Chapter 4

Stylometry

The contemporary audience for Gratian's *Decretum* seems to have been reasonably satisfied that a single author was responsible for the collection as they knew it, even if they knew so little about the historical Gratian that they felt the need to retroactively provide a largely fictional backstory for the eponymous author. Modern students of Gratian have been willing to entertain the possibility that the *Decretum*, at least in its final most widely-circulated form, was the product of collective authorship. Stephan Kuttner, as the first item on his 1984 agenda for Gratian studies, asked:

> was [the *Concordia discordantium canonum*] drafted and completed in one grandiose thrust, or did the original version go through successive redactions? And if the latter, was it Gratian himself, or Gratian with his disciples, or an early generation of canonists after him, who completed the final recension which from the mid-twelfth century on was used in the schools and in adjudging cases?[1]

Anders Winroth's 1996 discovery of the first recension of the *Decretum* provided a convincing answer to the first part of Kuttner's question: the original *did* go through

---

[1] Stephan Kuttner, "Research on Gratian: Acta and Agenda," in *Studies in the History of Medieval Canon Law*, Collected Studies CS325 (Aldershot, Hampshire, Great Britain : Brookfield, Vt., USA: Variorum ; Gower, 1990), 10.

successive redactions. Winroth used the name Gratian 1 to refer to the compiler or compilers of the first recension, and Gratian 2 to refer to the compiler or compilers of the second recension. He then reformulated the second part of Kuttner's question by asking whether Gratian 1 was the same person as Gratian 2. Over the last twenty years, Winroth's answer to the question posed in this form, that Gratian 2 was *not* the same person as Gratian 1, has been the focus of vigorous but so far inconclusive scholarly debate. Much of the debate has focused on competing evaluations of the Sankt Gallen Stiftsbibliothek 673 (Sg) manuscript of the *Decretum*, and more recently on secondary evidence, such as a twelfth-century Siena necrology in which the name Gratian appears.

It does not appear that the debate over the authorship of the *Decretum* can be settled on the basis of the currently available evidence. The goal of my dissertation project has been to find new evidence relevant to the question of the authorship the *Decretum* as it has been posed by both Kuttner and Winroth, by using computational stylometric methods to analyze the authorship of the *dicta* traditionally attributed personally to Gratian.

Preliminary observations

Stylometry is the measurement of style. "Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively."[2] While style has both qualitative and quantitative aspects, stylometry is concerned only with quantitative aspects of style. One well-established use of stylometry is to attribute authorship and for the purpose of authorship attribution, the formal linguistic features that stylometry measures are (typically) the frequencies of occurrence of common words.

Linguists draw a distinction between function words and content words. The more frequently a word occurs in a language, the more likely it is to be a function word than a content word. Function words are words like prepositions and conjunctions. Content words are words like adjectives, nouns, and verbs. Function words convey meaning by their use in grammatical structure. The Latin conjunction "*sed*" does not mean anything by itself. Rather it places two words or grammatical constructs into an adversative relationship with each other.

---

[2] J. Berenike Herrmann, Karina van Dalen-Oskam, and Christof Schöch, "Revisiting Style, a Key Concept in Literary Studies," *Journal of Literary Theory* 9, no. 1 (2015): 44.

Another way of thinking about the distinction between function and content words is to note that in a given language, function words constitute a closed class, while content words constitute an open class. Language-speaking communities can and do make up new adjectives, nouns, verbs all the time; content words are therefore an open class that can be added to at will. But new prepositions and conjunctions are almost never added to a language. Their usage changes very slowly over time, if it changes at all, and function words are therefore, for all practical purposes, a closed, finite, class.

Evidence from experimental psychology suggests that readers, and perhaps also writers, process function words at a subconscious level.[3] The frequency with which a given author uses particular function words is therefore considered to be more or less invariant, making it a reliable authorial signature.

Successes and Failures

Historians and literary scholars have long (if imprecise) memories about poorly-conceived attempts at authorship attribution (e.g., Tony Honoré's attempt to attribute Roman law texts to Ulpian, and R.A. Cooper and D.A. Pearsall's attempt to attribute

---

[3] Mike Kestemont, "Function Words in Authorship Attribution From Black Magic to Theory?" in *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLfL)* (Gothenburg, Sweden: Association for Computational Linguistics, 2014), 59–66.

authorship of the Gawain Poems).[4] Conversely, successful, methodologically rigorous, attributions of authorship are quickly forgotten when the author or the text in question is uninteresting to literary academia (e.g., Patrick Juola's successful attribution of *The Cuckoo's Calling*, a 2013 crime fiction novel, to J.K. Rowling of Harry Potter fame.)[5]

Bruce W. Frier of the University of Michigan Law School negatively reviewed Tony Honoré's *Ulpian*(1982) for its misuse of stylometric analysis for authorship attribution. Honoré generated lists of what would now be called Most Distinctive Words (MDWs) and Statistically Improbable Phrases (SIPs) from a concordance of the works of Ulpian, and used them to attempt to periodize and to determine authorship of texts attributed to Ulpian.

---

[4] My point is that at your defense a central point will be: "OK, Paul, you did an enormous amount of work inputting all the data, how can you and we trust what a computer program is telling us?" It would be very good if you can tell us about other medieval authors about which there is some consensus that the computer has given us firm conclusions. - Ken Pennington, 13 July 2020

[5] Patrick Juola, "The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions", *Digital Scholarship in the Humanities*, Vol. 30, Supplement 1, 2015.

The A.S.G. Edwards essay in the July 3, 2020 issue of *The Times Literary Supplement (TLS)*[6] appears to refer to "The Gawain Poems" by R.A. Cooper and D.A. Pearsall.[7]

Cooper and Pearsall conclude that the stylometric evidence they evaluate is consistent with common authorship for the four poems preserved in British Library MS Cotton Nero A.x (*Sir Gawain and the Green Knight*, *Pearl*, *Patience*, and *Cleanness*). Some of the stylometric measures they employ are only relevant to poetry, focusing specifically on the alliterative features of the texts. Cooper and Pearsall also analyze the frequencies of 11 function words, distinguishing between words occurring at the beginning of a line and words occurring in the middle of a line. (John Burrows and David Hoover in a series of papers starting in 2000 established 30 words as a best-practice minimum for this kind of word frequency analysis.) It is not, for the most part, very methodologically sophisticated work, even by the standards of 1988. The one exception is their analysis of the co-occurrence of parts of speech (POSs) within lines, which anticipates the kind of part of speech bigram and trigram analysis that is sometimes performed today when appropriately tagged electronic texts are available. (An example of part of speech

---

[6] A. S. G. Edwards, "Go Little Books: More Problems of Early Modern Attribution," *TLS. Times Literary Supplement*, no. 6118 (2020): 9.

[7] R. A. Cooper and D. A. Pearsall, "The Gawain Poems: A Statistical Approach to the Question of Common Authorship," *The Review of English Studies* 39, no. 155 (1988): 365–85.

bigram analysis would be to analyze the frequencies of occurrence of noun-adjective pairs versus adjective-noun pairs in Latin texts.)

Google Scholar indicates that the Cooper and Pearsall article has 12 citations as of August 2020, so it does not appear to have had much influence on the subsequent development of scholarship within its subfield. (To put that number in context, Google Scholar indicates that Pennington's "Bartolome de las Casas and the Tradition of Medieval Law" has 86 citations.)

This email is resumptive of our conversation of August 3 about recent attempts to apply stylometric methods to attribute medieval texts. Ken suggested citing attributions that have achieved some threshold of scholarly acceptance. Here's another one that I got from a conversation this week with Mike Kestemont, the University of Antwerp researcher whose work is a model for what I'm doing.

The text is a set of Latin love letters between an anonymous man and woman (Vir and Mulier or just V. and M.), extant in a single manuscript that was copied c.1470 under the title Ex epistolis duorum amantium. Constant Mews in The Lost Love Letters of Heloise and Abelard (2001) attributed the letters to Abelard and Heloise. Jan Ziolkowski of Dumbarton Oaks contested Mews's attribution, partly on stylometric grounds, in:

Ziolkowski, Jan M. “Lost and Not Yet Found: Heloise, Abelard, and the”Epistolae Duorum Amantium”.” The Journal of Medieval Latin 14 (2004): 171-202. (http://www.jstor.org/stable/45019598).

The stylometric component of Ziolkowski’s argument consists of a comparison of the relative frequencies of the function words autem, igitur, ergo, ita(que), quia, and quippe as they appear in Historia Calamitatum, Letters 3 and 5, and the letters attributed to Vir in the EDA collection. Ziolkowski’s rebuttal to Mews’s attribution is perhaps more relevant to my work than some of the other attributions we’ve discussed, in that it is a negative attribution. Like my claim that the case statements in the Decretum were not written by author(s) of the dicta, Ziolkowski is arguing that the texts were not written by the putative author. Furthermore, Ziolkowski’s negative attribution was made on the basis of function words drawn from discontinuous and relatively small samples of medieval Latin.

In keeping with the previous emails in this series, I’ll note that Google Scholar shows that Ziolkowski’s article has been cited 23 times, considerably less than Ken’s “Bartolomeo de las Casas and the Tradition of Medieval Law,” but better than the Cooper and Pearsall paper on Gawain that started this conversation.

The Federalist (Hamilton and Madison)

Stylometric analysis of the frequencies of common words for the purpose of attributing authorship has had a number of notable successes. The validity of this approach for textual scholarship was firmly established by the work on the *Federalist Papers* by Frederick Mosteller and David L. Wallace. The authorship of 12 of the *Federalist Papers*, 49-57 and 62-63, had been disputed since the early 19th century, with competing claims advanced on behalf of Alexander Hamilton and James Madison.[8] In 1944, Douglass Adair, using traditional scholarly methods, settled the dispute largely to the satisfaction of early American historians, determining that Madison was the author of all 12 of the disputed numbers.[9] In 1964, Mosteller and Wallace confirmed Adair's findings by conducting a stylometric analysis of the frequencies of 70 function words to compare the 12 disputed numbers with numbers securely attributed to Hamilton and Madison.[10]

---

[8] Frederick Mosteller and David L. Wallace, *Inference and Disputed Authorship: The Federalist*, Addison-Wesley Series in Behavioral Science: Quantitative Methods (Reading, Mass: Addison-Wesley Pub. Co, 1964), 14. See also Douglass Adair, "The Authorship of the Disputed Federalist Papers," *The William and Mary Quarterly* 1, no. 2 (1944): 104.

[9] Adair, "The Authorship of the Disputed Federalist Papers" and Douglass Adair, "The Authorship of the Disputed Federalist Papers: Part II," *The William and Mary Quarterly* 1, no. 3 (1944): 235–64.

[10] Mosteller and Wallace, *Inference and Disputed Authorship*.

Definition of first- and second-recension *dicta*

In the absence of good modern critical editions for the first and second recensions of the *Decretum*, a proxy for the first recension must be created using the first-recensions variants from Friedberg's text reported in the appendix to Winroth's *The Making of Gratian's "Decretum"*, "The contents of the first recension of Gratian's *Decretum*.[11] The proxy for the first recension can then be subtracted from Friedberg's text, leaving the text added in the second recension as the difference. The text samples from the first and second recensions of the *Decretum* that provide the basis for authorship attribution are built up by iteratively appending short units of text (the individual first- and second-recension *dicta*) that are non-contiguous in their original context in the *Decretum*.

Many of the examples in this chapter will distinguish between first- and second-recension *dicta*, so this is an appropriate point at which to introduce an explicit definition for the way in which those terms will be used in the following discussion. Because almost every word in the first-recension *dicta* corresponds to a word in the second-recension *dicta*, we could consider the first-recension *dicta* to be a subset of the second-recension *dicta*, and conversely, we could consider the second-recension *dicta* to be a superset of the first-recension *dicta*. While true enough from a commonsensical

---

[11] Anders Winroth, *The Making of Gratian's Decretum* (Cambridge: Cambridge University Press, 2000), 197–227.

point of view, this is not a useful definition for the kinds of questions we would like to answer, such as whether Gratian 1, the author of the first-recension *dicta*, is the same person as Gratian 2, the author of the second-recension *dicta*.

Instead, for the purpose of the following analyses, the second-recension *dicta* are defined as the ordered set of every word from the *dicta* in the text of Friedberg's 1879 edition of Gratian's *Decretum* for which there is not a one-to-one correspondence to a word in the first-recension *dicta* as defined by Anders Winroth's appendix "The Contents of the First Recension of Gratian's *Decretum*."[12] An alternative restatement would be to define the second-recension *dicta* as the difference left by subtracting all of the words of the first-recension *dicta* as defined by Winroth's appendix from the text in the Friedberg edition.[13]

This definition is implemented by passing sequentially through the *dicta* and applying three rules. First, if a *dictum* is listed in Winroth's appendix as being in the first

---

[12] Winroth, 197–227.

[13] Hence, for the purposes of this study, the text of the *dicta* of the second recension is the remainder of the *dicta* of the Friedberg text after the text of the *dicta* of the first recension has been subtracted. To compare the *dicta* of the two recensions using stylometry, we need samples of sufficient length. Subtracting the *dicta* of the first recension from the *dicta* of the Friedberg text produces two adequate samples. It is significant to note that the two sample texts do not have be continuous pieces of writing, such as the *Federalist Papers* are. The method only requires samples of sentences and phrases of sufficient length. (See below.)

recension of the *Decretum*, and as not having been added to or changed in the second recension, the text for that *dictum* is included in the first recension sample. This rule is applied on a per-*dictum* basis. Second, if a *dictum* is in the text of the Friedberg edition and is not listed in Winroth's appendix as being in the first recension, in either unmodified or modified form, the text for that *dictum* is included in the second recension sample. This rule is applied on a per-*dictum* basis. Third, if a *dictum* is listed Winroth's appendix as being in the first recension, but as having been added to or changed in the second recension, those words indicated by the appendix are included in the first recension sample, while those words in the text of Friedberg not corresponding to the words indicated by the appendix are included in the second recension sample. This rule is applied on a word-by-word basis.

Take D.54 d.p.c.23 as an example. The complete text of the *dictum* as it appears in the Friedberg edition (column 214) is:

> *Ecce, quomodo serui ad clericatum ualeant assumi, uel quomodo non admittantur. Liberti quoque non sunt promouendi ad clerum, nisi ab obsequiis sui patroni fuerint absoluti. Unde in Concilio Eliberitano:*

Winroth's appendix indicates that only the first sentence of the *dictum* appears in the first recension:

d.p.c. 23: **1** *Ecce quomodo serui* – **2** *quomodo non admittantur.*[14]

Therefore, "*Ecce, quomodo serui ad clericatum ualeant assumi, uel quomodo non admittantur.*" is included in the first recension text sample, and "*Liberti quoque non sunt promouendi ad clerum, nisi ab obsequiis sui patroni fuerint absoluti. Unde in Concilio Eliberitano:*" is included in the second recension text sample.

Note that the individual *dicta* are too short for direct analysis by the techniques discussed in this chapter.[15] The smallest unit of Latin prose for which computational stylometry works is about 2,500 words.[16] The longest first-recension *dictum* (*de Pen*. D.1 d.p.c.87) is 1,591 words, and the longest second-recension *dictum* (C.7 q.1 d.p.c.48) is 692 words. As a result, a first-recension sample long enough to be useful for the purpose of stylometric analysis has to be created by rolling up or concatenating the first-recension *dicta* as they occur sequentially but discontinuously throughout Gratian's text. The

---

[14] Winroth, *The Making of Gratian's Decretum*, 201. The numbers 1 and 2 refer to line numbers relative to the first line of the *dictum*, as opposed to the first line of the column, in the print version of the Friedberg edition.

[15] This paragraph may have to be moved to a separate section explaining the rationale for the various roll-ups of the *dicta* used in the following analyses, e.g., Gratian0, Gratian1, dePen, Gratian2, etc.

[16] Maciej Eder, "Does Size Matter? Authorship Attribution, Small Samples, Big Problem," *Literary and Linguistic Computing* 30, no. 2 (June 2015): 171.

corresponding second-recension sample is created in the same manner by concatenating second-recension *dicta* in the order of their occurrence in the *Decretum*.

Both the text of Friedberg's 1879 edition of Gratian's *Decretum* and Winroth's appendix are open to criticism. Although modern scholars admire Friedberg's learning and energy—the 1879 edition of the *Decretum* was only one of many such projects that he undertook—his editorial standards were those of 140 years ago. In particular, Friedberg's selection of eight German manuscripts considered unrepresentative by modern scholarship as the basis for his edition, and his particular reliance on two of them—Köln Erzbischöfliche Diözesan- und Dombibliothek 127 (Ka) and 128 (Kb)—are seen today as serious deficiencies.[17]

Winroth himself acknowledged the provisional nature of his appendix.[18] Furthermore, Pennington has pointed out that although Winroth's appendix includes D.100 d.a.c.1,

---

[17] Winroth, *The Making of Gratian's Decretum*, 9–11. See also Stephan Kuttner, "De Gratiani opere noviter edendo," *Apollinaris* 21 (1948): 118–28 (Latin), and Kuttner, "Research on Gratian," 10, 21–22, which mentions the deficiency of Friedberg's edition without offering a detailed critique.

[18] "The list is based on a collation of *incipits* and *explicits* of every canon and *dictum* in the first recension. Differences within the texts may very well have been overlooked, and minor differences have not normally been registered." Winroth, *The Making of Gratian's Decretum*, 197.

D.100 c.1, and D.101 d.p.c.1, in the Paris (P), Florence (Fd), and Barcelona (Bc) manuscripts, the text of the first recension ends with D.99 c.1.[19]

Nevertheless, in the absence of a critical edition for the first recension, applying the variants recorded in Winroth's appendix as a set of transformations to the text of Friedberg's edition to generate a stand-in or proxy for the text of the first recension is a workable approach.[20] This method is well-theorized in Digital Humanities as "deformance." The term, proposed by Lisa Samuels and Jerome McGann in "Deformance and Interpretation" (1999), conflates the words "deformation" and "performance" and describes a process through which a text is transformed by the application of a series of deformances to generate a "paratext."[21] The paratext is different from the original text but is defined by the deformances to the original text through which it is generated. The classic example of deformance in a literary context is reversing the order of lines in a poem. The paratext produced in this way can then be analyzed for insights into features of the original text that are otherwise undetectable.

---

[19] Kenneth Pennington, "The Biography of Gratian, the Father of Canon Law," *Villanova Law Review* 59 (2014): 685.

[20] Anders, my understanding is that you took a similar approach when you created the baseline text for the edition in progress of the first recension (although you adopted a set of orthographic conventions different from Friedberg's in the resulting text).

[21] Lisa Samuels and Jerome McGann, "Deformance and Interpretation," *New Literary History* 30, no. 1 (1999): 25–56.

The method for producing the text samples used in this project involves multiple stages of deformance. Starting with Friedberg's 1879 edition of Gratian's *Decretum* as the text, Winroth's appendix, which compactly encodes first-recension variants with respect to Friedberg, is used as a program (literally, as will be seen in the section on corpus preparation below) for deforming Friedberg's text to produce the first paratext, the proxy first recension *dicta*. The first paratext is then used as the basis for a second deformation, by which the first paratext is subtracted from Friedberg's text to create the second paratext representing second recension additions and changes to the *dicta*.

The approach of deriving all of the text samples used in this study using only Friedberg's text and the first-recension variants recorded in Winroth's appendix as sources has one final argument in its favor, which is that it enables reproducibility. Reliance on publicly available data means that those who wish to reproduce these results are not dependent on private decisions about the content of the text samples.

Depending on the nature of the analysis we wish to conduct, we may choose to either include or exclude the *dicta* from *de Penitentia*. Including the *dicta* from *de Penitentia*., there are 897 *dicta* represented in the first-recension text sample and 419 represented in the second-recension sample. Of those, 65 *dicta* are represented in both the first- and second-recension samples. Excluding the *dicta* from *de Penitentia*., there are 836 first-

recension and 398 second-recension *dicta*, of which 61 *dicta* are represented in both samples.

Corpus preparation

The most important and time-consuming aspect of any digital humanities project is corpus preparation. The availability of a suitable corpus of electronic texts is a baseline requirement for carrying out stylometric analysis. The ideal textual basis for a project of this nature would be a set of electronic texts of good modern critical editions of both the first and second recensions of Gratian's *Decretum*, which follow consistent orthographic conventions and adhere to a widely-accepted encoding standard such as the XML Text Encoding Initiative (TEI P5) format. The Mellon Foundation-supported effort directed by Anders Winroth to edit the first recension is ongoing, but work on Winroth's edition in progress had not reached a sufficiently advanced state of completion for it to be used for my project.[22]

---

[22] As of the 22 April 2019 version, eight case statements (for cases 1-3, 9, 15, 24, 30, and 34) appear to have a complete critical apparatus. An additional six case statements (for cases 4-7, 10, and 11) have an incomplete critical apparatus that records variant readings from Fd only. The critical apparatus for the case statement for case 35 records a single variant reading from Aa. The remaining 21 case statements (for cases 8, 12-14, 16-23, 25-29, 31-33, and 36) have no critical apparatus at all. The case statements are used as the example here because they are the focus of my dissertation.

As a result, this investigation depends for both the first and the second recension on the electronic text of the Friedberg edition that Timothy Reuter and Gabriel Silagi used to produce the *Wortkonkordanz zum Decretum Gratiani* for the MGH.[23] Anders Winroth and Lou Burnard of the Oxford Text Archive (OTA) each provided copies of the MGH e-text. The copies differed significantly, and the e-text had to be reconstructed through an editorial process quite similar to preparing a critical edition to restore it to a state as close as possible to what Reuter and Silagi presumably intended.

The establishment of good, usefully formatted, texts has been crucial to Digital Humanities research from the beginning. In 1946, Father Roberto Busa, SJ (d.2011) began work on what ultimately became the *Index Thomisticus*, a concordance of the works of St Thomas Aquinas. In 1949, Father Busa secured crucial support from Thomas J. Watson of IBM, allowing the generation of the concordance to be carried out by means of electro-mechanical and later electronic computers operating on punch-card data. The *Index Thomisticus* is recognized today as the first important humanities computing project and figures prominently in origin stories for Digital Humanities as a

---

[23] Timothy Reuter and Gabriel Silagi, eds., *Wortkonkordanz zum Decretum Gratiani*, Monumenta Germaniae historica. Hilfsmittel 10 (München: Monumenta Germaniae Historica, 1990).

discipline.[24] The success of Father Busa's project inspired a number of imitators, as well as the development of specialized software and data formats to support such efforts. Reuter and Silagi's *Wortkonkordanz* was probably the last major Busa-style concordance. The MGH e-text of the Friedberg edition they used was encoded in the now-obsolete Oxford Concordance Program (OCP) format.

The MGH e-text also introduced a small number of textual errors in addition to those it inherited from the printed version of Friedberg's edition. Appendix 3 lists all currently known errors in the MGH e-text.[25] An error in D.23 c.2 is particularly noteworthy. Contrary to the widespread belief that the MGH e-text was created by scanning a physical copy of the Friedberg edition using optical character recognition (OCR) technology and then correcting the results, it is the product of keyboard transcription.

---

[24] Susan Hockey, "The History of Humanities Computing," in *A Companion to Digital Humanities*, ed. Susan Schreibman, Raymond George Siemens, and John Unsworth, Blackwell Companions to Literature and Culture 26 (Malden, MA: Blackwell Pub, 2004), 4–6. The highest honor in the field of Digital Humanities is the Roberto Busa Prize, awarded by the Alliance of Digital Humanities Organizations (ADHO). A notable past recipient of the Busa Prize is John Burrows, who first introduced the fundamental stylometric technique now known as Burrows's Delta in a lecture he delivered on the occasion of receiving the award in 2001.

[25] Data current as of 18 May 2021. For more recent error reports, see the list I maintain for the Stephan Kuttner Institute on GitHub. Thanks to Anders Winroth for reporting the errors in D.6 d.p.c.3 (6 October 2019), D.15 c.2 (4 November 2020), D.16 c.9 (18 May 2021), and D.23 c.2 (23 August 2019).

Conclusively, the e-text contains at least one instance of homeoteleuton. The following four lines from the text of D.23 c.2 in the Friedberg edition (column 79):

> *Patre et Spiritu sancto omnium creaturarum; qui passus sit pro salute nostra uera carnis passione, mortuus uera corporis sui morte, resurrexit uera carnis suae receptione et uera animae resumptione, in qua ueniat iudicare*

were transcribed as the following three lines in the e-text:

> *Patre et Spiritu sancto omnium creaturarum; qui passus sit pro salute nostra uera carnis suae receptione et uera animae resumptione, in qua ueniat iudicare*

skipping over the words "*passione, mortuus uera corporis sui morte, resurrexit uera carnis*."[26]

Notwithstanding its textual flaws and the highly specialized and outdated requirements that constrained the choice of file format, the MGH e-text remains a useful tool for the study of Gratian's *Decretum*.[27]

---

[26] Thanks to Anders Winroth for bringing the instance of homeoteleuton at D.23 c.2 in the MGH e-text to my attention (August 23, 2019). Clemens Radl of the MGH confirmed to Winroth that the e-text was typed.

[27] Anders Winroth, "Uncovering Gratian's Original Decretum with the Help of Electronic Resources," *Columbia Library Columns* 46, no. 1 (1997): 26–31.

The deformance algorithm used to generate the paratexts described in the previous section on the definition of the first- and second-recension *dicta* was implemented in the form of a 201-line Python program.[28] The program reads the MGH e-text of the Friedberg edition and parses it to extract the *dicta*.

Most readers will have at least a passing familiarity with the syntax of the Hypertext Markup Language (HTML) documents that serve as the foundation for the World Wide Web. Elements in HTML documents like headers, paragraphs, and links are delimited by start and end tags. Tags in HTML documents are recognizable by the use of angle bracket characters surrounding an identifier representing an element type. For example, `p` is the identifier for the paragraph element in HTML, so `<p>` is the paragraph start tag and `</p>` is the paragraph end tag. HTML elements, defined as spanning start tag, content (if any), and end tag, wholly enclose or are wholly enclosed by other elements. Consider the following examples in which `<h1>` is the start tag and `</h1>` is the end tag for an HTML top-level section heading:

```
<h1>Valid</h1><p>This is valid HTML.</p>
<h1>Invalid<p></h1>This is invalid HTML.</p>
```

---

[28] Python is a widely-used general-purpose programming language. According to one frequently-cited industry metric, the TIOBE Index, Python was the third-most popular programming language worldwide as of July 2020, behind legacy languages C and Java. Python provides powerful features for performing operations on textual data.

The first example is syntactically valid HTML because the start tag, content, and end tag for the top-level section heading element come before the start tag, content, and end tag for the paragraph element. The second example is syntactically invalid HTML because the start tag for the paragraph element comes before the end tag for the top-level section heading element, with the result that neither of the elements wholly encloses or is wholly enclosed by the other.

The examples in the foregoing discussion were framed in terms of HTML for the sake of familiarity, but the syntax of the Extensible Markup Language (XML), the current standard for encoding electronic texts, exhibits the same attributes. A document composed of elements delimited by start and end tags in which the elements wholly enclose or are wholly enclosed by other elements is said to be tree-structured. XML documents are tree-structured, which make them easy to parse. Although the markup used in the MGH e-text of the Friedberg edition of Gratian's *Decretum* superficially resembles the Standard Generalized Markup Language (SGML), an ancestor of XML, the e-text is in fact encoded in the Oxford Concordance Program (OCP) format. OCP markup is extremely difficult to parse because it is not tree-structured—it has start tags

for textual elements such as canons and *dicta*, cases and distinctions, but (unlike XML) not end tags.[29]

The extraction engine captures every element of text between a *dictum* start tag (`<T A>` or `<T P>`) and the start tag for the next element that can possibly follow a *dictum*. The extracted *dicta* require considerable scrubbing before they can be used. Here, for example, is what D.54 d.p.c.23 looks like in its raw state:

```
[' -Gratian.+ Ecce, quomodo serui ad clericatum ualeant assumi,\n
uel quomodo non admittantur. Liberti quoque non sunt promouendi\n
ad clerum, nisi ab obsequiis sui patroni fuerint absoluti.\n
Unde in Concilio Eliberitano: -[c. 80.]+\n']
```

Each *dictum* is then processed into a key-value pair in a Python dictionary. A dictionary or associative array is a built-in Python data type that can be thought of as a set of key-value pairs. Dictionaries are unlike lists or arrays in that the values stored in a dictionary are accessed using a key rather than a numerical index. Keys are usually alphanumeric text strings, although numbers can be used. The only requirement is that each key in a dictionary must be unique. The key in the example below is the string `'D.54 d.p.c.23'`, and the value is a string representing the text of D.54 d.p.c.23 extracted from the MGH e-text of the Friedberg edition:

---

[29] Hockey, "The History of Humanities Computing."

```
{'D.54 d.p.c.23': 'Ecce, quomodo serui ad clericatum ualeant assumi, uel
quomodo non admittantur. Liberti quoque non sunt promouendi ad clerum, nisi
ab obsequiis sui patroni fuerint absoluti. Unde in Concilio Eliberitano:'}
```

The first recension variants from the Friedberg edition recorded in Winroth's appendix are then encoded as a list of dictionaries in which the `'pattern'` item is the variant represented as a Python regular expression. Regular expressions are a language for specifying arbitrarily complicated patterns of characters according to a rule. Once a regular expression for a pattern has been specified, it can be used to search for and replace units of text matching the pattern. (The use of the word *regular* in the term *regular expression* is analogous to its use in the term *canons regular*. In both cases a rule is being followed.) The regular expression `(Ecce, quomodo serui.*?quomodo non admittantur\.)` in the example below matches a text string starting with "*Ecce, quomodo serui*" and ending with "*quomodo non admittantur.*"

```
[{'key': 'D.54 d.p.c.23', 'pattern': '(Ecce, quomodo serui.*?quomodo non
admittantur\.)'}]
```

Finally, the deformance engine uses the variants encoded as regular expression patterns to generate the first and second paratexts corresponding the first- and second-recension *dicta*. For each *dictum*, the text matching the pattern is inserted into a dictionary representing the first recension paratext; then the text resulting when the text matching the pattern is replaced by the null string `''` is inserted into a dictionary representing the second recension paratext. Here is the resulting first recension paratext:

```
{'D.54 d.p.c.23': 'Ecce, quomodo serui ad clericatum ualeant assumi, uel
quomodo non admittantur.'}
```

and the corresponding second recension paratext:

```
{'D.54 d.p.c.23': 'Liberti quoque non sunt promouendi ad clerum, nisi ab
obsequiis sui patroni fuerint absoluti. Unde in Concilio Eliberitano:'}
```

Simplified two-dimensional visualization

Visualizing data from the *dicta* in a simplified two-dimensional form is a useful first step toward understanding how stylometric analysis works in practice. For the purpose of the following discussion, the *dicta* will be divided into four samples: i.) the hypothetical case statements (*dicta initialia* or *themata*) that introduce the thirty-six cases in Part II of the *Decretum*, ii.) the first-recension *dicta* from Parts I and II of the *Decretum*, iii.) the first- and second-recension *dicta* from *de Penitentia*, and iv.) the second-recension *dicta* from Parts I and II of the *Decretum*.

In the following code, tables, and plots, these four samples will be labelled Gratian0, Gratian1, dePen, and Gratian2. The Gratian1 and Gratian2 samples, representing the first- and second-recension *dicta* from Parts I and II of the *Decretum*, were generated according to the procedure detailed in the preceding section on corpus preparation.

Although it is theoretically possible to split the text of the hypothetical case statements (*themata*) and the *dicta* from *de Penitentia* into separate first- and second-recension samples following the same procedure used to produce the Gratian1 and Gratian2

samples, as a practical matter it is not useful to do so. The only case statement (*thema*) for which Winroth's appendix notes a textual difference is C.19 d.init.[30] The second-recension version of the text of C.19 d.init. adds a 13-word clause absent from the first-recension version, seemingly for the purpose of piling up descriptive detail. (*unus relicta propria ecclesia eo inuito, alter dimissa regulari canonica cenobio se contulit*). Gratian0, the sample containing the case statements, is therefore made up of 99.6% first-recension text. Similarly, the number of words added to the *dicta* in *de Penitentia* between the first and second recensions is relatively small, 556 words out of a total of 10,081; dePen, the sample containing the *dicta* from *de Penitentia*, is therefore made up of 94.5% first-recension text.

In addition to plotting a two-dimensional visualization of word frequency data from the *dicta*, this section will set the stage for a subsequent one, which introduces an authorship attribution technique known as Burrows's Delta. Burrows's algorithm calculates a metric for the distance between a sample of unknown authorship with a corpus of samples of known authorship. In that discussion, Gratian0, the sample containing the hypothetical case statements (*themata*), will be treated as the sample of unknown authorship. Gratian1, dePen, and Gratian2, the samples containing the first-

---

[30] Winroth, *The Making of Gratian's Decretum*, 216.

recension *dicta*, the first- and second-recension *dicta* from *de Penitentia*, and the second-recension *dicta*, will be treated as the corpus of samples of known authorship. Therefore, the values for means and standard deviations that provide the basis of comparison between the unattributed sample and the attributed corpus have to be calculated without taking the values from Gratian0 into account.

The demonstration of Burrows's Delta will make a point of the fact that the technique can be used at a higher number of dimensions ($n > 3$) than can be visualized in graphical form. Word frequency data for the four most frequent words (MFWs) will therefore be collected from the start, even though the data for the third- and fourth-most frequent words will not be used in this section. The first step is to identify the four most frequent words in the comparison text samples, Gratian1, dePen, and Gratian2.

The four most frequent words in the three comparison samples Gratian1, dePen, and Gratian2—the samples treated as being of known authorship—are *in*, *non*, *et*, and *est*. The selection of samples makes a difference to the order. Were Gratian0, the sample treated as being of unknown authorship, to be included, the four most frequent words would be *in*, *et*, *non*, and *est*. (The rank reversal between the second- and third-most frequent words is a result of the fact that *non* occurs quite infrequently in Gratian0; see the table below.) After identifying the four most frequent words in the three

comparison samples, next, count the number of occurrences of those words in each of the samples:[31]

	Gratian0	Gratian1	dePen	Gratian2
in	74	1450	252	411
non	24	1360	270	306
et	70	1293	260	345
est	13	965	182	167

After determining the number of occurrences of the MFWs, next, determine the length (total word count) for each of the samples:

	Gratian0	Gratian1	dePen	Gratian2
words	3605	56713	10081	14255

[31] Much of the analysis from this point forward will take advantage of the specialized capabilities of a Python software library called pandas. The name pandas is not a reference to the charismatic animal but an acronym derived from the term "panel data." The package is widely used in the field of data science and provides a dataframe abstraction that represents two-dimensional numerical word-frequency data in a much more natural way than native Python data structure like lists and dictionaries do. The pandas dataframe abstraction can be thought of as a close analog to the Excel spreadsheets that were such a ubiquitous feature of John Burrows's and David Hoover's early experiments in stylometry.

Finally, divide the number of occurrences of the MFWs in the samples by the sample length and multiply the quotient by 1,000 to determine the normalized frequency of occurrence per 1,000 words for each of the MFWs in each of the samples:

	Gratian0	Gratian1	dePen	Gratian2
in	20.5270	25.5673	24.9975	28.8320
non	6.6574	23.9804	26.7831	21.4662
et	19.4175	22.7990	25.7911	24.2020
est	3.6061	17.0155	18.0538	11.7152

In is the most frequently occurring word in the *dicta*. There are 1,450 occurrences of *in* out of 56,713 words in the first-recension *dicta* (25.5673 occurrences per 1,000 words), 252 occurrences of *in* out of 10,081 words in the *dicta* from *de Penitentia* (24.9975 per 1,000), and 411 occurrences of *in* out of 14,255 words in the second-recension *dicta* (28.8320 per 1,000). It is more convenient to characterize word frequencies in units of occurrences per 1,000 words than percentage, since at that scale most of the values we are concerned with are greater than 1.0. The overall mean frequency of occurrence of *in* for the combined *dicta* from the first recension, *de Penitentia*, and the second recension is therefore 2,113 occurrences out of 81,049 words or 26.0706 per 1,000. It is common, however, for authorship attribution algorithms to use the mean of the normalized frequencies of occurrence of a word for each of the samples rather than the overall mean

frequency of occurrence of the word in the corpus made up of all of the samples. The motivation for using the mean of normalized frequencies of occurrence for the individual samples rather than the overall mean frequency of occurrence is to ensure that the largest sample does not dominate the result. The value for the mean frequency of occurrence of *in* that will be required at subsequent stages of this demonstration, then, is the mean of the normalized frequencies 25.5673, 24.9975, and 28.8320, or 26.4656 occurrences per 1,000.

Non is the second most frequently occurring word in the *dicta*. There are 1,360 occurrences of *non* out of 56,713 words in the first-recension *dicta* (23.9804 occurrences per 1,000 words), 270 occurrences of *non* out of 10,081 words in the *dicta* from *de Penitentia* (26.7831 per 1,000), and 306 occurrences of *non* out of 14,255 words in the second-recension *dicta* (21.4662 per 1,000). The overall mean frequency of occurrence of *non* for the combined *dicta* from the first recension, *de Penitentia*, and the second recension is therefore 1,936 occurrences out of 81,049 words or 23.8868 per 1,000. The mean of the normalized frequencies of occurrence of *non* for each of the samples is the mean of 23.9804, 26.7831, and 21.4662, or 24.0765 occurrences per 1,000.

Comparing the frequencies of occurrence of *in* and *non* in the two outlying samples, dePen and Gratian2, reveals large variations for such common words. (Large variations would be less surprising with uncommon words for which small differences in number

could result in a large difference in percentage.) *In* occurs 15.3% more frequently in Gratian2 than in dePen, and 13.3% less frequently in dePen than in Gratian2. *Non* occurs 19.9% less frequently in Gratian2 than in dePen, and 24.8% more frequently in dePen than in Gratian2. Even compared to the mean, *non*, for example, occurs 11.2% more frequently in dePen and 10.8% less frequently in Gratian2 than the mean of normalized frequencies. It is clear then that against an overall background of "orderliness" in the word-frequency distribution, individual samples can display striking and potentially significant levels of variation.

Word count and sample length data were collected and used to calculate frequencies for Gratian0 above, but those values will not be used in this section. Disregard the Gratian0 column and use only the columns corresponding to the three comparison samples, Gratian1, dePen, and Gratian2, to calculate the means for the values in each of the rows in the frequency table representing the four most frequent words:

	Gratian1	dePen	Gratian2	mean
in	25.5673	24.9975	28.8320	26.4656
non	23.9804	26.7831	21.4662	24.0765
et	22.7990	25.7911	24.2020	24.2640
est	17.0155	18.0538	11.7152	15.5948

We can graph the number of occurrences of *in* and *non* per 1,000 words in the *dicta*, with the frequency of *in* plotted along the horizontal x-axis, and the frequency of *non* plotted along the vertical y-axis, to produce a simplified visualization of the total variation among the three samples. Means are provided for context: the vertical dashed line represents the mean of normalized frequencies for the horizontal (*in*) axis, and the horizontal dashed line represents the mean of normalized frequencies for the vertical (*non*) axis.

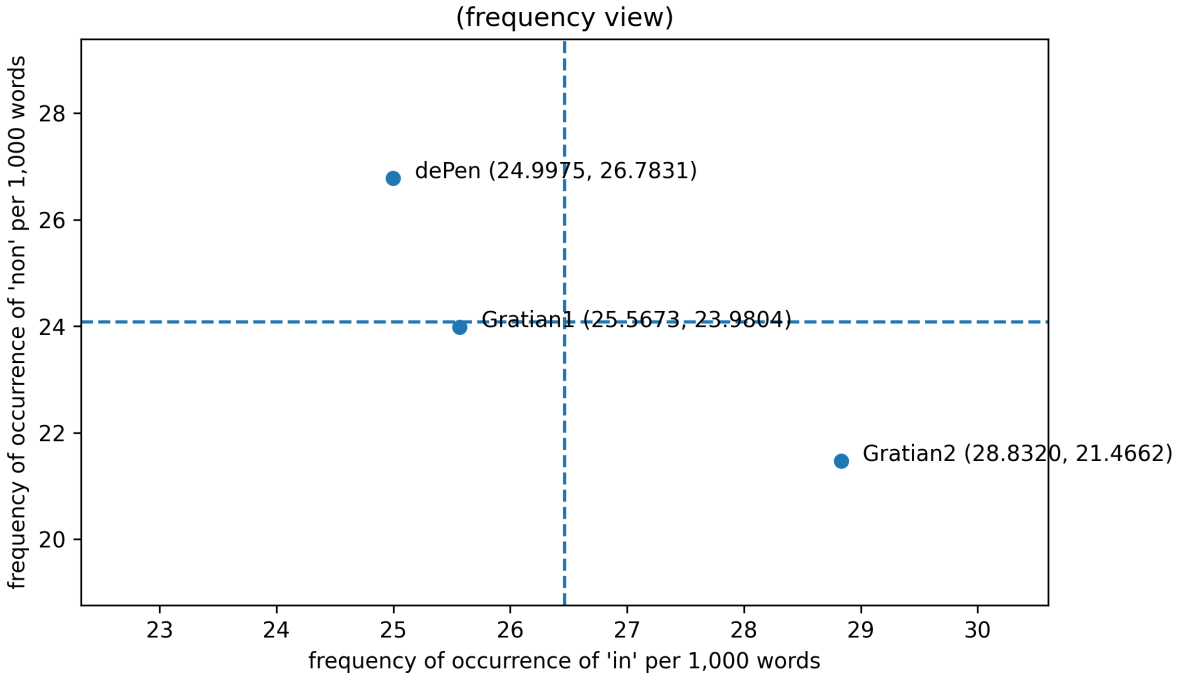

Figure 1: Figure 0a updated 14 May 2020[32]

[32] The actual generation of Figure 0a was deferred until after the sample standard deviations for *in* and *non* per 1,000 words had been calculated below. Framing the

Figure 0a introduces several conventions common to two-dimensional graphical representations of word frequency data that readers will encounter repeatedly throughout this chapter. The feature (in this case the frequency of occurrence of the word *in*) that explains more of the variation between the samples is plotted along the horizontal x-axis, while the feature (in this case the frequency of occurrence of the word *non*) that explains less of the variation between the samples is plotted along the vertical y-axis. Although this plot, produced by the Matplotlib Python two-dimensional plotting library, is rectangular and the axes are approximately to scale, many of the figures in this chapter were generated using stylo, an R package for stylometric analysis, which outputs square plots. Regardless of appearance, however, readers should bear in mind that the area plotted is always wider than it is tall, that is, that it displays greater variation between samples horizontally along the x-axis than it does vertically along the y-axis.

Figure 0a plots the values for the first-recension *dicta*, the second-recension *dicta*, the *dicta* from *de Penitentia*, the second-recension *dicta*, and (labelled Gratian1, dePen, and Gratian2 respectively), as well as the means (indicated by the dashed lines), for the frequencies of *in* and *non* per 1,000 words. It is more statistically meaningful, however,

---

dimensions of the plot to twice the standard deviation from the mean along both axes improves graphical layout and readability.

to measure and plot the differences between values and means in units of standard deviations rather than frequency per 1,000 words. The difference of a value from the mean divided by standard deviation is referred to as the value's z-score. A value that has a difference of one standard deviation from the mean is said to have a z-score of 1.0 or -1.0 depending on whether the value is greater or lesser than the mean. The formula used to calculate the sample standard deviation is:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

The formula is slightly daunting notationally, but it is not difficult to use it to calculate the desired results. The example immediately below shows all of the intermediate steps involved in using the formula to compute the sample standard deviation of the frequency of occurrence of the word *in* in the Gratian1, dePen, and Gratian2 samples. (The motivated reader can use a calculator to repeat the process for the frequency of occurrence of the word *non*.) Remember that for the purpose of calculating sample standard deviation, the value of the mean ($\bar{x}$) is not the overall mean frequency of occurrence of the word *in* across all of the samples, but rather the mean of the normalized frequencies of occurrence of the word *in* for each of the samples individually.

First, calculate the squared deviations from the mean of normalized frequencies for the frequency of *in* in the first-recension *dicta* (Gratian1):

$$(x_1 - \bar{x})^2 = (25.5673 - 26.4656)^2 = (-0.8983)^2 = 0.8069,$$

for the frequency of *in* in the *dicta* from *de Penitentia* (dePen):

$$(x_2 - \bar{x})^2 = (24.9975 - 26.4656)^2 = (-1.4681)^2 = 2.1553,$$

and for the frequency of *in* in the second-recension *dicta* (Gratian2):

$$(x_3 - \bar{x})^2 = (28.8320 - 26.4656)^2 = (2.3664)^2 = 5.5998.$$

Then, as indicated by the summation operator ∑, sum the three squared deviations from the mean of normalized frequencies, divide the sum by their number ($N = 3$) minus one, and take the square root of the quotient:

$$s = \sqrt{\frac{1}{2}(0.8069 + 2.1553 + 5.5998)} = \sqrt{\frac{1}{2}(8.5620)} = \sqrt{4.2810} = 2.0691$$

The units of s are the same as those used to calculate the mean, in this case, the frequency of occurrence of a word per 1,000 words.

Once again disregarding the Gratian0 column of the frequencies table, calculate the standard deviations for the rest of the rows representing the four most frequent words,

using only the values in the columns corresponding to the three comparison samples, and the means computed from them:

	Gratian1	dePen	Gratian2	mean	std
in	25.5673	24.9975	28.8320	26.4656	2.0691
non	23.9804	26.7831	21.4662	24.0765	2.6598
et	22.7990	25.7911	24.2020	24.2640	1.4970
est	17.0155	18.0538	11.7152	15.5948	3.3997

As noted above, the definition of a value's z-score is the difference of that value from the mean divided by the standard deviation. A z-score can be calculated for a value even if that value was not used to determine the mean and standard deviation to be used. That means that z-scores can be calculated for word frequencies in the unattributed sample Gratian0 using the means and standard deviations calculated using the corresponding word frequencies in the attributed samples Gratian1, dePen, and Gratian2. Just as word frequencies were calculated for Gratian0 above, z-scores will be calculated for Gratian0 here, which will be used in the next section to determine the value of Burrows's Delta. The formula used to calculate the z-score is:

$$z = \frac{x - \bar{x}}{s}$$

For the frequency of *in* in the case statements or *themata* (Gratian0):

$$z = \frac{x-\bar{x}}{s} = \frac{20.5270-26.4656}{2.0691} = \frac{-5.9386}{2.0691} = -2.8702,$$

for the frequency of *in* in the first-recension *dicta* (Gratian1):

$$z = \frac{x-\bar{x}}{s} = \frac{25.5673-26.4656}{2.0691} = \frac{-0.8983}{2.0691} = -0.4342,$$

for the frequency of *in* in the *dicta* from *de Penitentia* (dePen):

$$z = \frac{x-\bar{x}}{s} = \frac{24.9975-26.4656}{2.0691} = \frac{-1.4681}{2.0691} = -0.7095,$$

and for the frequency of *in* in the second-recension *dicta* (Gratian2):

$$z = \frac{x-\bar{x}}{s} = \frac{28.8320-26.4656}{2.0691} = \frac{2.3664}{2.0691} = 1.1437.$$

(Because both the numerator and the denominator of the formula for calculating z-scores have units of frequency of occurrence per 1,000 words, z is a dimensionless number.)

Calculate the z-scores for the remaining most frequent words, and then plot the coordinates of the attributed samples Gratian1, dePen, and Gratian2:

	Gratian0	Gratian1	dePen	Gratian2
in	-2.8702	-0.4342	-0.7095	1.1437

	Gratian0	Gratian1	dePen	Gratian2
non	-6.5491	-0.0361	1.0176	-0.9814
et	-3.2375	-0.9786	1.0201	-0.0414
est	-3.5264	0.4179	0.7233	-1.1412

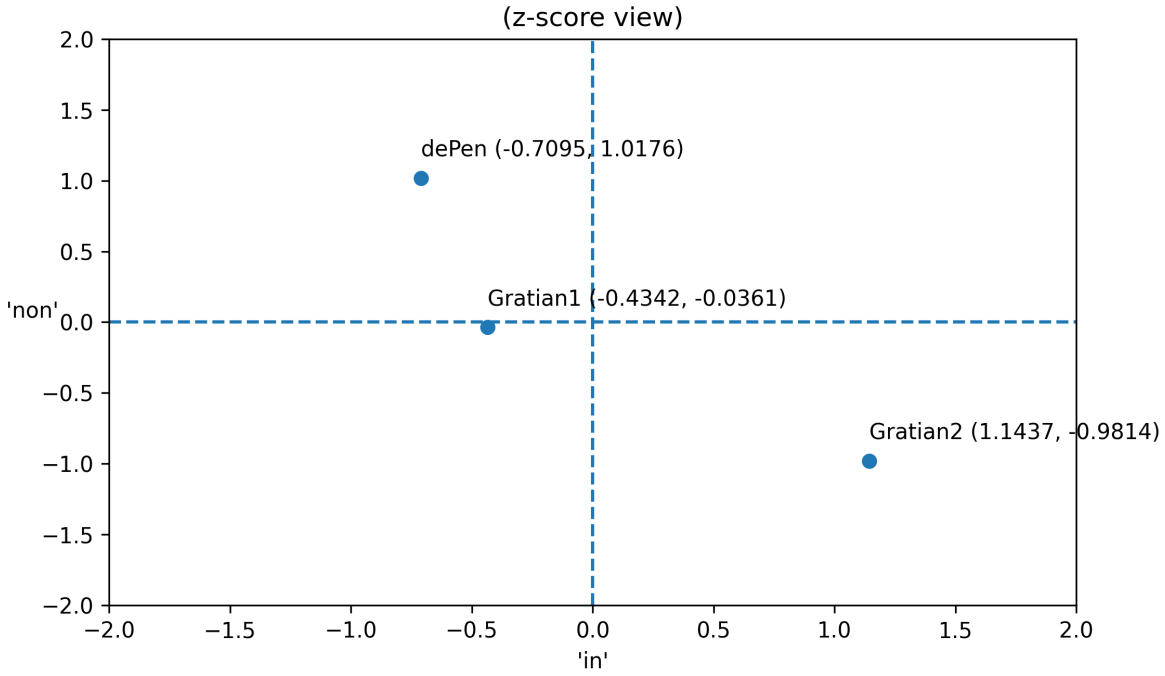

Figure 2: Figure 0b updated 15 May 2020

Labels on the axes of Figure 0b refer to standard deviations (values of z) away from the mean of normalized frequencies (represented by the dashed lines).

Figures 0a and 0b represent the axes as orthogonal (perpendicular) to one another. Although doing so is acceptable as a first-order approximation in a simplified representation of this kind, plotting the values along orthogonal axes invokes an

implicit assumption that the word frequencies (in this case, of *in* and *non*) are completely independent of one another, i.e., that there is no correlation or covariance relationship between the words' frequency of occurrence in the samples. This is not necessarily the case, and an advanced technique, introduced below, called principal component analysis (PCA), handles this problem in a more mathematically sophisticated way.

Now, we are obviously not going to make an attribution of authorship based on the frequencies of only two function words. Increasing the number of function words for which one collects frequency data increases the accuracy of stylometric analysis, up to a point. There is, however, a limit to the marginal value of each additional word included in an analysis, for two reasons. The first reason is that the frequency of each word in a corpus of text tails off inversely as the word's rank, a relationship (approximately) described by the equation $r \times f = C$. (The rank of the most frequent word is defined as 1, that of the second most frequent word as 2, and so on.) As a consequence, assuming words are weighted in proportion to their frequency of occurrence in the corpus, every marginal word has less value as evidence than the word before it. Stylometric techniques such as Burrows's Delta, which takes the mean of word frequency z-scores for an arbitrary number of words, have the effect of weighting each selected word from a corpus of texts equally. Such choices, however, need to be made with an informed

awareness of the trade-offs involved. The second reason is that every marginal word in a list, sorted from most to least frequent, of frequently used words is more likely than the previous word to be a content word instead of a function word. Twenty-four out of the thirty most frequently used words in Gratian's *dicta* are function words potentially suitable for use in stylometric analysis, but only 64 out of the 250 most frequent words.

Zipf's law

The observed empirical relationship that word frequencies in a corpus of text tail off inversely as the rank is known as Zipf's law, after the American quantitative linguist George Kingsley Zipf (d.1950). Though he does not seem to have claimed discovery, Zipf published the first extended discussions of the phenomenon.[33] The rank-frequency distribution of words in a corpus of any language is not merely a curiosity but rather provides a general background of what Zipf characterized as "orderliness" against which variations in frequency of occurrence of individual words are both detectable and meaningful. Thus, it is worthwhile to discuss theoretical predictions about the

---

[33] George Kingsley Zipf, *The Psycho-Biology of Language: An Introduction to Dynamic Philology* (Boston: Houghton Mifflin Company, 1935), 39–48; and George Kingsley Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (Cambridge, Mass: Addison-Wesley Press, 1949), 73–131. Zipf referred to the relationship as "the law of diminishing returns of words" (1949). He expressed the relationship as $ab^2 = k$ (1935) and r × f = C (1949).

rank-frequency distribution of words in a corpus and to demonstrate that Gratian's *dicta* display the expected orderliness of word distribution.

If the words in a corpus of text are rank-ordered from most to least frequent, Zipf's law posits that, as a first-order approximation, the frequency of the Nth word will be 1/N times that of the most frequent word.[34] In other words, the theoretical Zipf distribution predicts that the frequency of the second most frequent word in a corpus of text should be one half that of the most frequent word, the frequency of the third most frequent word should be one third that of the most frequent word, and so on. (See Figure Za below.)

[34] In Digital Humanities courses, 1/N is typically presented *as* Zipf's law. However, the reductionist 1/N representation of the rank-frequency relationship is misleading insofar as it ignores scaling considerations and elides the discrete rather than continuous nature of the variables representing rank and frequency.

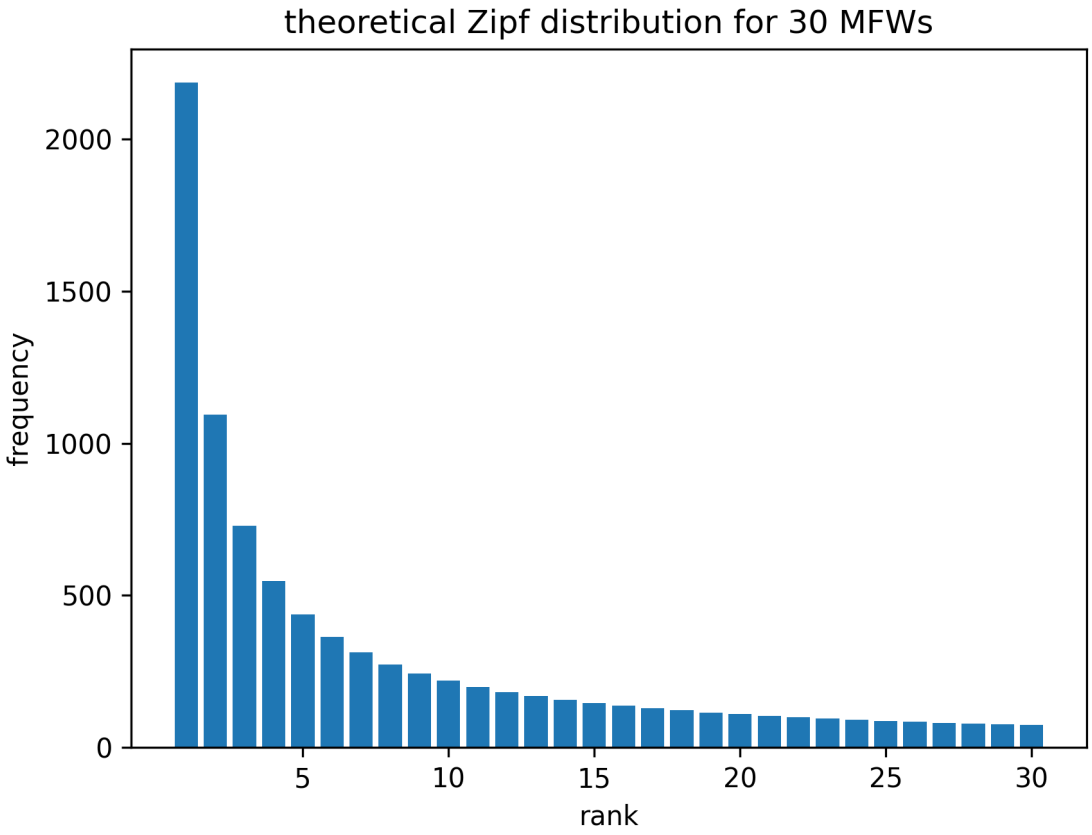

Figure 3: Figure Za updated 27 May 2020

Zipf tried several different approaches to the quantitative analysis of the distribution of words in corpora. His initial attempt, in 1935, sought to characterize the orderliness of word distributions by the relationship between the number of occurrences of a word, *b*, and the number of words *a*, a relationship Zipf expressed as $ab^2 = k$. For example, in Gratian's *dicta* there are 8,028 words (*a*) for which there is 1 (*b*) occurrence, 2,462 words (*a*) for which there are 2 (*b*) occurrences, 1,135 words (*a*) for which there are 3 (*b*) occurrences, and 660 words (*a*) for which there are 4 (*b*) occurrences. Plugging these values for *a* and *b* into the formula $ab^2 = k$ yields 8,028, 9,848, 10,215, and 10,560

respectively as values for the "constant" *k*. As the number of occurrences *b* increases, values of *k* for Gratian's *dicta* remain fairly stable, mostly falling between 10,000 and 11,000.

b	a	b^2	k
1	8028	1	8028
2	2462	4	9848
3	1135	9	10215
4	660	16	10560
5	423	25	10575
6	290	36	10440
7	225	49	11025
8	173	64	11072

As this example suggests, the explanatory power of the formula $ab^2 = k$ to characterize a word distribution is greatest for words that have a low number of occurrences. Zipf ultimately judged the formula $ab^2 = k$ to be an unsatisfactory model for the full spectrum of word distribution in a corpus, in part because it implied fractional values of

a for the most frequent words.[35] Zipf had a vivid awareness, ahead of its time for the pre-digital age in which he lived, of the fact that the variables in the formulas by which he sought to express these relationships represent discrete rather than continuous quantities.[36] Zipf also noted that value of the exponent of b is likely to differ from 2 depending on the size of the corpus[37]

Zipf's final attempt, in 1949, to give a quantitative account of the distribution of words in a corpus characterized the distribution in terms of rank and frequency as r × f = C. Values for the constant C differ between corpora, depending, among other things, on corpus size. As an example, the theoretical Zipf distribution plotted in Figures Za and Zb has been scaled to facilitate direct comparison with actual data from Gratian's *dicta* plotted in Figures Zc and Zd. In all four plots, the first data point has a rank of 1 and a frequency of 2187, corresponding to the 2,187 occurrences of the most frequent word *in*

---

[35] "Hence the $ab^2 = k$ relationship is valid only for the less frequently occurring words which, however, represent the greater part of the vocabulary in use, though not always a great majority of the occurrences." … "It is perhaps worth pointing out that the $ab^2 = k$ relationship which appears valid for the frequency distribution of the less frequent words would demand fractional words when applied to the speech-elements of highest occurrence, such as *the* in English." Zipf, *The Psycho-Biology of Language*, 42–43.

[36] Zipf used the term "integrality" to describe the discrete, discontinuous, nature of frequency and rank. Zipf, *Human Behavior and the Principle of Least Effort*, 31, 35.

[37] Zipf, *The Psycho-Biology of Language*, 43.

in the *dicta*. This scaling is equivalent to setting the value of C to 2187, and letting r × f = 2187.

Both the $ab^2 = k$ and r × f = C interpretations of Zipf's law can be restated with greater mathematical generality by noting that if the variables are plotted logarithmically, their relationships are linear, with the slope determined by the power (exponent) relationships involved. In the equation $ab^2 = k$, the number of occurrences b varies inversely as the square root of the number of words a (as $\frac{1}{\sqrt{a}}$, or equivalently as $a^{-1/2}$). The relationship of the logarithm of the number of occurrences to the number of words is linear, with a slope of -1/2 corresponding to the exponent (-1/2) of the number of words. The actual slope for data from Gratian's *dicta* for values of b (Number of Occurrences) from 1 to 30 is -0.5097, quite close to the predicted theoretical value of -0.5.

Similarly, in the equation r × f = C, the frequency f varies inversely as the rank r (as 1/r, or equivalently as r^{-1}). The relationship of the logarithm of frequency to the logarithm of rank is linear, with a slope of -1 corresponding to the exponent (-1) of the rank. (See Figure Zb below.) Note that the base of the logarithms does not matter, as long as the bases are the same for both axes. Regardless of whether we take base e (natural) or base 10 logarithms of rank and frequency, for example, the slopes will be the same: -1.0 for the theoretical Zipf distribution of word frequencies in Figure Zb, and -0.6518 for the actual frequencies of the thirty most frequent words in Gratian's *dicta* in Figure Zd.

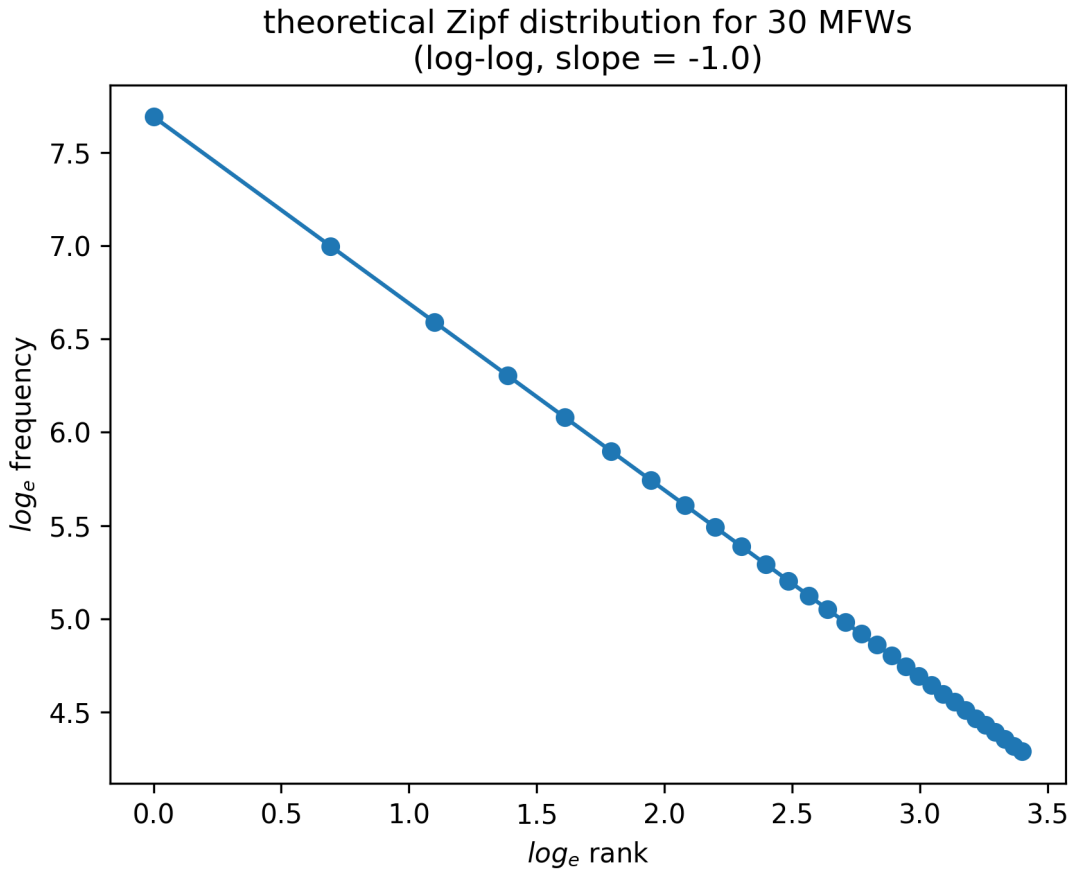

Figure 4: Figure Zb updated 27 May 2020

Figure Zc below plots the actual rank-frequency distribution of the thirty most frequent words (MFWs) in Gratian's *dicta*: *in* (2187), *et* (1968), *non* (1960), *est* (1327), *de* (925), *quod* (888), *ad* (832), *qui* (812), *sed* (736), *unde* (732), *uel* (705), *si* (669), *ut* (641), *cum* (589), *a* (588), *autem* (582), *ex* (501), *sunt* (428), *enim* (424), *que* (423), *uero* (411), *etiam* (405), *ab* (391), *ait* (349), *esse* (339), *ergo* (338), *quia* (336), *item* (327), *per* (304), *nec* (293).

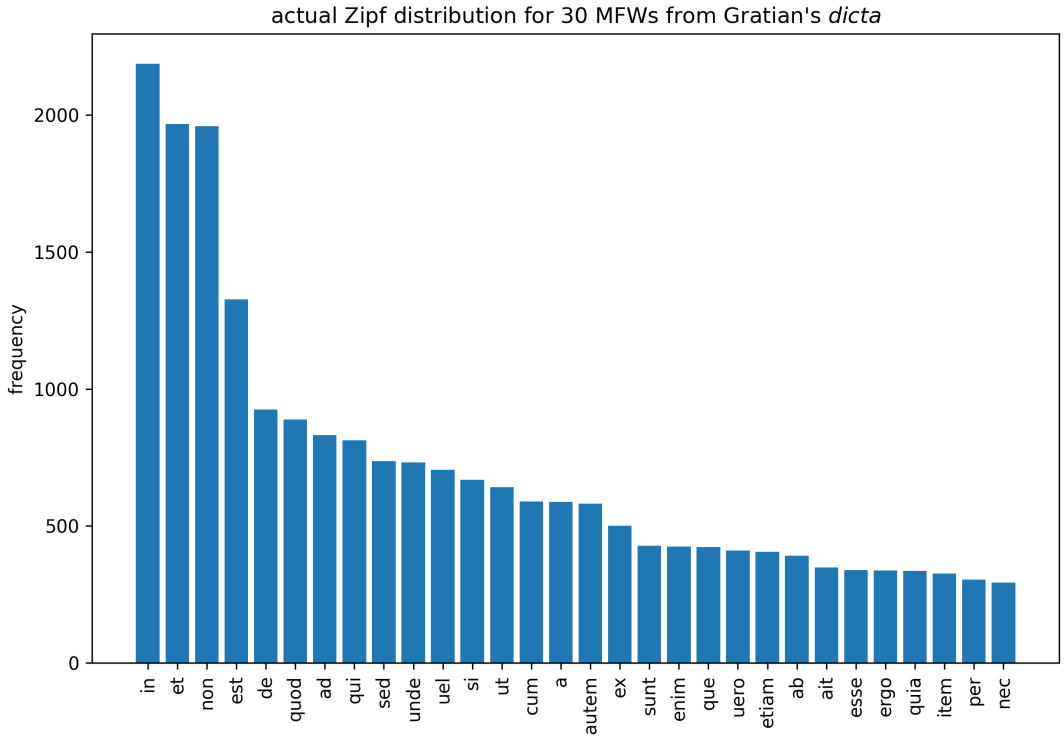

Figure 5: Figure Zc updated 27 May 2020

Zipf used word frequencies hand-tabulated from James Joyce's *Ulysses* as the data set for his exploration of the rank-frequency relationship, and it turns out that for English the r × f = C formulation holds up reasonably well.[38] The rank-frequency relationship does not on first inspection appear to hold up as well for Gratian's Latin as it does for Joyce's English, since the frequencies for the thirty most frequent words of the *dicta* do

---

[38] "we have found a clearcut correlation between the number of different words in the *Ulysses* and the frequency of their usage, in the sense that they approximate the simple equation of an equilateral hyperbola: r × f = C in which *r* refers to the word's rank in the *Ulysses* and *f* to its frequency of occurrence (as we ignore for the present the size of C)." Zipf, *Human Behavior and the Principle of Least Effort*, 24. See Zipf, 23–52, for Zipf's extended discussion of the rank-frequency distribution of words in Joyce's *Ulysses*.

not drop off quite as sharply as the r × f = C formulation of Zipf's law would predict. The frequency of *et*, the second most frequent word in Gratian's *dicta* is 0.8999 times that of *in*, the most frequent word, rather than 0.5 as Zipf's law would predict; and the frequency of *non*, the third most frequent word, is 0.8962 rather than 0.3333.

Plotting the data from Figure Zc on logarithmic axes and performing least-squares linear regression analysis lets us calculate the slope, -0.6518, for the rank-frequency tail-off of the thirty most frequent words from Gratian's *dicta*.[39] (See Figure Zd below.) Transposing that result back into the linear (as opposed to logarithmic) frame of reference used in Figure Zc, the expression $1/r^{0.6518}$ yields a better (though not perfect) fit to the actual rank-frequency data.

---

[39] $m = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$ or:

```
def regression_slope(data_points):
    n = len(data_points)
    x_values, y_values = zip(*data_points)
    x_bar = statistics.mean(x_values)
    y_bar = statistics.mean(y_values)
    xy_sum = 0
    x_squared_sum = 0
    for i in range(n):
        xy_sum += x_values[i] * y_values[i]
        x_squared_sum += x_values[i] ** 2
    return (xy_sum - n * x_bar * y_bar) / (x_squared_sum - n * x_bar ** 2)
```

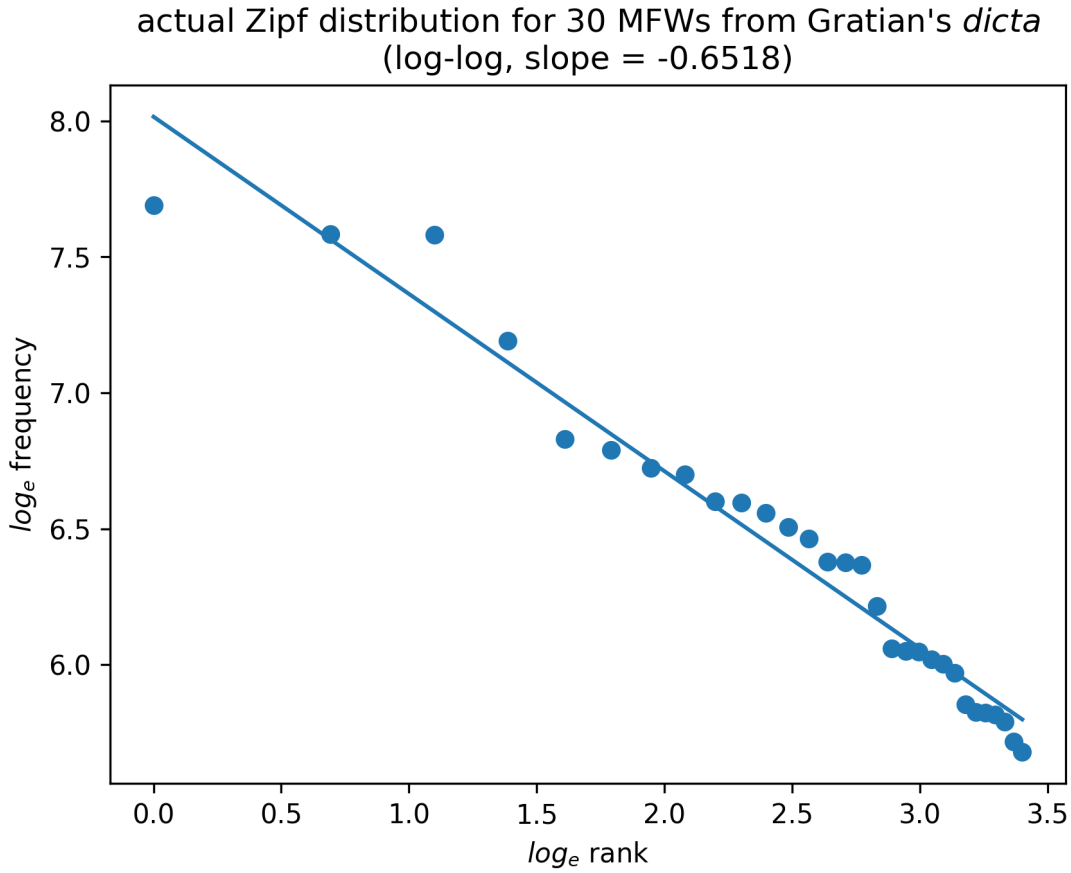

Figure 6: Figure Zd updated 27 May 2020

Burrows's Delta

The examples presented in the previous section are suggestive of ways in which differences between the frequencies of occurrence of common words in samples from a corpus of texts can be quantified in statistically meaningful units (standard deviations or values of z) and combined to represent the distance between those samples. This technique is, however, of limited value so long as we are restricted to the two, or at most three, dimensions the human mind is capable of visualizing. In 2001, John F. Burrows (d.2019) of the University of Newcastle, Australia, proposed a generalization

that gets around the limitation on the number of features to two or three by averaging z-score distance measurements of word frequency data for any number of features. This has the effect of collapsing distance measurements in an arbitrary number of dimensions into a single metric. Burrows called this metric the Delta, and it is now generally referred to as Burrows's Delta (Δ_B).[40] Expositions of Burrows's Delta sometime fail to make a clear enough distinction between the metric Δ_B and the authorship attribution method in which Burrows applied it. The metric is not the method.

Attempts to attribute authorship are typically undertaken in scenarios where there is a large (enough) number of texts securely attributable to a known author, and a text, or at most a small number of texts, of unknown authorship. The attempt is then made to attribute the unknown text to the known author, or to rule out such an attribution. Take the *Federalist* as an example. There are numbers of the *Federalist* of disputed or unknown attribution, a small and well-defined number of candidates for authorship—Hamilton,

---

[40] John Burrows, "Questions of Authorship: Attribution and Beyond: A Lecture Delivered on the Occasion of the Roberto Busa Award ACH-ALLC 2001, New York," *Computers and the Humanities* 37, no. 1 (February 2003): 5–32; and John Burrows, "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship," *Literary and Linguistic Computing* 17, no. 3 (September 2002): 267–87.

Jay, Madison—to whom those numbers might be attributed, and securely attributed samples from each of the candidates, conveniently enough from the same work.

Burrows's method assumes just such a scenario. He began by identifying the most frequent words (MFWs) in the corpus of comparison texts securely attributed to known authors. In Burrows's published descriptions of his method, he typically used all of the 30 most frequent words in the corpus of attributed comparison texts without distinguishing between function and content words. He then tabulated the number of occurrences of the most frequent words in each of the sample texts in the comparison corpus and normalized their frequency of occurrence as a percentage. Burrows then used the frequency data collected from the comparison texts to calculate a mean frequency of occurrence and sample standard deviation for each of the MFWs or features.

It is important to emphasize that the mean frequency of feature occurrence calculated at this stage of Burrows's algorithm and subsequently used to calculate the sample standard deviation for each feature is *not* the overall mean frequency across the corpus of attributed comparison text samples. Instead, the comparison corpus feature mean is calculated by averaging the normalized (percentage) frequency for each feature across all of the text samples in the attributed comparison corpus, without concern for differences in size (word count) between the samples. To refer back to the example

presented in the previous section as part of the two-dimensional visualization demonstration, we did not use the overall mean frequency of *in* across the three samples Gratian1, dePen, and Gratian2, (2,113 occurrences out of 81,049 words or 26.0706 per 1,000), but rather the mean of the normalized frequencies of *in* for each of the samples (the mean of 25.5673, 24.9975, and 28.8320, or 26.4656 occurrences per 1,000).

After calculating the mean of normalized frequencies and sample standard deviation for each of the features (MFWs), Burrows then converted the normalized (percentage) frequencies of occurrence for each feature in each sample in the comparison corpus to z-scores by subtracting the mean of the normalized frequencies from the frequency and dividing the positive or negative difference by the standard deviation for the feature. At this point, Burrows turned his attention to the unattributed text, tabulating all occurrences of the 30 MFWs for which data had been collected from the comparison texts, then normalizing the word counts by converting them to percentage frequencies of occurrence. Burrows then converted the normalized frequencies for each feature in the unattributed test sample to z-scores based on the values for the mean of normalized frequencies and sample standard deviation derived from the feature frequencies in the attributed comparison corpus samples.

With these preliminaries out of the way, Burrows then calculated the value of the Delta by taking the average (arithmetic mean) of the absolute value of the differences between the z-score for a given feature (MFW) for the unattributed test sample and each of the comparison samples in the corpus of attributed texts. In Burrows's interpretation, the comparison test sample from the attributed corpus with the lowest Delta with respect to the unattributed test sample was most likely to share a common author with it.

It is not possible to apply Burrow's method in the case of the *dicta* from Gratian's *Decretum* without modification. As the survey in Chapter 3 above indicated, near-contemporaries knew next to nothing about Gratian. Perhaps most notably, although Gratian was thought to have been a teacher, no one in the generation following made an unambiguous claim to have been his student. There are no other writings securely, or even insecurely, attributed to him. Fortunately, Burrows's Delta can be readily adapted to the particular situation in which we find ourselves, where there are no other texts attributed to Gratian with which we can compare, for example, the hypothetical case statements (*themata*) or second-recension *dicta*.

Although other delta methods of authorship attribution have been proposed since,[41] Burrows's Delta is widely accepted in the scholarly literature of the field of computational linguistics, and it will therefore be used as the basis for the demonstrations in this section.

The first experiment will be a comparison of four subcorpora, Gratian0 (the hypothetical case statements or *themata*), Gratian1 (the first-recension *dicta* excluding the *dicta* from *de Penitentia*), dePen (first- and second-recension *dicta* from *de Penitentia*), and Gratian2 (the second-recension *dicta* excluding the *dicta* from *de Penitentia*), using the frequencies of occurrence of the four most frequent words (MFWs) in Gratian's *dicta* as the basis for comparison. We will hypothesize that the subcorpus containing the hypothetical case statements (*themata*) is the work of an unknown author, and will treat the other three subcorpora as making up a corpus of works by a known author. Using four subcorpora and four features, where every feature analyzed is represented in a different dimension, demonstrates that z-score distance methods can be extended to cases in which the number of dimensions is greater than three. It also has the advantage

---

[41] Most notably Argamon's Delta, see Shlomo Argamon, "Interpreting Burrows's Delta: Geometric and Probabilistic Foundations," *Literary and Linguistic Computing* 23, no. 2 (June 2008): 131–47. For an overview of recent developments in the use of distance methods for the purpose of authorship attribution, see Stefan Evert et al., "Understanding and Explaining Delta Measures for Authorship Attribution," *Digital Scholarship in the Humanities* 32, no. suppl_2 (December 2017): ii4–16.

of making the solution compact enough to allow readers to follow along and reassure themselves of the mathematical validity of all of the intermediate steps leading to the final result.

The first experiment resumes directly where the two-dimensional visualization demonstration left off, so all of the function definitions and variable values in force at the conclusion of that demonstration are still valid. In particular, this experiment inherits the z-scores for all of the four most frequent words (MFWs). While we disregarded the data for the third and fourth most frequent words (*et* and *est*) for the purpose of the visualization demonstration, they will be fully taken into account here. (Remember that the values for mean and standard deviations used to derive the z-scores were calculated without reference to the Gratian0 sample here being treated as the unknown).

First, split the z-scores into two new dataframes, one for the test sample Gratian0, assumed for the purpose of this experiment to be the work of an unknown author:

	Gratian0
in	-2.8702
non	-6.5491
et	-3.2375

	Gratian0
est	-3.5264

the other for the comparison samples Gratian1, dePen, and Gratian2, assumed for the purpose of this experiment to represent the work of known authors:

	Gratian1	dePen	Gratian2
in	-0.4342	-0.7095	1.1437
non	-0.0361	1.0176	-0.9814
et	-0.9786	1.0201	-0.0414
est	0.4179	0.7233	-1.1412

The formula used to calculate Burrows's Delta is:

$$\Delta_B = \frac{1}{N}\sum_{i=1}^{N}|z_i(t) - z_i(c)|$$

It is easiest to deal with the formula in two steps, first evaluating the expression $|z_i(t) - z_i(c)|$. Note that because we take the absolute value of the result, the order of operands on either side of the subtraction operator '-' does not matter. For each of the three columns (Gratian1, dePen, and Gratian2) in the *corpus* dataframe, subtract the z-score in each row from the z-score in the same row of the *test* (Gratian0) dataframe, take the absolute value, and record the result in the corresponding column and row of the

differences dataframe. For example, the z-score for *non* in *test* (Gratian0) is -6.5491, the z-score for *non* in the Gratian1 column of *corpus* is -0.0361, so the absolute value of the difference recorded in the *non* row of the Gratian1 column of *differences* would be 6.5130.

	Gratian1	dePen	Gratian2
in	2.436	2.1606	4.0139
non	6.513	7.5667	5.5677
et	2.2589	4.2576	3.1961
est	3.9443	4.2497	2.3852

Given the layout of the *differences* dataframe in which we have stored the intermediate results, the part of the formula we deferred dealing with ($\frac{1}{N}\sum_{i=1}^{N}$) is simply a notationally exact way of indicating that we are to take the average (arithmetic mean) of the values in each of the columns and record the resulting value of Δ_B in the corresponding column of the *deltas* dataframe.

The seemingly simple act of taking the arithmetic mean (average) of the z-score distances between the samples for each feature has an interesting and non-intuitive implication. It was mentioned in passing in the previous section on visualization that plotting the z-score coordinates of word frequencies invokes the tacit assumption that the axes are in fact perpendicular to one another, an assumption that is at least

potentially open to challenge. Burrows's Delta generalizes this assumption into an arbitrary number of dimensions. The scholarly literature on authorship attribution methods describes distance metrics such as Burrows's Delta as measuring 'Manhattan Distance'. The analogy is to walking or driving from a starting to an ending point through a space in which the streets have been laid out at right angles to one another, like Manhattan.

	Gratian1	dePen	Gratian2
Gratian0	3.788	4.5586	3.7907

The Gratian1 subcorpus is just slightly closer than the Gratian2 subcorpus to the unknown Gratian0 test case, with values of Delta for both rounding to 3.79. A candidate is defined as being *closest* to the unknown when it has the lowest mean of the absolute values of the differences between the z-scores for the unknown and the candidate.But as Burrows pointed out, one candidate will always have the lowest Δ_B, so that in itself is not enough to make or to rule out an attribution of authorship. We will need further information before we can provide any kind of interpretation for the result. The most we can say based on this result is that the hypothetical case statements are less likely to have been written by the author of the *dicta* in *de Penitentia* than by the authors of either the first- or second-recension *dicta*.

The second experiment is a variation on the first, in which a 3881-word sample made up of seven extended passages from the pseudo-Augustinian *De vera et falsa penitentia* quoted by Gratian in *de Penitentia* are substituted for the 3605-word sample containing the hypothetical case statements.[42] As noted in Chapter 0 above, Gratian can be said with a high degree of confidence *not* to be the author of *De vera et falsa penitentia*. The authors are strongly distinguished by their choice of post-positive conjunctions: Gratian has a preference for *autem*, while pseudo-Augustine has an even stronger preference for *enim*. Substituting the pseudo-Augustinian sample in place of the case statements demonstrates the kinds of results to be expected from Burrows's Delta in a situation in which an attribution of authorship can reasonably be ruled out.

	Gratian1	dePen	Gratian2
psAug	2.6456	1.7373	3.4318

The third experiment extends the first by treating each of the subcorpora, Gratian0, Gratian1, dePen, and Gratian2 sequentially as the work of an unknown author, and the

[42] *de Penitentia* D.1 c.88 (R1), D.3 c.42 (R1), D.3 c.49 (R1), D.5 c.1 (R1), D.6 c.1 (R1), and D.7 c.6 (R1). These seven extended passages average 554.4 words in length. **See edF 1.XXXV, for a complete list of passages from *De vera et falsa penitentia* quoted in the *Decretum*. Explain rationale for omitting certain passages: D.25 c.5 (R2 or Palea), *de Penitentia* D.3 c.4.5 (what Friedberg means by 4.5 in this context is unclear), D.3 c.45 (R2). Acknowledge Karen Teresa Wagner, *De vera et falsa penitentia : an edition and study*, 1995.**

other three subcorpora as constituting a corpus of works by a known author. This is an attempt to demonstrate the adaptation of Burrows's technique in a circumstance in which there are no securely attributed comparison texts outside of the corpus, and in which there is some reason to suspect that there are multiple authors at work within the corpus.

	Gratian0	Gratian1	dePen	Gratian2
Gratian0	nan	3.788	4.5586	3.7907
Gratian1	1.4361	nan	0.3628	0.5453
dePen	1.9873	0.4515	nan	0.7673
Gratian2	1.7185	0.6278	0.7905	nan

Considering the results of the first three experiments together, we can start to form some very preliminary conclusions. Based on the values for Δ_B in the table above, the most likely attribution is that the first-recension *dicta* (Gratian1) and the *dicta* from *de Penitentia* (dePen) have the same author. It is less likely that the first-recension *dicta* (Gratian1) and the second-recension *dicta* (Gratian2) have the same author. It is still less likely that the *dicta* from *de Penitentia* and the second-recension *dicta* have the same author. It is much less likely that the case statements (Gratian0) have the same author as either the first- (Gratian1) or second-recension (Gratian2) *dicta*. Finally, the least likely

attribution is that the case statements (Gratian0) have the same author as the *dicta* from *de Penitentia*.

The fourth and final experiment will compare the thirty most frequent words (MFWs) across fourteen subcorpora: cases (C.1-36 d.init.), laws (D.1-20 R1 *dicta*), orders1 (D.21-80 R1 *dicta*), orders2 (D.81-101 R1 *dicta*), simony (C.1 R1 *dicta*), procedure (C.2-6 R1 *dicta*), other1 (C.7-10 R1 *dicta*), other2 (C.11-15 R1 *dicta*), monastic (C.16-20 R1 *dicta*), other3 (C.21-22 R1 *dicta*), heresy (C.23-26 R1 *dicta*), marriage (C.27-36 R1 *dicta*), penance (R1 and R2 *dicta* from *de Penitentia*), and second (all R2 *dicta*, excluding those from *de Penitentia*).[43] For each of the fourteen subcorpora, we will hypothesize each subcorpus in turn to be the work of an unknown author and will treat the other thirteen subcorpora as composing a corpus of works by a known author. The scale of the fourth experiment is similar to that of the experiments carried out by John Burrows and David Hoover, the pioneers of the technique, but makes it impractical to show intermediate results at every step in the process.

---

[43] The division of the first-recension (R1) *dicta* into twelve sections follows the division of Gratian's *Decretum* proposed in Alfred Beyer, *Lokale Abbreviationen des Decretum Gratiani: Analyse und Vergleich der Dekretabbreviationen "Omnes leges aut divine" (Bamberg), "Humanum genus duobus regitur" (Pommersfelden) und "De his qui intra claustra monasterii consistunt" (Lichtenthal, Baden-Baden)*, Bamberger theologische Studien ; Bd. 6 (Frankfurt am Main ; PLang, 1998), 17–18.

	cases	laws	orders1	orders2	simony	procedure	other1
cases	nan	2.2765	1.9247	2.0252	1.9637	1.9545	1.5714
laws	2.141	nan	1.249	1.502	1.4633	1.3147	1.4223
orders1	1.6184	1.0949	nan	1.1223	0.9685	0.8843	1.0499
orders2	1.8982	1.5244	1.2686	nan	1.382	1.684	1.4149
simony	1.6667	1.3491	0.9772	1.2195	nan	0.8878	1.1304
procedure	1.6187	1.1991	0.892	1.5095	0.8789	nan	1.079
other1	1.3353	1.3	1.0619	1.2722	1.1383	1.0753	nan
other2	1.9416	1.3233	1.0913	1.6291	1.1386	1.109	1.2963
monastic	1.4555	1.0451	0.8554	1.2676	1.0114	0.7986	0.93
other3	2.0705	1.3388	1.289	1.5146	1.1997	1.1057	1.3497
heresy	1.5177	1.031	0.7772	1.2182	0.5544	0.595	0.9485
marriage	1.5448	1.0263	0.9848	1.265	0.884	0.9667	1.0494
penance	1.5371	1.4473	0.7478	1.4005	0.9024	0.8781	1.3077
second	1.374	1.0852	0.7764	1.1717	1.0623	0.9634	0.7971

	other2	monastic	other3	heresy	marriage	penance	second
cases	2.2782	1.7622	2.3628	1.8717	1.8923	1.8589	1.6334

	other2	monastic	other3	heresy	marriage	penance	second
laws	1.4369	1.1931	1.4345	1.1875	1.1924	1.6218	1.2323
orders1	1.1109	0.8693	1.2397	0.8267	1.0124	0.7505	0.7777
orders2	1.6873	1.4492	1.6208	1.4198	1.4526	1.5523	1.3195
simony	1.1287	1.0413	1.1711	0.59	0.9166	0.9059	1.0863
procedure	1.1223	0.821	1.0726	0.6569	0.9993	0.8818	0.9852
other1	1.2792	0.9649	1.3054	0.996	1.0853	1.3272	0.8152
other2	nan	0.7979	1.0346	1.0592	0.654	0.8633	1.0961
monastic	0.7429	nan	1.0578	0.7602	0.6611	0.7999	0.7799
other3	0.9505	1.1229	nan	1.1209	0.7121	1.1521	1.3067
heresy	0.9839	0.7672	1.1395	nan	0.7783	0.6756	0.8484
marriage	0.6126	0.6577	0.6552	0.7672	nan	0.7974	0.8676
penance	0.9152	0.8101	1.0992	0.7609	0.8146	nan	0.9026
second	1.0674	0.7861	1.2408	0.877	0.8796	0.8927	nan

Because of the scale of the experiment, the results can be somewhat difficult to read, but they are entirely consistent with those obtained in the previous simplified experiments. They are divided into two tables to allow them to be represented on the printed page but should be imagined as a single table, with the second table extending the first table to the right. The first column of each row contains the name of the subcorpus

hypothesized to be the work of an unknown author. The previously obtained results from the simplified demonstration examples lead us to expect that the cases subcorpus corresponding to the 36 hypothetical case statements or *themata* will have highest value for Burrows's Delta in each row. Remember that the cases subcorpus having the highest Delta value in a given row indicates that it is the *least* likely to have the same author as the subcorpus indicated in the first column and hypothesized to be the work of an unknown author.

Disregard the first row—we are not interested in the Delta distance of the cases subcorpus from itself. Read each row starting at the second, comparing the value for the Delta distance between the subcorpus of unknown authorship and the cases subcorpus with the Delta values for each of the other subcorpora. Taking the second row of each table as an example, the value for the Delta distance between the laws and cases subcorpora is 2.141, which is greater than NaN, 1.249, 1.502, 1.4633, 1.3147, 1.4223 in the first table and, continuing on to the corresponding row in the second table, is also greater than 1.4369, 1.1931, 1.4345, 1.1875, 1.1924, 1.6218, and 1.2323. (In each row, the entry corresponding to the Delta distance between the subcorpus of unknown authorship and itself is undefined and is indicated by "NaN," a conventional abbreviation in numerical computing for "Not a Number.")

For each of the thirteen subcorpora, excluding cases, the value for the Burrows's Delta distance between that subcorpus and the cases subcorpus is the highest in the row. In only one row are there Delta values that are even close to the value for the cases subcorpus: for the other1 subcorpus (first-recension *dicta* from *Causae* 7-10), the Delta value for the cases subcorpus is 1.3353, while the Delta values for laws (first-recension *dicta* from the *Tractatus de legibus*), other3 (first-recension *dicta* from *Causae* 21-22), and penance (first- and second-recension *dicta* from *de Penitentia*) are 1.3, 1.3054, and 1.3272 respectively. Even in this case, the Delta value for the cases subcorpus indicates that it is the *least* likely of any of the subcorpora in the row to share an author in common with the other1 subcorpus.

Principal component analysis

Techniques such as Burrow's and Argamon's Delta (measuring Manhattan and Euclidean distance respectively), which collapse vector distance data for an arbitrary number of features or dimensions into a single scalar value interpreted as a nearest-neighbor classification metric, are one way of reducing feature distances to a tractable form. Principal component analysis (PCA) is an alternative to Delta metrics that projects vector distance information for numbers of features greater than three into a two- or three- dimensional space for convenient visualization. PCA therefore has the advantage

that it entails less loss of information than the Delta class of techniques that reduce data for all dimensions to a single metric.[44]

PCA first combines as many of the raw dimensions as possible into synthetic components on the basis of strong correlations, either positive or negative. For example, referring back to Figures 0a and 0b in the two-dimensional visualization section above, the two dimensions of the plot could be collapsed into a single axis or component that can be thought of as representing the frequency with which *in* does, and *non* does *not*, occur in a given sample. The effect would be to reconfigure the plots in such a way that the samples representing the *dicta* from *de Penitentia* (dePen), the first-recension *dicta* (Gratian1), and the second-recension *dicta* (Gratian2) would be placed from left to right along a single horizontal axis. PCA then displays the two components that contribute

---

[44] Earlier versions of this section were presented as conference papers. "Can Stylometry Provide New Evidence about the Identity of Gratian 1 and Gratian 2?" was presented to the session on Canon Law in the Twelfth and Thirteenth Centuries at the *Rem non novam nec insolitam aggredimur* conference and grand opening of the Stephan Kuttner Institute of Medieval Canon Law at Yale Law School, May 21-22, 2015. [Greta Austin, Thomas Bisson, Uta-Renate Blumenthal, Bruce Brasington, Melodie Eichbauer, Richard Helmholz, Eric Knibbs, Peter Landau, Kenneth Pennington, Edward Peters (University of Pennsylvania), Robert Somerville, and Anders Winroth.] "New evidence for the authorship of case statements and *dicta* in Gratian's *Decretum*" was presented to the Classical Sources III session at the Fifteenth International Congress of Medieval Canon Law (ICMCL) at Université Paris II Panthéon-Assas, July 17-23, 2016. [Gero Dolezalek, Anders Winroth (session chair).]

the most to the total variation between the samples, and graphically arranges the samples according to their probability relative to those two components.[45]

I used the Stylometry with R (stylo) package for computational text analysis developed by Maciej Eder, Jan Rybicki, and Mike Kestemont of the Computational Stylistics Group to generate all of the PCA plots in this section.[46] R is a statistically-oriented programming language.[47] In addition to his being one of the lead developers of the stylo R package, Kestemont is a researcher whose stylometric analysis of two visionary texts of Hildegard of Bingen was a useful example for this project.[48]

---

[45] For a general introduction to the use of principal component analysis (PCA) in literary stylometric analysis, see Hugh Craig, "Stylistic Analysis and Authorship Studies," in *A Companion to Digital Humanities*, ed. Susan Schreibman, Raymond George Siemens, and John Unsworth, Blackwell Companions to Literature and Culture 26 (Malden, MA: Blackwell Pub, 2004), 273–88 and Chapter 6 "Style" in Matthew Lee Jockers, *Macroanalysis: Digital Methods and Literary History*, Topics in the Digital Humanities (Urbana: University of Illinois Press, 2013).

[46] Maciej Eder, Jan Rybicki, and Mike Kestemont, "Stylometry with r: A Package for Computational Text Analysis," *R Journal* 8, no. 1 (2016): 107–21, https://journal.r-project.org/archive/2016/RJ-2016-007/index.html.

[47] R Core Team, *R: A Language and Environment for Statistical Computing* (Vienna, Austria: R Foundation for Statistical Computing, 2020), https://www.R-project.org/.

[48] See Mike Kestemont, Sara Moens, and Jeroen Deploige, "Stylometry and the Complex Authorship in Hildegard of Bingen's Oeuvre," in *Digital Humanities 2013: Conference Abstracts* (Lincoln, NE: University of Nebraska–Lincoln, 2013), 255–58, http://dh2013.unl.edu/abstracts/ab-126.html; and Mike Kestemont, Sara Moens, and Jeroen Deploige, "Collaborative Authorship in the Twelfth Century: A Stylometric

Stylometric analysis for the purpose of authorship attribution rests on the frequencies of occurrence of function words including conjunctions. It is therefore essential to properly account for the frequencies of enclitic endings representing conjunctions. Each word in the samples ending with *-que* where the ending represents an enclitic being used as a conjunction and is not simply part of the word has been mapped to a two-word sequence consisting of the word plus the pseudo-conjunction *xque*.[49] Other Latin enclitic

---

Study of Hildegard of Bingen and Guibert of Gembloux," *Literary and Linguistic Computing* 30, no. 2 (June 2015): 199–224. Kestemont was very generous in his technical advice during the early stages of this project.

[49] Following the example of Kestemont, Moens, and Deploige, "Collaborative Authorship in the Twelfth Century," 205: "To automatically isolate the clitic, we have stripped the suffix ('x*que*') from every word that did not occur in a list of words proposed by Schinke *et al*. (1996, p. 180-1)."

The list of words appears in Robyn Schinke et al., "A Stemming Algorithm for Latin Text Databases," *Journal of Documentation* 52, no. 2 (1996): 172–87. Schinke's article was published in a hard-to-find journal. The article is frequently referenced (55 citations in Google Scholar as of 18 March 2021), but I was unable to obtain a copy. My information about Schinke's stemming algorithm and pass list comes indirectly via Martin Porter, "The Schinke Latin Stemming Algorithm," accessed March 18, 2021, https://snowballstem.org/otherapps/schinke/.

In the case statements, 1st-, and 2nd-recension *dicta* from Gratian's *Decretum*, there are 747 occurrences of 79 unique words ending in *-que*. (This does not count 423 occurrences of the word *que* itself.) Of those, 498 are occurrences of 19 unique words from Schinke's 54-word pass list, while 249 occurrences of 60 unique words are not. It is from these 249 words that, according to Schinke, the *-que* ending should be detached as an enclitic.

However, the 249 words include 72 occurrences of 17 unique words ending with the adverbial enclitics *-cumque* or *-cunque*, from which the *-que* ending should not be

endings such as *-ne* and *-ve* occur infrequently enough in the samples that they can be disregarded for the purpose of pseudo-conjunction mapping.

Figure 1 below shows the PCA plot generated by a four-way comparison of the same samples used in the demonstration of Burrows's Delta in the previous section: the hypothetical case statements or *themata* (Gratian0)[50], the first-recension *dicta* excluding

---

detached. The 249 words also include a further 149 occurrences of 21 unique false positives:

cumque, eque (aeque), namque, pleraque, plerique, plerisque, plerumque, quinque, unamquamque, unaqueque, unicuique, uniuscuiusque, unumquemque, unusquisque, usquequaque, utramque, utraque, utrique, utrisque, utriusque, utrumque.

This leaves only 28 occurrences of 22 unique words from which the *-que* ending should actually be detached as an enclitic.

False positives over-represent the frequency of occurrence of the *-que* enclitic as a conjunction by an order of magnitude. Including all false positives makes *xque* the 37th most frequent word in the sample, while excluding them makes it the 376th most frequent word. There are 55 occurrences of the word *namque*, the most frequently occurring false positive. Detaching the *-que* ending from *namque* overstates the frequency of *nam*, making what is actually the 480th most frequent word appear to be the 130th, while making *namque*, which is actually the 176th most frequent word in the samples when false positives are excluded, disappear from the list altogether.

[50] As noted in the previous two-dimensional visualization section, the Gratian0 sample containing the hypothetical case statements or *themata* includes a thirteen-word clause added to C.19 d.init. between the first and second recensions of the *Decretum*. None of the wordlists used to perform principal component analysis include any of those thirteen words, so using the text of C.19 d.init. found in the Friedberg edition rather than a proxy first-recension version of the text has no effect on the outcome of any of the tests performed in this section.

the *dicta* from *de Penitentia* (Gratian1), first- and second-recension *dicta* from *de Penitentia* (dePen), and the second-recension *dicta* excluding the *dicta* from *de Penitentia* (Gratian2). The case statements are magenta (∆), the first-recension *dicta* are green (+), the *dicta* from *de Penitentia* are blue (○), and the second-recension *dicta* are red (×). Each of the texts has been divided into 1200-words samples. Principal component 1 along the horizontal axis is 10.3%. Principal component 2 along the vertical axis is 7.2%. That is, PC1 explains 10.3% of the total variation between the samples, and PC2 explains 7.2% of the total variation between the samples. This is good: as a general rule, we want to see a value for PC1 greater than 10% and we want to see a value for PC2 greater than 5%. The most visually striking aspect of this plot is the fact that the case statements are so far away from the *dicta*, and the next step is to take a look at which features produce that effect.

```
-----------------------------------------------
features (e.g. frequent words) actually analyzed
-----------------------------------------------

  [1]  in       et       non      de       quod     ad       sed      uel
  [9]  unde     si       ut       a        autem    cum      ex       enim
 [17]  uero     etiam    ab       ergo     quia     item     per      nec
 [25]  an       sicut    ita      nisi     tamen    pro      quam     sic
 [33]  quo      sine     aut      licet    post     contra   siue     quoque
 [41]  ante     ne       inter    super    atque    dum      apud     postea
 [49]  ideo     propter  ecce     quomodo
```

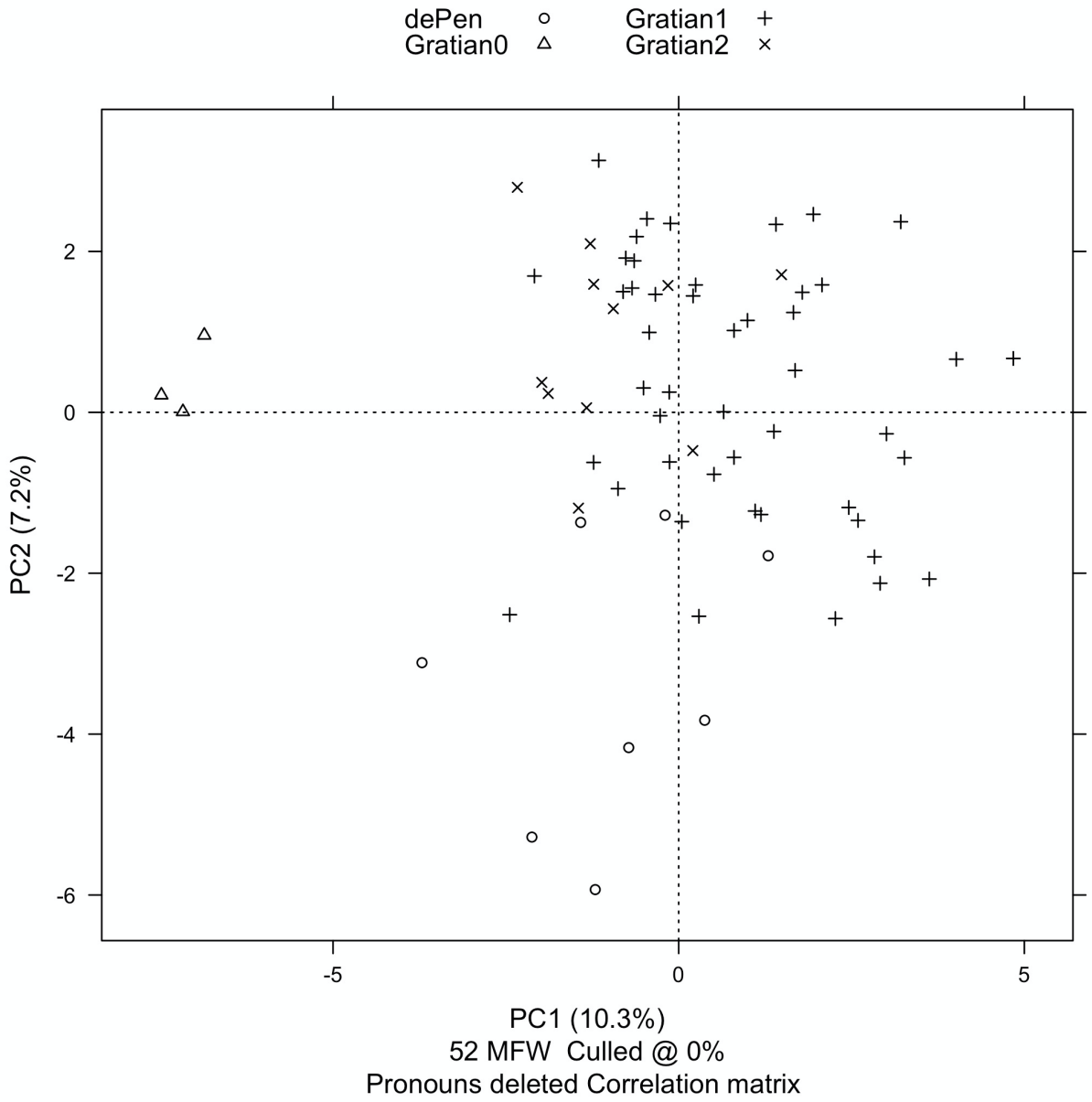

Figure 7: Figure 1 updated 28 May 2020

Turning on the stylo feature loadings option lets us see how strongly particular words influence the placement of text samples along the PC1 and PC2 axes; this is called the feature's discriminative strength. (See Figure 2 below.) For example, *sed* and *non* are located toward the right (positive) end of the PC1 axis, while *an* and *si* are located

toward the left (negative) end of the PC1 axis. Similarly, *uel* and *uero* are located toward the upper (positive) end of the PC2 axis, while *quomodo* is located toward the lower (negative) end of the PC2 axis. As we will see immediately below, *an* and *si* are closely associated with the hypothetical case statements (*themata*), while *quomodo* is closely associated with the first- and second-recesion *dicta* from *de Penitentia*.

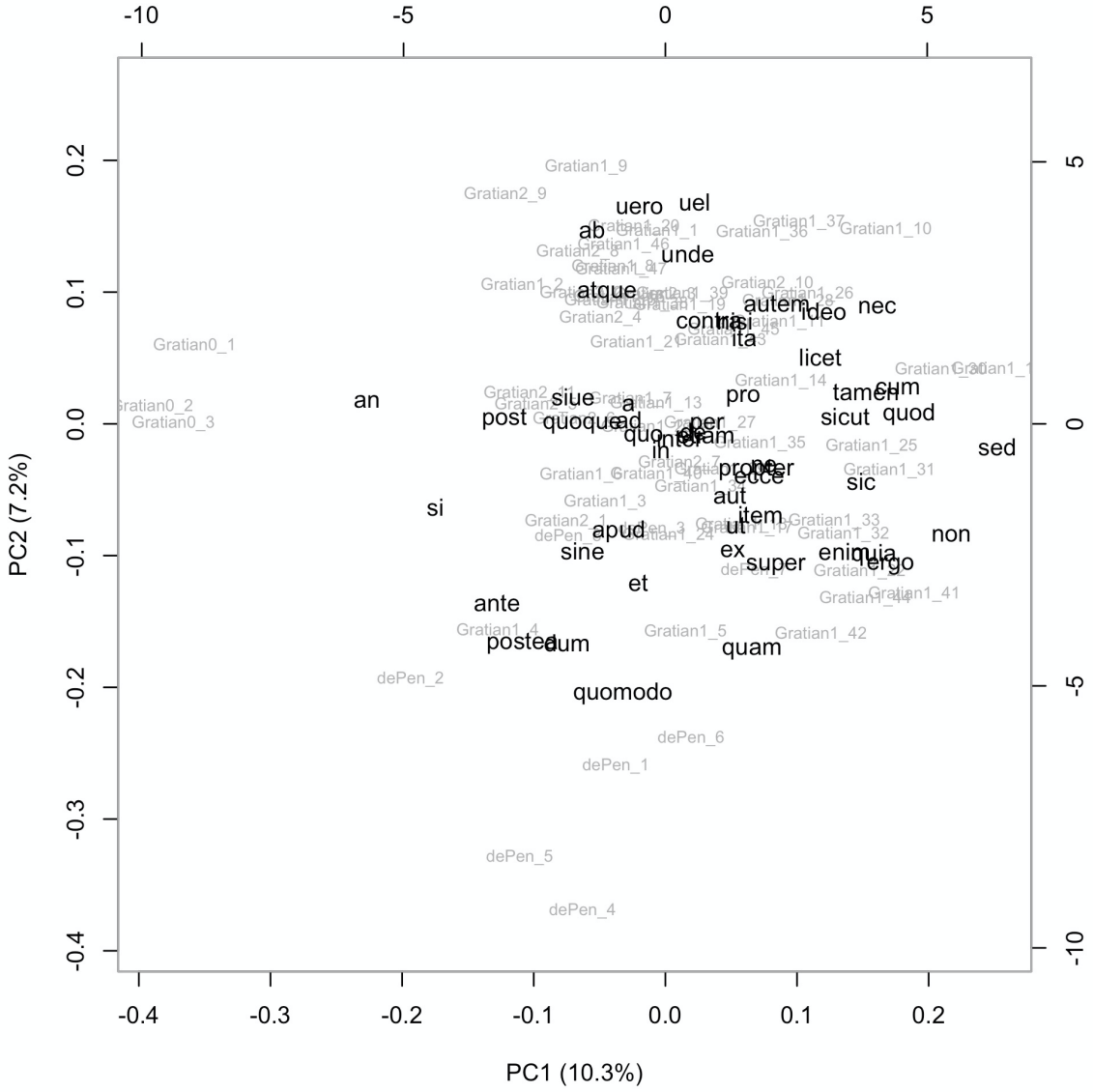

Figure 8: Figure 2 updated 22 Mar 2020

In the initial function word counting experiment, *non*, the second most common word in the samples, was strongly associated with the first-recension *dicta*.[51] In Figure 2, *non*

[51] **PLE**: This is true if only the first- and second-recension *dicta* are counted. If the case statements, the first- and second-recension *dicta*, and the *dicta* from *de Penitentia* are

appears far to the right, and in fact the samples from the first-recension *dicta*, but not those from the second-recension *dicta*, tend to spread out to the right. Note however that *in*, the most common word in the samples, is fairly close to the middle. To the extent then that *in* is more strongly associated with the second-recension than the first-recension *dicta*, it is a result of the fact that the word occurs less frequently in the first-recension *dicta* rather than that it occurs more frequently in the second-recension *dicta*.

The most visually striking feature of the function loadings plot in Figure 2 is the degree to which *an* and *si* cluster with the case statements, *an* very strongly so, *si* somewhat less strongly. This makes intuitive sense because indirect questions dominate the language of the case statements. It is a question of genre. There are two possible ways in which we might go about controlling for the vocabulary characteristic of the question genre in the *themata*: by editing the Gratian0 sample to remove the passages containing indirect questions from each of the case statements, and by editing the list of function words used by stylo to conduct the analysis to exclude individual words characteristic of indirect questions.

The case statements all follow a very regular formal pattern. They are introduced by a hypothetical narrative that is followed by an enumeration of the questions that Gratian

---

counted, *in* is the most frequent word, *et* is the second most frequent word, and *non* is the third most frequent word.

wants to investigate. C.27 d.init. (chosen for this purpose because it is the shortest case statement) demonstrates the pattern:

> *Quidam uotum castitatis habens desponsauit sibi uxorem; illa priori condicioni renuncians, transtulit se ad alium, et nupsit illi; ille, cui prius desponsata fuerat, repetit eam. Hic primum queritur, an coniugium possit esse inter uouentes? Secundo, an liceat, sponsae a sponso recedere, et alii nubere?*[52]

The transition between the narrative section and the enumeration of questions is clearly signalled in each of the case statements by the use of one of a small number of formulaic markers, of which *Hic primum queritur* is the most common.[53]

Running principal component analysis (PCA) after removing the enumerated questions from the cases statements (leaving all other samples unchanged) is not, however, a viable approach, because at 1,618 words, a Gratian0 *sine questionibus* sample would be

---

[52] A man having [made] a vow of chastity betrothed a wife to himself; she, renouncing her previous agreement, gave herself to another and married him; he to whom she had been first betrothed tried to get her back. Here it is first asked whether there is able to be a marriage between those vowing? Second, whether someone betrothed is allowed to abandon the person to whom they are betrothed and to marry another?

[53] The formulaic transition markers used in the hypothetical case statements are: *Hic primum queritur* (15), *Queritur* (8), *Modo primum queritur* (3), *Nunc primum queritur* (3), *Primo queritur* (2), *Primum queritur* (2), *Hic primo queritur* (1), *Modo queritur* (1), *Queritur autem* (1).

too far under the approximately 2,500-word minimum recommended for analysis of Latin prose.

The remaining alternative is to edit the list of function words used by stylo to conduct its analysis so that it excludes individual words characteristic of indirect questions, starting with the words *an* and *si* suggested by the stylo feature loadings.

The frequency of occurrence of the word *an* in the Gratian0 sample representing the thirty-six hypothetical case statements (*themata*) is a remarkably high 39.1123 occurrences per 1,000 words. By way of comparison, the mean frequency of occurrence of *an* across the three samples representing the first-recension *dicta* excluding *de Penitentia* (Gratian1), the first- and second-recension *dicta* from *de Penitentia* (dePen), and the second recension *dicta* (Gratian2) is 1.3815 occurrences per 1,000 words with a sample standard deviation of 0.5011. The frequency of occurrence of *an* in the Gratian0 sample is therefore 75.2996 standard deviations away from the mean frequency of occurrence of the same word in the Gratian1, dePen, and Gratian2 samples. The frequency of occurrence of the word *si* in the Gratian0 sample, on the other hand, is 14.4244 occurrences per 1,000 words. Given that the mean frequency of occurrence of *si* across the Gratian1, dePen, and Gratian2 samples is 9.2665 occurrences per 1,000 words and that the sample standard deviation is 2.6245, the frequency of occurrence of *si* in the

Gratian0 sample is far less of an outlier at 1.9653 standard deviations away from the mean than *an* was.

[Revised/Updated down to here!!!]

[Remove *an*, leave *si* in!!!]

We've now reached the final stage of the three-way comparison between the case statements, the first-recension *dicta*, and the second-recension *dicta*. We are now using the 49 most frequent words on our function list instead of the 51 most frequent words, having commented out *an* and *si*. And even without *an* and *si*, PC1 still explains 10.5% of the total variation between the samples, down slightly from 11.2%. PC2 still explains 7.3% of the total variation between the samples. So, even controlling for genre, the distance between the case statements and the *dicta*—both first- and second-recension—is still quite striking.

```
------------------------------------------------
features (e.g. frequent words) actually analyzed
------------------------------------------------

  [1]  in       et       non      de       quod     ad       sed      uel
  [9]  unde     si       ut       a        autem    cum      ex       enim
 [17]  uero     etiam    ab       ergo     quia     item     per      nec
 [25]  sicut    ita      nisi     tamen    pro      quam     sic      quo
 [33]  sine     aut      licet    post     contra   siue     quoque   ante
 [41]  ne       inter    super    atque    dum      apud     postea   ideo
 [49]  propter  ecce     quomodo
```

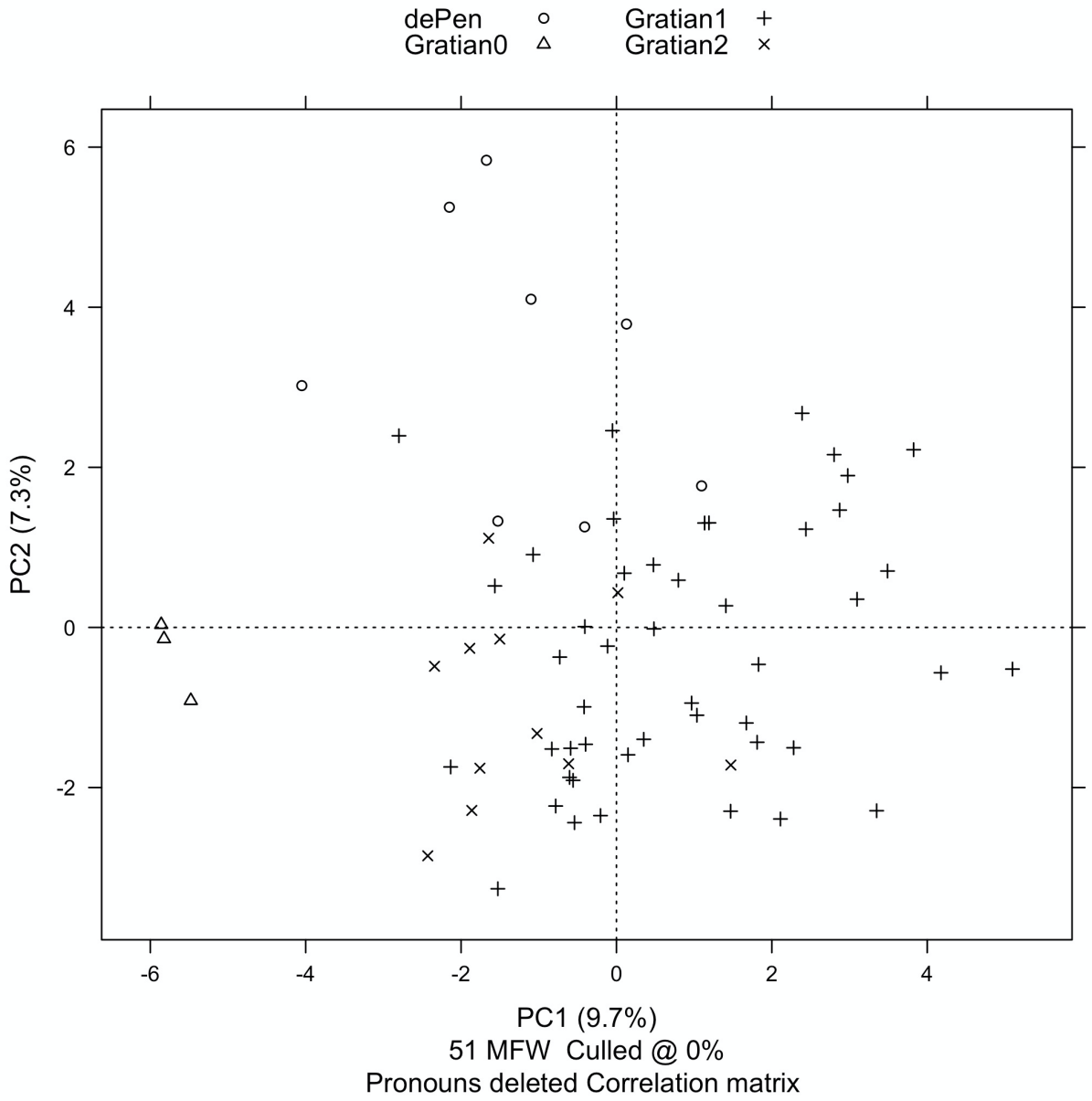

Figure 9: Figure 3 updated 2 Jul 2020

To turn to the other interesting aspect of the three-way comparison, you'll note that the second-recension *dicta* in blue cluster strongly to the upper-left quadrant. Now, Mike Witmore, a member of my dissertation committee who isn't an insider with respect to

debates about Gratian's *Decretum*, but is very experienced in the use of stylometry with the plays of Shakespeare, was somewhat optimistic on the basis of this evidence that the first- and second-recension *dicta* might be statistically distinguishable.

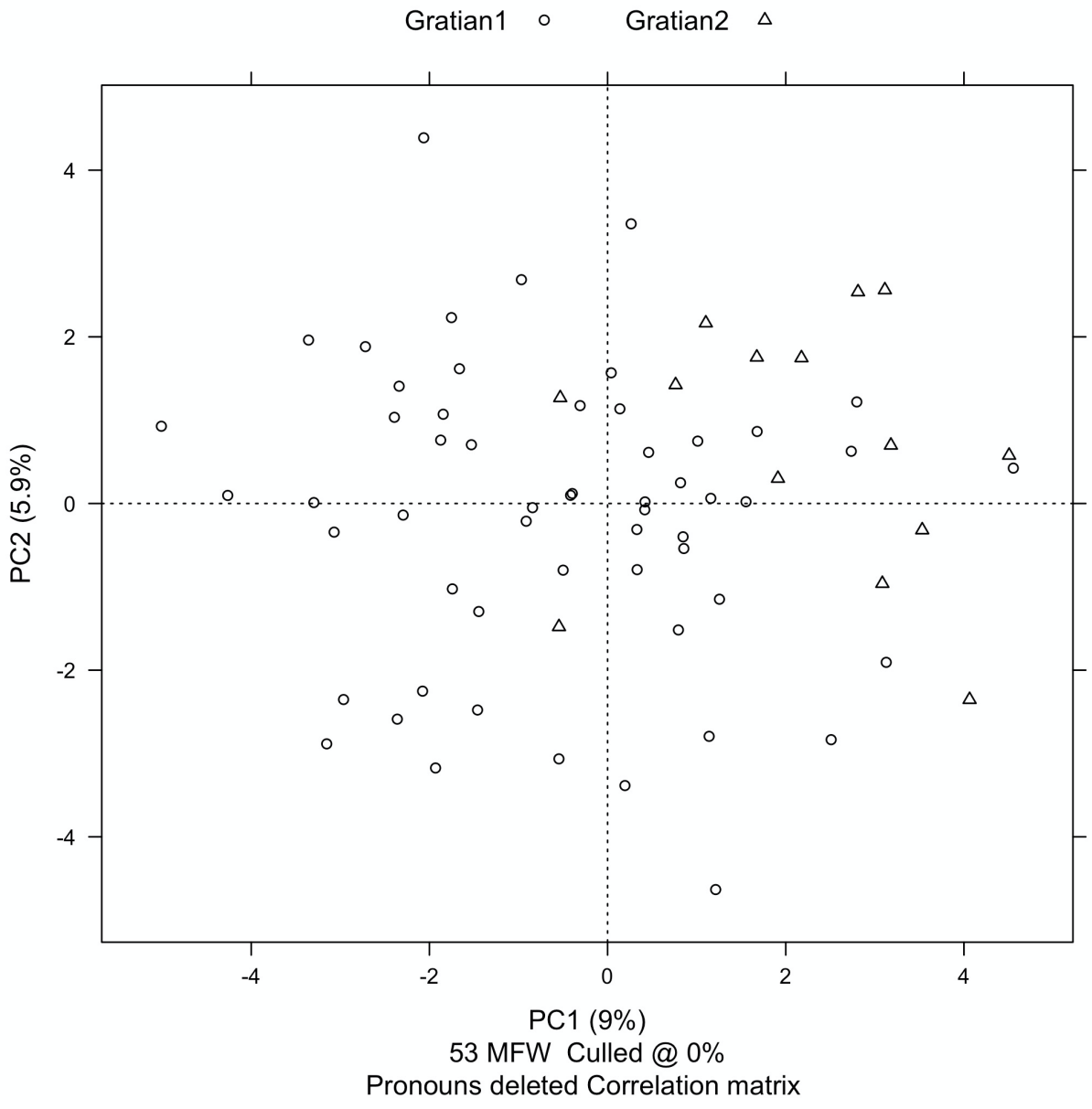

Figure 10: Figure 7 updated 19 Mar 2020

So, in an attempt to take a closer look at the *dicta* by themselves, I removed the case statements and ran a two-way comparison of 1000-word samples of just the first- and second-recension *dicta*, again, excluding the *dicta* from *de Penitentia*. (Stylo changes the color assignments depending on the number of samples, so in this plot the first-recension *dicta* are red and the second-recension *dicta* are green.) And the results are ambiguous. The PC1 axis is 9%, somewhat under the 10% threshold we would like to see. Also, although we see the second-recension *dicta* clustering mostly to the right of the PC1 axis, the two sets of samples are not separated as cleanly as we'd like to see, and certainly nowhere near as cleanly as the case statements were from the *dicta*.

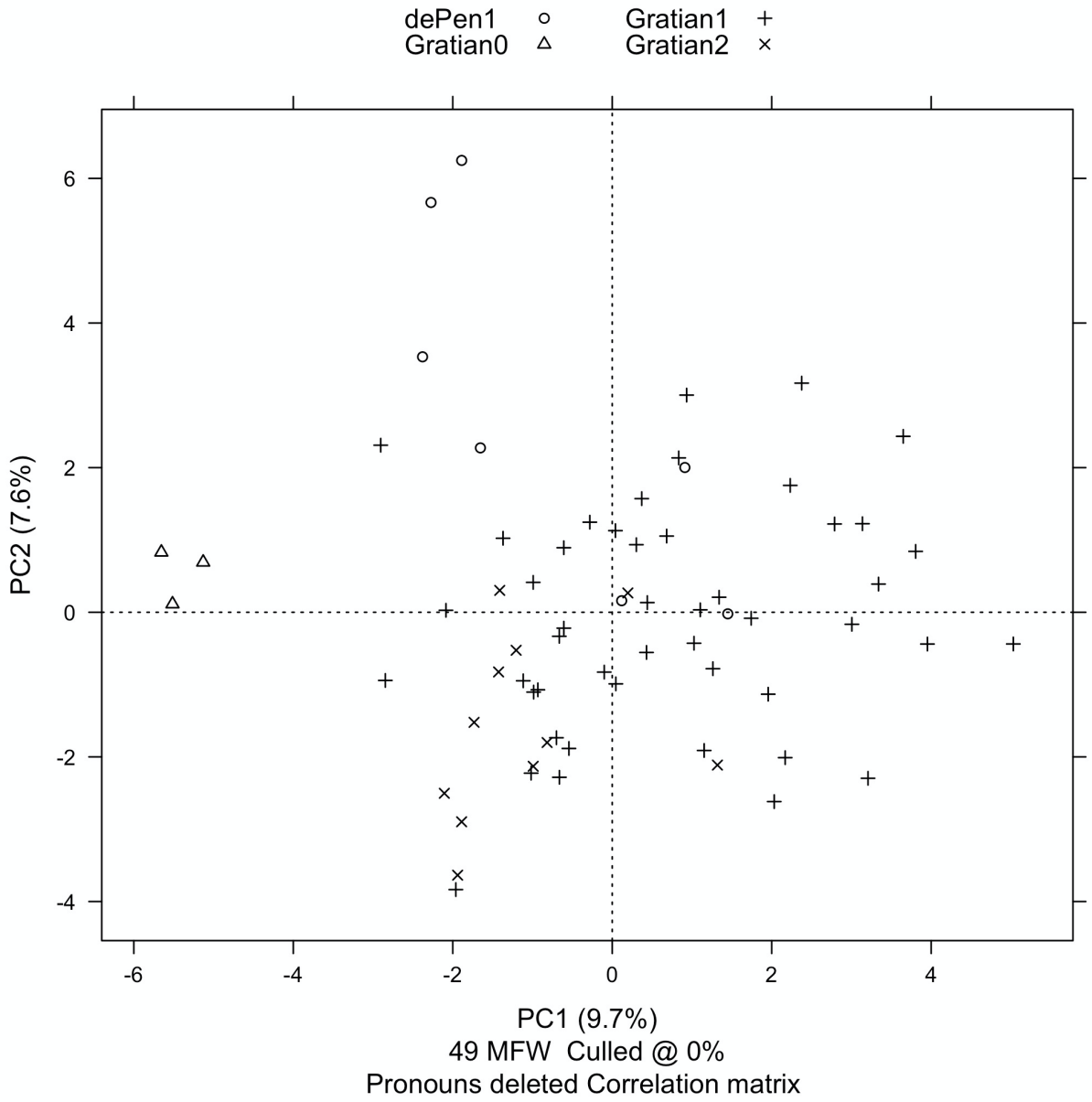

Figure 11: Figure 8 updated 19 Mar 2020

All of the slides we've seen so far exclude the *dicta* from *de Penitentia,* so before moving on to my conclusion, I do want to quickly show you what the results look like when we include the first-recension *dicta* from *de Pen.* (there are not enough words in the second-recension *dicta* in *de Pen.* to be statistically significant—9,525 vs. 556). Many scholars

have observed that *dicta* and canons are poorly separated in *de Pen.* I believe that the unusual dispersion of the samples that you see in this plot is a result of that feature.

Conclusion

Principal component analysis (PCA) of the frequencies of function words (prepositions and conjunctions) in the texts strongly suggests that the author of the case statements was not the same person as the authors of either the first- or second-recension *dicta*. PCA also suggests (less strongly) that the first- and second-recension *dicta* were not the work of either one or two authors, but are more likely to have been the product of collaborative authorship.