Twenty-One* Pseudo-Chrysostoms and More: Authorship Verification in the Patristic World

Thibault Clérice[1,*,†], Anthony Glaise[2,‡]

[1]ALMAnaCH, Inria Paris, France

[2]CESR, Université de Tours, France

Abstract

As the most prolific of the Church Fathers, John Chrysostom (344–407 CE) has a vast textual mass and theological importance that has led to a significant misattribution of texts, resulting in the existence of a second corpus known as the pseudo-Chrysostomian corpus. Like many Greek-language Church Fathers' works, this corpus comprises anonymous texts, which scholars have attempted to reattribute or group together based on factors such as the person's function, biography, ideology, style, etc. One survey conducted by Voicu in 1981 explored potential groupings of such texts and produced a critical list of 21 Pseudo-Chrysostom works identified by scholars, including Montfaucon (1655–1741), one of the first modern editors of Chrysostom's writings. In this paper, we present a novel approach to addressing pseudonymous work in the context of Chrysostomian studies. We propose to employ Siamese networks within an authorship verification framework, following the methodology commonly used in recent computational linguistic competitions. Our embedding model is trained using commonly used features in the digital humanities landscape, such as the most frequent words, affixes, and POS trigrams, utilizing a signal-to-noise ratio distance and pair mining. The results of our model show high AUCROC scores (84.5%). Furthermore, the article concludes with an analysis of the pseudo-Chrysostoms proposed by Voicu. We validate a significant portion of the hypotheses found in Voicu's survey while also providing counter-arguments for two Pseudo-Chrysostoms. This research contributes to shedding light on the attribution of ancient texts and enriches the field of Chrysostomian studies.

Keywords

Patristic Studies, Ancient Greek, Stylometry, Siamese Networks, Authorship verification

1. Introduction

Late Antiquity literature in Latin and Ancient Greek bears a profound influence from Christian literature, shaping the era's literary landscape. Patristic studies focus on the examination of the earliest Christian authors, spanning from the 1st century CE to the 7th century CE, with some scholars extending the period up to Jean Damascene in the middle of the 8th century

CHR 2023: Computational Humanities Research Conference, December 6 – 8, 2023, Paris, France

[*] Corresponding author.

[†] Contribution on all domain.

[‡] Contribution on the pseudo-chryostomian context and evaluation of results in regards to the "traditional" patristic academic work.

[§] All authors participated in the writing of the paper.

thibault.clerice@inria.fr (T. Clérice); aglaise@laposte.net (A. Glaise)
https://ponteineptique.github.io/ (T. Clérice)
0000-0003-1852-9204 (T. Clérice); 0000-0003-4715-5184 (A. Glaise)

CE. Within the realm of Church Fathers, two prominent figures stand out for their enduring impact on the corpora: Augustine (354–430) in Latin and John Chrysostom (344–407) in Ancient Greek. Despite their significance, both Augustine and John Chrysostom encountered a common challenge in Late Antiquity. Numerous works were falsely attributed to them, possibly with the intention of elevating the popularity of these writings, whose actual authors were less renowned. This misattribution has led to the creation of an extensive body of texts with uncertain authorship. According to Voicu [39], there are "more than a thousand" pseudo-Chrysostomian works whose true originators remain unknown[1]. To address this issue, scholars of patristics have attempted to categorize some of these spurious works, associating them with a single person, who may be anonymous or identified with a specific historical figure.

The task of distinguishing anonymous authors within the pseudo-Chrysostomian corpus is a significant and essential endeavor. By doing so, historians, patristicians, and philologists can work with a more manageable collection of texts. This process offers valuable insights into the construction of Christian theological frameworks, allowing the identification of ideological or thematic clusters within authorship groups. Establishing authorship for texts with at least one proposed date enables scholars to study the evolution of Christian ideology more accurately. Moreover, in rare cases, dating or attributing a specific anonymous text (especially when milestones are present) can provide new dates to other texts, as they might be cited or reused [7]. This effort holds immense potential for advancing our understanding of the historical and intellectual landscape of Late Antiquity, contributing to a deeper comprehension of the development of the Christian tradition.

In traditional patristics, authorship attribution largely relies on the identification of rare patterns [33], as well as extra-textual clues, such as historical events serving as *terminus post quem* or *ante quem* markers, and the analysis of ideological traits or the use of specific quotations. Authors' tendencies to rely on particular parts of the *Bible* or cite other authors also serve as *de facto terminus ante quem.* In 1981, Voicu [39] presented a comprehensive analysis of attributions using such methods in the context of the pseudo-Chrysostomian corpus. He identified 21 text groups for which a common authorship has been proposed by himself or other scholars. However, for some of these pseudo-Chrysostoms, Voicu and the scholars he cites sometimes express hesitancy regarding the attribution or hold differing perspectives on the matter.

In such a context, automatic authorship verification can offer new insights that either support or challenge previous hypotheses. Two other fields that are actively engaged in authorship verification are computational linguistics and digital humanities (DH). In the realm of stylometrical analysis, DH scholars tend to treat texts as bag-of-words and utilize statistically significant features, such as the most frequent words of a corpus [9], part-of-speech 3-grams, and character n-grams [5]. On the other hand, computational linguistics (CL) has recently shown a preference for treating text as a sequence of words, employing masked language models [36]. In terms of approach, DH papers lean towards classification (utilizing SVM or similar models) or unsupervised clustering methods. In contrast, CL uses both DL classification methods and siamese neural networks, regardless of the type of input [17].

In our research, we propose a novel methodology that combines approaches from both computational linguistics and digital humanities to analyze the 21 potential Pseudo-Chrysostoms

[1] Voicu has been working on a database for these texts, https://www.trismegistos.org/pseudo-chrysostomica/.

identified by Voicu and his predecessors. Specifically, we aim to evaluate the effectiveness of different features, including most frequent words, affixes, and POS 3-grams, within the context of a Siamese network using a linear projection in N dimensions. To maximize the potential of our corpus, we introduce Easy-SemiHard Pair Mining[34] to our batch learning process and utilize a signal-to-noise ratio distance[43] to distinguish and separate our texts.

In summary, the contributions of our paper are as follows:

- Introducing a new approach to authorial verification for ancient texts, incorporating Siamese networks and easy-semihard pair mining into the landscape of stylometry within computational humanities.
- An analysis of various size for three types of features (POS, MFW, Affixes), which shows consistency in terms with previous studies.
- An evaluation of common distances (Manhattan and Euclidian distance [L2]) and a newly introduced one (signal-to-noise ratio distance), which shows that the latter out-performs the best performing Manhattan distance in terms of stability within our settings.
- Conducting an in-depth analysis of the results obtained from applying our approach to the 21* pseudo-Chrysostoms' texts identified by Voicu.

The remaining of the paper is organised as follows. Section 2 provides background on the article of Voicu and in general on the issue of the pseudo-Chrysostomian corpus, as well as a deeper background on authorship verification and stylometry in general. Section 3 (Proposed Methods) provides details about the architecture used for the experiments. Section 4 (Experimental Setup) provides insight on the corpus, features selection and metrics. Section 5 provides an evaluation of the results on independent test sets. Section 6 reuses the models built and to provide insight on the PC corpus.

2. Background and Related Work

2.1. Background

Regarding the pseudo-Chrysostomian corpus, a non-specialist might notice the lack of recent discussions concerning various attribution hypotheses. Most of these "old" hypotheses trace their origins back to the 18th century, thanks to the diligent efforts of Bernard de Montfaucon, a Benedictine monk who dedicated his work to publishing the works of Athanasius (1698) and John Chrysostom (1718). Montfaucon's editions and commentary laid the groundwork for later comprehensive editions of numerous patristic texts [30], including the renowned *Patrologia Graeca* compiled by Migne, which is still widely used today as it is now in the public domain. Fortunately, in 1981, Voicu [39] provided the most recent comprehensive summary of the authorship hypotheses within the pseudo-Chrysostomian corpus. This summary encompassed both refutations and ongoing debates surrounding various authorship attributions. Since then, some new hypotheses have emerged, but there hasn't been a comparable effort to Voicu's in terms of summarizing the existing scholarship.

In his paper, Voicu provides a summary regarding authorship clustering of 88 texts, grouped under the pseudo-identity of 21 different authors by various scholars (PC1 to PC21). On top of the 21 base PC, another group of text is hypothetically attributed to one of them: PC20 is accompanied by PC20b, for which he drafts a possible ensemble without confirmation that they might be of the same authors (see Table 1). The arguments regarding the attributions can vary, and we summarise them in four different categories:

- **Theological arguments** are based on ideological (in)compatibilities between texts, in the context of a non-unified Christian religion in which subgroups can be found. *E.g.* PC16 is categorized as a moderate Antiochian, which makes them incompatible with an Alexandrian ideology.
- **Sequential arguments** are based on the obvious continuity between two texts, with established narrative links in both senses (Text A builds on Text B and vice-versa).
- **Stylistic arguments** are based on the style of the authors. They range from the use of what is perceived as ”bad Greek” (PC6 and 7) to citation habits regarding the scriptures.
- **Extra-textual arguments** are mostly based on events or texts referenced within the texts, which gives them a common *terminus ante quem* or *post quem*, or actually makes them incompatible. It also refers to transmission proofs, such as the constant grouping of the same texts in their transmission history (PC17).

Table 1
List of pseudo-Chrysostoms and the scholars behind the hypotheses. Confidence concerning the grouping of texts under a single pseudonymous author is provided based on Voicu's commentary of each scholar's analysis, including his own. We present a simple typology of the arguments when provided. All the cited scholarship, except for Montfaucon, was published between 1940 (Marx [21]) and 1981 (Voicu).

Cluster	Number of texts	Original Hypothesis	Confidence	Additional analysis	Confidence	Argument type
1	2	Montfaucon		Altendorf	Refuted	theological
2	2	Montfaucon		Voicu	Confirmed	continuity
3	3	Montfaucon		Voicu	Refuted	
4	7	Montfaucon		Voicu	Low	
5	3	Montfaucon		Voicu	Partially refuted	
6	3	Montfaucon		Voicu	Low	
7	7	Montfaucon		Voicu	Possible	
8	2	Montfaucon		Voicu	Refuted	stylistic
9	5	Marx		Voicu	Refuted	
10	2	Weyer				
11	3	Nautin				
12	2*	Liébart				theologic
13	3	Leroy		Voicu	Refuted	
14	4	Voicu	High			
15	5	Voicu	Mostly high			
16	5	Voicu	High	Wenger	Partially Possible	
17	2	Voicu	High			
18	2	Voicu	Possible			stylistic,theologic
19	2*	Rilliet	High			
20	5	Datema	High	Voicu		
20b	12	Datema	Possible	Voicu		
21	2	Voicu	High			extra-textual

Most of the cited texts are available in digital formats, specifically in the *Thesaurus Lingua Graeca*, which is unfortunately closed access[2]. However, one PC could not be tested in our

[2] All the feature extraction process is shared within the repositories, and the features themselves are made available.

framework: PC19 only refers to one text in Ancient Greek, the other text being a Syriac translation. PC12 has been produced using the OCR available on Google Books, which has been lightly post-corrected.

2.2. Related work in stylometry and authorship verification

The present work is related to two fields: computational linguistics (CL) and digital humanities (DH). The existing literature reveals distinct approaches to the problem of authorship verification or authorship identification, primarily influenced by the attributes of each field's corpus and expectations for explainability. We focus on the latest approaches applied in both fields, emphasizing the common technical approaches they have shared in the past.

Computational Linguistics CL offers a clear landscape thanks to PAN, a "series of scientific events and shared texts on digital text forensics and stylometry." PAN's shared tasks have provided recurring competitions since as early as 2011, focusing on authorship attribution (2011, 2012, 2018, 2019) or verification (2013–2015, 2020–2023). As in most other computational linguistics tasks, deep learning has seen a rise in popularity, leading to improved scores but lower explainability. An insight into some of the approaches taken in authorship verification since 2015 reveals the following methods and features used:

- In 2015, Word-based n-grams, sentence length, word frequencies, punctuation frequencies, POS frequencies, and POS n-grams were shared features across different papers [28, 27]. These features were eventually fed into standard classifiers such as Random Forest or SVMs.
- In the 2018 PAN authorship attribution task [19], features continued to focus on characters and word n-grams, with various weighting and normalization methods. While SVMs were commonly used, some early neural network approaches also appeared.
- In the 2020 PAN authorship verification task [16], methods employing neural networks made an appearance, particularly in the form of Siamese neural networks. Features mainly remained classical stylometric features, such as normalized frequencies of tokens or POS. The winning paper utilized a Siamese network with extracted "linguistic embedding vectors" using an LSTM network with attention to produce document embeddings.
- In 2021 [17], features and approaches from 2019 were carried over, with a paper using BERT and another one using Siamese networks ranking first and second place, respectively, in the large dataset competition on overall scores, with the first one winning the competition on all provided metrics.
- In 2022 [36], all competitors moved away from SVM and instead used either masked language models' embeddings or previous existing features with Siamese network approaches or fully connected neural networks.

With this summary, we observe that CL is gradually moving away from explainability and increasingly employing text sequences (using RNNs, CNNs, or transformers like BERT or T5) rather than treating texts as bags of words (BoW). Scoring is conducted using various metrics, including AUROC and a modified F1-Score $F_{0.5u}$ [3], which "emphasizes correctly-answered same-author cases and rewards non-answers". Siamese networks gained prominence in authorship verification tasks in 2020 and 2021.

In contrast to previous competitions, 2022 departed from 2021 in the type of dataset used. While 2021 and previous shared tasks utilized datasets sharing a common domain (such as fanfictions for 2021), 2022 focused on *cross-discourse authorship verification*, encompassing "emails, essays, texts messages, and business memos." This shift resulted in a significant performance drop compared to previous years. Additionally, regarding the dataset, each competition relied on fixed pairings of texts, with limited or no information about the author in the metadata of the documents.

Digital Humanities Unlike CL, DH has largely preferred using feature selection and treats most texts in an authorship attribution framework rather than authorship verification. The three most frequently recognized features are: most frequent words [8, 32], function words, affixes (especially for languages with significant variations in spelling or flexions) [6], and POS, with the latter often used as n-grams [5]. Some other features, such as rhymes [5], meters [24], and even treebank syntactic tags for Ancient Greek [14], have demonstrated usefulness and stylometrical importance but are less commonly used[3].

In terms of technologies, tools offering explainability (particularly feature weight) are highly favored, particularly SVMs or, in an unsupervised setting, hierarchical clustering using distances such as Manhattan, Cosine, Burrows' Delta [4], or Eder's Delta [10], etc. We hypothesize that the need for explainability arises due to the fundamentally different objectives pursued by computational linguistics (CL) and digital humanities (DH) in the context of authorship verification and identification. While CL is primarily concerned with developing robust models at scale, DH seeks to employ the results as potential evidence in scholarly research. Consequently, the ability to interpret the decisions made by a model becomes invaluable for detecting "invisible biases to the human eye," such as the presence of unique characters in specific author editions, which might lead the model to overfit and yield erroneous conclusions. This risk is significantly amplified in the context of the frequent diachronic peculiarities prevalent in most DH inquiries or the constraints posed by a limited pool of available historical documents. Conversely, CL typically concentrates on short-term, cohesive, and large corpora, such as social network messages or fan fiction, where interpretability might not be as crucial.

In 2016, General Imposters (GI) methods [20], an authorship verification method, were introduced to DH [18]. Unlike authorship verification methods of CL, which rely on learning to recognize a true pair of texts from the same author and a pair of texts from different authors, GI introduces impostor authors alongside candidate authors in a classification task. It randomly removes features (50% by default in the *stylo* package [11]) over *n* experiments (100 by default in *stylo*) and proposes, for each of these experiments, the author of the closest text as a potential match. If one of the candidates has an overwhelming presence in the closest texts over all experiments, it is proposed as the verified author of the text.

[3] Probably due to the cost of said annotations and the lack of well-performing automatic models for such tasks.

3. Proposed Method

We propose to bridge the gap between DH and CL practices by introducing supervised authorship verification using text pairing validation through a siamese network. However, Voicu's anonymous PCs should not be used in the context of supervised training and authorship attribution. This is because supervised training relies on having ground truth data to identify the anonymous authors, and for some of Voicu's PCs, they may not even be processed through the general impostors method as they only offer a single pair of texts, which is insufficient for effective training and verification.

Features Following most of the DH literature and PAN approaches until 2020, we select the relative frequency of most frequent words (MFW), POS tri-grams (MFP), and trigram character affixes (MFT) within each text. We reject the use of any punctuation-based features, as punctuation in our texts is the result of the editorial task and can be editor or period dependent (an editor from the 18th century might not punctuate like a contemporaneous one). In order to test these, we tested a range of different values for each features:

1. 0, 100 and 200 most frequent POS-trigrams
2. 0, 250, 500, 750, 1000 most frequent words[4]
3. 0, 250, 500, 750, 1000 most frequent affixes.

MFP(100), MFW(1000), MFT(1000) provided the lowest standard deviation and highest AUCROC with two different distances (cf. Appendix, section B). These numbers, at least for MFW, are in line with current literature, most notably the recent paper from Rebora's [31]. We concatenate the relative frequencies of these features into a single vector, denoted as features T_i, consisting of 2100 dimensions representing the text sample i.

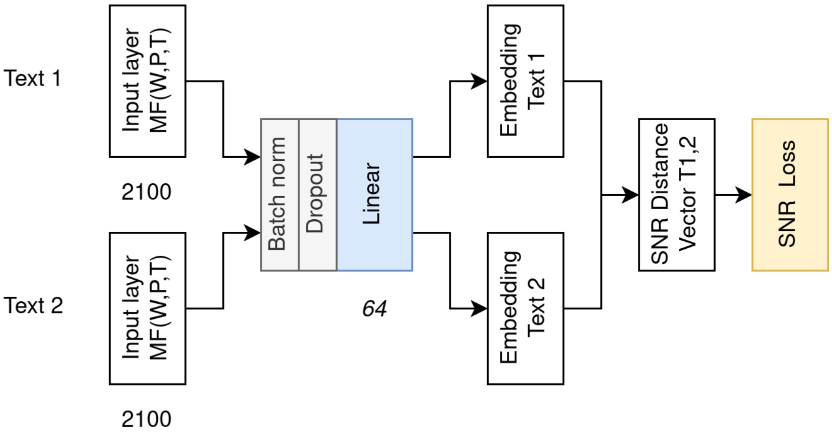


Figure 1: Architecture of our models: vectors are passed in parallel in the embedding projection layer and a distance vector is then computed.

[4] We maxed our value at the size of our sample, 1000 words.

Model structure The model follows a typical architecture of linear layers in the context of Siamese networks (see Figure 1): feature vectors are passed through the same projection layers, paired, and a distance-based decision is proposed. We use linear layers to reduce T_i to a given dimension m, resulting in an embedding representation E_i of the text. For the loss and distance metric, we tested Manhattan, Euclidean (L2) and "Signal-to-Noise Ratio" distance [43] (SNR, SNRD for the distance) with contrastive loss. SNR-D considers the noise between an anchor embedding E_i and a compared embedding E_j, represented as $N_{ij} = E_j - E_i$, and uses its variance as an informative measure (variance of values across dimensions). The SNR for pair i, j is defined as $SNR_{i,j} = \frac{var(E_i)}{var(N_{ij})}$, and the SNR-Distance is $\mathrm{SNRD}_{ij} = \frac{1}{\mathrm{SNR}_{ij}}$. Unlike distance metrics such as the Euclidean or Manhattan distance, $\mathrm{SNRD}_{ij} \neq \mathrm{SNRD}_{ji}$, which can provoke unmirrored attributions (cf. our analysis of PC corpora).

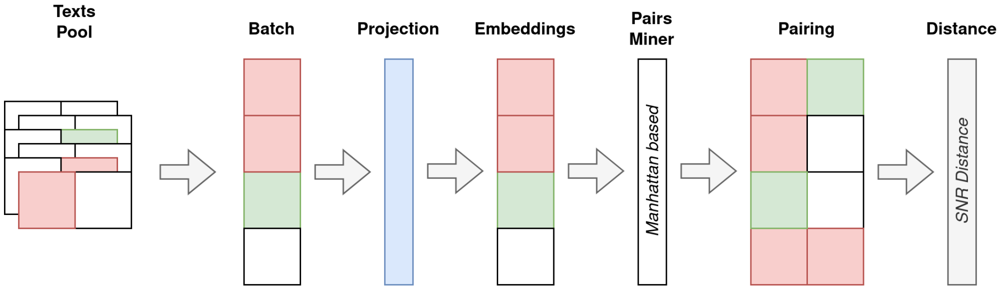


Figure 2: Learning pipeline using miner to provide pairs within the model.

Learning methods Unlike PAN tasks, we can build our own corpus without pre-established pairs, allowing us to utilize pair mining methods, which are particularly useful in the context of Siamese networks (see Figure 2). We use Easy–Semi-Hard (ESH) triplet mining [42, 34] using the aforementioned distance. Given embeddings E and their classes K, $ESH(E, K)$ computes the distance between all embeddings and provides all positive pairs ($K_i == K_j$), as well as semi-hard and hard negative pairs. Negative pairs are considered semi-hard when "they are further away from the anchor than the positive exemplar, but still hard because the squared distance is close to the anchor-positive distance". We use this miner for both the training step and the evaluation step. In their recent paper on the state of authorship verification, Tyo, Dhingra, and Lipton [38] showed that ESH mining could help feature-based models be as good as sequential models using transformers or similar layers.

4. Experimental Setup

Datasets Unlike social networks content or fan-fictions, or even 19th-century literature, Ancient Greek literature spans from 9 BCE to at least 9 CE in the context of our problem. We are dealing with variation in corpus genre, linguistic changes, ideological changes, etc. In order to address this at the training and evaluation step, we design a corpus that is focused on Christian and theological texts using the TLG CD-ROM [2] and their XML export through Diogenes [15].

Unfortunately, the raw dataset is not shareable, and no current open dataset provides a quantitatively large enough dataset for this period of Ancient Greek[5]. For POS tagging, we use Singh, Rutten, and Lefever [35]'s Ancient Greek BERT-based tagger, which is, to our knowledge, the only Ancient Greek tagger trained with both Medieval Greek and classical Greek.

Within the available corpus, we removed:

- any *dubia* or *spuria* (de-attributed texts) or anonymous texts;
- any text from John Chrysostom, as the sheer mass of his work and his style seemed to impact the model too much in early experiments, as well as both the *Old* and *New Testaments*;
- texts marked as commentaries, *scholia*, fragments, codices transcriptions, or specific *recensio*;
- any text containing a single line of poetry[6].

However, all of our most frequent features were extracted from the full Christian corpus.

To counter the reduced mass of texts and address the imbalanced state of the dataset, we produced 1000-word samples for each text using the following rules:

- if the text was shorter than 2000 words, we used the 1000 words in the middle of the text, in order to avoid introductions and conclusions.
- If the text is larger than 2000 words, we used up to 5 samples of 1000 words, randomly selected within the text except for the first 500 words and the last 500 words.

We then split the corpus in an 80-10-10 ratio along the titles of the various works, so that authors can span in different sets but samples of the same title remain inside the same text (see Figure 3). The final corpus is diverse in genre but is composed of two genres which make up around two-thirds of its texts (Homelia and Treatise) while Oratio, Letters, Hagiography, and other non-frequent genres rapidly decrease in numbers of titles (see Figure 4, full list of authors in appendix, Table D).

The PC corpus was not sampled, but relative frequencies were issued using the same process.

Metrics We chose to evaluate the general performance of the model based on the Area Under the Curve of the Receiver Operating Characteristics (AUCROC), which allows for measuring the relationship between the False Positive Rate (FPR) and the True Positive Rate (TPR). This metric also enables us to select a task-oriented threshold by minimizing the FPR while maintaining a high enough TPR.

Evaluating stability Adding to this metric, we chose to also evaluate the stability of the precision, after seeing in our first experiments important variation between models using the same parameters. We expect the randomness of training neural networks to be responsible for producing latent spaces focusing on different features, and as such wanted to provide a full picture of these potential variation. To evaluate the variation, we evaluate

[5] We hope, however, that the Patristic Text Archive [37] will reach one day this kind of milestone for reproducibility purposes. We provide the annotated and processed samples as well as the processing pipeline.

[6] Detected through the presence of TEI-l tags in the XML files.

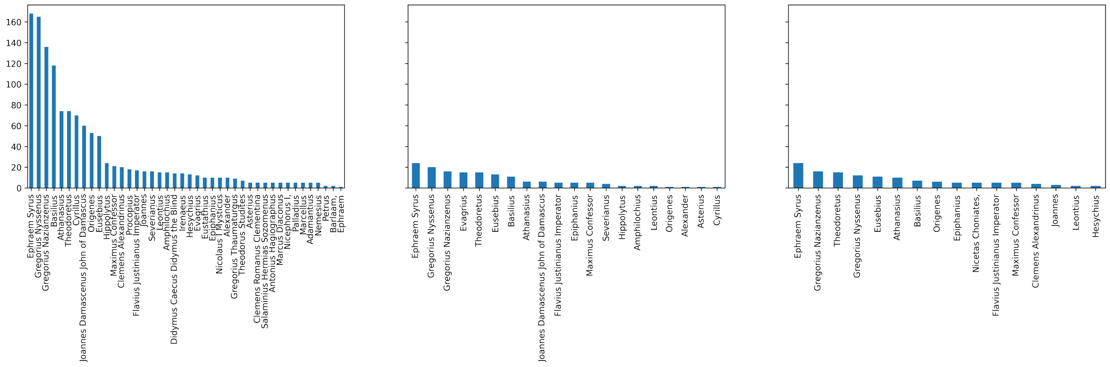


Figure 3: Sample count per author in the training (left), development (centre), and testing set (right).

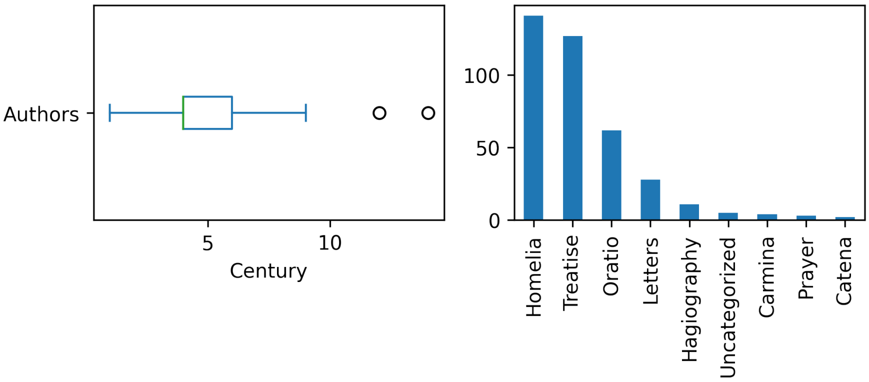


Figure 4: On the left, authors' century of activity, on the right, distribution of genres across works. Each author's century of activity is either provided by the century they are known to live in (uncertain dating) or the century for which they lived most of their adulthood (adulthood being considered as 20 years old and more).

1. the standard deviation of their AUCROC, independently from the other hyperparameters. Each models was trained three times for each parameters combination, we look at their average standard deviation across distances.
2. using the framework of inter-annotator agreement through the Fleiss Kappa (κ, Fleiss and Cohen [13]) – which allows for $N > 2$ annotators – and the test corpus, we look at the k of the two best distances, Manhattan and STN (cf. appendix, section B). We train 100 models for each distance. To make binary decision on the test set, we use a threshold based on the development set precision.
3. using the corpus for qualitative analysis, we look at the variation of pairing percentage (described below) within each pseudo-Chrysostomian sub-corpus.

Qualitative analysis method and pairing percentage To apply our model to the pseudo-Chrysostomian corpus, we need to provide a confidence score to interpret the distance between pairs of texts. We choose to compute the precision of our model at a given distance threshold

δ_{AB} for each pair A-B (and its reverse B-A) on both the development set and the test set. This threshold produces a "precision threshold" (PT) that allows us to estimate the probability of a false positive at any given distance. For example, if $\delta_{AB} = 0.4$ and the precision of the model for distances less than or equal to 0.4 is 0.9, then $PT_{AB} = 0.9$, indicating a high probability that A and B are from the same author.

In order to have a quick overview for the pseudo-Chrysostomian corpus as a whole, we evaluate the "pairing percentage" at a variety of precision threshold. Pairing percentage indicates how many positive pairing are operated at a given PT.

Hyper-parameters The hyperparameters used for the experiment were as follows: Adam optimizer, learning rate of $1e^{-4}$, embedding size of 64, batch size of 64, 30% dropout of features, class sampling of 2, and a minimum of 100 epochs for training. Training was stopped after 20 consecutive bad epochs, using the dev loss as an indicator.

Software implementation The experiment was implemented using the following software and hardware: it was run on an nVidia RTX 3090 GPU with 24GB of RAM, using `torch` [29], `torchmetrics` [26] for AUROC metrics, `pytorch-lightning` [12] for training, and `pytorch-metric-learning` [23] for the mining operations. The same process was also successfully run on a CPU (AMD Ryzen 5700 with 32GB RAM).

5. Results

5.1. Best distances and models' consistency

On top of the features selection, unlike most traditional DH approaches, training linear models to produce latent spaces can introduce potential randomness due to the initialization of model weights, the choice of optimization algorithms, etc. This possibility is exacerbated by the fact that we are using a changing development pool produced through ESH mining of the full development set.

Looking at the standard deviation of our models across all parameters (Table 2a), STN seems to have the most stable standard deviation moving from dev to test, and the lowest on the test set. This would indicate that STN based model are more stable across runs from a score perspective, and that, at least in our context, that its performances translate in a similar fashion from the dev to the test set. This is further confirmed by our analysis of the variation of $\text{AUCROC}_{dev} - \text{AUCROC}_{test}$ across our parameter search, with a very stable distance between dev and test sets (cf. Table 2b).

Looking at our models randomness further, we can see that the models using STN distance (Table 3) are much more in agreement across 100 different models and training seed than those using Manhattan distance, despite higher AUCROC scores. We can see that STN has more than 2.5 times the κ of Manhattan, and Manhattan only reaches the .5 κ at 80% of PT. In a philological set-up, this advocates even more for the use of the most stable model.

Looking at the qualitative dataset and pairing percentage, we can see the same phenomenon, with much more unstable pairing using Manhattan rather than STN (Figure 5). At 100 PT, the

Table 2
Variation across the parameters sweep of models according to the distance (and therefore loss) used.

(a) Mean of standard-deviation across the parameters sweep depending on distances.

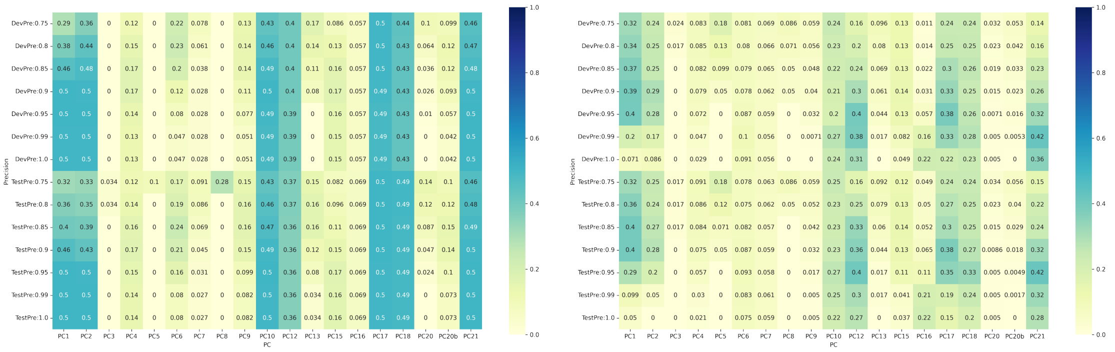
	Dev Mean	±	Test Mean	±
Distance				
L2	1.04	0.55	1.82	1.02
Manhattan	1.09	0.59	1.84	1.29
STN	1.18	0.68	1.14	0.59

(b) Absolute difference between dev and test scores across models depending on the distance used.

	Mean	±	min	25%	Median	75%	max
Distance							
L2	6.63	2.28	0.29	5.26	6.37	7.91	11.66
Manhattan	6.80	2.68	1.08	5.22	6.87	8.42	12.43
STN	1.87	1.04	0.01	1.10	1.83	2.61	4.91

Table 3
Fleiss Kappa (FK, inter-annotator agreement) using 100 models as annotators on the test set. Predictions on test are thresholded using the development distance found at the given precision.

Development Precision	100%	99%	95%	90%	85%	80%
			Fleiss Kappa			
Manhattan	0.18	0.18	0.20	0.29	0.40	0.50
STN	0.50	0.60	0.69	0.72	0.72	0.72



(a) Using Manhattan

(b) Using Signal-to-Noise Ratio distance

Figure 5: Standard deviation of the pairing percentage in clusters depending on the PT. Darker means more variation.

standard deviation is 50% for more 6 clusters across for the Manhattan distance, a number unseen using Signal-to-noise at the same threshold. Only lower PT achieve such a high standard deviation (See appendix for more figures across 100 models).

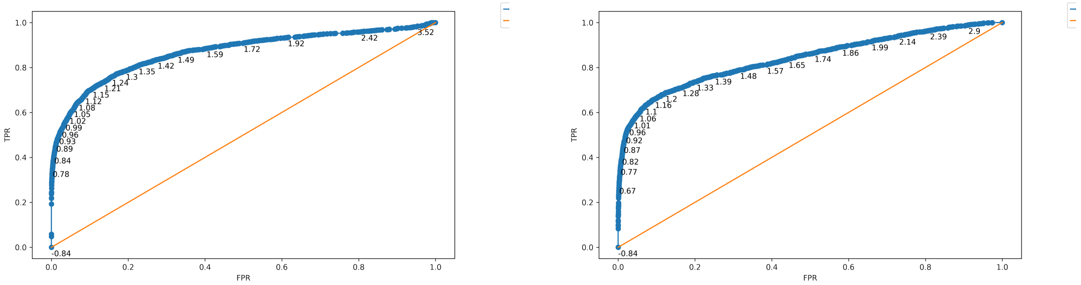
5.2. Analysis of the results

The results achieved in this study fall within the range of the current state of authorship verification. Tyo, Dhingra, and Lipton [38]'s implementation of the N-Grams based model of Weerasinghe, Singh, and Greenstadt [40] and ESH mining achieved between 77% to 91% AU-

Table 4
Details on the train, dev, and test sets along with their corresponding scores. "First FP Dist." stands for First False-Positive distance, which indicates the distance at which the first false positive is encountered. The AUCROC is computed on all pairs, such that most text have more negative pairs than positive pairs overall. The metrics are computed using A-B and B-A pairs.

	Authors	Texts	Samples	AucROC	1st FP Dist.	1st FP Classes
Train	42	384	1294	92.35	0.519	Ephraem Syrus - Cyrillus
Dev	20	47	155	86.75	0.405	Leontius - Amphilochius
Test	16	47	132	83.76	0.478	Gregorius Nazian. - Nicetas Chon.

CROC[7]. The proposed model stopped converging after 219 epochs on the ESH mined pairs and provides an 86.75% AUCROC on all pairs of the development set (see Table 4). The model reaches a slightly lower score on the test set with an 83.76% AUCROC. This is inline with the oberserve differences during the parameter sweep. We also would like to note that due to ESH mining, the pairs used in the dev set for loss might change from one iteration to the others, as the embeddings are used to decide which pairs to use, resulting in a different kind of over-fitting than with a traditional fixed development set.



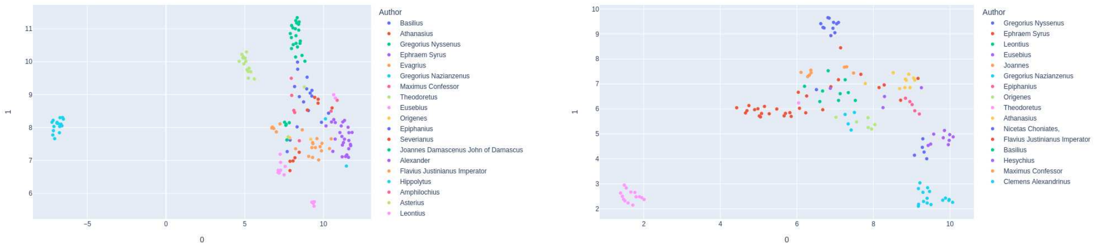
(a) Development set

(b) Test set

Figure 6: ROC curves on both the development set and the test set. Y axis is the True Positive Rate, X axis is the False Positive Rate. Marks on the curve represent the distance threshold used to determine the positives.

The ROC curves (see Figure 6) exhibit a compelling shape, with a significant slope in the initial percentages of the False Positive Rate (FPR), followed by a more gradual increase until reaching 100% True Positive Rate (TPR). This suggests that the model effectively distinguishes between positive and negative pairs, especially in the early stages of classification. UMAP [22] (Figure 7), while still serving as a proxy for the complexity of the embeddings' dimension, provides a clear depiction of the discrepancy between both sets. In the test set, embeddings of texts from the same authors are notably more distant from each other, indicating that the model's representations are better disentangled in this set. On the other hand, the development set displays more overlap and mixing of embeddings, suggesting that the model's representations

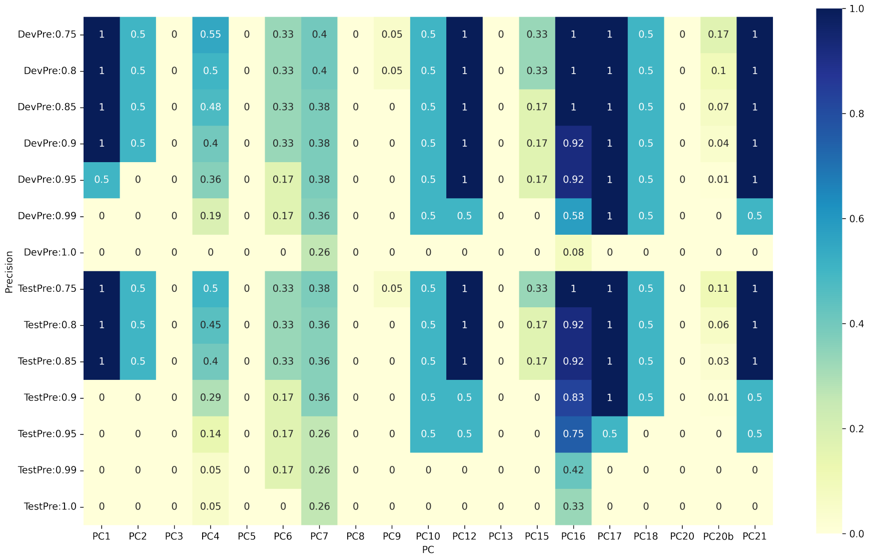
[7] Note that our corpus are different. Numbers from Tyo et al. are given to provide some form of context, given the lack of application of this method on corpus similar to ours.



(a) Development set

(b) Test set

Figure 7: UMAP 2 dimension projections of the embeddings.



Precision	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC12	PC13	PC15	PC16	PC17	PC18	PC20	PC20b	PC21
DevPre:0.75	1	0.5	0	0.55	0	0.33	0.4	0	0.05	0.5	1	0	0.33	1	1	0.5	0	0.17	1
DevPre:0.8	1	0.5	0	0.5	0	0.33	0.4	0	0.05	0.5	1	0	0.33	1	1	0.5	0	0.1	1
DevPre:0.85	1	0.5	0	0.48	0	0.33	0.38	0	0	0.5	1	0	0.17	1	1	0.5	0	0.07	1
DevPre:0.9	1	0.5	0	0.4	0	0.33	0.38	0	0	0.5	1	0	0.17	0.92	1	0.5	0	0.04	1
DevPre:0.95	0.5	0	0	0.36	0	0.17	0.38	0	0	0.5	1	0	0.17	0.92	1	0.5	0	0.01	1
DevPre:0.99	0	0	0	0.19	0	0.17	0.36	0	0	0.5	0.5	0	0	0.58	1	0.5	0	0	0.5
DevPre:1.0	0	0	0	0	0	0	0.26	0	0	0	0	0	0	0.08	0	0	0	0	0
TestPre:0.75	1	0.5	0	0.5	0	0.33	0.38	0	0.05	0.5	1	0	0.33	1	1	0.5	0	0.11	1
TestPre:0.8	1	0.5	0	0.45	0	0.33	0.36	0	0	0.5	1	0	0.17	0.92	1	0.5	0	0.06	1
TestPre:0.85	1	0.5	0	0.4	0	0.33	0.36	0	0	0.5	1	0	0.17	0.92	1	0.5	0	0.03	1
TestPre:0.9	0	0	0	0.29	0	0.17	0.36	0	0	0.5	0.5	0	0	0.83	1	0.5	0	0.01	0.5
TestPre:0.95	0	0	0	0.14	0	0.17	0.26	0	0	0.5	0.5	0	0	0.75	0.5	0	0	0	0.5
TestPre:0.99	0	0	0	0.05	0	0.17	0.26	0	0	0	0	0	0	0.42	0	0	0	0	0
TestPre:1.0	0	0	0	0.05	0	0	0.26	0	0	0	0	0	0	0.33	0	0	0	0	0

Figure 8: Heatmap of the precision thresholds for all pseudo-Chrysostomian corpora. The percentage given in each cell gives the amount of pairs (A-B and B-A) of each PC that are found to be positive given the precision threshold.

are less distinct and might have captured some biases from the data splitting process.

6. Application on PCs

6.1. General results

Using a visualization of pairing percentage using precision threshold (Precision $\geq$ 0.75, see Figure 8), we see that our overall results are remarkably consistent with the summary proposed by Voicu. Notably, the PC 3, 4, 5, 8, 9, and 13 are not confirmed as single authors by our model and are either indicated as low probability clusters or refuted groupings in Voicu's article. On the other hand, only the 20 and 20b clusters are completely ignored by our model despite having a high chance of being from the same authors according to both Datema[7] and Voicu. Among the highly connected clusters, PC 1, 2[8], 12, 16, 17, 18, and 21 are all marked as highly possible or confirmed by Voicu and are found at a precision threshold of 85% or higher, either on dev or test or both.

6.2. Closer look at some pseudo-Chrysostoms

We turn our attention to the pseudo-Chrysostoms whose scores are not relating to the philological studies: PC 6–7, 10–11, 14–15, 20, and 20b.

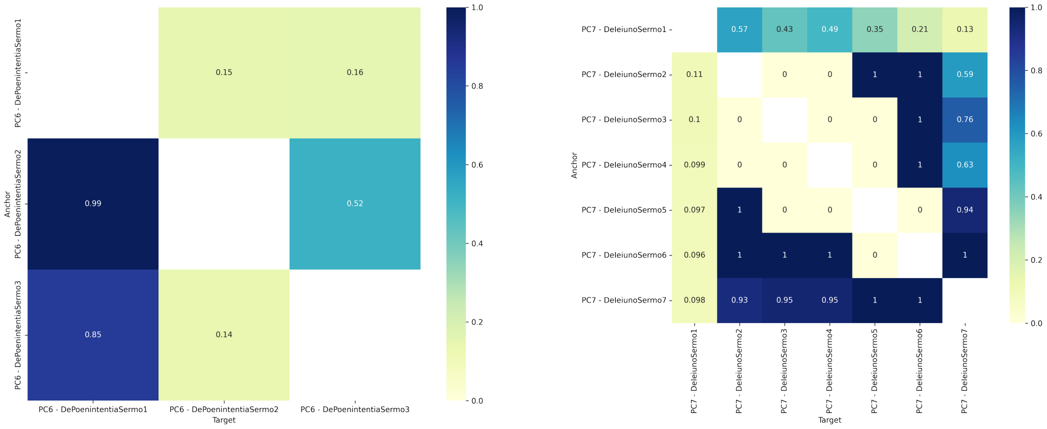
Montfaucon's pseudo-Chrysostoms 6 and 7 Both pseudo-Chrysostom 6 and 7 are hypotheses made by Montfaucon in the 18th century. PC6 is considered to be written by bad (and stupid) ancient Greek speakers (*inepti Graeculi*), with *ad nauseam* use of repetitions and epithets, while ignoring common rules of grammar (PG, 60, 681-682). Regarding PC7 and the *De jejunio* sermons, the only arguments of Montfaucon (PG, 60, 711-712) were again his absence of knowledge of basic Greek (*imperiti Graeculi*) and his ability to write mostly nonsensical things (*plerumque nugacis*).

Looking at the precision threshold matrices in Figure 9, we see that the matrix behaves in a non-reciprocating way, such that some texts are deemed to be of the same author in the sense A–B but not in the sense B–A. Moreover, for PC7, we see that consecutive sermons are unconnected between each other (*e.g. De jejuno 3* and *De jejuno 4* are not deemed to be of the same author at any precision) but are connected with subsequent or previous sermons (*e.g. De jejuno 3* and *De jejuno 4* are deemed to be of the same author as *De jejuno 5*). Given the scores, we might be tempted to confirm a single author for each of the PC6 and PC7 clusters, but we cannot confirm this without reasonable doubt. Our hypothesis, though untested, suggests that if PC6 and PC7 employ such an unconventional form of Ancient Greek compared to the majority of the corpus, they could prove challenging to categorize uniformly.

The case of Pseudo-Chrysostoms 13, 14, and 16 PC13, 14, and 16 are a particular case in Voicu's survey, as PC13 is refuted by Voicu and its texts are dispatched into two other sub-corpora: two texts (*Contra Iudaeos, Gentiles et Haereticos* and *In uenerabilem*) are thought to go with PC14's *De Eleemonsyna*, while *De Epiphania* would go with PC16. Our precision matrix threshold does confirm the unity of PC16 except for the inclusion of PC13's text (see Figure 11a). However, while the pairs connections are low for PC13 and PC14, they seem to be rather

[8] The situation is less clear for PC2. It seems that it is well detected by our Manhattan models, see 13.

(



(a) Pseudo-Chrysostom 6

(b) Pseudo-Chrysostom 7

Figure 9: Matrix of the precision thresholds for authors whose bad prose is considered a proof of authorship.

high when connected to PC16. Our model seems to recommend the following composition for PC16:

- *De Eleemosyna* (PC14)
- *In Psalmum Homilia 50 homiliae*, 1–2 (PC16)
- *In Illud: Sufficit Tibi Gratia Mea* (PC16)
- *In Illud: Si Qua Christo Nova Creaturqa* (PC16)
- (Lower probability) *Contra Iudaeos, Gentiles et Haereticos* (PC13*)
- (Lowest probability) *In Venerabilem Crucem* (PC13*)

Pseudo-Chrysostom 20 and 20b: What happened? The most significant disagreement between traditional scholarship and our model arises with PC20, which is classified as unconfirmed by our model (see Figure 10). Voicu mentions PC20 as the outcome of ”on-going research” and comprises 6 base texts (PC20) and 12 potential others (PC20b). Unfortunately, some of these texts are unedited, and we did not have access to them. As of now, none of the texts in PC20 and PC20b form prominent clusters.

We can see two pairs that might be interesting to focus on, specifically:

- 0.91 *In Drachmam ...–In Parabolam...*, with a .86 PT in the other direction. Both are within a high range of PT and comes from the same PC corpus;
- *In Illud: Ascendit* and *In Parabolam de Ficu*, with .60 and .74 scores.

However, the connection between *In Drachmam* and *In Illud: Ascendit* is not strong (.3).

Moreover, the precision thresholds for most pairs with high values are not reciprocated when considering the inverse direction:

- 0.82 *In Rachelem...–De Non Judicando*, with a 71-point drop in the other direction;

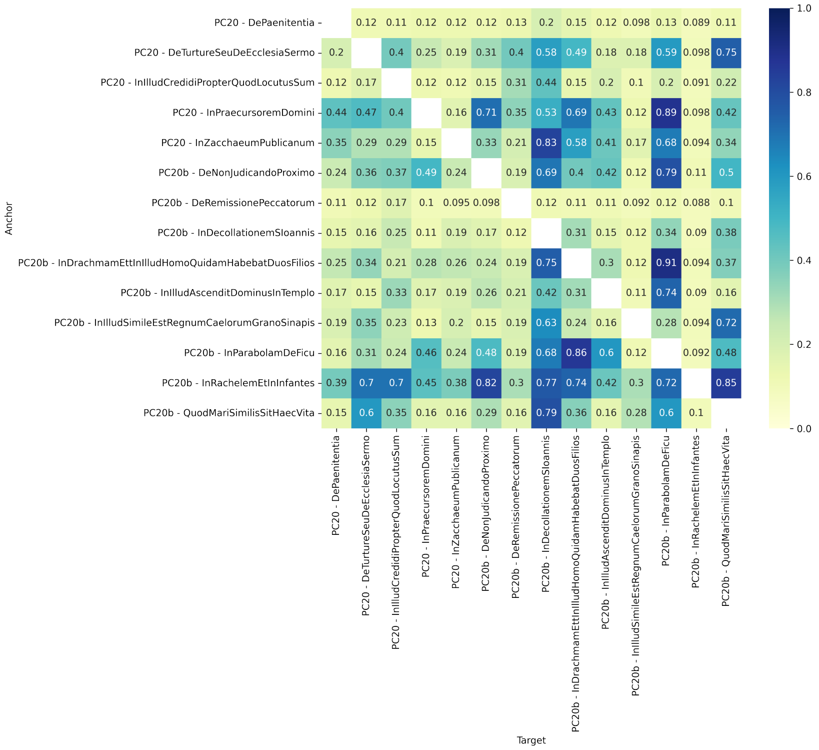


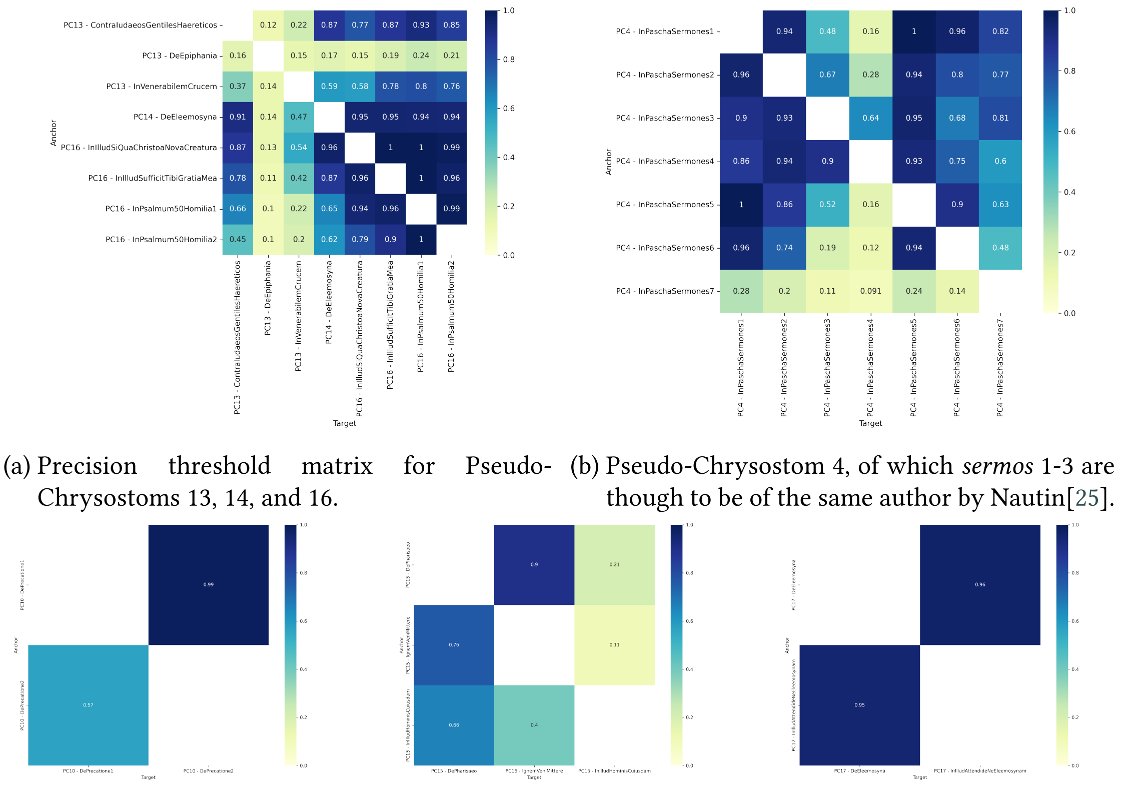
Figure 10: Matrix of the precision thresholds for the biggest unconfirmed cluster: PC20 and its extension PC20b.

- 0.85 *In Rachelem et in Infantes–Quod Mari Similis*, with a 75-point drop in the other direction.

The lack of clear clustering and the significant differences in precision threshold values between pairs in opposite directions raise questions about the coherence and authorship attributions within PC20 and PC20b. It is essential to recognize that authorship verification is a complex task, and discrepancies between our model's findings and traditional scholarship can be attributed to various factors, including the linguistic features used for analysis, the dataset's size and quality, and the nature of the texts themselves. We also want to stress that we have a low recall, which can lead to having many false negatives.

Others For other Pseudo-Chrysostom, we propose the following analysis:

- Pseudo-Chrysostom 10 (Figure 11c): This cluster falls into the group of Pseudo-Chrysostoms for which the A–B and B–A distances differ. As it consists of a single pair of texts, and that the confidence is strong in at least one direction, we would be inclined to confirm the assumption of Weyer[41] regarding their common authorship.
- Pseudo-Chrysostom 11 (Figure 11b): This cluster is a subset of Pseudo-Chrysostom 4 and comprises 7 sermons, of which the first three are believed to be of a single author



(a) Precision threshold matrix for Pseudo-Chrysostoms 13, 14, and 16.

(b) Pseudo-Chrysostom 4, of which *sermos* 1-3 are though to be of the same author by Nautin[25].

(c) Pseudo-Chrysostom 10

(d) Pseudo-Chrysostom 15

(e) Pseudo-Chrysostom 17

Figure 11: Matrix of the precision thresholds.

according to Nautin[25]. While the unity of the PC4 corpus is not confirmed through our model, our results are more aligned with the PC11 hypothesis, but we would not go as far as confirming it, as the cluster suffers from unmirrored pair distances.

- Pseudo-Chrysostom 15 (Figure 11d): In this cluster, we observe a strong reciprocal connection between *De Pharisaeo* and *Ignem Veni Mittere*, while *In Illud Hominis...* shows high variation depending on the pair directions.

7. Conclusion

Authorship verification is a crucial task in the field of computational humanities and the humanities in general, as it offers a new approach to validate older hypotheses made using traditional philological methods, such as transmission study or stylistic analysis. This is particularly important in patristic studies, where pseudo-author corpora, such as those attributed to Augustine or John Chrysostom, are significant. The challenge lies in dealing with pseudonymous corpora where most authors are unknown.

To address this, we proposed exploring the use of Siamese Networks with embeddings based

Table 5
Analysis summary. C stands for confirmed, HP high probability, LP low probability, R refuted. On the second line, U stands for cluster unconfirmed, which differs from Refutation as our model is not trained for recall. PC19 could bot been studied because of the Syriac content.

PC	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	20b	21
In Voicu, 1981	R	C	R	LP	LP	LP	LP	R	R	C	C	C	R	C	C	C	C	C	*	C	C	C
Our model	HP	HP	U	U	U	LP	HP	U	U	HP	LP	C	U	U	LP	C	HP	C	*	U	U	C

on known stylometric features, such as the relative frequencies of most frequent words, affixes, and POS 3-grams. To optimize our approach, we leveraged pair mining and signal-to-noise ratio distance, both originally designed for Siamese network architectures. The results we obtained on our development and test sets are very promising, and we argue that, unlike authorship attribution problems, authorship verification problems are less susceptible to overfitting when using such architectures.

Our findings mostly align with the survey conducted by Voicu in 1981, which identified 21 potential unique pseudo-authors in the corpus of the pseudo-Chrysostomian texts (see Table 5). Some pseudo-Chrysostoms are highly probable under both the framework of classical philological studies and our approach. However, we encountered discrepancies with two pseudo-Chrysostoms. First, PC1, hypothesized by Montfaucon and refuted by Altendorf [1], is mostly confirmed within our framework. Second, we were not able to confirm PC20. It's important to note that our method is designed for precision rather than recall, so this result does not necessarily refute Voicu's hypothesis but still stands out compared to the rest of our results.

In future work, we are interested in extending our approach without relying on a learned embedding space and instead using probabilistic tools for better explainability. The instability of the models we saw with Manhattan distances and the lower – yet existing – instability of STN based models argue for more interpretable and stable models. This kind of work has already been explored by Weerasinghe, Singh, and Greenstadt [40] and would be a valuable addition to the digital humanities landscape. Further analysis on PC20 should be produced, specifically by looking at the research produced by Voicu since 1981 on this particular topic.

References

[1] H. Altendorf. "Untersuchungen zu Severian von Gabala". PhD thesis. Tübingen, 1957.

[2] L. Berkowitz, K. A. Squitier, and M. Pantelia. *Thesaurus Linguae Graecae. Canon of Greek Authors and Works.* 2020.

[3] J. Bevendorff, B. Stein, M. Hagen, and M. Potthast. "Generalizing Unmasking for Short Texts". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 654–659. DOI: 10.18653/v1/N19-1068.

[4] J. Burrows. "'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship". In: *Literary and Linguistic Computing* 17.3 (2002), pp. 267–287. DOI: 10.1093/llc/17.3.267.

[5] F. Cafiero and J.-B. Camps. “Why Molière most likely did write his plays”. In: *Science advances* 5.11 (2019), eaax5489. DOI: 10.1126/sciadv.aax5489.

[6] J.-B. Camps, T. Clérice, and A. Pinche. “Noisy medieval data, from digitized manuscript to stylometric analysis: Evaluating Paul Meyer’s hagiographic hypothesis”. In: *Digital Scholarship in the Humanities* 36.Supplement_2 (2021), pp. ii49–ii71. DOI: 10.1093/llc/fqab033.

[7] C. Datema. “An unedited homily of Ps. Chrysostom on the birth of John The Baptist (BHG 843k)”. In: *Byzantion* 52 (1982), pp. 72–82. URL: http://www.jstor.org/stable/44170752.

[8] M. Eder. “Rolling stylometry”. In: *Digital Scholarship in the Humanities* 31.3 (2015), pp. 457–469. DOI: 10.1093/llc/fqv010.

[9] M. Eder. “Style-markers in authorship attribution: a cross-language study of the authorial fingerprint”. In: *Studies in Polish Linguistics* 6.1 (2011). URL: http://www.ejournals.eu/SPL/2011/SPL-vol-6-2011/art/1171/.

[10] M. Eder. “Taking stylometry to the limits: Benchmark study on 5,281 texts from Patrologia Latina”. In: *Digital humanities 2015: conference abstracts*. 2015, pp. 1919–1924.

[11] M. Eder, J. Rybicki, and M. Kestemont. “Stylometry with R: a package for computational text analysis”. In: *The R Journal* 8.1 (2016). DOI: 10.32614/rj-2016-007.

[12] W. Falcon and The PyTorch Lightning team. *PyTorch Lightning*. Version 1.4. 2019. DOI: 10.5281/zenodo.3828935. URL: https://github.com/Lightning-AI/lightning.

[13] J. L. Fleiss and J. Cohen. “The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of reliability”. In: *Educational and Psychological Measurement* 33.3 (1973), pp. 613–619. DOI: 10.1177/001316447303300309.

[14] R. Gorman. “Author identification of short texts using dependency treebanks without vocabulary”. In: *Digital Scholarship in the Humanities* 35.4 (2019), pp. 812–825. DOI: 10.1093/llc/fqz070.

[15] P. Heslin. *Diogenes*. 2023.

[16] M. Kestemont, E. Manjavacas, I. Markov, J. Bevendorff, M. Wiegmann, E. Stamatatos, M. Potthast, and B. Stein. “Overview of the cross-domain authorship verification task at PAN 2020”. In: *Working notes of CLEF 2020-Conference and Labs of the Evaluation Forum, 22-25 September, Thessaloniki, Greece*. Vol. 2696. 2020, pp. 1–14. URL: https://ceur-ws.org/Vol-2696/paper%5C%5F264.pdf.

[17] M. Kestemont, E. Manjavacas, I. Markov, J. Bevendorff, M. Wiegmann, E. Stamatatos, M. Potthast, and B. Stein. “Overview of the cross-domain authorship verification task at PAN 2021”. In: *Working notes of CLEF 2021-Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania*. Vol. 2936. 2021, pp. 1743–1759. URL: https://ceur-ws.org/Vol-2936/paper-147.pdf.

[18] M. Kestemont, J. Stover, M. Koppel, F. Karsdorp, and W. Daelemans. “Authenticating the writings of Julius Caesar”. In: *Expert Systems with Applications* 63 (2016), pp. 86–96. DOI: 10.1016/j.eswa.2016.06.029.

[19] M. Kestemont, M. Tschuggnall, E. Stamatatos, W. Daelemans, G. Specht, B. Stein, and M. Potthast. “Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection”. In: *Working notes of CLEF 2018-Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.* Vol. 2125. 2020, pp. 1–14. URL: https://ceur-ws.org/Vol-2125/invited%5C%5Fpaper%5C%5F2.pdf.

[20] M. Koppel and Y. Winter. “Determining if two documents are written by the same author”. In: *Journal of the Association for Information Science and Technology* 65.1 (2014), pp. 178–187. DOI: 10.1002/asi.22954.

[21] B. Marx. “Procliana: Untersuchung über den homiletischen Nachlass des Patriarchen Proklos von Konstantinopel”. In: *Münsterische Beiträge zur Theologie* 23 (1940).

[22] L. McInnes, J. Healy, and J. Melville. “UMAP: Uniform Manifold Approximation and Projection”. In: *Journal of Open Source Software* 3.29 (2018), p. 861. DOI: 0.21105/joss.00861.

[23] K. Musgrave, S. J. Belongie, and S.-N. Lim. “PyTorch Metric Learning”. In: *ArXiv* abs/2008.09164 (2020). DOI: 10.48550/arXiv.2008.09164.

[24] B. Nagy. “Metre as a stylometric feature in Latin hexameter poetry”. In: *Digital Scholarship in the Humanities* 36.4 (2021), pp. 999–1012. DOI: 10.1093/llc/fqaa043.

[25] P. Nautin. *Homélies pascales: II. Trois homélies dans la tradition d'Origene.* Vol. 36. Sources chrétiennes, 1953.

[26] Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. *TorchMetrics - Measuring Reproducibility in PyTorch.* 2022. DOI: 10.21105/joss.04101. URL: https://github.com/Lightning-AI/torchmetrics.

[27] S. Nikolov, D. Tabakova, S. Savov, Y. Kiprov, and P. Nakov. “SU PAN'2015: Experiments in Author Verification.” In: *CLEF (Working Notes).* 2015. URL: https://ceur-ws.org/Vol-1391/151-CR.pdf.

[28] M. L. Pacheco, K. Fernandes, and A. Porco. “Random Forest with Increased Generalization: A Universal Background Approach for Authorship Verification.” In: *CLEF (Working Notes).* 2015. URL: https://ceur-ws.org/Vol-1391/87-CR.pdf.

[29] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32.* Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[30] J.-L. Quantin. “Du Chrysostome latin au Chrysostome grec: une histoire europèenne (1588-1613)”. In: *Chrysostomosbilder in 1600 Jahren* (2008), pp. 267–346. DOI: 10.1515/9783110207309.3.267.

[31] A. Baillot, T. Tasovac, W. Scholger, and G. Vogeler, eds. *Short texts with fewer authors. Revisiting the boundaries of stylometry.* Zenodo, 2023, pp. 191–193. DOI: 10.5281/zenodo.7961822.

[32] S. Rebora, J. B. Herrmann, G. Lauer, and M. Salgaro. "Robert Musil, a war journal, and stylometry: Tackling the issue of short texts in authorship attribution". In: *Digital Scholarship in the Humanities* 34.3 (2018), pp. 582–605. DOI: 10.1093/llc/fqy055.

[33] M. Schatkin. "The authenticity of St. John Chrysostom's de Sancto Babyla, Contra Iulianum et gentiles". In: *Festschrift Johannes Quasten.* Ed. by P. Granfield and J. A. Jungmann. Vol. 1. Kyriakon. Aschendorff, 1970, pp. 474–489.

[34] F. Schroff, D. Kalenichenko, and J. Philbin. "FaceNet: A unified embedding for face recognition and clustering". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2015, pp. 815–823. DOI: 10.1109/cvpr.2015.7298682.

[35] P. Singh, G. Rutten, and E. Lefever. "A pilot study for BERT language modelling and morphological analysis for ancient and medieval Greek". In: *5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, co-located with EMNLP 2021.* Association for Computational Linguistics. 2021, pp. 128–137. DOI: 10.18653/v1/2021.latechclfl-1.15.

[36] E. Stamatatos, M. Kestemont, K. Kredens, P. Pezik, A. Heini, J. Bevendorff, B. Stein, and M. Potthast. "Overview of the authorship verification task at PAN 2022". In: *CEUR workshop proceedings.* Vol. 3180. 2022, pp. 2301–2313. URL: https://ceur-ws.org/Vol-3180/paper-184.pdf.

[37] A. von Stockhausen. "Die Modellierung kritischer Editionen im digitalen Zeitalter". In: *Zeitschrift für Antikes Christentum* 24.1 (2020), pp. 123–160. DOI: 10.1515/zac-2020-0019.

[38] J. Tyo, B. Dhingra, and Z. C. Lipton. *On the State of the Art in Authorship Attribution and Authorship Verification.* 2022. DOI: 10.48550/arXiv.2209.06869. arXiv: 2209.06869 [cs.CL].

[39] S. J. Voicu. "Une nomenclature pour les anonymes du corpus pseudo-chrysostomien". In: *Byzantion* 51.1 (1981), pp. 297–305. URL: http://www.jstor.org/stable/44170685.

[40] J. Weerasinghe, R. Singh, and R. Greenstadt. "Feature Vector Difference based Authorship Verification for Open-World Settings". In: *CLEF (Working Notes).* 2021, pp. 2201–2207. URL: https://ceur-ws.org/Vol-2936/paper-197.pdf.

[41] J. Weyer. "De homiliis quae Joanni Chrysostomo falso attribuuntur". PhD thesis. Bonn, 1952.

[42] H. Xuan, A. Stylianou, and R. Pless. "Improved embeddings with easy positive triplet mining". In: *2020 IEEE Winter Conference on Applications of Computer Vision (WACV).* 2020, pp. 2474–2482. DOI: 10.1109/wacv45572.2020.9093432.

[43] T. Yuan, W. Deng, J. Tang, Y. Tang, and B. Chen. "Signal-To-Noise Ratio: A Robust Distance Metric for Deep Metric Learning". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2019, pp. 4810–4819. DOI: 10.1109/cvpr.2019.00495.

A. Online Resources

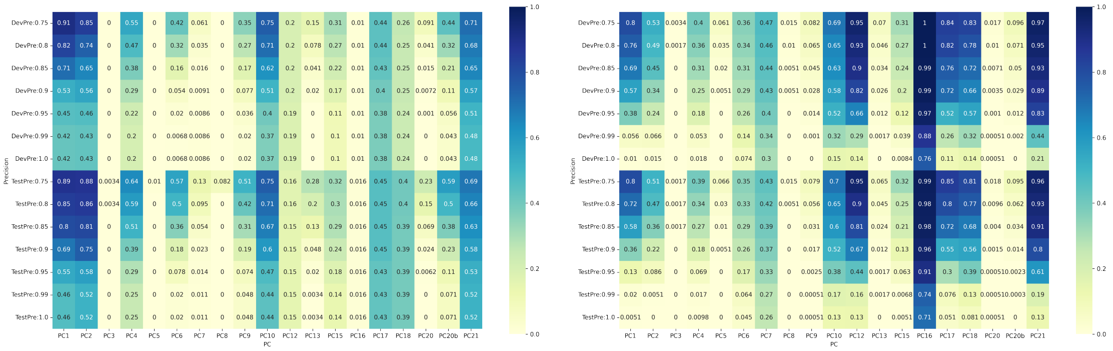
Code and data for the Voicu experiment are available at https://github.com/PonteIneptique/Chryso-Voicu.

B. Parameters search results

Features and parameters search and their respective AUROC (in %). Each configuration is run three times (seed 42, 128, 256), average and standard deviation are provided. Tested parameters: MFP (0, 100), MFW (0, 250, 500, 1000), MFT (0, 250, 500, 1000), Distance (Manhattan, L2 and STNR). Scores are ranked according to their test AUROC. Only the top 10 are shown, the remaining scores are on the online github repository.

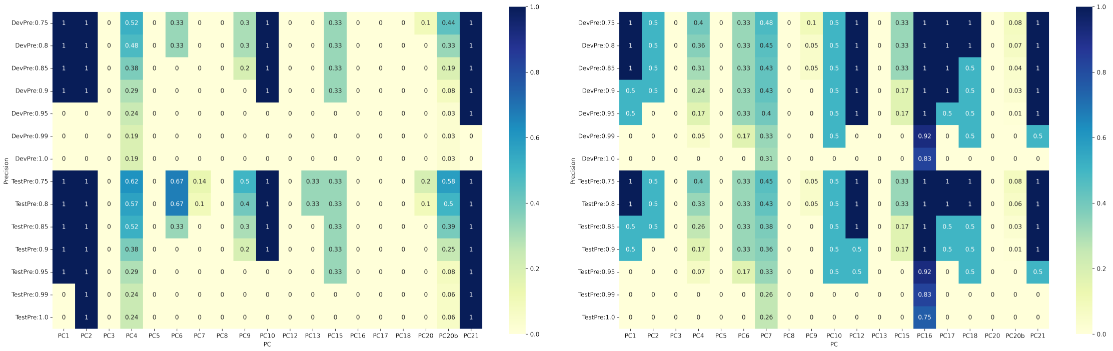
Parameters				Dev		Test	
POS	FW	Affixes	Distance	mean	std	mean	std
100	1000	1000	Manhattan	88.04	0.58	86.01	0.53
200	1000	750	Manhattan	87.63	0.83	85.88	2.03
100	750	1000	Manhattan	86.75	0.80	85.12	1.67
100	1000	750	STN	85.77	1.91	84.51	0.93
100	1000	1000	STN	87.03	0.65	84.42	1.31
100	1000	750	Manhattan	87.19	1.25	84.33	4.10
100	750	1000	STN	85.38	0.33	84.23	0.63
200	1000	750	STN	85.81	0.15	84.15	0.81
100	1000	500	Manhattan	86.85	1.44	83.68	4.48
200	750	750	Manhattan	86.64	0.51	83.63	1.08

C. Variation of prediction on PC corpora using Manhattan or STN



(a) Using Manhattan (b) Using Signal-to-Noise Ratio distance

Figure 12: Mean of the pairing percentage in clusters depending on the PT.



(a) Using Manhattan (b) Using Signal-to-Noise Ratio distance

Figure 13: Median of the pairing percentage in clusters depending on the PT.

D. Samples for the corpora

Number of samples per author depending on the train, dev or test set.

	Train	Dev	Test
Ephraem Syrus	172	18	26
Gregorius Nyssenus	145	30	21
Gregorius Nazianzenus	130	13	24
Basilius	105	15	15
Theodoretus	89	3	10
Athanasius	66	12	9
Eusebius	66	2	5
John of Damascus	56	5	5
Origenes	55	4	0
Cyrillus	46	9	15
Maximus Confessor	26	0	5
Clemens Alexandrinus	24	0	0
Hippolytus	21	5	0
Flavius Justinianus Imperator	20	1	6
Severianus	18	2	0
Evagrius	17	4	5
Procopius	17	0	1
Leontius	16	1	0
Epiphanius	15	3	0
Amphilochius	15	2	0
Hesychius	14	0	1
Irenaeus	14	0	0
John of Caesarea	14	5	0
Nicolaus I Mysticus	10	0	0
Didymus the Blind	10	0	4
Eustathius	10	0	0
Gregorius Thaumaturgus	9	0	0
Theodorus Studites	7	0	0
Marcellus	5	0	0
Marcus Diaconus	5	0	0
Nicetas Choniates,	5	0	0
Dio Chrysostomus	5	0	0
Asterius	5	0	1
Antonius Hagiographus	5	0	0
Salaminius Hermias Sozomenus	5	0	0
Clemens Romanus Clementina	5	0	0
Adamantius	5	0	0
Nicephorus I	5	0	0
Nemesius	5	0	0
Palladius	5	0	0
Barlaam,	2	0	0
Ephraem	1	0	0
Alexandros, monachos	1	5	5
Peter, Patriarch of Alexandria	0	2	0