A Stylometric Study of Nicholas of Montiéramey's Authorship in Bernard of Clairvaux's *Sermones de Diversis*

Jeroen De Gussem
jedgusse.degussem@ugent.be
Universiteit Gent, Belgium

Case Study

This short paper revisits the authorship of Bernard of Clairvaux's *Sermones de Diversis* (c. 1090 – 1153) through computational stylistics. Bernard's *De diversis* corpus comprises an assembly of unpolished and rudimentary sermons found in various, heterogeneous manuscripts. Bernard never disseminated the *De diversis* sermons himself, they have been first assembled, enumerated and published by Jean Mabillon in the 17th century (Callerot, 2006). Since Bernard of Clairvaux usually collaborated with secretaries, the obscure context of the corpus' composition and constitution has often made its sermons subject to some debate when it comes to Bernard's authorship. By 1145, the abbot's acclaim as the icon and figurehead of the Cistercian movement had brought along such a considerable administrative workload that the assistance of a group of secretaries was indispensable. These secretaries acted as Bernard's stand-ins, and spared him the time and effort it would cost of having to take up the quill himself at every single occasion. The *reportatio*, as it was called, entailed that the contents of Bernard's letters or sermons were engraved on wax tablets in a tachygraphic fashion. The cues, keywords and biblical references which Bernard had spoken aloud provided a framework that captured the gist of his diction. Afterwards, the scribe reconstructed what he had heard to a text on parchment which could pass for Bernard of Clairvaux's in its literary allure (Rassow, 1913; Leclercq, (2)1962; Constable, 1972). Amongst these amanuenses, Nicholas of Montiéramey († 1176 / 78) was as a focal figure, and a highly skilled imitator of his master's writing style. The influence of Nicholas' mediation on several particular text instances within Bernard's *De diversis*, and more generally on his entire oeuvre, has fallen subject to much debate.

Nicholas began serving Bernard as an emissary around 1138-41, carrying letters concerning Abelard's heresy to Rome (Turcan-Verkerk 2015). His literary qualities, likely to have been acquired through his education in the Benedictine abbey of Montiéramey, enabled him to immediately enter the scriptorium and officially become Bernard's closest secretary. Their friendship, however, knew an abrupt and painful ending in the final years of Bernard's life, around 1151-2, when Nicholas must have severely breached his master's trust by sending out letters without his permission. The scandal has for a long time upheld Nicholas' portrayal as a disreputable Judas by Bernard's side, a condemnation which has shimmered through on a scholarly level as well, and has resulted in highly subjective and speculative attributions. For instance, Nicholas has been found sending out Bernard's *De diversis* 6, 7, 21, 62, 83, 100 and 104 in a letter to count Henry the Liberal, claiming that they were "of [his] invention, of [his] style, aside from what was taken from others in a few places" (Leclercq, (1)1962). Also *De diversis* 40, 41 and 42 have been found within Nicholas' oeuvre (Rochais, 1962). Nevertheless, Nicholas' reputation as a fraud and a plagiarist has withheld 20th-century scholars such as Leclercq and Rochais to believe that his claim to authorship is any sense warranted, and has maintained the sermons' authenticity as uncontested, even despite the fact that later scholars have seriously doubted their views on Nicholas of Montiéramey's alleged deceitfulness and falsification (Jaeger, 1980; Constable, 1996). The temptation for historians to draw lines in between imitation and plagiarism in order to categorize writings and collate them in attributed editions, valuable as it is, can also be rather anachronistic or even unbefitting in a medieval context (Nichols, 1990; Cerquiglini, 1999). Perhaps Nicholas felt himself to be a rightful partaker in the composition of these works, a participation which might disclose itself stylistically.

Stylometry

The texts of Bernard of Clairvaux are edited in the *Sancti Bernardi Opera* (SBO), Jean Leclercq's edition published in 8 volumes. Nicholas' letters and sermons, on the other hand, still lack a modern edition, and can only be found in Migne's Patrologia Latina (see table 4). Although experiments and debates as to which textual features best capture stylistic difference are still ongoing, many state-of-the-art studies employ function words, which still prove to be the most robust discriminators for writing styles (Burrows, 2002). Function words are usually short and insignificant words that pass unnoticed, such as pronouns, auxiliary verbs, articles, conjunctions and particles, whose main

advantage is their frequent occurrence, less conscious use by authors and content- or genre-independent character. Their benefit and success for stylometry in Latin prose have been convincingly demonstrated before, although the methodology still raises acute questions which keep stylometrists on the lookout for alternatives.

Because medieval Latin is a synthetic language with a high degree of inflection, the texts required some preprocessing (Mantello and Rigg, 1996). For instance, enclitica such as *-que* and *-ve* had to be separated from the token in order to be recognized as a feature. Secondly, texts are more easily mined for information when the lexemes are lemmatized (which means that the instance of the word is referred to its headword) and when its words (tokens) are classified according to grammatical categories (parts of speech). For this purpose we applied the Pandora lemmatizer-tagger on the texts, a piece of software developed by Kestemont and De Gussem that is equipped to achieve specifically this (Kestemont and De Gussem, forthcoming).

token	lemma	PoS-tag (*simplex*)
harum	hic	PRO
imo	immo	ADV

Figure 1. Excerpt from table showing tags applied to the texts

The third column, the part-of-speech-tag (PoS), allowed to immediately restrict the culling of most frequent words to those word categories that make up the collection of function words: conjunctions (CON), prepositions (AP), pronouns (PRO) and adverbs (ADV). This likewise filtered out some noise caused by ambiguities or homonyms like *secundum*. Once procedures of this sort were carried out in full, we arrived at a list of the 150 most frequent function words (MFFW) of the corpus (Figure 2)

1-25	26-50	51-75	76-100	101-125	126-150
et	nos	nam	uterque	iuxta	seipse
in	per	quoniam	aliquis	quisquis	item
qui	ex	inter	tunc	videlicet	quicumque
non	autem	denique	solum	apud	an
hic	noster	magis	sane	profecto	donec
is	que	nunc	quando	scilicet	certe
sed	vel	unde	igitur	prius	vere
ad	ergo	quidam	ante	nemo	quisque
ille	quidem	sine	talis	parve	absque
quod	tamen	propter	post	porro	interim
ut	iste	quasi	bene	plane	unquam
de	pro	tam	nullus	ibi	numquam

Figure 2. Excerpt from contents of a table showing most frequently occurring function words.

Each of the corpora was segmented into samples. This yields the advantage of "effectively [assessing] the internal stylistic coherence of works," (Eder et al., 2016) which answers directly to the primary goal of the present study. The sermons were segmented into 1500 word-samples (Figures 3-4 present aexcerpts from tables describing the texts contained in each sample).

sample (1500 words)	contents
sample_*n*	*SBO* index and paragraph
di_1	sm. 1.1-7
di_2	sm. 1.7ff., 2.1-6
di_3	sm. 2.6ff., 3.1-4
di_4	sm. 3.4ff., 4.1-2
di_5	sm. 4.2ff., 5.1-4
di_6	sm. 5.4ff., 8.1

Figure 3. Excerpt from a table describing the sample contents (1500 words) for Bernard's Sermones de Diversis as shown in figures 5-7.

sample (1500 words)	contents
sample_*n*	*PL* (vol: col.)
ep_1	ep. 1 (196: 1593a-1594b)
	ep. 2 (196: 1594b-1596a)
	ep. 3 (196: 1596b-1597b)
ep_2	ep. 4 (196: 1597b-1598c)
	ep. 5 (196: 1598d-1600a)
	ep. 6 (196: 1600b-1601b)

Figure 4. Excerpt from a table describing the sample contents (1500 words) for Nicholas' sermons and letters as shown in figures 5-7.

It should be noted that 1500 word-samples run the risk of increased imprecision, a consideration which should nuance any interpretation of the results (Kestemont et al., 2013). Once the corpus was divided, each of the text samples was vectorized to document vectors. The raw counts were TF-IDF-normalized (*Term frequency inverse document frequency*), a procedure which divides the function word frequencies by the amount of text samples that respective function word appears in (Manning, 2008; Kestemont et al., 2016). As a consequence, less common function words received a higher weight which prevents them from sinking away (and losing statistical significance) in between very common function words. Once the data was preprocessed and regulated, two statistical techniques were applied to visualize its dynamics.

The first is a *k* Nearest Neighbors network in GEPHI (hereafter abbreviated to *k*-NN) (Jockers, 2013; Eder, 2015; Jacomy et al., 2014), the second is principal components analysis (hereafter PCA) (Binongo et al., 1999). Their respective results will prove to be similar in a general sense, yet crucially different in the details. We argue that such an

additional statistical validation provides for a more accurate, nuanced interpretation and a better intuition of the data. In the first visualization, the k-NN networks, we first calculated the 5 closest text samples to each text sample by applying k-NN on the frequency vectors. Accordingly for each text the 5 most similar or closest texts were calculated, weighted in rank of smallest pairwise distance (Minkowski metric, a Eucledian metric) and consequently mapped in space through force-directed graph drawing (algorithm Force Atlas 2). The weights were directly calculated from the distances. The intuition is then that the distances should be normalized to a (1,0) range (as a higher distance responds to a lower weight). Secondly, PCA is a technique that allows to reduce a multivariate or multidimensional dataset of many features, such as our function word frequencies, to merely 2 or 3 principal components which disregard inconsequential information or noise in the dataset and reveal its important dynamics. The assumption is that the main principal components, our axes in the plot, point in the direction of the most significant change in our data, so that clustering and outliers become clearly visible. Each word in our feature vector is assigned a weighting or loading, which reflects whether or not a word correlates highly with a PC and therefore gains importance as a discriminator in writing style. In a plot, the loadings or function words which overlap with the clustered texts of a particular author are the preferred function words of that author (see Figure 7 under "Results").

Results

Figure 5 (*k*-NN) and Figure 6 (PCA) feature the results of matching up Bernard's *Sermones de Diversis* benchmarked against the latter's *Sermones Super Cantica Canticorum* (his literary masterpiece) and the sermons and letters of his secretary Nicholas of Montiéramey.

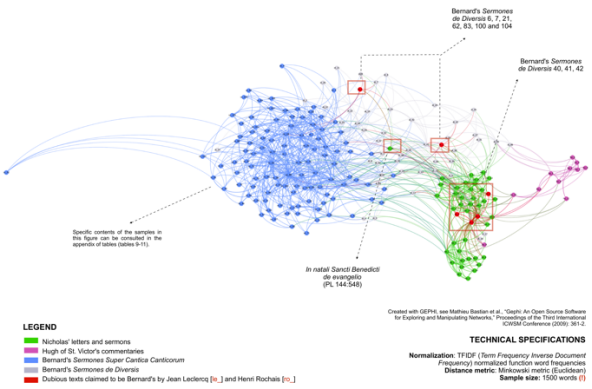

Figure 5: k-NN Networks

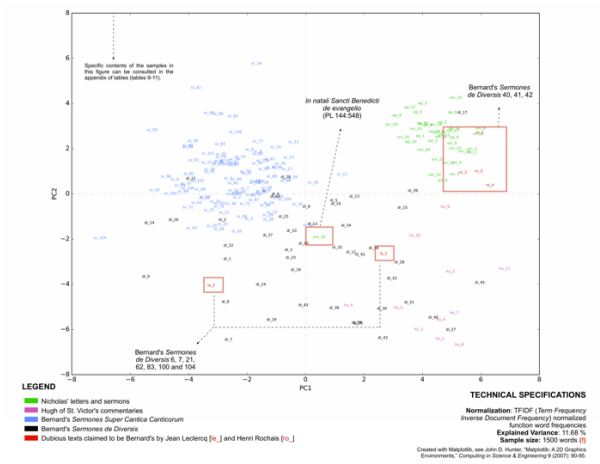

Figure 6. Principal Components Analysis (PCA)

Figure 7. PCA Loadings

Firstly, when examining the visualizations, it is striking how – indeed – the diversity of Bernard's *De diversis* is captured. Especially PCA demonstrates the discernible stylistic incoherence as the samples burst open all over the plot (especially along the vertical axis of the second principal component), at times suggesting the interference of other writers than Nicholas or Bernard in their composition. Other samples gravitate in between Nicholas and Bernard, and in some cases Nicholas' influence on the style is undeniable. *De diversis* 6, 7, 21, 62, 83, 100 and 104, which Nicholas included in the letter to count Henry the Liberal (they are split up in two red samples labeled with le_ of Leclercq), do not betray an obvious affinity to Nicholas' style (although le_1 is not far off). Neither are they unambiguously Bernard's. Both samples diverge strongly and seem too hybrid in nature to be restrained. The case rather ostensifies how difficult it is to defend concepts such as "single authorship" or "text theft" in a medieval context: the le_ samples are clearly not of a "singular" style (nor of Nicholas's style, nor of Bernard's), but defy classification. In fact, if we compare both *k*-NN and PCA, Nicholas' influence in sample le_1 seems considerably larger than Bernard's.

It has by now become an untenable simplification to argue that Nicholas has stolen these sermons, especially if we review the results of the second case, that of *De diversis* 40, 41 and 42 (four red samples labeled with ro_ of Rochais): although the sermons emanate from bernardian thought, PCA and *k*-NN unambiguously cluster all three sermons together with those written by Nicholas, not Bernard.

Bibliography

Binongo, J.N.G., and Smith, M.W.A. (1999). The Application of Principal Components Analysis to Stylometry, Literary and Linguistic Computing 14(4): 446-66.

Burrows, J.F. (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship, Literary and Linguistic Computing 17(3): 267-87.

Callerot, F. (2006). Introduction. In Leclercq, J., Rochais, H.M. and Talbot, C.H. (eds.), Bernard of Clairvaux, Sermons Divers (3 vols). Paris: Sources Chrétiennes, pp. 1:21-60

Cerquiglini, B. (1999). Wing, B. (transl.) In Praise of the Variant: A Critical History of Philology. Baltimore: John Hopkins University Press.

Constable, G. (1996). Forgery and Plagiarism in the Middle Ages. In Culture and Spirituality in Medieval Europe. Aldershot: Variorium, pp. 1-41.

Constable, G.. (1967). Nicholas of Montiéramey and Peter the Venerable. In The Letters of Peter the Venerable (2 vols.). London: Harvard University Press, pp. 2: 316-330.

Constable, G.. (1994). The Language of Preaching in the Twelfth Century, Viator 25: 131-52.

Eder, M. (2015). Visualization in Stylometry: Cluster Analysis Using Networks, Digital Scholarship in the Humanities: 1-15.

Eder M., Rybicki J. and Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis, The R Journal 16(1): 1-15.

Jacomy, M. et al. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software, PLoS ONE 9(6): e98679. doi:10.1371/journal.pone.0098679 (accessed 6 April 2017).

Jaeger, S.C. (1980). Prologue to the Historia Calamitatum and the "Authenticity Question", Euphorion 74: 1-15.

Jockers, M.L. (2013). Macroanalysis: Digital Methods and Literary History. University of Illinois Press.

Kestemont, M., Moens, S. and Deploige, J. (2013). Collaborative Authorship in the Twelfth Century: A Stylometric Study of Hildegard of Bingen and Guibert of Gembloux, Digital Scholarship in the Humanities 30(1): 199-224.

Kestemont, M. and De Gussem, J. (forthcoming). Integrated Sequence Tagging for Medieval Latin Using Deep Representation Learning. Journal of Data Mining and Digital Humanities.

Kestemont, M., Stover, J., Koppel, M., Karsdorp, F. and Daelemans, W. (2016). Authenticating the writings of Julius Caesar, Expert Systems with Applications 63: 86-96.

Leclercq, J. ((1)1962). Les collections de sermons de Nicolas de Clairvaux. In Recueil d'études sur saint Bernard et ses écrits (4 vols.). Rome: Edizioni di storia e letterature, pp. 1:47-82.

---. ((2)1962). Saint Bernard et ses secrétaires. In Recueil d'études sur saint Bernard et ses écrits (4 vols.). Rome: Edizioni di storia e letterature, pp. 1: 3-25.

---. (1969). Notes sur la tradition des épitres de S. Bernard Recueil d'études sur saint Bernard et ses écrits (4 vols.). Rome: Edizioni di storia e letterature, 3:307-22.

Mantello, F. A. C. and Rigg, A. G. (eds.) (1996). Medieval Latin: An Introduction and Bibliographical Guide. Washington (D.C.): Catholic University of America Press.

Nichols, S.G. (1990). Introduction: Philology in a Manuscript Culture, Speculum 65(1): 1-10.

Rassow, P. (1913). Die Kanzlei St. Bernhards von Clairvaux. Studien und Mitteilungen zur Geschichte des Benediktiner-Ordens 34. London: FB&c Ltd.

Rochais, H.M. (1962). Saint Bernard est-il l'auteur des sermons 40, 41 et 42 «De diversis»? Revue Bénédictine 72(3-4): 324-345.

Turcan-Verkerk, A.-M. (2015). L'introduction de l'ars dictaminis en France: Nicholas de Montiéramey, un professionel du dictamen entre 1140 et 1158. In Turcan-Verkerk A.-M. and Grévin B. (eds.) Le dictamen dans tous ses états: perspectives de recherche sur la théorie et la pratique de l'ars dictaminis (xie-xve siècles). Turnhout: Brepols, pp. 63-98.