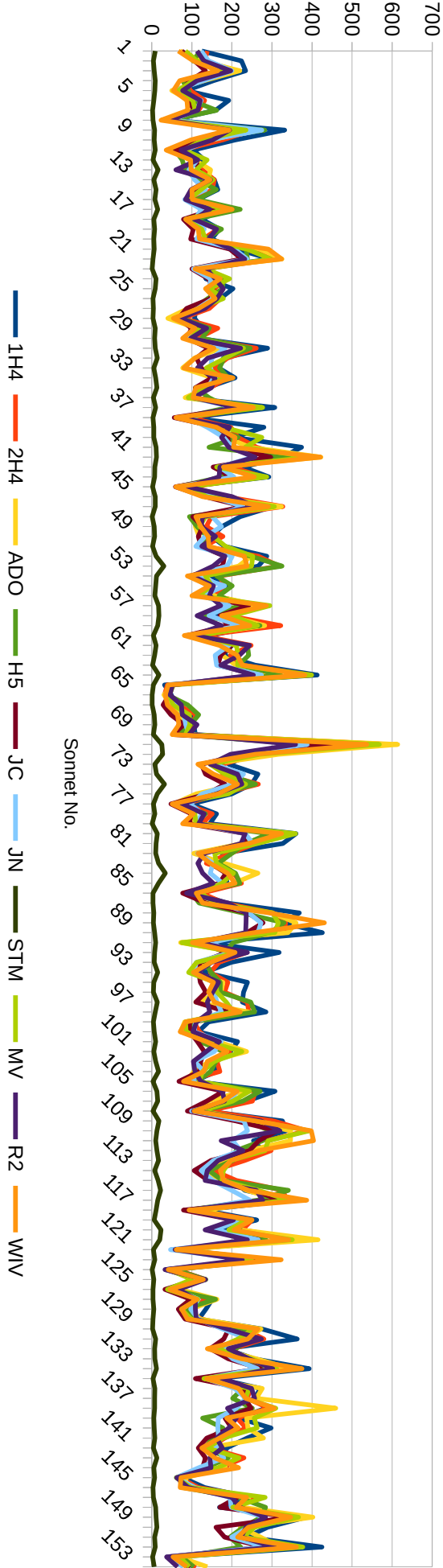
Appendix A

N-Grams in Sonnet Sequence

2- 3- 4- and 5-grams of Period 3 in Sonnets

Sonnet No.

AWW — AYL — HAM — MFM — OTH — TN — TRO — PHT — LC
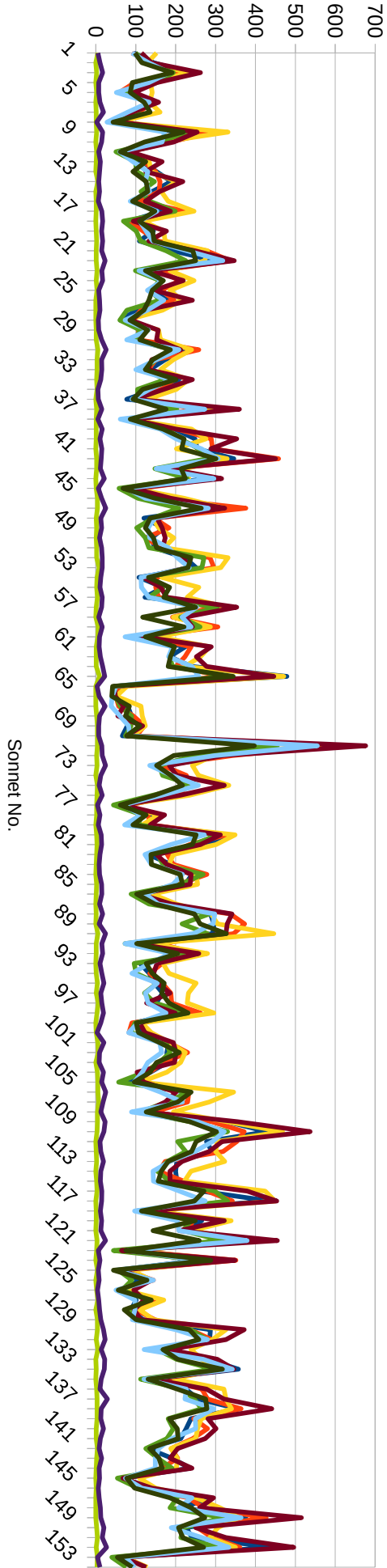
2- 3- 4- and 5-grams of Period 4 in Sonnets

Sonnet No.

ANT — COR — TIM — CYM — H8 — LR — MAC — PER — TMP — TNK — WT

Appendix A

N-Grams in Sonnet Sequence

Shared Sonnet N-Grams Relative to Period Size

Shared N-grams / Period Size (words)

Sonnet No.

Period 1
Period 2
Period 3
Period 4
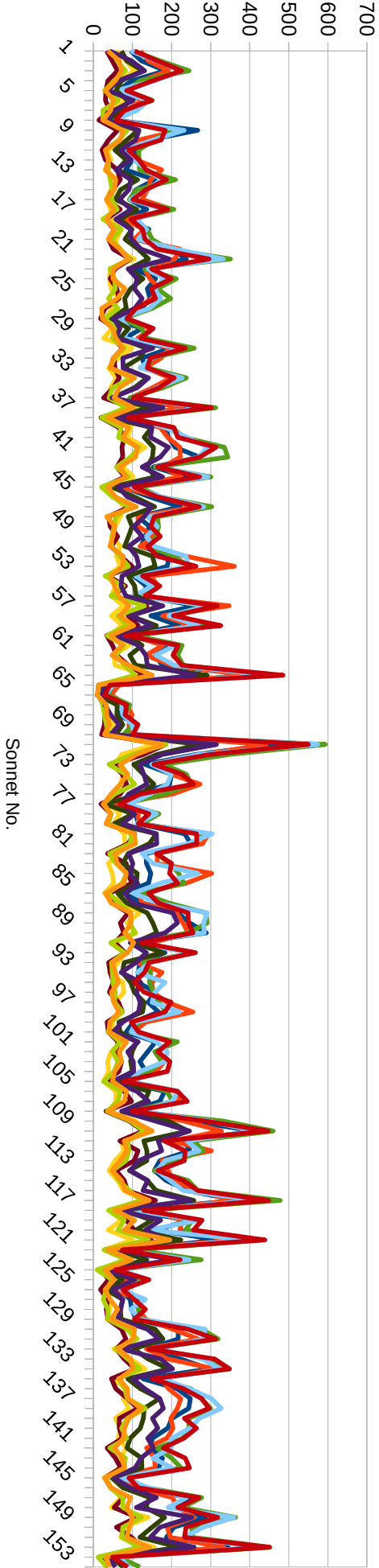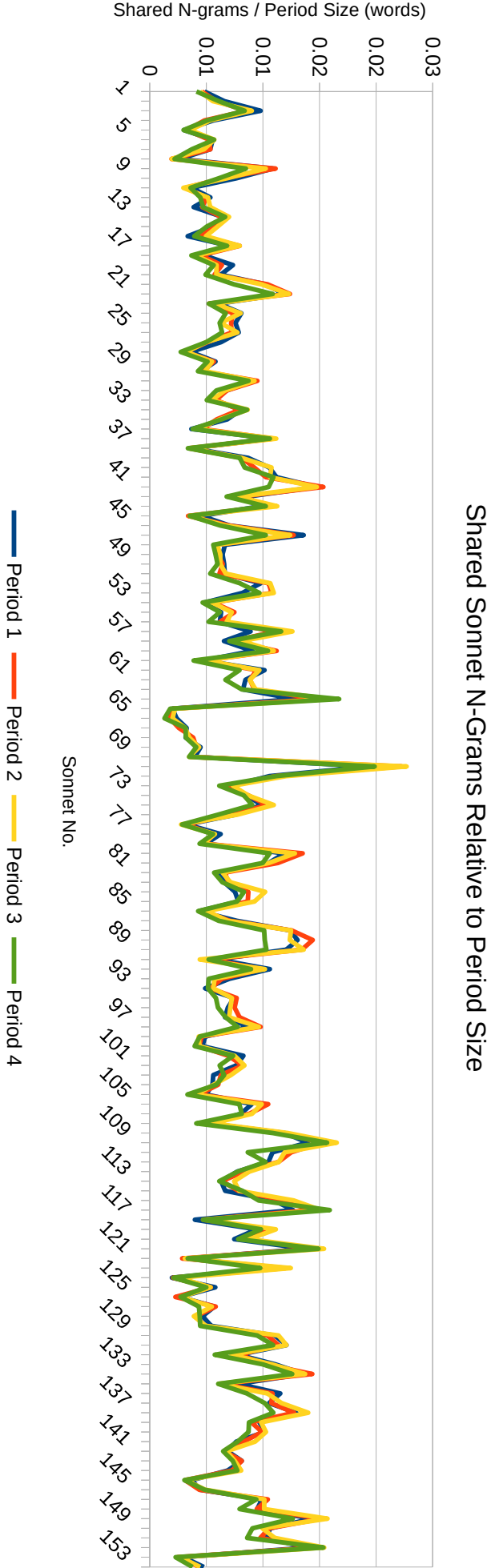
Appendix A

Appendix B
Metrical Data

Headings	Meaning
OL	Open Lines, Enjambment.
F End	Feminine Line Endings.
Most/10K	Chronological feature identified by Elliott and Valenza. Instances of token "most" per 10,000 words.
Colloq	Colloquialisms and colloquialisms per 20,000 words.
Arch	Archaisms and archaisms per 20,000 words.
MLE	Midline Endings.
Light End	Light Line Endings (unstressed monosyllabic line ending, pronouns and auxiliaries).
Weak End	Weak Line Endings (unstressed monosyllabic line endings, prepositions and conjunctions).
ULE	Unpunctuated Line Endings.
MTP	Midline Terminal Punctuation.

Sources

Hall.	F. E. Halliday, *A Shakespeare Companion* (London: Gerald Duckworth & Co. Ltd. 1968), 164.
Tarl.	Marina Tarlinskaja, *Shakespeare's Verse: Iambic Pentameter and The Poet's Idiosyncrasies* (New York; Bern; Frankfurt am Main; Paris: Lang, 1987), 182-4.
Foster	Donald W. Foster, *Elegy by W.S.: A Study in Attribution* (University of Delaware Press. 1989).
Wells	Stanley Wells and Gary Taylor with John Jowett and William Montgomery, *William Shakespeare: A Textual Companion* (Oxford: Clarendon Press, 1987), 102-4.
Elliott	Ward E. Y. Elliott and Robert J. Valenza, "Oxford by the Numbers: What Are The Odds The The Earl of Oxford Could Have Written Shakespeare's Poems and Plays?" *Tennessee Law Review* 72 (Fall 2004) 451-60.

Notes:

The first table shows plays ordered according to the Riverside chronological sequence. In sonnet data, OL, ULE, FE, LE, WE are derived from hand counts using New Penguin edition of *Sonnets*, edited by John Kerrigan. MTP, Colloq and Arch in sonnets derive from machine counts of the Riverside *Sonnets*. All other, non-original data attributed.

Appendix B

Title	OL Foster	OL Hall.	OL Elliott	F End Tarl.	F End Hall.	F End Elliott	Most /10K Colloq Wells	Colloq /20K Elliott	Arch Wells	Arch/ 20K Elliott	MLE Hall.	Light end. Hall.	Weak end. Hall.	FE+ OL	OL+ MLE .1 arch	OL+ MLE- arch
1H6	14	10	11	7	8	7	32	31	146	143	1	3	1	22	15	1
3H6	12	10	13	14	14	2	47	40	87	75	1	3	0	26	13	6
2H6	13	11	12	12	14	6	57	55	86	83	1	2	1	28	15	7
R3	16	13	16	20	20	11	21	15	126	90	3	4	0	37	20	11
TIT	14	12	10	8	9	5	11	17	71	108	3	5	0	24	18	7
ERR	9	13	16	14	17	9	24	38	52	82	1	0	0	29	13	5
TGV	15	12	17	19	18	5	32	50	51	80	6	0	0	34	22	14
SHR	11	8	15	19	18	3	63	75	68	81	4	1	0	29	15	7
R2	21	20	9	10	11	8	21	19	71	65	7	4	0	34	30	23
LLL	14	18	9	8	8	25	31	43	40	56	10	3	0	22	24	18
JN	18	18	9	5	6	7	28	27	132	130	13	7	0	29	36	23
ROM	16	14	9	8	8	7	49	47	103	99	15	6	1	24	31	21
MND	15	13	8	5	7	12	20	31	63	98	17	0	1	22	32	22
1H4	28	23	12	5	5	5	17	26	75	115	14	5	2	34	43	32
WIV	19	20	17	10	27	4	17	24	17	60	21	1	0	43	37	31
MOV	25	22	17		16	9	19	24	75	98	21	1	1	45	49	39
2H4	27	21	14	16	16	16	8	13	62	100	18	2	0	44	44	34
H5	26	22	16	14	21	16	51	69	87	117	18	2	0	47	45	39
AYL	22	17	18	23	26	17	27	59	26	57	22	2	0	49	39	39
JC	20	19	15	19	20	15	44	50	72	82	20	2	0	42	42	34
ADO	20	19	15	21	23	12	10	35	24	84	21	1	1	42	40	32
HAM	24	23	19	23	23	27	167	156	56	52	52	8	0	50	79	74
TN	25	15	17	21	26	17	34	92	17	46	36	3	1	49	59	54
TRO	27	27	15	22	24	13	75	84	42	47	31	6	0	50	57	52
AWW	34	28	19	25	29	18	123	205	26	43	74	11	2	61	106	102
MFM	28	23	20	26	26	25	104	159	36	55	51	7	0	56	81	75
OTH	24	20	17	27	28	22	192	183	44	42	54	2	0	52	78	74
LR	30	29	19	27	29	20	149	164	47	52	61	5	1	60	92	87
MAC	31	37	18	25	26	15	178	232	22	29	77	21	2	61	112	109
ANT	41	43	17	26	27	25	273	251	23	21	78	71	28	68	119	117
TIM	28	33	16	25	22	20	105	222	24	51	63	16	5	53	94	89
COR	44	46	18	30	28	19	358	345	30	29	79	60	44	74	125	122
PER	32	25	14	20	22	19	63	164	25	65	71	15	5	56	105	98
CYM	46	46	20	29	31	17	286	250	21	18	85	78	52	78	132	130
WT	47	38	21	32	33	20	276	286	15	17	88	57	43	74	136	134
TMP	46	42	22	35	35	26	148	148	16	25	85	42	25	81	131	129
H8	52	39	22		32	22	211	254	25	30	72	45	37	83	123	120
TNK	61	30	20		30	17	140	335	8	19	92	50	34	76	138	136

Metrical Data in Sonnets

Sonnet No.	OL	ULE	FE	*Most*	Colloq	Arch	LE	WE	MTP	FE + OL	OL + MTP
1	0	3	0	0	2	2	0	0	0	0	0
2	3	3	0	0	6	6	0	0	0	3	3
3	3	4	4	0	2	2	1	0	0	7	3
4	4	4	0	0	3	3	0	0	0	4	4
5	4	5	0	0	4	4	0	1	0	4	4
6	4	4	0	0	5	5	0	0	1	4	5
7	3	3	0	0	2	2	0	0	0	3	3
8	0	0	4	0	4	4	0	0	0	4	0
9	2	4	2	0	2	2	0	0	0	4	2
10	0	0	2	1	6	7	0	0	0	2	0
11	2	2	4	0	4	4	0	1	0	6	2
12	0	1	0	0	2	2	0	0	0	0	0
13	4	4	0	0	1	1	0	0	3	4	7
14	1	2	0	0	1	1	0	0	0	1	1
15	2	2	2	1	1	2	0	0	0	4	2
16	4	4	0	0	0	0	0	0	0	4	4
17	1	3	2	1	4	5	0	0	0	3	1
18	0	0	0	0	5	5	0	0	0	0	0
19	1	1	0	1	2	3	0	0	1	1	2
20	1	1	14	0	1	1	0	0	0	15	1
21	2	4	0	0	2	2	0	0	0	2	2
22	1	2	2	0	1	1	1	0	0	3	1
23	1	2	0	0	4	4	0	0	0	1	1
24	2	3	0	0	3	3	1	0	0	2	2
25	2	3	0	1	3	4	0	0	0	2	2
26	4	5	2	0	0	0	0	0	0	6	4
27	2	3	2	0	1	1	0	0	0	4	2
28	2	3	4	0	6	6	0	0	0	6	2
29	1	2	2	1	3	4	0	0	0	3	1
30	2	2	0	0	7	7	0	1	0	2	2
31	2	3	0	0	2	2	0	0	0	2	2
32	2	2	2	0	1	1	0	0	0	4	2
33	3	3	2	0	1	1	0	0	0	5	3
34	3	4	0	0	2	2	0	0	0	3	3
35	2	2	0	0	1	1	0	0	0	2	2
36	2	2	0	0	0	0	0	0	0	2	2
37	1	3	0	0	2	2	0	0	0	1	1
38	6	6	0	0	2	2	2	0	0	6	6
39	4	4	0	0	2	2	0	0	0	4	4
40	1	1	4	0	1	1	0	0	0	5	1
41	0	1	0	0	1	1	0	0	0	0	0
42	0	0	2	0	2	2	0	0	1	2	1
43	2	2	2	1	0	1	0	0	0	4	2
44	3	3	0	0	1	1	0	0	0	3	3
45	4	4	1	0	2	2	0	0	0	5	4
46	2	3	0	0	1	1	0	0	1	2	3
47	1	2	2	0	1	1	0	0	0	3	1
48	1	2	0	0	1	1	0	0	0	1	1
49	2	2	0	0	2	2	0	0	0	2	2
50	4	5	0	0	2	2	0	1	0	4	4
51	2	2	0	0	1	1	0	0	0	2	2

Appendix B

Metrical Data in Sonnets

Sonnet No.	OL	ULE	FE	*Most*	Colloq	Arch	LE	WE	MTP	FE + OL	OL + MTP
52	1	1	2	0	2	2	0	0	0	3	1
53	1	1	0	0	0	0	0	0	0	1	1
54	4	5	2	0	2	2	0	0	1	6	5
55	6	6	0	0	1	1	0	0	1	6	7
56	7	7	2	0	2	2	0	0	1	9	8
57	6	6	0	0	0	0	0	0	0	6	6
58	5	5	2	0	3	3	0	0	0	7	5
59	6	6	0	0	1	1	0	0	0	6	6
60	0	0	0	0	1	1	0	0	0	0	0
61	3	3	2	0	1	1	1	0	0	5	3
62	0	0	0	0	2	2	0	0	0	0	0
63	5	5	0	0	6	6	0	0	0	5	5
64	3	3	0	0	0	0	0	0	0	3	3
65	1	1	2	0	1	1	0	0	1	3	2
66	0	0	0	0	6	6	0	0	0	0	0
67	3	5	0	0	1	1	0	0	0	3	3
68	2	2	0	0	1	1	0	0	0	2	2
69	3	3	0	0	1	1	0	0	0	3	3
70	2	2	0	0	9	9	0	0	0	2	2
71	6	6	0	0	0	0	0	1	0	6	6
72	5	5	0	0	2	2	1	0	0	5	5
73	4	4	0	0	5	5	0	0	0	4	4
74	2	3	0	0	0	0	0	0	0	2	2
75	3	3	2	0	3	3	0	0	0	5	3
76	1	1	0	0	0	0	0	0	0	1	1
77	2	2	0	0	2	2	0	0	0	2	2
78	4	5	0	1	1	2	0	0	0	4	4
79	4	4	0	0	1	1	0	0	2	4	6
80	1	2	0	0	1	1	0	0	1	1	2
81	1	1	2	1	1	2	0	0	0	3	1
82	6	6	0	0	6	6	0	0	1	6	7
83	3	3	0	1	0	1	0	0	0	3	3
84	4	4	2	1	0	1	0	0	0	6	4
85	2	3	2	1	4	5	1	0	0	4	2
86	3	3	0	0	2	2	0	0	0	3	3
87	0	0	12	0	3	3	0	0	0	12	0
88	1	3	2	0	2	2	0	0	0	3	1
89	1	2	0	0	0	0	0	0	0	1	1
90	1	1	2	0	4	4	0	0	0	3	1
91	1	1	2	1	0	1	0	0	0	3	1
92	1	1	0	0	0	0	0	0	0	1	1
93	3	3	0	0	1	1	0	0	0	3	3
94	0	1	2	1	0	1	0	0	0	2	0
95	2	2	0	0	1	1	0	0	0	2	2
96	3	3	0	0	4	4	0	0	0	3	3
97	3	3	0	0	2	2	1	0	0	3	3
98	1	1	0	0	3	3	0	0	0	1	1
99	3	3	0	0	2	2	0	0	1	3	4
100	2	2	0	0	3	3	0	0	0	2	2
101	4	4	0	0	5	5	0	0	1	4	5
102	2	2	0	0	2	2	0	0	0	2	2
103	4	4	0	0	0	0	0	0	0	4	4

Appendix B

Metrical Data in Sonnets

Sonnet No.	OL	ULE	FE	*Most*	Colloq	Arch	LE	WE	MTP	FE + OL	OL + MTP
104	2	2	0	0	5	5	0	0	1	2	3
105	0	1	0	0	3	3	0	0	0	0	0
106	4	4	0	0	2	2	0	0	0	4	4
107	3	3	0	1	5	6	0	0	0	3	3
108	3	3	2	0	2	2	0	0	0	5	3
109	4	4	0	0	5	5	0	0	1	4	5
110	2	2	0	2	4	6	0	0	1	2	3
111	4	4	2	0	2	2	1	0	0	6	4
112	5	5	0	0	4	4	0	0	0	5	5
113	2	2	0	2	2	4	0	1	0	2	2
114	2	2	4	1	2	3	0	0	0	6	2
115	2	2	2	1	5	6	0	0	0	4	2
116	3	3	0	0	3	3	0	0	1	3	4
117	3	3	0	0	1	1	0	0	0	3	3
118	4	4	4	0	1	1	0	0	0	8	4
119	3	3	6	0	3	3	0	0	0	9	3
120	4	4	4	0	3	3	0	0	0	8	4
121	2	4	4	0	0	0	0	0	0	6	2
122	5	5	0	0	3	3	0	0	0	5	5
123	3	3	0	0	0	0	0	0	1	3	4
124	2	2	2	0	5	5	0	0	0	4	2
125	3	3	2	1	4	5	0	0	0	5	3
126	3	3	3	0	4	4	0	0	0	6	3
127	1	1	2	0	3	3	0	0	0	3	1
128	6	6	0	0	3	3	0	0	0	6	6
129	3	4	0	0	5	5	0	0	0	3	3
130	3	3	0	0	1	1	0	0	0	3	3
131	2	3	0	1	1	2	0	0	0	2	2
132	3	3	0	0	1	1	0	0	0	3	3
133	1	1	2	0	2	2	0	0	0	3	1
134	2	2	0	0	4	4	2	0	0	2	2
135	1	2	2	0	0	0	0	0	0	3	1
136	1	2	0	0	1	1	0	0	0	1	1
137	2	2	0	0	1	1	0	0	0	2	2
138	1	1	0	0	2	2	0	0	0	1	1
139	3	3	0	0	4	4	0	0	1	3	4
140	3	3	0	0	0	0	0	0	0	3	3
141	2	2	2	0	2	2	0	0	0	4	2
142	1	2	2	0	4	4	1	0	0	3	1
143	3	4	0	0	1	1	1	0	0	3	3
144	2	3	2	0	2	2	0	0	0	4	2
145	1	2	0	0	7	7	0	0	0	1	1
146	0	0	0	0	0	0	0	0	0	0	0
147	2	2	0	0	2	2	0	0	0	2	2
148	1	1	0	0	2	2	0	0	1	1	2
149	3	3	0	0	3	3	0	0	0	3	3
150	2	5	0	0	2	2	0	0	0	2	2
151	3	3	2	0	0	0	0	0	2	5	5
152	1	2	4	1	3	4	0	0	2	5	3
153	4	4	0	0	2	2	0	0	1	4	5
154	5	5	0	0	4	4	0	0	0	5	5

Appendix B

Feminine Endings in Sonnets
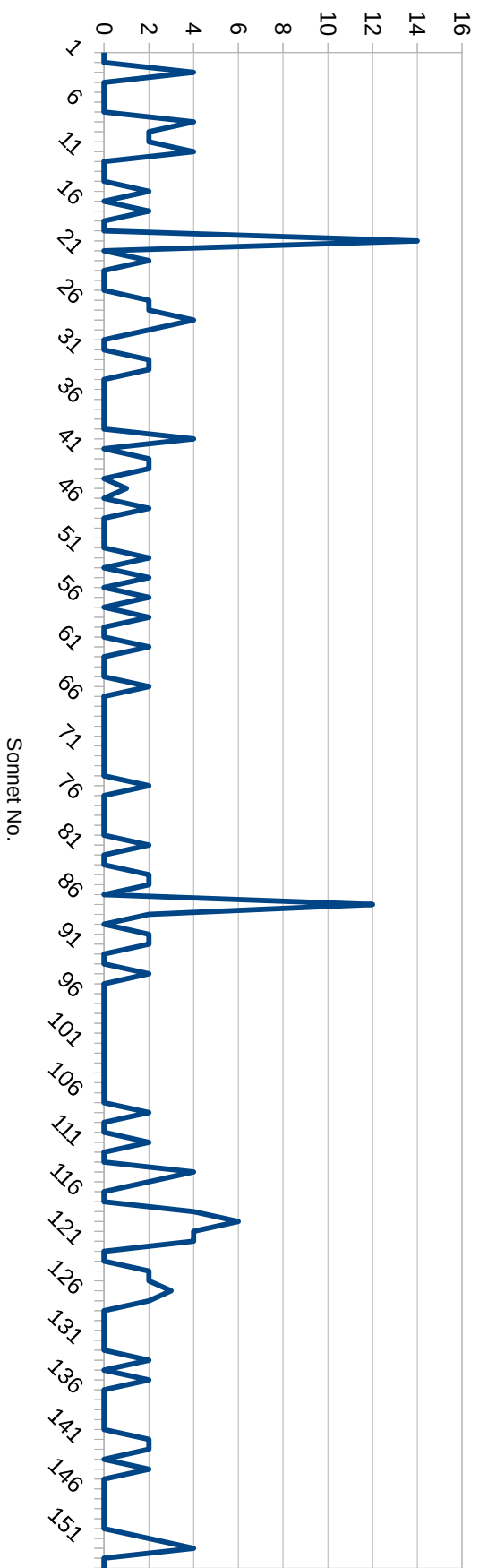
Sonnet No.

Unpunctuated Line Endings, Open Lines in Sonnets

Sonnet No.

UOL
U
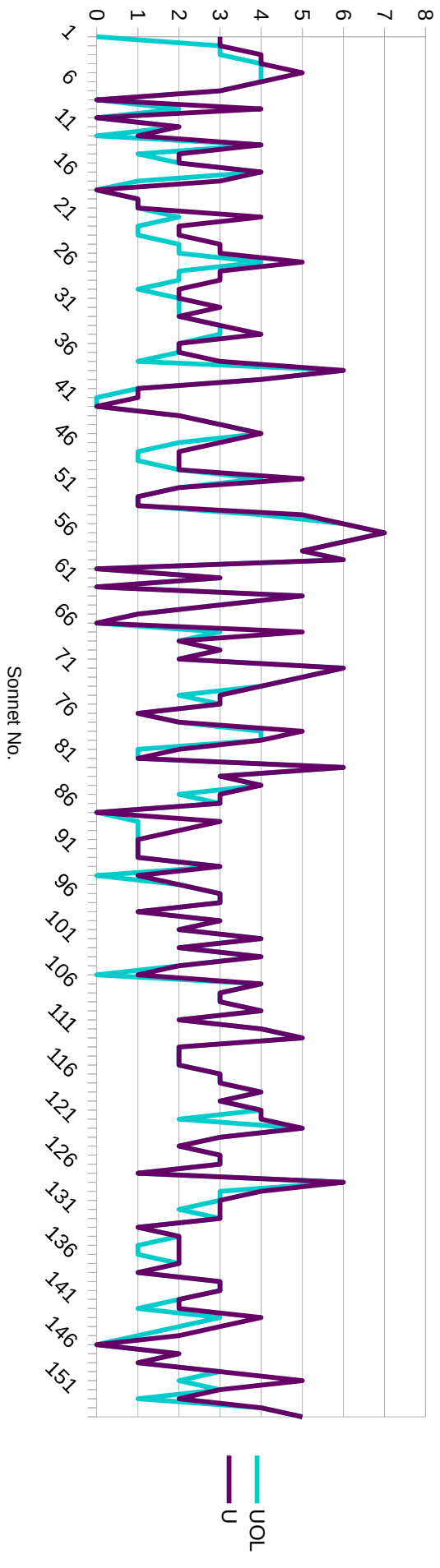
Appendix B

Midline Terminal Punctuation in Sonnets
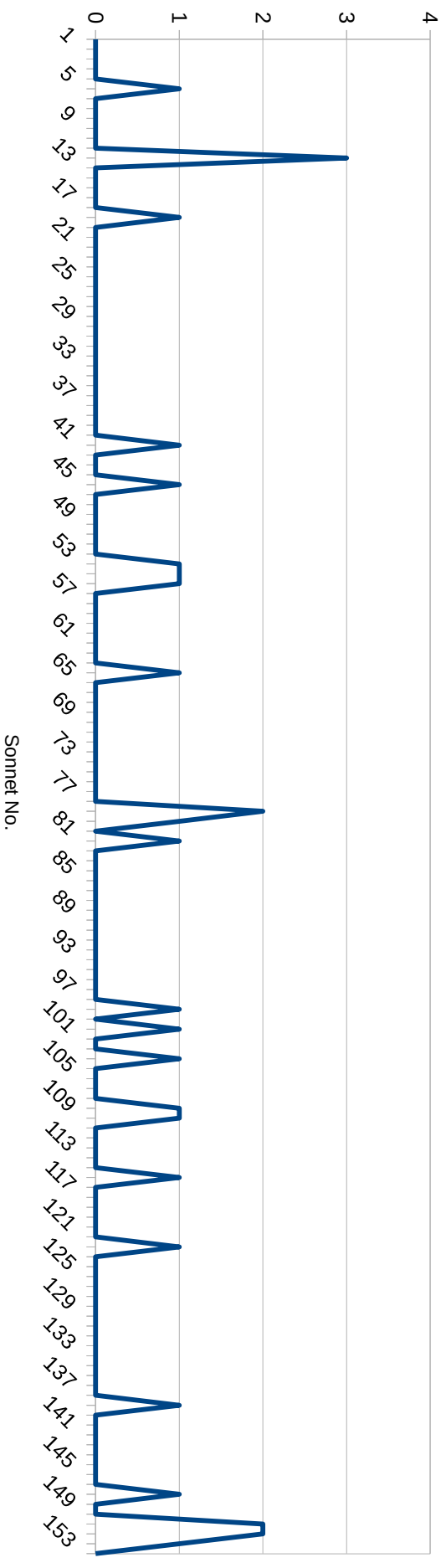
Sonnet No.

Light Endings, Weak Endings in Sonnets

Sonnet No.

LE
WE
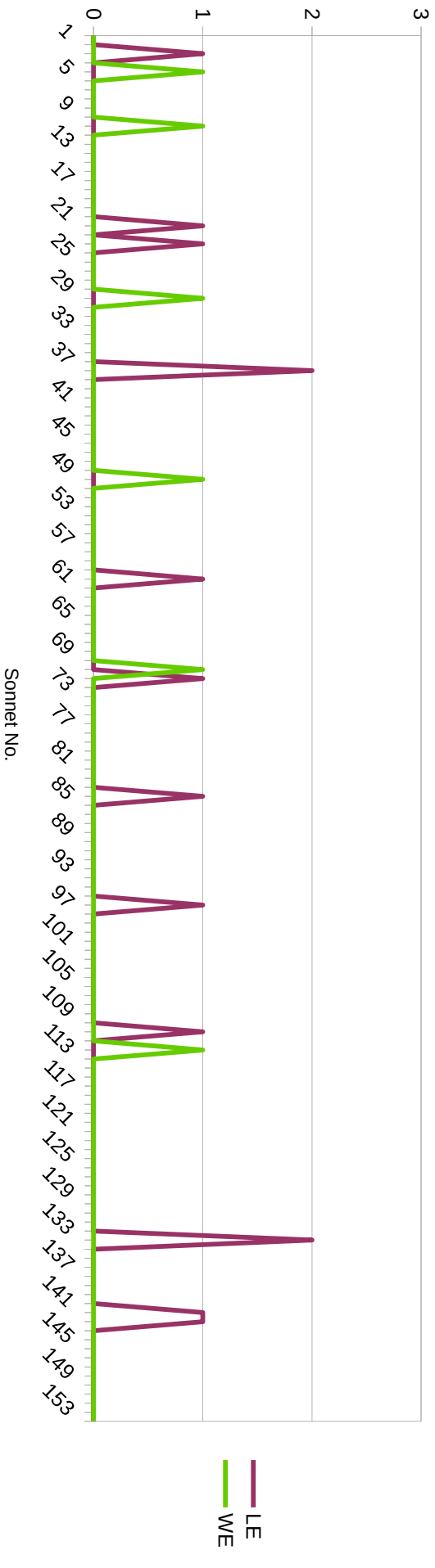
Appendix B

Colloquialism in Sonnets

Sonnet No.

Legend:
- 't
- 'th'
- 're
- 'd/'ld
- st/'ve
- o'
- 's (is)

Archaisms in Sonnets

Sonnet No.

Legend:
- eth
- ion
- ed
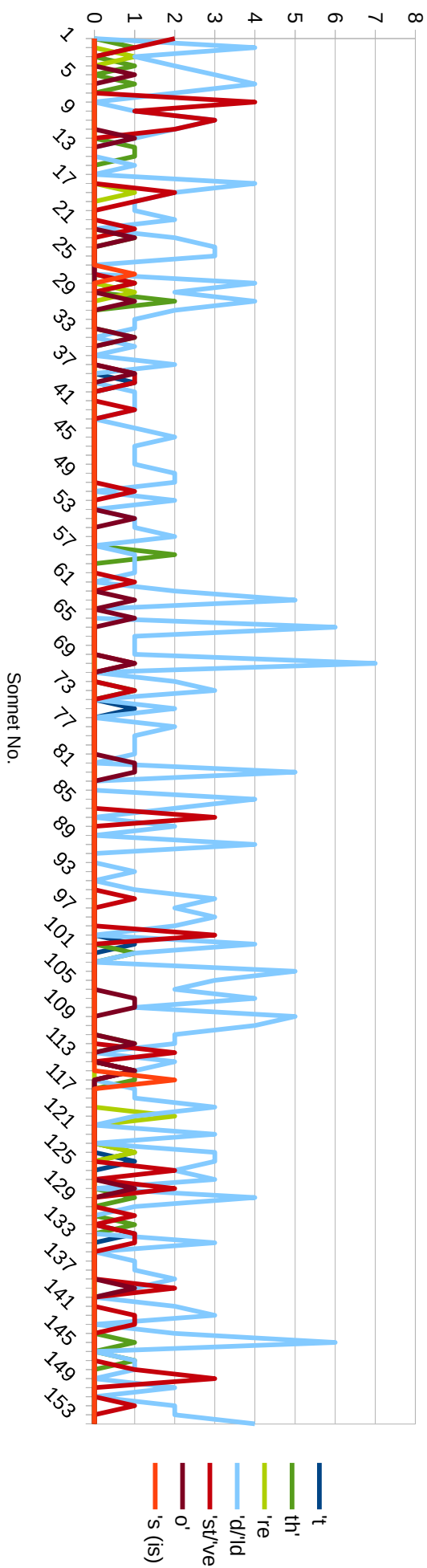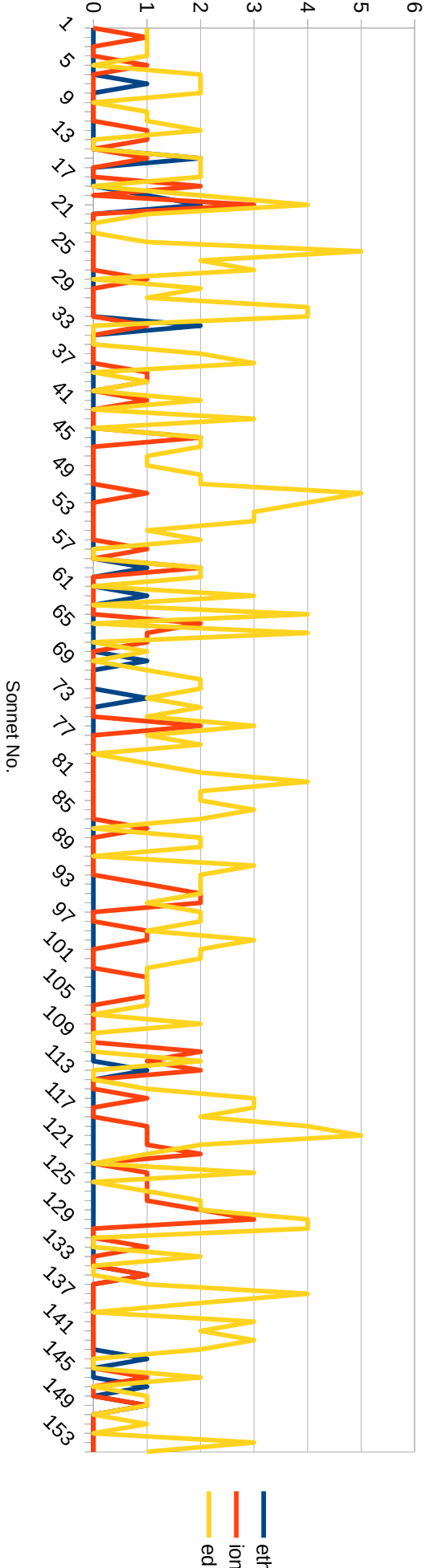
Appendix B

Total Archaisms, Total Colloquialisms in Sonnets

Sonnet No.
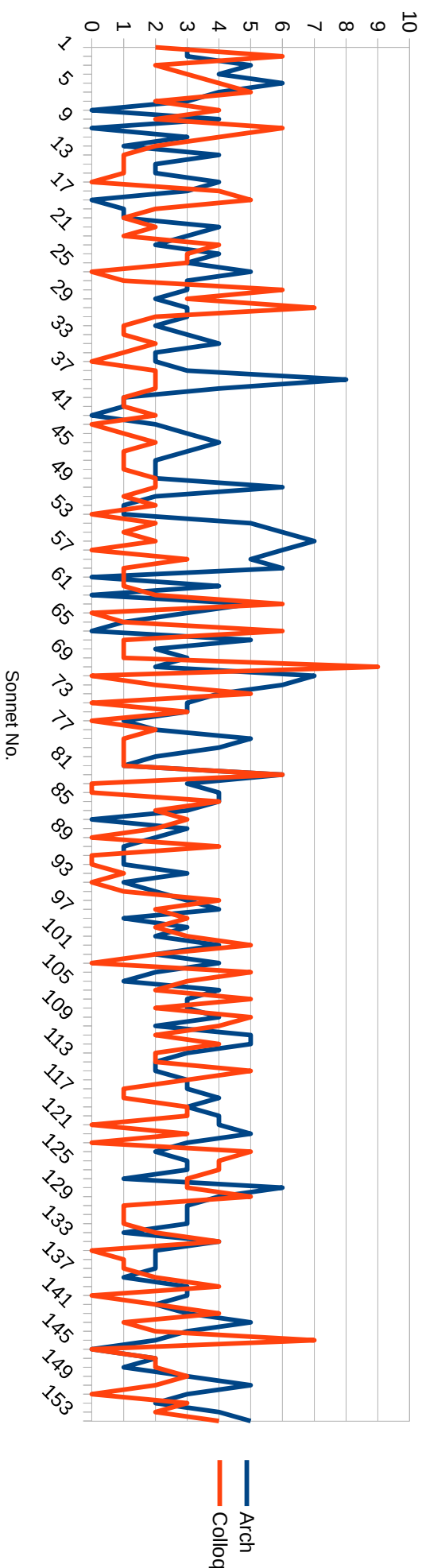
Arch
Colloq

'd (Colloquialism) and -ed (Archaism) in Sonnets

Sonnet No.

'd/'ld
ed

Appendix B
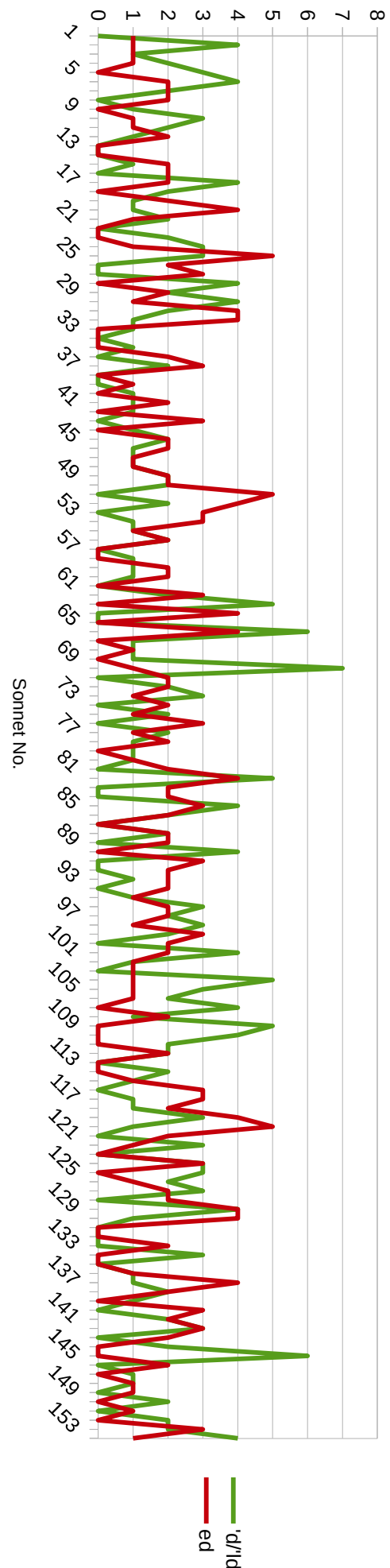
Rhyme Pair	Texts
how now	101 LLL LLL MFM OTH R2
lie by	1H4 73
report sort	36 96 MND
delight white	130 98 LUC VEN VEN
perish cherish	11 LUC
forbid hid	65 LLL
away day	145 1H4 1H4 2H6 73 75 E3 E3 H5 H5 H5 JN LLL MFM MND PP PP R2 R3 ROM TGV TNK LUC
affords words	105 85 LUC
new you	15 53 76 ERR MOV
eyes despise	141 149
age rage	17 64 LC
remedy by	154 MND
annoy joy	8 R3 LUC VEN VEN
brow now	106 2 33 63 68 H8 JN LLL LLL LLL LLL
sway day	150 1H4
me free	134 134 AWW LUC
glass was	5 LUC
thine mine	134 142 2 26 92 COR E3 R3 LUC VEN VEN
will still	134 2H6 ERR ERR ERR JC PP ROM ROM LUC LUC VEN VEN
near there	136 MND
remains contains	74 LC
depend end	92 ROM
is his	2H6 67 80 AYL COR LLL LR MFM MND R2 WIV WIV LUC
behind kind	143 HAM HAM LUC
brow allow	112 19
grief chief	42 VEN
fled dead	71 MND PHT PP R2 R2 R2 VEN
dear there	110 2H4 WT
life knife	100 63 74 LLL PER TN LUC LUC
sun done	24 35 59 CYM ERR E3 MAC LC LUC VEN VEN VEN
kind mind	10 69 ADO AYL H5 MOV TGV LC VEN
twain remain	36 39 MND
it sit	103 JN
trim him	98 TRO LC VEN
gate state	29 E3 LUC
love above	110 ADO AWW AYL TN TRO
increase decease	1 97
thought sought	30 LUC
state fate	29 LLL WIV
part art	48 AWW MAC LC
about out	113 ADO ROM ROM WIV WIV LUC
heart art	125 131 139 22 24 24 41 E3 H8 MAC MND MND R2 LC LUC
argument spent	100 105 76
gone moan	44 71 HAM MND
amiss is	151 59 HAM MOV
blindness kindness	152 ERR TGV
love remove	116 AYL PP ROM LC VEN VEN
rhyme time	106 107 16 17 32 55 LLL PER WIV
offense defense	89 JN
hence defense	12 LLL
play away	98 AWW HAM HAM TRO TRO
lack black	127 132 LLL OTH
me fee	120 AWW MND
mind find	27 77 92 MOV PER TN TRO LC LC LUC LUC LUC
grows shows	15 LLL
frame same	5 59
taken shaken	116 120
sell well	21 TRO
graces faces	17 94 TGV LUC
need speed	2H4 51 JN JN
spring thing	98 TN
skill kill	126 LLL
head fled	148 LR MAC MND VEN
strange change	123 76 89 93 HAM LLL
state gait	128 TMP
true anew	119 PP
eyes arise	55 CYM CYM MND MND ROM LUC
pen men	16 19 32 81 H5 LUC
place face	131 137 93 JN LLL MND LC LUC LUC LUC
horse force	91 VEN
hold bold	122 LUC
compare are	35 E3 SHR VEN
swear bear	131 LUC
light fight	60 88 R2 WIV
grave have	1H6 81 AWW CYM R2 R2 R2 R2 R2 ROM VEN
down frown	117 LR VEN VEN
even heaven	132 28 AWW AYL VEN
head dead	2H4 3H6 68 E3 HAM MND PER R3 VEN
bough now	102 VEN
die lie	2H4 81 92 ADO ARD COR CYM ERR ROM TN VEN
flower power	65 MND ROM LC LC VEN
say way	50 CYM ERR LUC

skill still	16 24 MND PER LUC LUC LUC
hours flowers	16 LLL
outworn born	68 LUC
bad mad	140 ERR LUC
derive thrive	14 AWW
scope hope	29 52 HAM
best rest	115 3H6 91 ARD PP TGV
crave slave	58 LUC
sound ground	130 MND MND
case place	108 JN LC
reproving loving	142 LUC
light bright	43 LLL LLL LUC VEN
strong long	73 LUC VEN
pleasure treasure	126 20 52 75
there where	100 5 R2 TN
born forsworn	66 LLL LLL
her her	42 ADO ADO ADO ADO ADO ANT ARD AWW AYL AYL AYL AYL CYM H8 H8 H8 H8 LLL LR OTH PER PP PP ROM SHR SHR SHR SHR SHR TGV TGV TGV TGV TGV TNK TNK TN WIV WIV WIV WT WT LUC LUC LUC VEN VEN
pride aside	76 E3
ear bear	8 LUC
suppose those	57 AWW
still hill	7 VEN
stand hand	104 128 60 99 MND R2 ROM ROM LC LUC LUC LUC LUC
sweet meet	5 94 AWW HAM HAM LLL MND R2 ROM
pay say	79 MAC MAC
chest breast	48 R2
delight sight	47 75 MND LUC
alone groan	131 133 HAM VEN
day stay	3H6 3H6 43 MND MOV LUC
eye majesty	7 JN LUC
tongue song	102 17 LLL
delight spite	36 37
use abuse	134 4 ROM VEN
fire expire	2H4 73
i lie	72 LLL MND ROM TMP
cover lover	32 ROM
hearts parts	31 AYL MND SHR
sake awake	61 MND MND R3
die eye	25 9 ERR R2 ROM ROM TRO WIV LUC LUC LUC
bereft left	5 ERR TIM LUC
pilgrimage age	7 AYL R2 LUC
face chase	143 VEN
seen been	97 R2
none one	136 8 AWW MAC PHT PHT R3 ROM TGV TRO LUC
lies eyes	1 137 153 17 2 24 2H4 46 CYM LLL LLL LLL MND MOV R2 ROM ROM TMP LC LC LUC LUC LUC LUC VEN VEN VEN
verse rehearse	21 38 71 81
end fiend	145 PHT
behind mind	50 9 LR OTH PER LUC LUC LUC
hell well	129 58 ERR ERR JN LLL MND R2 R3 TN
might sight	123 139 150 MND PER
so grow	115 LLL R3 TN
are prepare	13 LLL
morrow sorrow	90 R2 ROM LUC LUC LUC
bright night	147 28 ROM ROM
blind mind	113 149 MND
hue true	67 MND
so show	105 43 54 LLL LLL MND MND LUC LUC
wear bear	77 VEN
it it	1H6 26 2H6 3H6 9 ADO ADO ADO ADO ADO ADO ANT ARD AWW AWW AWW AWW AWW AWW AWW AWW AYL AYL CYM CYM ERR ERR ERR H5 H5 H5 H8 HAM HAM HAM JC JN LLL LLL LLL LLL LLL LLL LLL LLL LR LR MND MND MOV MOV OTH OTH R3 R3 R3 SHR TMP TNK TN TN WIV WIV WIV WT WT WT WT WT WT LUC LUC LUC LUC LUC VEN VEN VEN VEN VEN VEN VEN VEN VEN VEN
most lost	152 LLL
choose lose	64 AWW ERR MOV R2
seen green	104 33 63 68 VEN
another mother	3 8
tears fears	119 LC
youth truth	110 138 37 41 54 60 AWW PP TN LC
bad had	67 E3
end friend	110 30 50 ANT AYL CYM LLL MND ROM TGV LUC LUC LUC
sweet greet	145 TN
weep keep	9 LLL ROM
bear were	13 H8
took look	47 75 ERR LUC
smell tell	98 PP
friend fiend	144 PP
you adieu	1H6 57 LLL LLL LLL MAC MFM MND ROM TNK VEN
end spend	146 9 MAC

Appendix C – Shared Rhymes in Sonnets and Other Works

me thee	10 122 133 143 150 1H4 1H4 22 24 26 28 2H4 2H6 2H6 2H6 2H6 31 36 37 38 39 3H6 3H6 41 43 45 47 50 61 74 88 97 ADO ADO ANT ANT ARD ARD ARD ARD ARD AWW AWW AWW AWW AYL ERR ERR ERR ERR ERR ERR E3 E3 E3 H5 H5 H8 HAM HAM HAM JN LLL LLL LLL LR LR LR LR LR MAC MND MND MND MND MOV OTH PER PP PP R2 R2 R3 R3 R3 R3 R3 R3 ROM ROM ROM ROM SHR SHR TGV TGV TGV TGV TIM TIT TIT TIT TMP TMP TMP TMP TMP TN TN WIV WIV LUC LUC LUC LUC VEN VEN VEN VEN VEN	bred dead	108 112 LUC LUC VEN
		foes shows	40 TNK
		cure endure	153 ADO VEN
		taste waste	77 ROM
		cries eyes	29 LLL LUC LUC LUC
		doubt out	144 JN JN LLL LLL PP VEN
		eyes spies	121 LUC
		grief thief	40 48 OTH LUC
		earth dearth	146 VEN
		hour power	126 LC LUC
		use muse	21 78 82
		this amiss	35 ERR HAM MOV PP
		good blood	109 121 1H4 AWW LLL MAC MAC MND R2 R2 TIT LC LUC LUC VEN
		be decree	93 LLL
crime time	120 124 19 58 WT LUC LUC LUC	away stay	143 1H6 2H6 74 92 ERR ERR E3 E3 E3 JN JN LLL LLL LR MND MOV MOV R2 R3 ROM ROM ROM SHR TN LUC
land stand	44 CYM LR LUC		
break speak	34 HAM LLL MAC TRO LUC LUC VEN		
not blot	2H6 92 LLL LUC	grace face	132 ERR LLL LLL LLL MND LC LUC VEN
this kiss	128 AWW OTH ROM ROM ROM TGV VEN VEN	beside pride	103 LC
sight night	113 15 27 30 61 63 ERR MAC PP ROM VEN VEN	reason treason	151 VEN
		tomb come	17 ROM
content spent	119 AYL CYM MAC MND	can man	141 1H6 ANT MND R2
right sight	117 2H6 46 PHT	calls falls	124 MND
convert art	14 LUC	wrack back	126 MAC LUC LUC VEN
lie deny	46 CYM MND	control soul	107 125 LUC LUC
thee free	125 TGV	delight night	102 MND PP ROM ROM LUC LUC LUC LUC LUC VEN
made shade	43 53 PP PP LUC		
curse worse	84 MND R2 R3		
despair fair	144 PP ROM VEN VEN	old told	123 138 76 PP
spring sing	102 LLL PP LUC LUC	sing bring	39 TNK
day decay	13 R2 LUC	now bow	120 90 LR VEN VEN
doom room	3H6 55 R2 WIV	live give	13 2H6 31 37 39 4 54 79 AWW AWW R2 ROM ROM TRO WT LUC LUC
alone anon	75 E3 WIV		
begin sin	114 R2 LUC		
shown own	121 69 MND PP ROM ROM	wit writ	23 84 LLL PER
new view	110 27 56	deeds breeds	111 LUC
tongue wrong	112 139 89 ARD LLL LLL MND MND PP R2 R2 R2 TIT LUC LUC LUC VEN VEN VEN VEN	life strife	75 MND OTH R2 ROM ROM TIM TIM LUC LUC LUC VEN VEN VEN
		so no	148 ERR ERR LLL LLL LLL LLL MFM MFM MND MOV R2 TRO VEN
be see	137 1H4 56 ANT ANT AYL COR E3 H5 H5 HAM LLL MAC MFM MND MND MND MND MOV PP ROM TN TRO WIV WIV LC LC LUC LUC VEN		
		tomb womb	3 ROM
		owe show	70 LUC
		this is	1H4 1H4 2H4 72 AWW AYL LR LR MFM MOV MOV PER R3 ROM TIM TRO WT VEN
taste last	90 LUC VEN		
brow mow	60 COR		
lie eye	152 31 E3 E3 LLL ROM VEN VEN	sounds confounds	128 8
		life wife	9 AWW ERR ERR MOV PER R2 TN LUC LUC LUC LUC

Appendix C – Shared Rhymes in Sonnets and Other Works

then men	146 H5 MND MOV OTH PP PP ROM
deem'd esteem'd	96 E3
thought nought	57 MAC R3
heart part	113 122 132 23 46 46 47 53 62 AYL AYL ERR ERR JC LLL LLL LLL LLL LLL MND MOV MOV PP R2 R2 R2 R2 ROM ROM LC LUC LUC LUC VEN VEN
fire desire	154 45 WIV WIV LUC LUC LUC LUC VEN VEN VEN VEN VEN VEN
respect defect	149 LUC
new true	68 93 MOV R2
thought brought	32 44 H5 PER LUC
disgrace face	103 127 33 34 R2 R2 LUC LUC LUC
still thy	135 135 143 1H4
cloak smoke	34 LUC
state hate	124 142 145 150 LUC
prime time	12 3 70 97 AYL R2 LUC VEN
shame name	127 1H6 2H6 36 95 AWW ERR E3 LLL LLL LLL R2 TGV TRO LUC LUC LUC
you grew	84 86
king spring	3H6 63 R2 LUC
breast guest	153 LUC
be thee	1 101 123 126 136 140 141 142 151 1H4 1H4 3 4 44 78 CYM CYM ERR H8 HAM MOV R2 R2 R2 SHR TGV TIT TNK TN WT LUC LUC VEN
rage stage	23 ROM
tomb dumb	101 83 ADO AWW MND
light might	100 WIV
divine thine	108 LUC
hide side	50 MFM
are care	112 147 3H6 48 MAC R2
confound crown'd	60 69
story glory	84 88 H5 ROM LUC
appear bear	80 ERR MND TRO
spend friend	149 PP
note dote	141 ERR LLL LC LUC VEN
buried dead	31 ROM ROM TGV TGV
ill still	144 147 MAC PP LUC LUC LUC
feast guest	1H4 47 ERR ROM ROM VEN
sake make	145 MOV LC
not forgot	149 71 MFM MND
on gone	5 ANT MND OTH PER R3 VEN
pain disdain	132 140 PP LUC
devise eyes	83 MND
read indeed	62 R2

turn'd burn'd	104 AYL
cheek seek	67 LLL
long song	100 1H4 PP
loss dross	146 TRO
deeds proceeds	131 AWW
more before	40 85 ERR H8 PP R2 TN
leave deceive	39 4 MND TRO
is still	136 VEN
oppress'd rest	28 CYM
book o'erlook	82 MND
pride hide	52 E3 ROM TN LUC
will ill	22 57 89 AWW ERR E3 H5 JN ROM ROM LUC LUC LUC
sake take	134 AWW MND ROM LUC
alone one	105 36 39 42 MND PP R3 ROM SHR LUC
offense thence	51 TGV LUC LUC
doom come	107 116 145 R2 LUC
go know	130 MFM R2
thee melancholy	1H4 45 E3
white night	12 1H6 SHR WIV LUC
kings things	115 LUC LUC VEN
breath death	1H6 2H4 99 CYM JN LLL MAC MND PP R2 R2 R2 R2 R2 R3 TN TRO TRO LUC LUC LUC VEN VEN VEN VEN
again slain	1H6 22 ROM VEN VEN VEN
stronger longer	28 LUC
ride tied	137 LC
forth worth	103 38 72 AWW HAM TRO LC VEN
broken open	61 VEN
behold cold	73 LUC
cost boast	91 E3
find wind	14 51 LLL MND PP PP LC
score more	122 LR MAC
fair air	21 70 LLL MAC MND PP TNK LUC VEN
was pass	49 WT
brand hand	111 154
know show	53 77 AYL CYM ERR H8 LLL MAC MAC MND MND PP R2 R2 LUC
name fame	80 PER R3 LUC
woe flow	30 H8
go grow	12 R2
shine mine	33 LLL LLL PP TN
away decay	11 64 80 LUC
woe so	127 129 3H6 71 90 ARD ERR E3 LLL MFM MFM R2 R2 R2 TIT LUC LUC LUC VEN VEN VEN VEN VEN
eternity memory	122 77

Appendix C – Shared Rhymes in Sonnets and Other Works

|	
me be	132 133 138 1H4 1H6 2H4 2H4 2H6 35 91 ADO ADO ADO ANT ARD ARD AYL AYL AYL AYL COR ERR E3 E3 H5 H5 H5 H5 HAM JC LR MND PER PP R2 R2 ROM ROM SHR TGV TGV TMP TN TRO TRO WIV WT WT WT LC LUC LUC LUC VEN
perpetual thrall	154 LUC
compare rare	130 21
make take	81 91 AYL LUC
all call	109 117 40 AWW MND PER
do too	88 CYM HAM LLL MND MND MND R2 ROM TRO WIV
light sight	38 7 LUC LUC
so know	13 140 1H4 ANT ERR ERR HAM HAM LLL LLL LLL LLL LLL LLL MND MND MND MND OTH R2 R2 R2 TIT TN LUC VEN
drink think	111 LLL
hair despair	99 LUC
feel steel	120 LUC VEN
blame shame	129 1H6 HAM R3 LUC LUC LUC LUC LUC
head bed	27 MND MND MOV ROM LUC LUC LUC
fashion passion	20 AYL LLL LUC
beguil'd child	59 MND
days praise	106 2 38 59 62 70 82 95 LLL LLL
head red	130 LUC VEN
hooks looks	137 ROM
contents monuments	55 LUC
move love	47 AYL ERR HAM HAM HAM LLL MND PP PP ROM TN TN VEN
tongue young	138 PP PP PP PP
love prove	10 117 136 151 153 154 32 39 72 AWW E3 HAM LLL LLL LLL MND PP PP PP ROM SHR SHR TGV TN TN TN TRO LUC VEN VEN
seen queen	96 MND MND MND PHT R2 R2 TN LUC VEN VEN
mud bud	35 LUC
alone gone	31 4 45 66 JN TGV TNK VEN
many any	10 LR VEN
me ye	111 42 H8 H8 LR TGV TGV TIM
true view	148 JN MOV LUC
make forsake	12 AWW LUC
pain gain	141 LUC
behold bold	131 MOV
woe know	50 LC LUC LUC
say lay	101 MAC LUC
store more	11 135 146 37 84 ROM LUC
thee see	18 27 3 43 95 99 AYL JN MOV MOV OTH TRO LUC LUC VEN VEN
behold old	22 LC LUC
pride side	144 151 E3 PP R2 R3 LUC
hell tell	144 2H6 PP LUC
loss cross	1H6 34 42
blind find	148 LUC
place grace	79 LLL LLL LR MAC MND PER R2 TMP LC LC LUC
grow show	69 83 93 MND LUC
heir fair	127 6 AWW CYM OTH ROM
deeds weeds	69 94
found ground	153 75 ERR PP
unjust trust	138 PP
book look	59 77 LLL LLL LLL LUC
well tell	103 14 ERR LR LR MND WT LC
thee thee	152 2H6 ADO ARD CYM CYM CYM CYM CYM HAM HAM JN LR LR OTH PP PP R3 R3 R3 R3 R3 TGV TGV TIM TIM TIT TMP TMP TNK LUC LUC LUC LUC LUC LUC VEN VEN VEN VEN VEN
love approve	147 70 LR MND
cold old	104 2 MOV LUC VEN
heart depart	109 MAC VEN
skill ill	150 66 91 LUC LUC
remedy eye	62 MND
made fade	54 CYM TMP
respect effect	36 85
you true	114 118 85 CYM ERR H8 H8 JN LLL LLL MFM MFM MFM MFM MND MND MOV MOV ROM TMP TN TN TN TRO TRO TRO TRO WT WT
dwell tell	84 89 93
may day	18 CYM HAM LLL LLL MAC MAC PP PP R2
away gay	68 PP
way stay	44 48 LLL R2 LC LUC LUC VEN VEN
day way	34 7 ERR ERR MND MND R2 WT LUC
breast best	110 3H6 MND
name same	108 76 CYM LLL LLL PHT LUC
repair fair	16 ERR LLL PHT TGV
devil evil	144 LLL LLL PP TN LUC LUC LUC LUC LUC
you untrue	113 72 AWW CYM JN LLL
catch latch	113 LUC

Appendix C – Shared Rhymes in Sonnets and Other Works

Period 1 Shared Rhyme Pairs
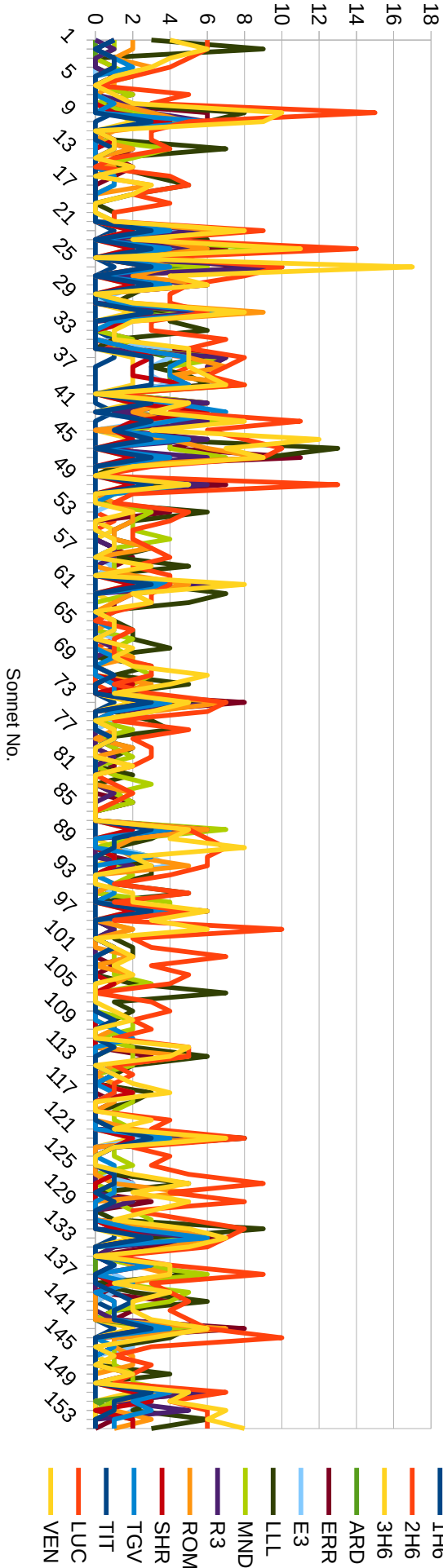
Period 2 Shared Rhyme Pairs

Sonnet No.

Appendix C
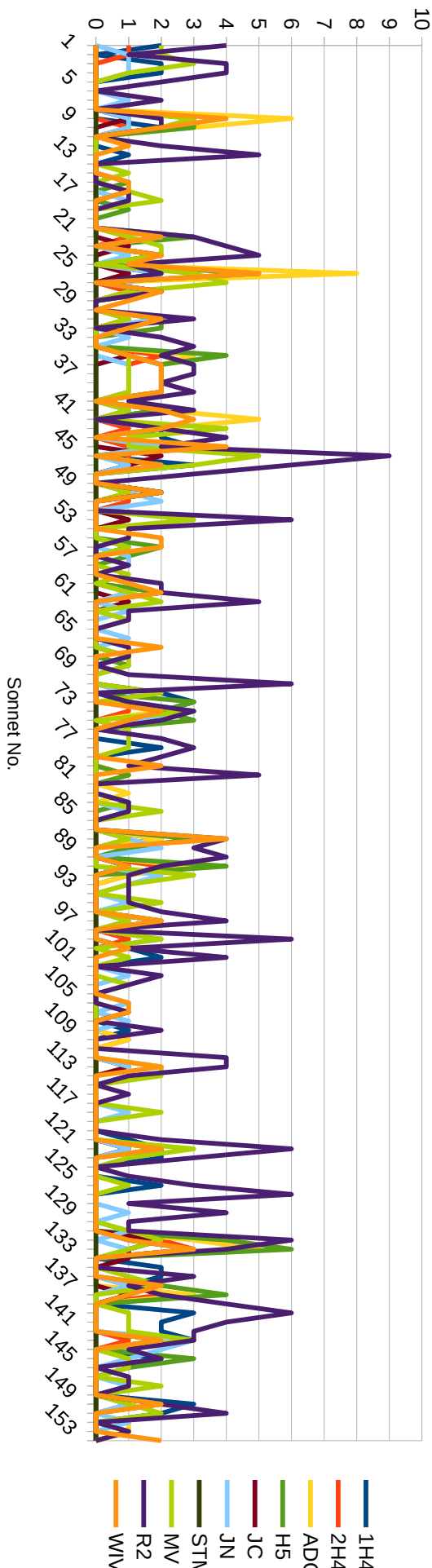
Period 3 Shared Rhyme Pairs
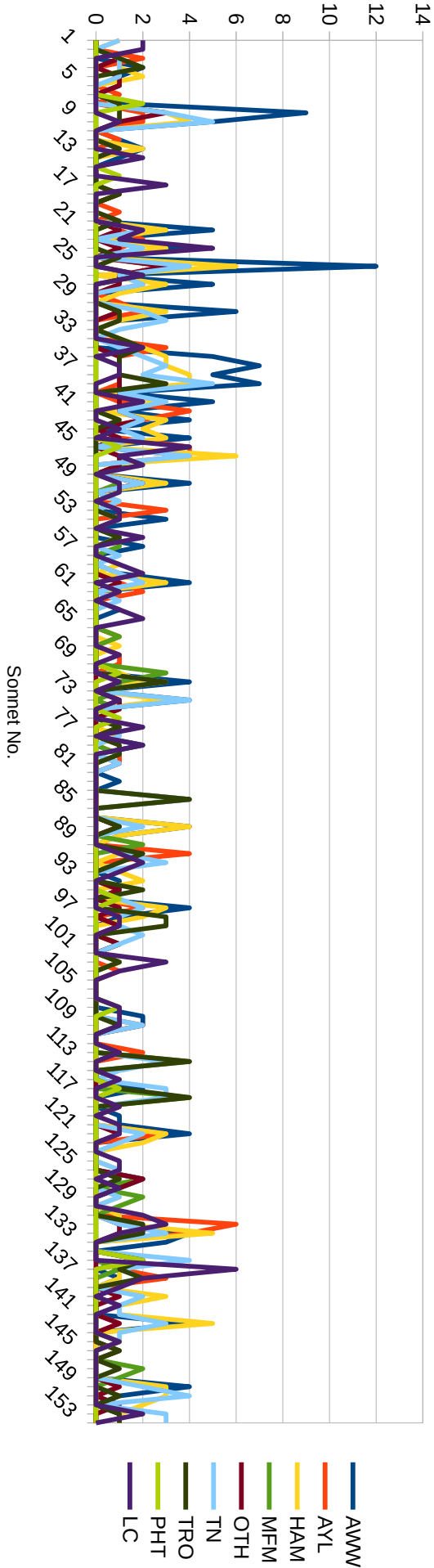
Period 4 Shared Rhyme Pairs

Sonnet No.

Sonnet No.

Appendix C
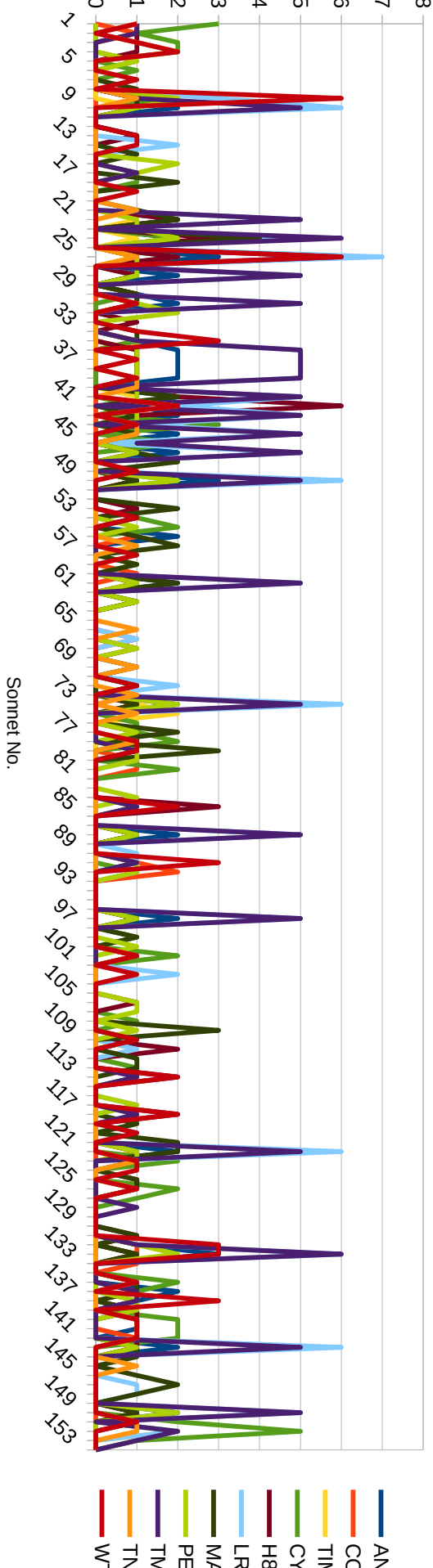
Shared Rhymes by Period
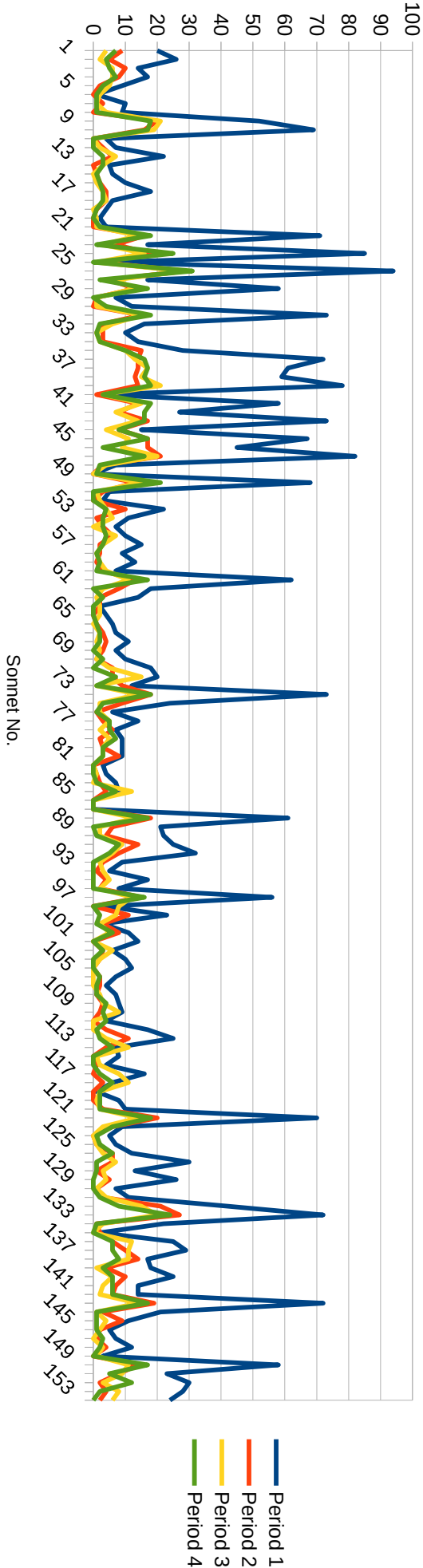
Sonnet No.

Ratio of Shared Rhymes in Sonnet to All Shared Rhymes in Period

Sonnet No.

Appendix C
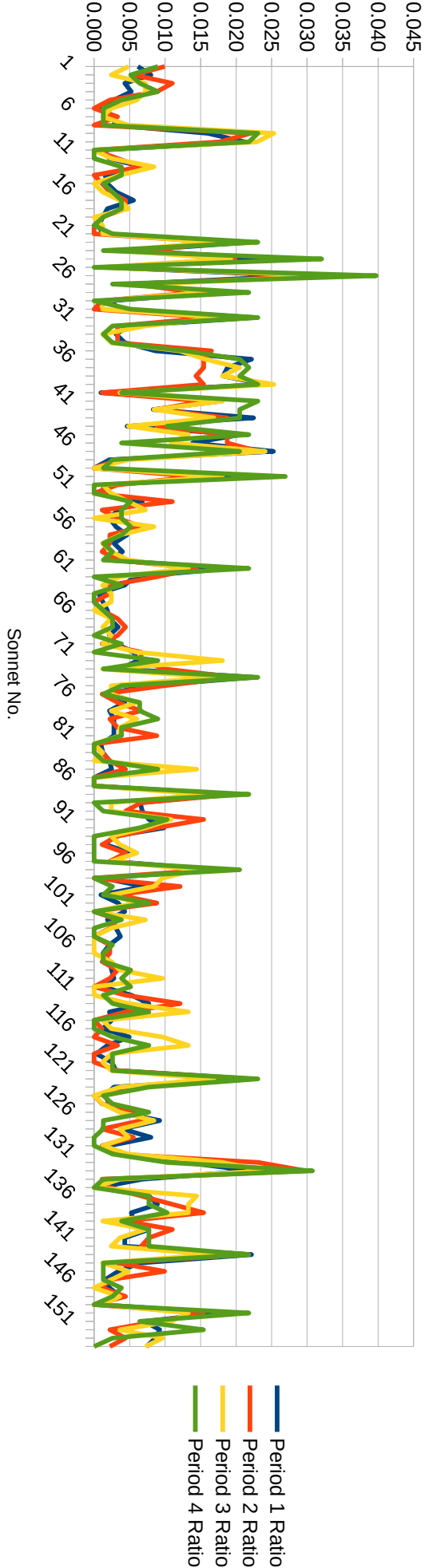
	Wells	Booth	Evans
1	ANT CYM LUC 1H4 VEN JC VEN	ROM AYL	VEN ROM TIT CYM VEN JN ROM VEN
2	1H4 1H4	TIM	VEN R2
3	LUC ANT MFM VEN LUC LC	WT LR CYM PER AWW ANT	VEN ROM LUC CYM
4		LLL MOV	MOV
5	VEN WIV	WIV	MND
6		JN	PER JN MAC
7	WT TIM	ANT	ROM ROM ROM ROM ROM ROM WT PP OTH PP R3 ROM TIM
8	AYL	MND AWW MFM AWW	MOV ROM
9		H5	
10			ERR TGV
11	JN	2H6 AYL	JN WT
12	STM TN	LLL	VEN HAM MND VEN
13	WIV	1H4 AYL LUC	MAC TRO AWW WIV
14			LLL ROM JN PHT
15	AYL TGV LUC WT	ANT H8 H5	LUC MOV LR TMP H8 R3
16	TIM MOV	MOV TIM	VEN LUC ROM ROM
17	TIM	1H4	MND JN
18	R2 JN 2H4 VEN	R2 AYL	MND LUC R2 ANT 1H6
19	TIT		PHT PHT TMP 1H6
20	VEN TRO 2H4	VEN 2H4 LR HAM	LC TRO TIT PER 2H4 TIT
21		HAM HAM HAM AYL PER LR ROM	JN LR ROM MOV MAC LLL
22			LLL R3
23	LLL VEN VEN LUC	MND AWW CYM LLL	COR HAM VEN LUC 1H4 CYM COR MND JN
24	LUC HAM	R2 SHR OTH SHR	TIT AWW JN R2 COR
25	VEN	R2 CYM LUC	ROM ROM H8 HAM JN 1H6
26	LUC	LUC R2 AYL H8 MFM HAM JC MND 2H4 H8 CYM	LUC LUC LUC E3
27	ROM	MAC TRO 1H4 TIT	ROM HAM JN ROM TIT HAM H5
28	LUC	TIT	VEN MND
29		1H4 ROM	JN CYM
30		JC AYL WIV	OTH 2H6
31	R2	VEN LLL HAM AWW TIT	TIT LC LC TIT
32	H5	TN	HAM VEN LUC
33	JN VEN OTH 1H4 HAM	LUC MOV HAM	ERR VEN JN E3 TGV 1H4 MAC MAC ROM ANT
34	LUC	ADO	LUC LUC JN LUC OTH TIM
35	LUC ANT	AYL R2 TGV WIV TIM LLL TGV ADO MAC TIM AYL	R2 LUC ROM TGV LUC JC
36	TRO	1H4 TRO TMP PHT MAC PHT	TRO VEN LUC
37	LR		LR 1H6
38	1H4	TRO HAM ROM	JN
39	R2 LUC	R2 ERR JC ANT	LUC ANT
40	ADO	MND LR TRO OTH ADO	TGV ANT
41	1H4 TIT OTH	LLL 1H4 TIT OTH	E3 1H6 TIT R3 OTH H5 JC

Appendix D - Editorial Citations in Sonnets

42	OTH	R2 JN 1H6 OTH	LLL
43		MAC WT R3 MFM LR LUC ERR ADO	MAC LUC
44	ERR H5 2H4	LLL 1H6 TIT OTH	H5 H5 ANT LC LC
45	MFM	PER TGV ROM	VEN TGV TGV JN
46	LUC 1H4 LUC	ROM ROM LLL HAM HAM JN R3	LUC LUC
47	LUC VEN	TMP MFM	ERR 2H6 LC JN
48	R3 VEN VEN	VEN LLL MOV TRO CYM TIM ROM LUC HAM R2 ROM VEN	VEN VEN VEN
49	MFM LUC	HAM JN 2H4 TIM WT H5 LR	JN JC
50		LLL MND ERR R3 AYL LLL VEN AYL H5 3H6 LR LR	R2 H5
51	VEN STM	TIM 1H6 R3 SHR H5 ERR TGV	H5 2H4 JN MAC VEN R2 JN E3 TGV
52	1H4	CYM TRO TIM LR MOV TGV R3 LUC	1H4 1H4
53	AYL	LUC VEN TMP AYL	VEN ANT
54		H5 VEN OTH ANT R2	ADO LLL PP
55	LUC	MND H8 AYL TRO	JN COR TRO
56	OTH	ERR MND MOV VEN	TN
57		HAM TGV LLL TN H8 LUC	LLL LLL
58		2H4 TGV AYL H8 AWW AYL TRO TIM TN	
59		HAM TN TMP MFM MND MND TGV AYL	LC
60		TRO WT	MOV LUC
61	LUC ADO HAM	VEN R3 MFM HAM	VEN VEN
62		LR SHR MND WT SHR LLL	JN LR 1H6 R3
63			ROM LLL
64	VEN	2H4 WT H8	2H4 R2 2H4 AWW
65		R3 R3 1H4 LC OTH	TRO
66	LUC HAM	CYM PER TMP 2H4	LUC MOV HAM CYM
67	H5	VEN TGV 1H4 MAC LUC R2 H5	TIM
68	MOV	LUC AYL ERR MOV	LUC TRO MOV TIT
69	LLL	HAM MND MAC TGV 1H4 OTH AWW	VEN
70		R2 MOV HAM	ERR
71		TMP	2H4 2H4
72		1H4 ADO 1H6	
73	CYM MAC	MAC	JN MAC CYM TIM CYM HAM MAC MAC
74	HAM	OTH HAM R3	LC HAM 3H6 LUC COR MAC 2H4
75	LUC	2H6 VEN HAM	VEN
76	ERR 1H4 CYM	OTH LUC PHT WIV H5 1H4 AYL	CYM JN
77	3H4 HAM HAM	1H4 ADO 1H4 HAM	3H6
78	LLL	MOV MOV LR AWW TMP	OTH LLL LLL 1H6
79		TIM MFM AYL TIT 2H4 AYL	
80	TRO	JC AWW MFM LR 1H4	VEN LUC MOV TRO

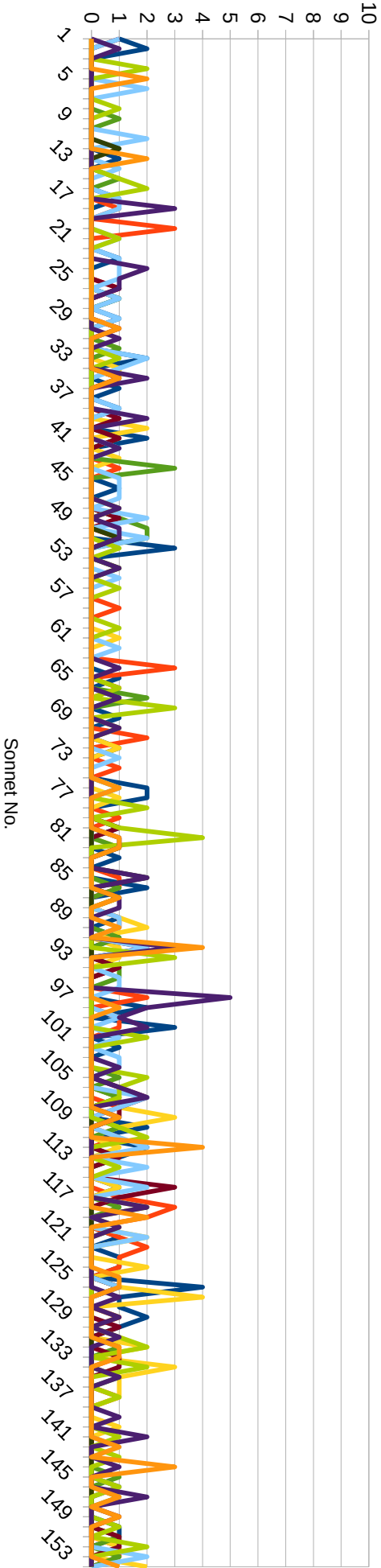| 81		MOV MFM WT MOV R2 TNK CYM TMP MOV	MOV HAM WIV
82	VEN R2 LUC	WIV VEN COR LUC H5 ADO LLL LLL	LUC LLL HAM LLL
83		1H4 HAM MND MFM TRO AYL AWW VEN	LLL LLL LR
84	MND	CYM LR MAC MND COR	
85	R2	TN TN R2 AYL PER HAM TGV OTH 2H4	STM STM COR MAC TRO
86	MND 2H4 1H4	1H6 MND 3H6 LUC H5 1H4	MND ROM HAM LUC
87	MOV	R2 WIV OTH AWW LLL TN OTH HAM JC	ROM
88		R2 HAM LR	2H6 OTH HAM LR MND
89	WT	AYL HAM LLL JN MND 1H6 VEN ROM 1H4	TGV TGV HAM ANT MND 2H4 ADO
90	ANT LUC LUC	TMP JN TGV R3 CYM SHR TRO 2H4 TIT WIV LUC	ANT ADO LUC ADO
91	H5	TGV SHR 1H4 COR PER	CYM
92	ERR	TGV MND TN WT TRO HAM TGV WIV WIV ERR JN JC R3 TRO WIV TN WIV VEN TGV JC H8 R2 TN ADO 2H6 TGV R2 CYM R2 VEN	LLL JN LUC
93		TN 2H6 ADO TMP HAM MOV MOV TIM 3H6	MAC CYM LLL MOV
94	CYM	H8 MND 1H4 2H4 1H6 JC VEN MFM E3	LC LLL COR H5 HAM E3 ROM LUC E3 LUC E3 E3 E3 E3
95		OTH TIT ROM LLL ANT ROM HAM H5 ROM	OTH ANT ROM JN
96		JN MND HAM	HAM LUC MND
97		2H4 WT R2 WT COR MND MND R2 2H4 ANT MND	R2 R2 MND R2
98	WT	1H4 OTH AWW LLL H8 WT PER 1H4 TIM R2 TGV CYM WT MND R2 R3 LUC LLL WIV	ROM VEN LLL OTH
99		TN MAC 2H4 R2 WT MND LLL SHR	VEN VEN JN 1H6 ROM
100	2H4 1H4	MND TIT R2 R3 CYM 1H4 TIM R2 LUC	1H4 TGV H8
101	MOV	SHR LUC TN VEN TN JN MOV COR TN ROM	
102			LLL WT LLL 1H6 TIT HAM LR ROM PP 1H4
103	LR	H8 ROM H8 LUC JN SHR VEN TGV LR	LR LUC
104	MFM	R2 SHR TRO JN AYL AYL	ROM MAC
105		PER HAM MOV TGV TMP H5 HAM LLL TGV AYL	ROM MOV
106	COR LUC	JC OTH MOV 3H6	R2 PP TN HAM 3H6
107	ANT LUC LR VEN	ADO HAM JC 1H4 LUC H5 LUC JN LR ANT MOV JN R2 JC COR OTH OTH TRO	HAM ANT LR LR 1H6 R2

Appendix D - Editorial Citations in Sonnets

108	2H4 VEN	HAM 3H6 ADO WT 3H6 WT HAM JN JC	
109	MAC	AWW LUC JC WIV ADO TN WT R3 LLL TGV WT ADO AYL ROM ERR R2 HAM ADO	MAC MAC LC
110	OTH	MOV AYL TRO H5 1H4 VEN SHR MFM MFM MND TN MAC R3 ROM 1H4 ADO MAC	TRO CYM
111		AYL 2H4 ERR WT LR ROM ERR 2H6 MOV HAM LR OTH ANT MOV	OTH CYM HAM JN HAM
112	HAM	2H6 LR LLL 1H6 MND WIV WIV 1H4 AWW MAC TRO WIV JN OTH TRO MFM WIV ADO LUC MFM R3 1H4 HAM R3	JN TRO LUC HAM TRO H8 PER
113	MAC	H8 OTH PER MAC JC AYL LUC 1H4	LUC HAM MAC MAC LUC H5
114	JN	COR MOV H8 OTH TMP LC	JN VEN TN
115		H8 R3 TN WT WT	ROM HAM HAM LLL
116	COR	MFM COR JC ADO TMP AWW HAM LR HAM JC LLL AYL	LR JN COR JC JN ANT 1H4 COR AWW
117		JC TIM MFM LLL WT	OTH 2H4 OTH CYM
118		HAM R2 R3 MND LR 2H4 MOV PP ROM R2 2H6 MAC 2H4 H5 MOV	2H4 HAM TRO TN ANT R3
119	LC	ERR HAM WT HAM AWW ANT ANT COR 1H6 2H4 WIV 2H6 WIV	LC LR HAM HAM 2H4 OTH
120	MAC	TMP LLL MOV TMP HAM CYM	LUC LUC OTH TRO R2
121	R3 LUC	LLL H8 JN 3H6 JN LUC LLL H8 1H4 ADO OTH 3H6 ANT 2H4 HAM ADO	HAM LC H8 LC 3H6 OTH HAM HAM
122	HAM	2H4 HAM HAM LR 2H6 AWW 2H4 AWW TIM TGV R3	HAM HAM
123		TN HAM 2H4 ERR 1H4	LUC
124	OTH	3H6 LR LR OTH TN COR ADO ADO SHR	2H4 COR LC ANT 1H4 TN COR MAC HAM
125		OTH HAM WIV MFM	OTH LC LC
126	1H4 MND AWW WT 1H4	TRO JC COR WIV TIM TRO H5 ANT 1H4 TMP 1H4 ADO TIT	WT JN LC WT HAM
127	MAC LLL R2 MND	LLL TGV AYL LUC ERR LLL 1H4 AWW LUC ERR ADO VEN	LLL ADO ADO ADO LLL MAC WT LLL OTH LLL LLL ANT
128	WT	ROM TIT TNK WT OTH WT ROM LLL COR R3 SHR 1H4 PER CYM MND	TIT SHR R3 WT OTH LLL

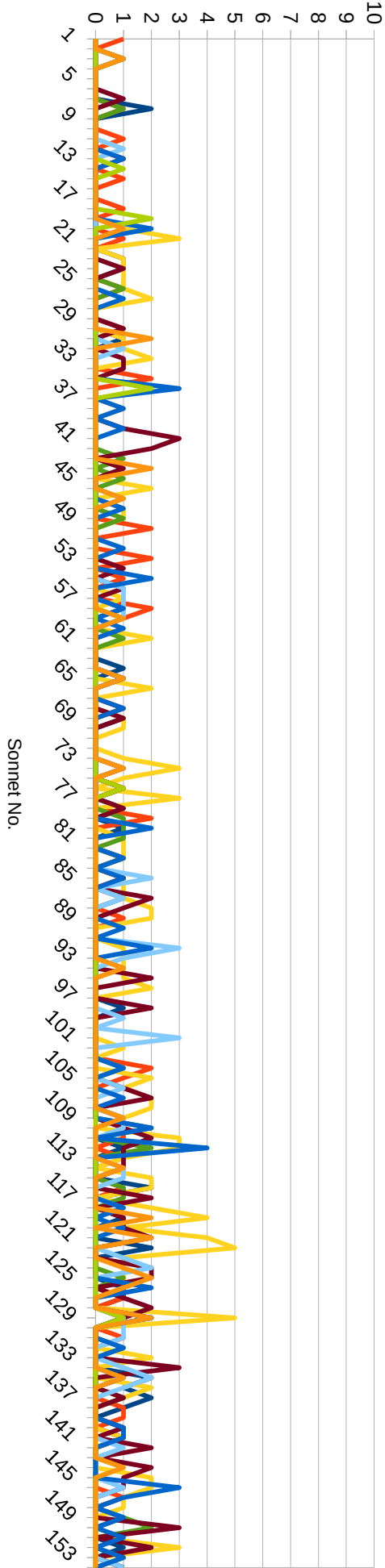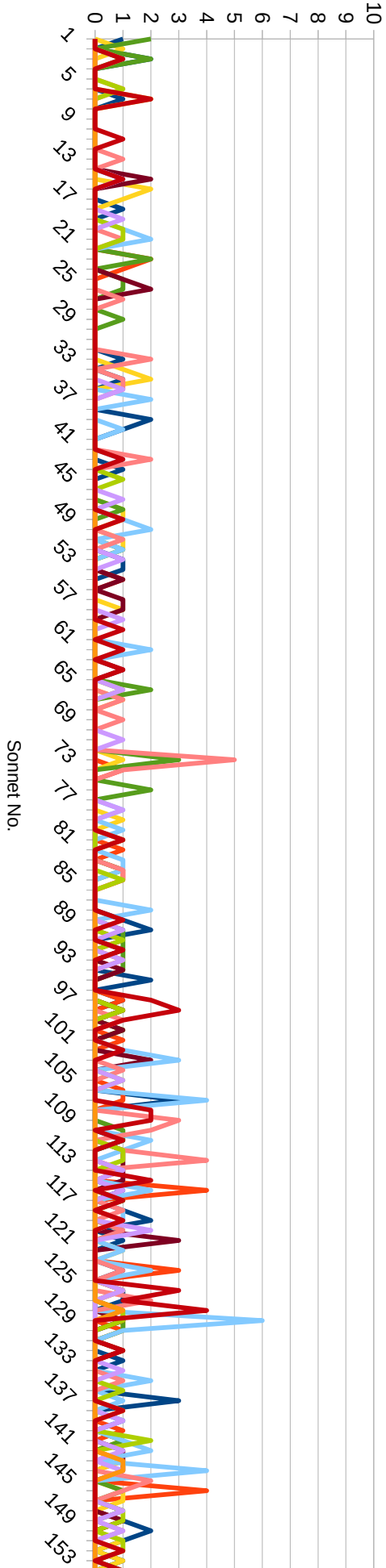|			
129	ROM LUC LUC	TRO VEN LR AWW 1H4 ANT LLL OTH HAM HAM MND MFM PER H8 ROM AYL AWW TN ADO VEN LUC 2H4 VEN CYM R2 LLL MND	LUC HAM TIM LR LR VEN LUC LUC LC HAM HAM LR LUC VEN LUC TRO LC 1H4 LR VEN LUC LR PHT
130	VEN 1H4 LUC VEN	TGV LUC CYM LR TN	VEN LUC COR JC TIM VEN VEN
131	TN	R2 MOV AYL ROM	ERR
132	MOV	LLL VEN ADO H5 TRO MOV SHR TIM WIV H8 VEN MFM WT ADO	ROM ROM
133	2H4	ROM JC HAM PP HAM ANT	MND WIV
134	MOV	OTH OTH OTH ADO TMP 2H6 ADO MOV TN LLL	JC ADO ROM
135		LUC TRO AWW LR LLL TIT MAC LUC 1H6 TN H5 R2 LC	AWW LR TIT ADO TN
136	TN LUC	HAM ADO AWW TGV LLL PER ROM HAM	
137	ANT	ROM AWW LR ANT ANT OTH PP VEN LLL 2H4 AWW MFM	ADO ROM MOV R3 HAM
138	PP LUC 3H4 AWW	PP LLL AYL ROM PP TIM PP MND PP PP WT PP MND PP	PP PP 2H6 LUC PP PP PP PP PP PP
139		TMP 1H6 ROM LLL AYL ANT 1H6 R2	ROM ROM LR
140	LUC COR	OTH MFM TRO ADO	
141	VEN CYM MFM MOV TIM	ROM TRO ERR PP HAM R2 AYL AWW LR 2H6 LLL PER	R2 PER LLL ROM
142		TMP JN OTH 1H6 H8 TGV VEN LR R3 LR SHR WIV TN	E3 LUC TGV VEN OTH
143	COR	TIM TNK MOV H8 3H6	
144	PP WIV	PP LC R2 H8 VEN OTH PHT PP PP LR TNK LR WIV	PP TMP WIV PP ROM PP OTH PP PP LR LR
145		H5 HAM MAC	LUC HAM MAC SHR LUC LUC LUC OTH
146	LUC OTH LUC TRO	HAM SHR ROM JN HAM TRO JC 2H6 TN MAC TRO TIM	ROM ROM LUC LUC COR LLL MOV CYM COR COR COR
147		LUC H5 TRO R2 TIM R2 LLL AYL	WIV LLL VEN LLL HAM
148		2H6 HAM TMP PER CYM	
149	JC PER AWW	ANT WIV LUC MFM	H8 ROM COR TRO
150	ANT MFM OTH	TMP MOV 1H4 MAC SHR MFM OTH	ROM ANT OTH
151	TGV	H5 R3 LLL MND SHR JC ROM LUC VEN MND ANT TIM PER	ROM ROM ROM LUC TRO 1H4
152	SHR	JC HAM AWW ROM LLL WT MOV OTH LLL 2H4 SHR ADO TGV MOV HAM	OTH LUC LLL LLL LLL PER HAM MAC
153	JN	TIM MFM VEN ANT H5 2H4 ROM	LUC LUC TRO JN VEN VEN VEN
154		AWW WIV OTH CYM WT ADO ADO TN	

Citations of Period 1 Plays in Sonnets

Sonnet No.

1H6 — 2H6 — 3H6 — ARD — ERR — E3 — LLL
R3 — ROM — SHR — TGV — TIT — LUC — VEN
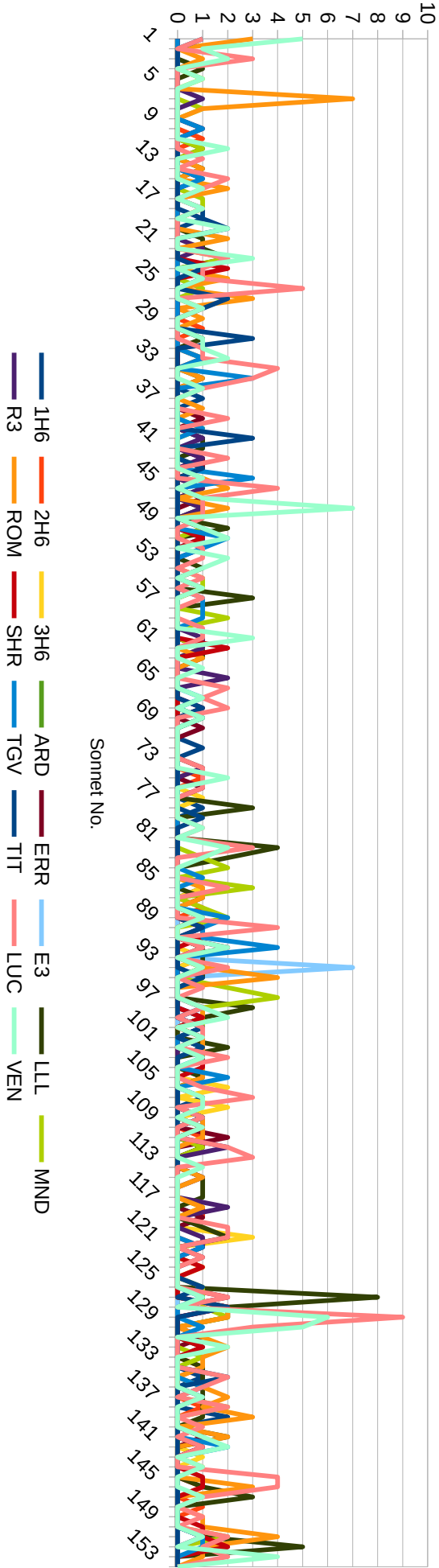
Citations of Period 2 Plays in Sonnets

Sonnet No.

1H4 — 2H4 — ADO — H5 — JC — JN — STM — MOV — R2 — WIV

Appendix D

Citations of Period 3 Plays in Sonnets

Sonnet No.

AWW —— AYL —— HAM —— MFM —— OTH —— TN —— TRO —— PHT —— LC

Citations of Period 4 Plays in Sonnets

Sonnet No.

ANT —— COR —— TIM —— CYM —— H8 —— LR —— MAC —— PER —— TMP —— TNK —— WT

Appendix D

Total Citations in Sonnets by Period

Sonnet No.

Period 1 ——— Period 2 ——— Period 3 ——— Period 4

Appendix D
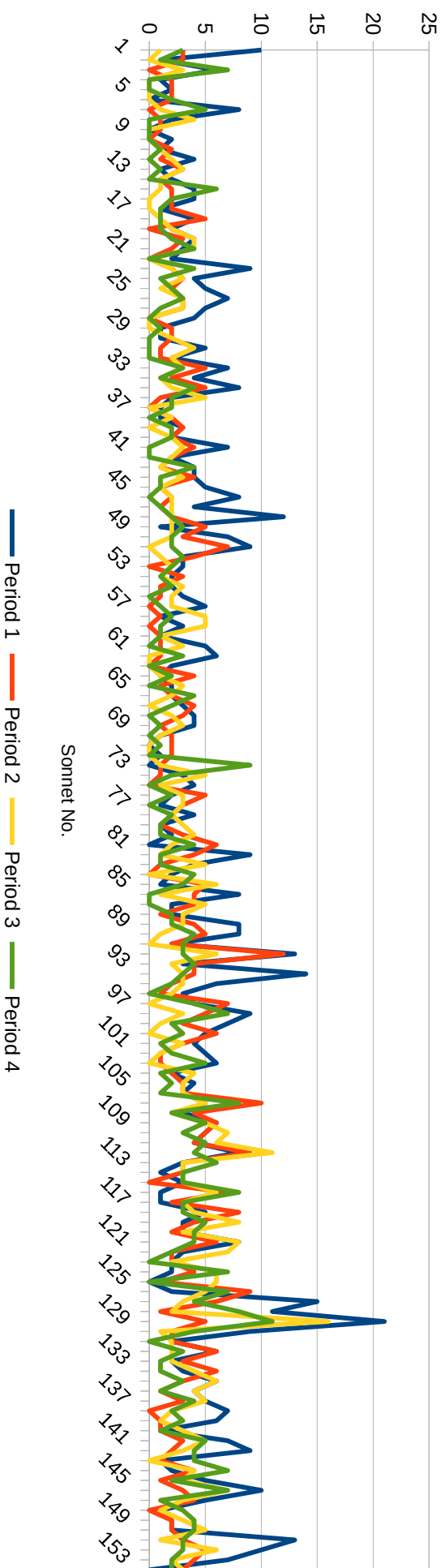
Dislegoma in Sonnets

essay	110 LR	steepy	63 TIM	maturity	60 TRO
couplement	21 LLL	augur	107 PHT	hardest	95 WIV
insconce	49 ERR	eas'd	28 R2	willfulness	117 H5
receivest	40 JC	receiv'st	8 8	believed	140 ARD
highmost	7 ROM	hied	153 VEN	restful	66 R2
unfather'd	124 2H4	hundreth	59 HAM	utt'ring	69 LUC
sickle	116 126	reproving	142 LUC	annex'd	99 ANT
enclose	95 TNK	vassalage	26 TRO	unfolding	52 OTH
anchor'd	137 R3	gainer	88 WIV	proved	116 MV
bow'r	127 1H4	featur'd	29 ADO	crowning	115 R3
enfeebled	86 H5	intitled	37 LLL	languish'd	145 WT
effectually	113 TIT	groans	131 VEN	intermix'd	101 R2
recounting	45 LR	composed	59 TGV	distillation	5 WIV
honoring	125 PER	imitated	53 HAM	polish'd	2H4 85
viewest	3 LLL	unthrifts	13 R2	entombed	81 LUC
world-without-end	57 LLL	new-appearing	7 R3	excusing	35 JN
stell'd	24 LUC	hearsay	21 ADO	re-survey	32 H5
play'st	128 H5	tottered	26 R2	deserv'st	39 WIV
loan	6 HAM	possessing	75 87	paws	19 TIT
strumpeted	66 ERR	loveliness	4 OTH	allege	49 TRO
equipage	32 WIV	hammered	120 WT	endeared	31 JN
oblation	125 LC	decrease	15 WIV	churls	69 CYM
accumulate	117 OTH	moiety	46 LUC	siren	119 ERR
glazed	24 R2	seething	153 MND	sheeds	34 LUC
twilight	73 ARD	cheered	15 R3	perfect'st	51 MAC
new-found	76 TGV	subscribes	107 TRO	bristly	12 VEN
graced	78 TGV	compeer	86 LR	engrossed	133 2H4
betraying	151 ARD	unrespected	43 54	onwards	126 ARD
gentlest	113 PER	wrongs	41 LR	tann'd	62 HAM
indigest	114 JN	int'rim	56 TN	perusal	38 HAM
alt'ring	115 WT	informer	125 VEN	youngly	11 COR
quicker	45 MND	gilding	20 33	shorn	68 WT
exchang'd	109 1H4	carcanet	52 ERR	describe	53 MV
o'ersways	65 HAM	sharp'st	115 H8	unused	48 OTH
issueless	9 WT	tend'red	120 VEN	clears	148 JN
expiate	22 R3	brav'ry	34 SHR	ingraft	15 OTH
impeach'd	125 R2	hoisted	117 ERR	quietus	126 HAM
befriends	120 TRO	acceptable	4 E3	staineth	33 VEN
incertainties	107 WT	releasing	87 VEN	blunting	52 LC
mus	85 MND	proving	2 LUC	extern	125 OTH
bett'ring	32 R3	steep-up	7 LLL	expired	27 ARD
afloat	80 JC	lameness	89 LR	pictur'd	24 CYM
unseeing	43 TGV	girded	12 H5	possesseth	62 JN
masked	54 ADO	masonry	55 AWW	spend'st	100 MV
widowed	97 COR	stick'st	10 MV	singleness	8 ROM
wiry	128 JN	soundless	80 JC	rud'st	113 CYM
destroys	9 ANT	mouthed	1H4 77	insufficiency	150 MND
ever-fixed	116 OTH	ingrafted	37 JC	alcumy	114 33
strained	2H4 82	unsway'd	141 R3	defaced	64 LUC
disdaineth	33 TIT	riches	87 94	impair	83 MND
		shallowest	80 MND	totter'd	1H4 2
		rhymer	38 ANT	hugely	124 AYL
		compile	78 LLL	dulling	103 TMP

Appendix E – Sarrazin Links in Sonnets and Other Works

convertest	11 E3	
eisel	111 HAM	

Trislegoma in Sonnets

deface	6 H5 MV
untold	136 LUC PER
tincture	54 JC WT
randon	147 ERR VEN
swerving	87 ANT H5
disabled	66 AYL MV
maketh	113 LLL LUC
dark'ning	100 H8 TRO
deem'd	96 E3 H8
budding	1H4 95 SHR
dreading	97 CYM JN
granting	87 MFM TGV
sheaf	12 AYL TIT
climb'd	3H6 7 JC
dressing	123 76 MFM
unswept	55 COR WIV
freezing	97 CYM LUC
rased	25 64 R3
sourly	35 41 COR
flatt'ry	114 TIT VEN
feed'st	1 SHR TIM
inviting	124 CYM OTH
brightness	150 ROM TRO
slavery	133 OTH TMP
put'st	134 3H6 LR
meadow	33 LLL TIT
princes'	25 2H6 MV
askaunce	110 LUC SHR
usher	132 ANT COR
forget'st	100 MFM TMP
show'st	126 COR LR
eyes'	137 LR ROM
wire	130 130 ANT
balmy	107 OTH OTH
expressing	105 MV TMP
robb'ry	2H4 40 99
self-loving	62 COR VEN
thinly	52 OTH ROM
remembered	74 ARD H5
withering	126 MND MNI
love-suit	136 CYM H5
shifting	20 LUC MND
outcast	29 2H6 SHR
profitless	4 ADO OTH
feel'st	2 TNK TNK
untutor'd	138 2H6 3H6
resty	100 CYM TRO
cured	118 2H4 LUC

universe	109 E3 H5
spending	76 AWW LLL
corrupting	35 H5 R3
ruinate	10 3H6 LUC
suppress'd	138 H5 LUC
fil'd	85 MAC WT
liker	16 JN LLL
rebuk'd	119 H5 MAC
gazeth	20 LUC VEN
unthrift	9 MV TIM
taker	129 ADO ROM
ordering	8 JN WT
alters	116 116 WT
selling	146 2H6 OTH
assailed	41 ARD JN
governs	113 MFM TN
dully	50 ADO TGV
invocate	38 E3 R3
vowing	152 TRO TRO
decrepit	37 LLL VEN
perfumed	1H4 54 ANT
toil'd	25 2H4 R2
crawl	60 LR MND
applying	119 LC VEN
sullied	15 1H6 E3
gazers	3H6 96 E3
riper	1 102 AYL
situation	128 2H4 H5
vengeful	2H6 99 TIT
accusing	2H4 58 ADO
dedicated	82 TIM TMP
pyramid	123 ANT MAC
wane	11 MND MND
tress	68 JN TNK
unlearned	138 AWW LLL
limping	66 ROM TIM
grow'st	11 126 18
new-fangled	91 AYL LLL
alien	1H4 78 MV
despising	29 COR TRO
unlettered	85 LLL LLL
divert	115 H5 TRO
abysm	112 ANT TMP
flattered	138 JC R3
o'erpress'd	139 COR PER
variation	1H4 76 H5
enlarg'd	1H4 70 HAM
preposterously	109 H5 WIV
hateth	149 2H6 MND
sympathiz'd	82 LLL LUC
purging	45 HAM HAM
builded	124 ANT LC

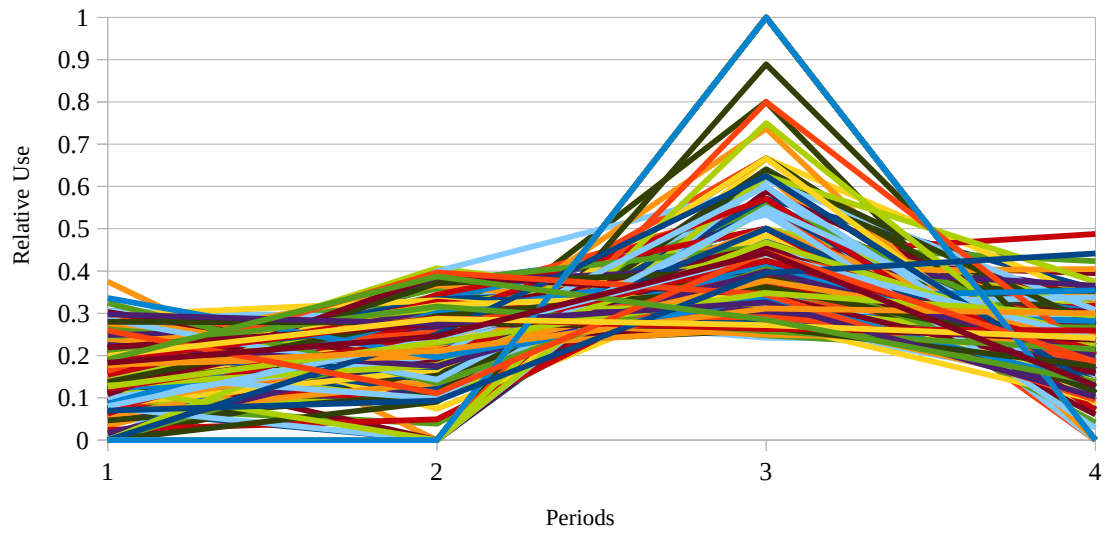Total Dislogmena in Sonnets: 158
Total Trislogmena in Sonnets: 100
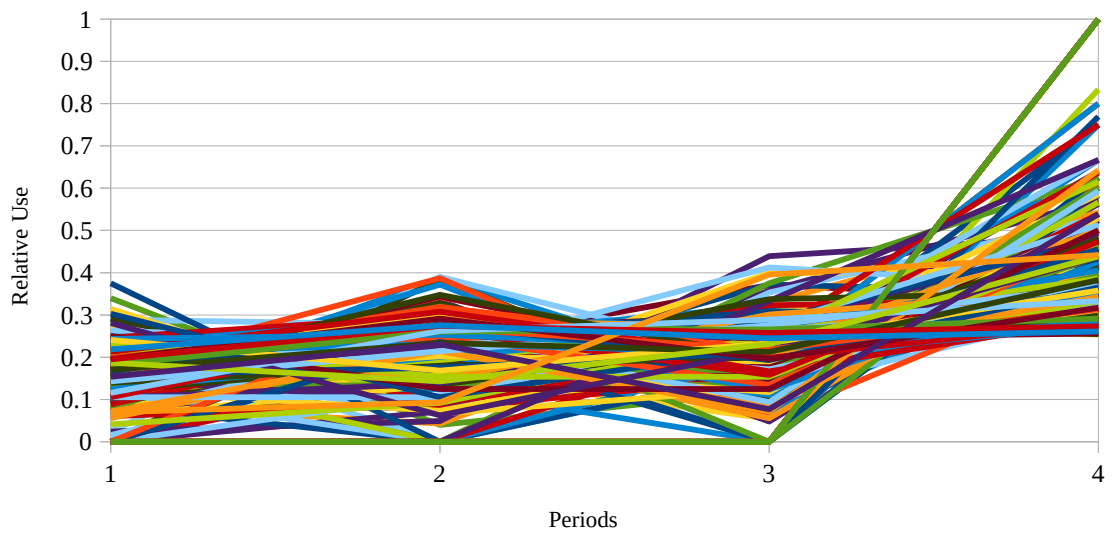
Frequencies of Period 1 Distinctive Words
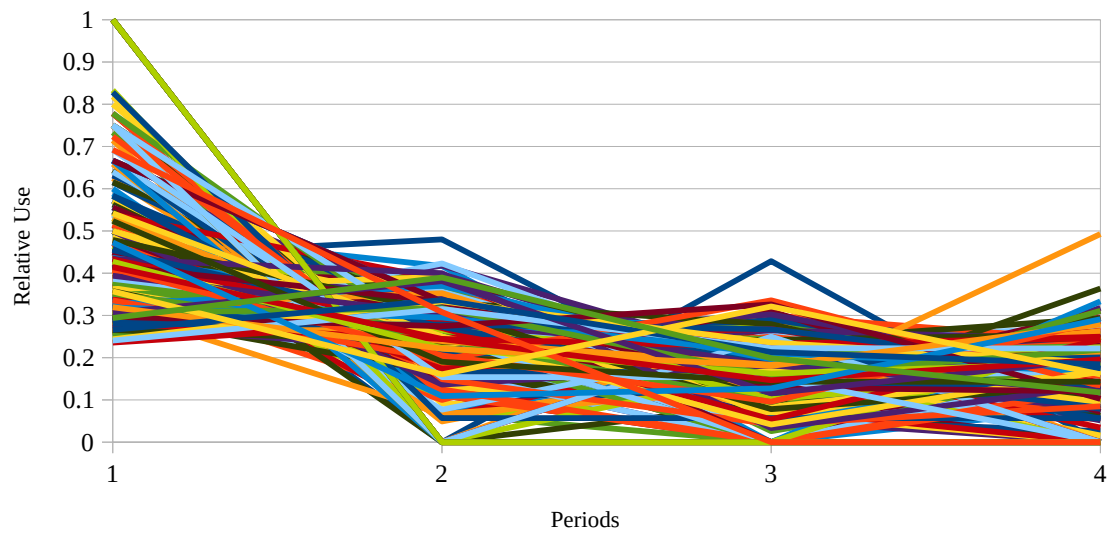
Frequencies of Period 2 Distinctive Words

Appendix F – Distribution of Distinctive Words over Periods
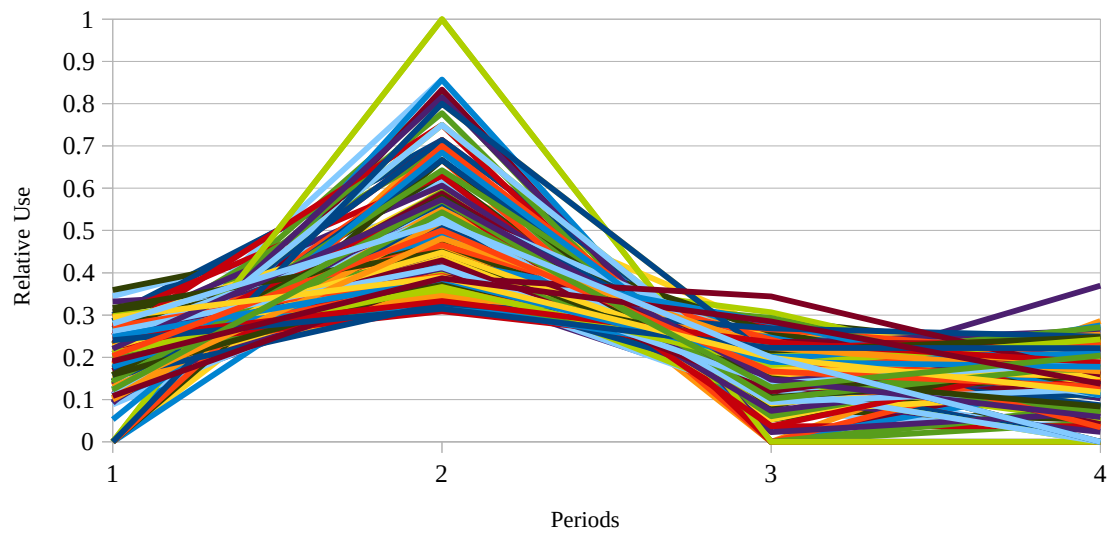
Frequencies of Period 3 Distinctive Words

Frequencies of Period 4 Distinctive Words

Appendix F – Distribution of Distinctive Words over Periods

Hieatt Zones

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154

Clustering on Universal Attributes (Letter, Digram and Word Frequencies)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154
Silhouette 0.3638
Inertia 997.2308

Clustering on Metrical Attributes (MLE, FE, OL, LE, WE, Col, Arch)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154
Silhouette 0.3546
Inertia 171.9548

Clustering on Shared Period Attributes (Shared N-grams, Shared Rhyme, Period Citation, Shared Sarrazin Links, Shared Distinctive Words)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154
Silhouette 0.3803
Inertia 241.7500

Clustering on Sarrazin Links Alone

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154
Silhouette 0.4482
Inertia 104.4367

Appendix G - Clusterings

Clustering on Metrical and Period Attributes

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154
Silhouette 0.3880
Inertia 269.8768

Clustering on All Attributes

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154
Silhouette 0.3570
Inertia 1099.8855

Works Cited

Antonia, Alexis, Hugh Craig, and Jack Elliott. "Language chunking, data sparseness, and the value of a long marker list: explorations with word n-grams and authorial attribution." *Literary and Linguistic Computing* 29, no. 2 (2014).

Arthur, David and Sergei Vassilvitskii, "kmeans++: The advantages of careful seeding" *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, Society for Industrial and Applied Mathematics (2007).

Booth, Stephen, ed. *Shakespeare's Sonnets*. New Haven; London: Yale University Press, 1977.

Bruster, Douglas, and Geneviéve Smith. "A New Chronology for Shakespeare's Plays." *Digital Scholarship in the Humanities*, December 2014.

Burrow, Colin, ed. *William Shakespeare: The Complete Sonnets and Poems*. Oxford: Oxford University Press, 2002.

Calle-Martín, Javier, and Antonio Miranda-Garía. "Stylometry and Authorship Attribution: Introduction to the Special Issue." *English Studies* 93, no. 3 (2012).

Cox, D. R., and L. Brandwood. "On a Discriminatory Problem Connected with the Works of Plato." *Journal of the Royal Statistical Society* 21 (1959).

Craig, D. H., and Arthur F. Kinney, eds. *Shakespeare, Computers and the Mystery of Authorship*. Cambridge: Cambridge University Press, 2012.

Duncan-Jones, Katherine, ed. *Shakespeare's Sonnets* (London: Arden Shakespeare 1997).

———. "Was the 1609 *Shake-speare's Sonnets* Really Unauthorized?" *The Review of English Studies* 34 no. 134 (May, 1983).

Eliot, T. S. "*Hamlet* and his Problems," *The Sacred Wood and Major Early Essays* (New York: Dover Publications 1998).

Elliott, Ward E., and Robert J. Valenza. "Oxford by the Numbers: What Are The Odds The The Earl of Oxford Could Have Written Shakespeare's Poems and Plays?" *Tennessee Law Review* 72 (Fall 2004).

Evans, G. Blakemore, ed. *The Sonnets*. Cambridge: Cambridge University Press, 1996.

Evans, G. Blakemore, and J. J. M. Tobin, eds. "The Electronic Riverside Shakespeare." Last modified 2014. Digital file.

Forsyth, Richard S., and David I. Holmes. "Feature-Finding for Text Classification." *Literary and Linguistic Computing* 11, no. 4 (1996).

Fineman, Joel, *Shakespeare's Purjured Eye: The Invention of Subjectivity in the Sonnets* (University of California Press, 1986).

Halliday, F. E. *A Shakespeare Companion*. London: Gerald Duckworth & Co., 1968.

Hieatt, Kent A., Charles W. Hieatt, and Anne L. Prescott. "When Did Shakespeare Write Sonnets 1609." *Studies in Philology* 52, no. 205 (2001).

Leslie Hotson, *'Shakespeare's Sonnets Dated' and Other Essays* (New York: Oxford UP, 1949).

Jackson, MacDonald P. *Determining the Shakespeare Canon: Arden of Faversham and Lover's Complaint*. Oxford: Oxford University Press, 2014.

———. "Punctuation of the Compositors of Shakespeare's Sonnets 1609." *The Library* 30, no. 1 (1975).

———. *Shakespeare's 'A Lover's Complaint':Its Date and Authenticity*, University of Auckland Bulletin 72, English Series 13 (Auckland, 1965)

———. "Vocabulary and Chronology: The Case of Shakespeare's Sonnets." *The Review of English Studies* 52, no. 205 (2001).

Jowett, John. Review of *The Riverside Shakespeare*, edited by G. Blakemore Evans. *Modern Language Review* 94, no. 4 (1999): 1079.

Kerrigan, John, ed. *The Sonnets and A Lover's Complaint* (London: Penguin, 1986).

Ledger, G. R. *Re-Counting Plato*. Oxford: Oxford University Press, 1989.

Ledger, G. R., and T. V. N. Merriam. "Shakespeare, Fletcher, and the Two Noble Kinsmen." *Literary and Linguistic Computing* 9 (1994).

Mosteller, Frederick, and David Wallace. "Inference in an Authorship Problem." *Journal of the American Statistical Association* 58, no. 302 (1963).

Mosteller, Frerderick, and David Lee Wallace. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. New York: Springer-Verlag, 1984.

Pedregosa *et al.* "Scikit-learn: Machine Learning in Python" *Journal of Machine Learning Research* 12 (2011).

Sarrazin, Gregor. "Wortechos bei Shakespeare." *Shakespeare Jarhbuch* 33 (1897).

Slater, Eliot. *The Problem of the Reign of King Edward III: A Statistical Approach*. Cambridge: Cambridge University Press, 1988.

———. "Shakespeare: Word Links between Poems and Plays," *NQ* 220 (1975).

Spevack, Marvin. *Complete and Systematic Concordance to Shakespeare*. Hildesheim: G. Olms Verlagsbuchhandlung, 1968.

Stamatotos, Efsanthios. "A Survey of Modern Authorship Attribution Methods." *Journal of the American Society for Information Science and Technology* 60, no. 3 (2009).

Tweedie, F. J., and D. I. Holmes. "Neural Network Applications in Stylometry: The Federalist Papers." *Computers and the Humanities* 30, no. 1 (1996).

Ule, L. "Recent Progress in Computer Methods of Authorship Attribution." *ALLC Bulletin* 10 (1982).

Vendler, Helen. *The Art of Shakespeare's Sonnets*. Cambridge, MA: Harvard University Press, 1997.

Vickers, Brian. "Incomplete Shakespeare: Or, Denying Coauthorship in 1 Henry VI." *Shakespeare Quarterly* 58, no. 3 (2007): 311-52.

———. *Shakespeare: Co-author: A Historical Study of Five Collaborative Plays*. Oxford: Oxford University Press, 2002.

Wells, Stanley and Gary Taylor with John Jowett, and William Montgomery. *William Shakespeare: A Textual Companion*. Oxford: Clarendon Press, 1987.

Wentersdorf, Karl. "Shakespearian Chronology and the Metrical Tests." *Shakespeare-Studien Festschrift für Heinrich Mutschmann*, 1951.