Acta Poloniae Historica
112, 2015
PL ISSN 0001–6892

Maciej Eder
Institute of Polish Studies, Pedagogical University of Cracow
Institute of Polish Language, Polish Academy of Sciences, Cracow

IN SEARCH OF THE AUTHOR OF *CHRONICA POLONORUM* ASCRIBED TO GALLUS ANONYMUS: A STYLOMETRIC RECONNAISSANCE

Abstract

The article deals with the question of authorship of the thirteenth-century *Chronica Polonorum* (or *Gesta principium Polonorum* [*The Deeds of the Princes of the Poles*]), also known as *The Polish Chronicle*. It seeks to verify the hypothesis, recently reproposed by Tomasz Jasiński, whereby the author was of Venetian origin. The hypothesis is namely based on the textual similarities observed between *Translatio Sancti Nicolai* by an author referred to as the 'Monk of Lido' (Monachus Littorensis) and the *Chronica*. The attribution attempt put forth by M. Eder is based upon stylometric methods that measure the frequencies of the most frequent words in the texts under research (mainly, conjunctions, prepositions, pronouns, and particles) which are subsequently subjected to cluster analysis, multidimensional scaling, or principal components analysis. The outcome of the experiment in question has demonstrated a strong resemblance between the *Translatio Sancti Nicolai* and the *Polish Chronicle*, which may be regarded as an substantial argument in support of the Venetian background hypothesis.

Keywords: Gallus Anonymus, *Chronica Polonorum*, authorship attribution, stylometry, multidimensional methods, Monachus Littorensis

I
INITIAL REMARKS

This essay (or, rather, an outline or germ of a larger study) does not seek to discuss in detail the decades-long discussion on the authorship of *Chronica Polonorum*. Neither will I attempt to propose a critical evaluation on the arguments contributing to the discussion. My point is to scrutinise the hypothesis of a Venetian origin of the chronicler – a conjecture that was first put forward some time ago and gained popularity recently. To be specific, Danuta Borawska was the first to have observed certain similarities between the *Chronicle* and the

http://dx.doi.org/10.12775/APH.2015.112.01

Translatio Sancti Nicolai,[1] which was subsequently confirmed by Marian Plezia[2] and thereafter subjected to thorough critical analysis and deep reflection in a cycle of dissertations by Tomasz Jasiński.[3] The arguments amassed by the researchers lead to the conclusion – which is particularly powerful in Jasiński's studies – that the anonymous author of the *Translatio* (known in the literature as Monachus Littorensis – the 'Monk of Lido') was identical with the anonymous author of *Chronica Polonorum*.

Among the arguments gathered by Jasiński, the most intriguing one is the outcome of statistical analysis that has been applied to the clausulae of which the *Chronica*'s sentences are structured.[4] Jasiński has compared the use of the *cursus* (the rhythm of the clausulae) in a considerable number of medieval texts, amongst which the *Translatio* and the *Chronica* appeared to quite strongly resemble each other, mainly in terms of the use of the so-called veloxes. An inventive method elaborated by Jasiński, albeit using no mathematical apparatus or achievements of contemporary statistics, contributes to the venerable tradition of stylometry, that is, arguing for authorship of (literary) texts on the grounds of statistical analysis of their linguistic characteristics.

---

[1] See in this issue of APH, Danuta Borawska, 'Gallus Anonymus, or, Italus Anonymus', on pp. 313–326.

[2] Marian Plezia, 'Nowe studia nad Gallem-Anonimem', in Helena Chłopocka and Brygida Kürbis (eds.), *Mente et litteris. O kulturze i społeczeństwie wieków średnich* (Poznań, 1984), 111–20.

[3] Tomasz Jasiński, 'Czy Gall Anonim to Monachus Littorensis?', *Kwartalnik Historyczny*, cxii, 3 (2005), 69–89; *idem*, 'Rozwój średniowiecznej prozy rytmicznej a pochodzenie i wykształcenie Galla Anonima', in Dariusz A. Sikorski and Andrzej M. Wyrwa (eds.), *Cognitioni gestorum. Studia z dziejów średniowiecza dedykowane Profesorowi Jerzemu Strzelczykowi* (Poznań and Warszawa, 2006), 185–93; *idem*, *O pochodzeniu Galla Anonima* (Kraków, 2008); *idem*, 'Die Poetik in der Chronik des Gallus Anonymus', *Frühmittelalterliche Studien. Jahrbuch des Instituts für Frühmittelalterforschung der Universität Münster*, 43 (2009), 373–91; *idem*, 'Jak Gall Anonim tworzył veloxy? Przyczynek do poznania rytmiki "Kroniki polskiej"', in Anna Odrzywolska-Kidawa (ed.), *Klio viae et invia. Opuscula Marco Cetwiński dedicata* (Warszawa, 2010), 17–23.

[4] Tomasz Jasiński, *"Kronika polska" Galla Anonima w świetle unikatowej analizy komputerowej nowej generacji*. Wykłady inauguracyjne Instytutu Historii Uniwersytetu im. Adama Mickiewicza, semestr letni 2010/2011 [i.e. as part of the Opening Lectures cycle at the Adam Mickiewicz University's Institute of History, summer 2010/11], VI (Poznań, 2011).

http://dx.doi.org/10.12775/APH.2015.112.01

A direct motivation to undertake the present study was the fact that the experiment proposed by T. Jasiński was carried out, as it were, in isolation from the existing attribution methods. The research query that sprang out quite naturally in this context, can be phrased thus: Will the conclusions regarding the authorship of the *Chronica Polonorum*, when referring to a 'unique computerised analysis' (to quote the title of one of Jasiński's studies[5]), prove to be verifiable with use of the recognised attribution methods that have been developed for at least several decades now, with their ever-growing accuracy? Hence, the research assumption behind this essay was fairly simple: In the event that the results of two different stylometric methods have independently identified Monachus Littorensis as the probable author of the chronicle, the 'Venetian background' hypothesis will thus become reliably reconfirmed.

The thus outlined subject-matter of research is reflected in the rather sparse scholarly apparatus this essay builds upon. The reader who expects to come across an abundant bibliography on the origin or background of Gallus or the mathematical details of stylometric attribution will certainly be disappointed. That this particular essay is not embedded in adequately numerous bibliographical references is deliberate, since I believe the other authors have done the job better.

II
STYLISTICS AND STATISTICS

Authorship attribution based upon statistical analysis of style has a long history behind it, dating back to at least the nineteenth century; in any case, certainly a time before the computer age. The history of the discipline as well as a series of attempts to apply the attributive methodology – more or less successful, depending on the case – are extensively discussed elsewhere,[6] and hence I shall confine myself

[5] *Ibidem*.

[6] Harold Love, *Attributing Authorship: An Introduction* (Cambridge, 2002), Chap. 2: 'Historical survey', 14–31; Efstathios Stamatatos, 'A Survey of Modern Authorship Attribution Methods', *Journal of the American Society for Information Science and Technology*, 60 (2009), 538–56; Maciej Eder, 'Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu – przykład stylometrii', *Teksty Drugie*, 2 (2014), 90–105.

http://dx.doi.org/10.12775/APH.2015.112.01

here to mentioning just a few studies that refer to attribution methods in the research in ancient and medieval texts.

The classic studies – as a broad concept, medieval and neo-Latin studies included – is usually regarded as quite a conservatively-inclined area of research, one that is reluctant towards new methodological paradigms. Quite contrarily to this claim, however, 'digital humanities', so much in vogue these days, were conceived around the classical languages. This holds true for the first electronic corpora – that is, large collections of texts supplemented with morphosyntactic annotation (the Perseus Project being a pioneering initiative of the sort) – as well as the early statistical research in linguistic domain. Let us remark that the term 'stylometry' was coined by the classicist Wincenty Lutosławski in the late nineteenth century.

Taken in the nineteenth century, the first reliable attempts at attribution that made use of statistical methods referred to Latin and, primarily, Greek texts. Among the early studies that lay the foundations for stylometry, one comes across not only the frequently evoked debate between August de Morgan and Thomas Mandelhall on the Shakespearian canon[7] but also the studies of William Benjamin Smith (who published under the penname Conrad Mascol) on the stylistic uniformity of the Pauline Epistles. In 1867, Lewis Campbell, professor of Greek literature, carried out several statistical tests in order to determine a relative chronology of Plato's *The Sophist* and *The Politician*; he primarily analysed the sequence of words, the rhythm, avoidance of hiatus (occurrence of two vowel sounds without pause or intervening consonantal sound) and 'originality of word-inventory' measured by the number of *hapax legomena*. Campbell's studies remained unnoticed over the subsequent thirty years; in any case, Constantin Ritter's 1888 dissertation on the chronology of Plato's dialogues, which made use of very similar methods, was compiled probably independently of Campbell. The aforementioned Wincenty Lutosławski thoroughly expanded the scope of research of his predecessors and introduced a 'stylometric method', as he himself named it. The dating of Plato's dialogues, an exercise commenced by the

[7] David I. Holmes, 'The Evolution of Stylometry in Humanities Scholarship', *Literary and Linguistic Computing*, 13 (1998), 112; Joseph Rudman, 'The State of Authorship Attribution Studies: Some Problems and Solutions, *Computers and the Humanities*, 31 (1998), 354.

http://dx.doi.org/10.12775/APH.2015.112.01

Polish scholar in a 1897 monograph[8], has generally remained accepted till our day.[9]

The more recent stylometric studies regarding attribution of Latin or Greek texts are too numerous to be mentioned here. They encompass, i.a., the problems of authorship of the individual books of the New Testament,[10] the attribution of the collection of Roman emperors' lives entitled the *Historia Augusta*;[11] two medieval visions traditionally attributed to Hildegard of Bingen;[12] the famous treatise *De consolatione*, published under the name of Cicero but actually penned by Carlo Sigonio in 1583;[13] or, the finding that it was John Milton, in fact, who authored the treatise *De doctrina Christiana*.[14] These and similar studies have shown that the attributive methods, as once elaborated to support analysis of English-language texts, are of use also for languages with a different grammatical structure. It is, thence, legitimate to believe that the anonymous *Chronicle* by Gallus is no less adequate a research material.

---

[8] Wincenty Lutosławski, *The Origin and Growth of Plato's Logic: With an Account of Plato's Style and of the Chronology of His Writings* (London, 1897).

[9] Adam Pawłowski and Artur Pacewicz, 'Wincenty Lutosławski (1863–1954). Philosophe, helléniste ou fondateur sous-estimé de la stylométrie?', *Historiographia Linguistica*, xxxi, 2/3 (2004), 423–47.

[10] H.H. Greenwood, 'Common Word Frequencies and Authorship in Luke's Gospel and Acts', *Literary and Linguistic Computing*, 10 (1995), 183–7.

[11] Penelope J. Gurney and Lyman W. Gurney, 'Authorship Attribution of the *Scriptores Historiae Augustae*', *ibidem*, 13 (1998), 119–31; Joseph Rudman, 'Non-traditional Authorship Attribution Studies in the *Historia Augusta*: Some Caveats', *ibidem*, 151–7.

[12] Mike Kestemont, Sara Moens, and Jeroen Deploige, 'Collaborative Authorship in the Twelfth Century: A Stylometric Study of Hildegard of Bingen and Guibert of Gembloux', *Digital Scholarship in the Humanities*, xxx, 4 (2013), <http://dsh.oxfordjournals.org/content/30/2/199> [Accessed: July 15, 2015].

[13] Richard S. Forsyth, David I. Holmes, and Emily K. Tse, 'Cicero, Sigonio, and Burrows: Investigating the Authenticity of the *Consolatio*', *Literary and Linguistic Computing*, 14 (1999), 375–400, <http://www.richardsandesforsyth.net/pubs/CONSol99.pdf> [Accessed: July 15, 2015].

[14] Fiona J. Tweedie, David I. Holmes, and Thomas N. Corns, 'The Provenance of *De Doctrina Christiana* Attributed to John Milton: A Statistical Investigation', *ibidem*, 13 (1998), 77–87.

http://dx.doi.org/10.12775/APH.2015.112.01

III
A LITERARY DACTYLOSCOPY

The methodological assumptions behind stylometry are twofold: in brief, they would be describable as linguistic and mathematical. As regards the former type, attribution employs the concept referred to as 'stylistic fingerprint', which can be defined as a cohort of linguistic features that are unique for particular authors: despite freedom of choice of words and grammatical structures, every user of language tends to use certain traits of language in his or her peculiar, individualised and unique way. In his pioneering study, Lutosławski assumed that these individual stylistic features clearly define the writing person and appear to be even more strongly determined than the individual graphological characteristics.[15] Scholars nowadays avoid posing unambiguous statements highlighting stylistic uniqueness, but the very conviction about linguistic habits or idiosyncrasies (less distinct once, and somewhat more evident another time) that disclose the author has not been challenged. What makes contemporary attribution methods contrary to, perhaps, every traditional definition of style is the assumption – intuition-contradicting as it is – that the author's individuality is determinable based on the prepositions, conjunctions, and other function words.

The commonsensical concept of style as a combination of individual linguistic features suggests that certain rare phrases, formulaic wordings, or even single words, would bear the author's stamp. This perspective is mostly taken advantage of in studies using methods other than quantitative;[16] it is visible as well in the treatises penned by the nineteenth-century pioneers of stylometry. The statistical method used by T. Jasiński is based on a similar principle, since it has been

[15] Lutosławski, *The Origin and Growth*, 66.

[16] Such an approach, almost in a model form, has been applied by Maria Karpluk in her falsification of the hypothesis claiming Mikołaj Rej to be the author of *Historia w Landzie*; see *eadem*, '"Historia w Landzie" nie jest utworem Reja', in Tadeusz Bieńkowski, Janusz Pelc, and Krystyna Pisarkowa (eds.), *Mikołaj Rej w czterechsetlecie śmierci* (Wrocław, 1971), 211–20. Adam Karpiński assumed a similar arguing method in his essay on erotic poems ascribed to Mikołaj Sęp Szarzyński; see *idem*, 'Filologiczne pytania do Mikołaja Sępa Szarzyńskiego na marginesie rękopiśmiennych wierszy z "Rytmów"', in Juliusz A. Chrościcki *et al.* (eds.), *Corona scientarum. Studia z literatury i kultury nowożytnej ofiarowane Profesorowi Januszowi Pelcowi* (Warszawa, 2004), 71–87.

http://dx.doi.org/10.12775/APH.2015.112.01

applied to the elaborate rhythm patterns of Gallus and Monachus Littorensis, easy for a trained ear to grasp. For style markers of this kind, a considerably restrictive factor is, however, that phrases and words which seem, at first glance, to be characteristic of the author are very easy imitable, forgeable or plagiarisable. In the case of Gallus, there is clearly no point speaking about fakery; yet, the method used in operating the *cursus*, including the saturation of the prose text with veloxes, is a stylistic trait which is used deliberately; thoroughly learned, it is perhaps indicative of Gallus's education or writing school he belongs to, rather than of some actual individual characteristics of the author.

At this point, we have approached the essence of what is the stylometric authorship attribution: namely, the distinguishing features of style ought to be sought where they are not expectable by the author – or, to be more specific, where the author is not in the position to control them and use them subconsciously. In a groundbreaking study on attribution, Frederick Mosteller and David Wallace have demonstrated that function words, such as articles, conjunctions, prepositions, particles, and certain personal pronouns, form an excellent gauge enabling to differentiate the authors.[17] No author, especially if not sensitive to subtleties of language, can possibly be in control of a more or less frequent use of, for instance, the words such as 'et' or 'si'; controlling the frequencies of several dozens of such words at a time would be all the more awkward.

The discoveries made by George Zipf, who has formulated the fundamental linguistic law rendering a given word in the corpus dependent on its position in the ranking of the most frequent words,[18] make us aware that it is function words (articles, conjunctions, etc.) that form the most frequent lexemes of any natural language; therefore, the entire modern stylometry is based, in a vast majority, on the analysis of occurrence of the most frequent words in the authors under research. This solution is quite convenient, as it suffices to count the words appearing within a corpus and arrange them in the sequence from the most frequent to the least frequent one: the stylistic fingerprint

[17] Frederick Mosteller and David L. Wallace, *Inference and Disputed Authorship: The Federalist* (Reading, Mass., 1964).

[18] George K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (Cambridge, MA, 1949).

http://dx.doi.org/10.12775/APH.2015.112.01

will hide somewhere in the upper area of the list. A separate problem that appears with highly inflected languages, Latin included, is the role of individual word forms as style markers: should the word forms such as, for example, 'sunt', 'est', 'fuit', 'eram', 'estote', etc., be lemmatised, that is, represented by their basic form (i.e. 'sum'/'esse'), or analysed in their original inflected form? Preliminary tests for the Latin language have shown that lemmatisation, being a labour-intensive task, does not increase the attributive efficiency.[19] For this reason, the present study uses non-lemmatised texts only. An exemplary matrix of the frequencies of the most frequent words for Gallus's *Chronicle*, *Translatio* of Monachus Littorensis and a few other texts is shown in Table 1.

Table 1. Frequencies (occurrences, per cent) of the most frequent words

	Gallus, ***Chronica***	**Monachus,** ***Translatio***	**Benedictus,** ***Regula***	**Alanus,** ***Planctus***	**Abelardus,** ***Consolatio***
et	3.646	4.756	4.159	0.778	3.540
in	1.840	2.144	2.445	2.849	2.189
non	0.894	0.709	1.698	0.696	0.998
est	0.475	0.392	1.057	0.242	0.772
ut	0.298	0.325	1.185	0.684	1.065
ad	0.852	1.176	0.898	0.572	1.418
cum	1.150	1.235	0.755	0.543	1.040
quod	0.532	0.668	0.906	0.100	0.856
qui	0.460	0.776	1.132	0.372	0.738
sed	0.886	0.517	0.634	0.478	0.403
quam	0.290	0.259	0.257	0.136	0.923

Now, apart from the fingerprint concept, we have come up to the other element of the attribution methodology: the aforementioned mathematical assumptions (these are merely indicated in this essay). One needs no advanced statistical method to detect, with the naked eye, a single (superficial) numerical regularity. The conjunction 'et', the most frequent word in the medieval Latin texts corpus, appears to

[19] Maciej Eder, 'Taking Stylometry to the Limits: Benchmark Study on 5,281 Texts from "Patrologia Latina"', Digital Humanities 2015 conference abstracts, <http://dh2015.org/abstracts/> [Accessed: July 12, 2015].

http://dx.doi.org/10.12775/APH.2015.112.01

be in use unexpectedly seldom in Alan of Lille's *Complaint of Nature* (*De planctu naturae*) (merely 0.7% of all the occurrences); the converse is the case with the Rule of St Benedict and, primarily, Monachus's *Translatio* (the occurrences of 'et' account for almost 5% of the whole word-hoard). Gallus's *Chronicle* definitely differs from the *Translatio* in terms of use of this particular conjunction, which obviously means that the possible conclusions with regards to the homogeneity of these two texts are backed by no facts whatsoever. Moreover, the preposition 'in' is also distributed otherwise in the *Chronicle* than in the *Translatio*, which is also true for 'ad' and 'sed'. A comparison of the distributions of other words, however, shows a substantially different picture and seems to come in support of the stylistic homogeneity argument: as regards the uses of the forms 'non', 'est', 'ut', or 'cum', the mutual similarities between Monachus and Gallus prove very strong, if not striking – whilst there appear considerable differences between these two texts and the other works from the same corpus.

A question quite clearly comes up: Which of the frequencies specified in the table ought, consequently, to be regarded significant, since some of them seem to be indicative of a homogeneity of the *Chronica* and the *Translatio*, whilst the others show something completely opposite? The other question is, what technique should be applied to browse through the table in order to find the possible regularities (if any)? A laborious comparing of numerical values becomes apparently troublesome when one deals with a dozen-or-so, or several dozen, texts in a corpus. In case the table comprises more columns (texts) and, moreover, a considerable number of rows (linguistic features subject to comparison), the task becomes, plainly, undeliverable.

Multidimensional statistical methods can remedy both above-outlined problems. The way they operate is that, rather than being analysed separately from one another, individual words are analysed *en bloc*, even if there are several hundred or thousand of them. A remarkable advantage of this technique is that it helps find regularities that tend to be omitted when using the naked eye. If the aforementioned 'in' and 'ad' seem to have been distributed dissimilarly in the *Chronica* and in the *Translatio*, whilst the opposite holds true for 'est' or 'ut', then the multidimensional analysis is capable of precisely defining which of the resemblances are essential and which are casual; furthermore, it can measure the general (averaged) degree of similarity between the texts under research. The following attempt to verify the hypothesis

http://dx.doi.org/10.12775/APH.2015.112.01

of a Venetian background behind *Chronica Polonorum* will use a few such methods, including multidimensional scaling and cluster analysis.

The efficiency of multidimensional methods in tracking the authorial fingerprint has several times been revalidated experimentally, by means of so-called controlled (blind) authorship tests.[20] In a nutshell, such experiment consists in amassing a collection of texts of known authorship to subsequently check whether the computer has properly 'guessed' their actual authors. The assumption is that the right answer is known to the researcher while the software is 'unaware' of who the author is and its task is to carry out a classification based strictly on style analysis. As it turns out, the quantity of authorial signal extractable in this way appears to be astonishingly considerable. For instance, for nineteenth-century English novels, the efficiency is close to 100 per cent (indeed!); for contemporary English novels, the rate is approximately 80 per cent. As regards Polish, Latin or German text corpora, they generally perform below the standard of their English peers. A particularly disappointing outcome has been produced by the corpus that included several dozen Polish novels.[21] For Latin texts, the results oscillated around 90 per cent of correctly recognised authors.[22]

Before passing on to the analysis of the *Chronica*, yet another important factor needs to be mentioned. While the classification methods in use in exact sciences over the decades and gradually adapted in stylometry are increasingly precise, it should not mean that one can blindly believe in every single result they produce.[23]

---

[20] Matthew L. Jockers and Daniela M. Witten, 'A Comparative Study of Machine Learning Methods for Authorship Attribution', *Literary and Linguistic Computing*, 25 (2010), 215–23; Moshe Koppel, Jonathan Schler, and Shlomo Argamon, 'Computational Methods in Authorship Attribution', *Journal of the American Society for Information Science and Technology*, lx, 1 (2009), 9–26; Jan Rybicki and Maciej Eder, 'Deeper Delta across Genres and Languages: Do We Really Need the Most Frequent Words?', *Literary and Linguistic Computing*, 26 (2011), 315–21; Maciej Eder, 'Style-Markers in Authorship Attribution: A Cross-Language Study of the Authorial Fingerprint', *Studies in Polish Linguistics*, 6 (2011), 99–114.

[21] Maciej Eder and Jan Rybicki, 'Do Birds of a Feather Really Flock Together, or How to Choose Training Samples for Authorship Attribution', *Literary and Linguistic Computing*, 28 (2013), 229–36.

[22] Eder, 'Style-Markers', 108–12.

[23] The problem of reliability of attribution tests and the various factors contributing to the final outcome of the attribution are discussed in detail in the other studies (co-)authored by the undersigned (*op. cit.*).

http://dx.doi.org/10.12775/APH.2015.112.01

The search for the authorship of an anonymous text is, namely, based on identification of what is called the nearest neighbour and thus may carry a considerable risk of error in case that the comparative corpus does not comprise the samples of all the possible authors.

In determining the authorship of an anonymous literary (or some other) text, the stylometrist basically faces the task of gathering the maximum possible number of texts written by the 'candidates', or the authors who could have authored the anonymous text in question. (Usually, a few other texts from the period are added to the corpus, to serve as a control group.) Regardless of how sophisticated is the method actually used, the stylometric test always consists in finding the 'nearest neighbour', which is the stylometrically closest text identified among all those gathered within the corpus. The so-called open-set attribution problem occurs when the researcher cannot be wholly certain whether the reference corpus contains the samples of all the 'candidates'. In some cases, one deals with the open-set problem *ex definitione*. For instance, in a comparative analysis of the anonymous *Batrachomyomachia*, where the text is tested against the other extant Greek epic poems – those by Homer, Hesiod, Apollonius, Aratos, and Nonnus, one of these poets will be indicated as the most plausible option; such indication will, however, be obviously erroneous (the possible authorship of Pigres of Halicarnassus cannot be verified as there is, simply, no comparative material available).

The *Chronicle* by Gallus is a typical example of the open-set problem: while Monachus Littorensis is the candidate author, one cannot possibly ascertain whether he is the only one. This issue will be fundamental when it comes to interpreting the results of the analysis.

IV
GALLUS, OR ITALUS?

It is time now to pass on to the experimental section. It would perhaps be a good idea to start with a slightly unusual approach, that is, by directly comparing the *Translatio* and the *Chronica* against each other, with no reference yet to the comparative corpus. Such an initial test would of course not seek to resolve the authorship question but to show the mutual relationship between Gallus and Monachus. To this end, both texts have been divided into sections of 10,000 words each.

http://dx.doi.org/10.12775/APH.2015.112.01

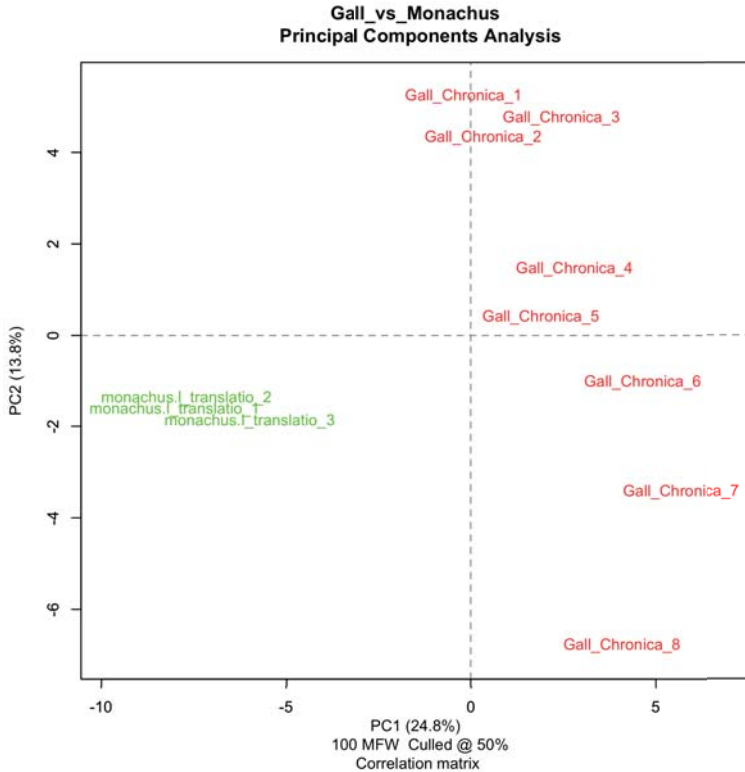

Fig. 1: Gallus's Chronicle vs. Monachus's Translatio: a comparison analysing crucial constituents, based on 100 most frequent words.

As a result, *Translatio* is represented by three samples (of which the first two contain the *Translatio* itself and the third, the *Miracula*, an addendum thereto attached). The Chronicle was divided into eight samples: the first three almost exactly coinciding with Book I, samples 4, 5, 6 and a part of 7 belong to Book II, whereas a lion's share of sample 7 and the whole of 8 form Book III. The outcome of the principal components analysis for the eleven samples thus generated is shown in Fig. 1. The method is based on the placing of all the text samples in a multidimensional space, the number of its dimensions equalling that of the words being analysed. Since a space of multiple (several dozen) dimensions eludes human perception, the next step is to mathematically transform the whole arrangement in a manner

http://dx.doi.org/10.12775/APH.2015.112.01

so as to reduce the dimensions to merely two, whilst preserving as much of the original information on the spatial differentiation of the samples as possible. A space thus reduced can be shown in an easy-to-interpret diagram (Fig. 1): the more condensed the samples, the more they resemble one another; and, conversely: the more distant they are, the stronger the differences. Putting the thing briefly, the point is to identify the areas in the diagram that encompass the samples getting grouped.

The outcome that immediately stands out is the quite concentrated distribution of the *Translatio* samples, which speaks in favour of considerable stylistic homogeneity of the *Translatio* and the *Miracula* attached, against Gallus's samples appearing remarkably spread, which clearly testifies to a differentiation of the *Chronicle*'s style. Most interestingly, however, one has to do here with an evolution – from the earliest to the latest fragments of the work, and with a fairly clear division into Book I (samples 1–3, drifting in the upper area), Book II (the three samples at the centre) and Book III (the less clearly separated samples 7 and 8). While the outcome is apparently not quite critical, it may provide an impulse for further stylometric research, potentially focused on examination of Gallus's language and style. This might obviously bring about new findings with regard to the time when the respective sections of the *Chronicle* were written.

The above-outlined preliminary results have contributed to the problem of authorship nothing of much relevance. Hence, the next step was to gather a reference corpus that would be composed – with no actual candidate authors available – of texts written by authors chronologically close to Gallus. The research assumption was that in case the *Translatio* and the *Chronica* consistently appeared close to each other, a non-coincidental stylistic correspondence would be the case.

Cluster analysis, with its end result pictured as a tree of similarities between the samples, so-called dendrogram (Fig. 2), is based on calculation of the measure of similarity between the samples and, subsequently, identification of the nearest neighbours: the stylistically closest texts will appear close to one another in the dendrogram. While cluster analysis is rather efficient, it still proves to be not-quite-stable a method: depending on the experiment's parameters, the final results of the classification may differ substantially. Hence, in order to extract from the cluster analysis any robust groupings and filter out accidental ones, a considerable number of tests were carried out

http://dx.doi.org/10.12775/APH.2015.112.01

to check various combinations of the number of words, measures of similarity, and algorithms for the building of dendrograms. An exemplary dendrogram for 100 most frequent words, the Delta distance measure and Ward's linkage algorithm are shown in Fig. 2.

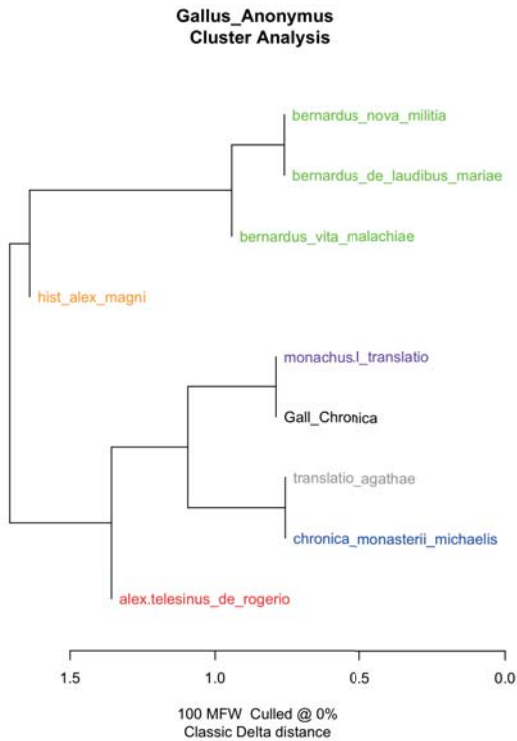

Fig. 2: A comparison of nine chronologically close texts: cluster analysis, Delta distance measure, and 100 most frequent words.

Dendrograms are interpreted by finding the 'leaves' and 'branches' of the tree, the assumption being that the most similar texts would appear on shared branches. The diagram above shows rather clearly that such a common branch hosts three treatises by Bernard de Silvestris – properly recognised as written by the same author – whereas on another branch, the *Chronicle*'s closest neighbour is Monachus's *Translatio*. The other parameters contributed, in essence, to a similar picture, in spite of various turbulences and mutual regroupings, there

http://dx.doi.org/10.12775/APH.2015.112.01

were two clusters permanently appearing: Bernard on the one side and the pair Gallus–Monachus on the other. This is an important, though clearly not decisive, argument in support of the supposed stylistic uniformity of the latter-said two works.

The subsequent stylometric tests were performed using a pretty large corpus of 159 Latin texts, and this in order to check whether the similarities demonstrated in the previous experiment would be confirmed with a higher degree of difficulty in the classification. The selection of texts in the corpus was, to some extent, random rather than systematic, and it contained prose texts by classical authors such as Cicero, Tacitus, Seneca, or Livy, along with those penned by Church Fathers (St Augustine, St Ambrose, the Venerable Bede), plus – quite obviously – a handful of medieval authors (Dante, Anselm of Canterbury, Alan of Lille), and some early modern writers (Erasmus, Thomas More, Piccolomini, Pico). The idea was to check whether, in the context of a large and multifaceted background of diverse writing and stylistic traditions, Gallus Anonymus would still choose the neighbourhood of Monachus Littorensis, and whom of the authors would he point to as his neighbours. To this end, a new stylometric method has been used which combines the nearest neighbour techniques and network analysis.[24] The method in question is based on calculation of the resemblances between the texts under investigation and, subsequently, displaying them in the form of a network of interrelations between the most similar texts.

Network analysis is quite a comfortable technique used in the modelling of complex phenomena, since it reduces a complicated reality to two elements: the nodes and their (inter)connections. This is how various physical or chemical phenomena, social behaviours, and so on, can be shown. Used in stylometry, the method seeks to represent

[24] The theoretical assumptions of the method and the first attempts to apply it are discussed in Maciej Eder, 'Visualizing Stylometry: Cluster Analysis Using Networks', *Digital Scholarship in the Humanities*, xxx (2015); for more on the applications in the research of Polish and Latin literature, see *idem*, 'Metody ścisłe w literaturoznawstwie', 100–4; *idem*, 'A Bird's Eye View of Early Modern Latin: Distant Reading, Network Analysis and Style Variation', in Michael Ullyot, Diane Jakacki, and Laura Estil (eds.), *Early Modern Studies and the Digital Turn* (forthcoming); Jan Rybicki, 'Pierwszy rzut oka na stylometryczną mapę literatury polskiej', *Teksty Drugie*, 2 (2014), 106–27.

http://dx.doi.org/10.12775/APH.2015.112.01

a textual reality in a shortened form, provided that individual literary texts become the nodes of a network, and the 'nearest neighbour' relationships are the connections. This new method – or, in fact, the embedded algorithm used to calculate textual similarities – is quite powerful as it enables to reveal the most apparent similarities whilst also giving an insight into some not-too-well-visible intertextual relations. Between every text examined and its stylometric neighbours, several connections become established, of which the strongest one links a given literary work with its nearest neighbour whereas the slightly weaker one connects with the second most similar text, and the weakest link joins the piece with the third most similar text. Next, a subsequent algorithm is used whose task is to arrange the nodes on a plane by the strength and number of individual connections. In brief, the strongly connected nodes tend to arrange close to one another. The final outcome is illustrated in Fig. 3.

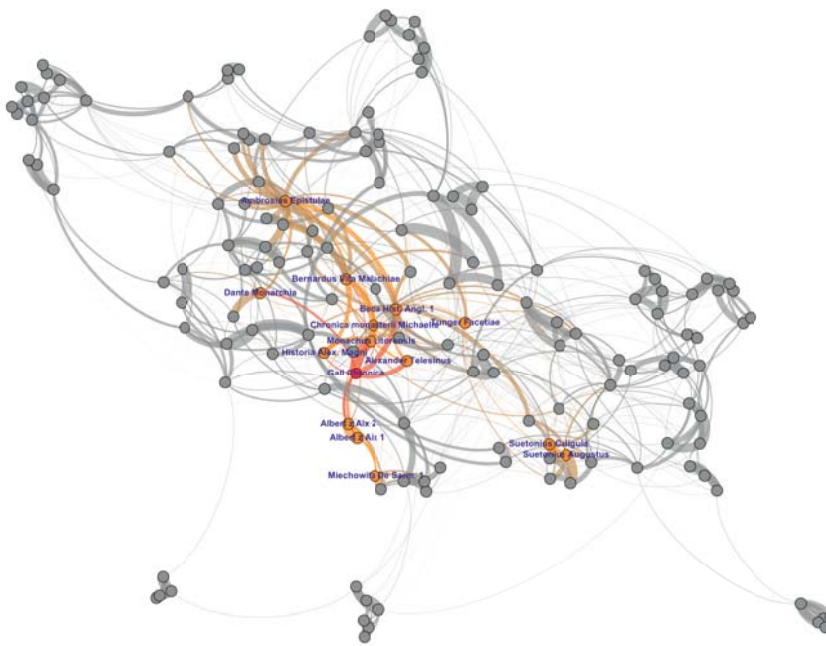

Fig. 3: A network of stylometric connections between 159 Latin texts from various epochs (prose works only). The colour is used to mark the texts showing similarities to Gallus's *Chronica*.

http://dx.doi.org/10.12775/APH.2015.112.01

The interpretation of the network of 159 texts might commence with the observation that the Antique texts tend to be grouped in the outer areas of the graph, its central area being occupied by, primarily, medieval and Late-Antique texts (only a few samples are subscribed, for the sake of legibility). The right side of the diagram displays, virtually, historical works only – those by Julius Caesar, Tacitus, Sallust, Florus, and others. In terms of the authorship of the *Chronicle*, fundamental are the works which are directly connected to the Gallus's work, whether stronger or weaker. All such similarities are marked in colour in the diagram, the relevant nodes being subscribed. These include, starting from the most distant relationships: Sallust's *Caligula* and *Res Gestae Divi Augusti*, Ambrose's *Letters* (these being very strongly connected with a number of other nodes), Dante's *Monarchia*, St Bernard of Clairvaux's *Life of St Malachy of Armagh*, Albert of Aachen *Historia Ierosolimitana*, and Augustin Tünger's *Facetiæ*. Gallus's *Chronicle* demonstrates some stylometric affinity with these works, yet none of them appears to be its closest neighbour. Such neighbours are seen best in a close-up, as shown in Fig. 4.

This close-up shows further relationships, including: *Chronicon Coenobii Sancti Michaelis de Clusa* (11th c.), *The Alexander Romance* by Pseudo-Callisthenes, *The Ecclesiastical History of the English People*

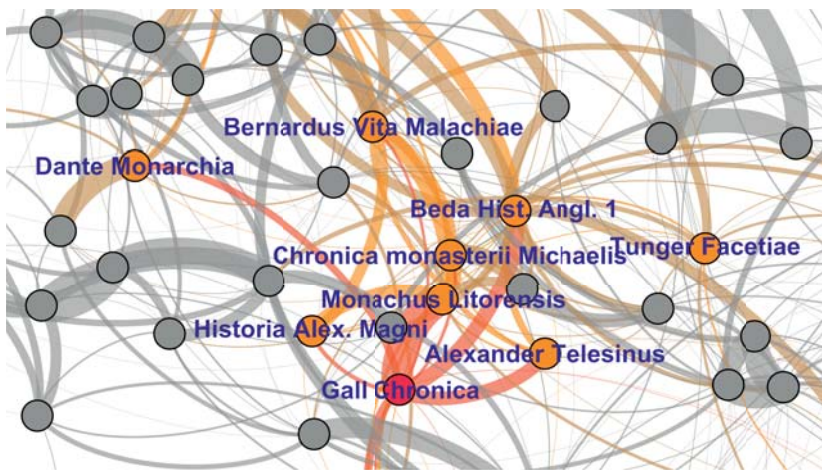

Fig. 4: A network stylometric connections between 159 selected Latin texts: a fragment.

http://dx.doi.org/10.12775/APH.2015.112.01

by Beda Venerabilis, Alexander of Telese's *The Deeds Done by King Roger of Sicily*, and, lastly, the *Translatio* by Monachus Littorensis, which is associated with the Gallus's work due to an extremely strong stylistic resemblance. The diagram shows quite clearly how strong the *Chronica–Translatio* connection is; in the other sections of the graph, similarly strong connections are the case only with the works written by the same author (the instances of Cicero, Vitruvius, or Varro).

An obvious conclusion comes to mind: the similarity between Gallus and Monachus is so evident that no accidental concurrence can be the case. The clear implication is that both anonymous authors are, in fact, one and the same author; however, let us recall the aforementioned caveat described as the open-set problem. Bearing in mind our incomplete knowledge on medieval literary output, one cannot be absolutely certain that Gallus was identical with the 'Monk of Lido'; however, it can be reasonably argued that of all the authors known to us, Monachus is definitely the closest to Gallus, the type of similarity being one that is usually characteristic of texts written in one and the same hand. As long as no better prospective author is proposed, and as long as his work shows no stronger stylistic affinity with the *Chronica* compared to Monachus, the Venetian background hypothesis will be difficult to undermine.

V
CONCLUSIONS

The stylometric analysis of Gallus Anonymus's *Chronica Polonorum*, carried out with use of three various 'nearest neighbour' methods, has shown that of all the texts having been liable to analysis, Monachus Littorensis shows the strongest affinity with Gallus, in quite a resolute manner. This is obviously not conclusive as far the actual authorship of the *Chronicle* is concerned; still, such very strong resemblance implies that one cannot be indifferent to the Venetian background hypothesis. The evidential value of the experiment described herein remains considerable, in spite of its reconnaissance character, also because it has reconfirmed the stylometric observations made by Tomasz Jasiński, which were based on a thoroughly different research method.

The general conclusion and, in parallel, the research postulate implied by the analyses described in this essay is the need to collect

http://dx.doi.org/10.12775/APH.2015.112.01

a larger comparative corps that would comprise, above all, a considerable, and possibly complete, representation of eleventh- and twelfth-century texts – historical works and pieces of rhythmic prose in the first place. It cannot be precluded that some other interesting (and unobvious) prospects might appear amidst Gallus's stylometric neighbours. Further analysis ought, moreover, to focus on linguistic characteristics other than those researched herein – one example being Gallus's syntactic structures. Lastly, worth revisiting and being critically reanalysed is, probably, the chronological evolution of style as noticed in the *Chronicle*. Such issues, however, are beyond the scope of this essay, which has focused on the authorship question.

trans. Tristan Korecki

SELECTED BIBLIOGRAPHY

Borawska Danuta, 'Gallus Anonim czy Italus Anonim', *Przegląd Historyczny*, lvi (1965), 111–19.

Eder Maciej, 'A Bird's Eye View of Early Modern Latin: Distant Reading, Network Analysis and Style Variation', in Michael Ullyot, Diane Jakacki, and Laura Estil (eds.), *Early Modern Studies and the Digital Turn* (forthcoming).

Eder Maciej, 'Style-Markers in Authorship Attribution: A Cross-Language Study of the Authorial Fingerprint', *Studies in Polish Linguistics*, vi (2011), 99–114.

Jasiński Tomasz, 'Czy Gall Anonim to Monachus Littorensis?', *Kwartalnik Historyczny*, cxii, 3 (2005), 69–89.

Jasiński Tomasz, *"Kronika polska" Galla Anonima w świetle unikatowej analizy komputerowej nowej generacji* (Wykłady inauguracyjne Instytutu Historii Uniwersytetu im. Adama Mickiewicza, 6: Semestr letni 2010/2011, Poznań, 2011).

Jasiński Tomasz, *O pochodzeniu Galla Anonima* (Kraków, 2008).

Koppel Moshe, Schler Jonathan, and Argamon Shlomo, 'Computational Methods in Authorship Attribution', *Journal of the American Society for Information Science and Technology*, lx, 1 (2009), 9–26.

Love Harold, *Attributing Authorship: An Introduction* (Cambridge, 2002).

Plezia Marian, 'Nowe studia nad Gallem-Anonimem', in Helena Chłopocka and Brygida Kürbis (eds.), *Mente et litteris. O kulturze i społeczeństwie wieków średnich* (Poznań, 1984), 111–20.

Stamatatos Efstathios, 'A Survey of Modern Authorship Attribution Methods', *Journal of the American Society for Information Science and Technology*, lx (2009), 538–56.

Maciej Eder – early Polish literature and scholarly editing; associate professor at the Institute of Polish Studies, Pedagogical University of Cracow and at the Institute of Polish Language, Polish Academy of Sciences, Cracow; e-mail: maciejeder@gmail.com

http://dx.doi.org/10.12775/APH.2015.112.01