Improved distance measures for 'fixed-content miscellanies': an adaptation for the collections of sayings of the desert fathers and mothers[1]

Elisabet Göransson
Centre for Languages and Literature and Centre for Theology and Religious Studies, Lund University, Sweden

Luke Maurits
Max-Planck-Institute for Evolutionary Anthropology Department of Developmental and Comparative Psychology, Germany

Britt Dahlman
Centre for Theology and Religious Studies, Lund University, Sweden

Karine Åkerman Sarkisian
Department of Modern Languages, Uppsala University, Sweden

Samuel Rubenson
Centre for Theology and Religious Studies, Lund University, Sweden

Michael Dunn
Department of Linguistics and Philology, Uppsala University, Sweden

Abstract

Collections of sayings of the desert fathers and mothers are extant in manuscripts in many languages and are organized differently. They are 'fixed-content miscellanies' (FCM): they include material that belongs to the same genre, but is variable both when it comes to appearance and order. Distance measurement methods are particularly suitable for large text traditions including variable content in the so-called mixed-content miscellanies, such as recipes, anthological compilations of shorter text passages, or catalogues, but can also be suitable for text genres like collections of sayings, that are equally variable in appearance and order of sayings, even though the genre is fixed; hence 'fixed-content miscellanies'. In the article, collections of sayings in seven languages were compared using four distance measures methods. Each segment of the sayings was given a unique id to be comparable. The first method used, the Jaccard distance measure, disregards the linear order of items

Correspondence:
Elisabet Göransson, CTR, LUX, Box 192, 221 00 Lund, Sweden.
E-mail:
elisabet.goransson@ctr.lu.se

Digital Scholarship in the Humanities
https://doi.org/10.1093/llc/fqac025

and instead considers each collection compared only as a 'bag of stories'. In two other methods used (Birnbaum and Levenshtein methods), the order in which the narratives of each saying appear is compared. All three methods yielded interesting results, but the collections that were apparently closely related were clustered together so tightly that it was not possible to make more nuanced analyses. In order to remove false negatives, particulars concerning lacunes in the material were taken into account in the proposed modified Levenshtein method, the fixed-content miscellanies (FCM)-Levenshtein method. By applying the FCM-Levenshtein method, previously unknown relations between collections witnessed in different languages could be detected.

1 Introduction

Textual scholarship within Classical philology has traditionally focused on texts which may be considered 'literary' or 'canonized', i.e. which are fixed and stable, are written by a known author, can be directly associated to one specific context, and which have been reproduced in manuscripts with the ambition to render the text exactly as it stands without additions, deletions, or transpositions of parts of the texts. However, many ancient and medieval texts having a long and complicated reception history do not fit this mold. For instance, monastic works, such as those treated in this study, were sometimes produced in several versions, have been subject to multiple revisions, and are preserved in a variety of redactions. They could be extensively adapted, recombined, translated, or otherwise changed over time in order to best fulfil an intended purpose or to fit new cultural, social, or educational settings. Traditional goals of textual scholarship do not apply to such works, since there is no reason to assume a single archetype ever existed.

Since these monastic texts have been subject to frequent and repeated translation, adaptation, compilation, and general 'remixing', they are ideal for studying the various factors and processes at play in the evolution of texts over time. Especially valuable for this purpose are the so-called mixed-content miscellanies, defined as 'manuscript books that consist of an arbitrary set of texts (articles) selected and arranged without the application of any particular organizational principle' (Birnbaum, 2003). In this article, we will focus on such manuscripts when considering the development of methods for studying the evolution of texts. The subject of this case study is collections of the *Apophthegmata Patrum* (AP), which consist of short sayings and anecdotes mostly attributed to the desert fathers and mothers of the early Egyptian monastic communities. These texts are, however, not specifically mixed-content but rather fixed-content miscellanies (FCM): they belong to a genre that is more or less fixed in its content but with variable appearance and order of the sayings. (For a discussion on the difference between mixed-content and fixed-content miscellanies, see Birnbaum (2003), referring to Miltenova (1986a, b, 1987, 2001).)

Cultural evolution is increasingly studied via the use of quantitative and computational methods, often taking inspiration from tools originally developed for the study of biological evolution. The sequential nature of mixed-content miscellanies (or fixed-content miscellanies) makes them especially amenable to this approach, since biological evolution is also concerned with sequential arrangements of values from a fixed 'alphabet' (e.g. DNA bases, amino acids). However, despite this structural similarity between biological and cultural datasets, the underlying evolutionary processes can be expected to vary substantially. Therefore, equally substantial adaptation of existing tools must be expected to make them appropriate for textual datasets and to maximize their utility there. Here we contribute to this development by considering the adaptation of a fundamental class of quantitative methods used in bioinformatics, namely distance methods, to the study of fixed-content miscellanies. The methods discussed in this article are equally applicable to mixed-content miscellanies, which combine, e.g. collections of recipes, catalogues, anecdotes, or anthological compilations of shorter text passages with texts from other genres.

2 The Subject of the Study: *Apophthegmata Patrum*

The monastic sayings normally consist of a statement by a desert father or mother and/or a brief narrative, sometimes introduced with a question from a monk. Shorter and in some cases longer tales are also found in the collections of sayings. The sayings are sometimes 'wrapped up' with a conclusive moral statement, or with a contrast that focuses on what to avoid, as in this example (the Pelagius and John collection, chapter II.11):

> Abba Nilus said: 'The arrows of the enemy cannot touch him who loves quiet. But he who moves in a crowd will be often wounded.'

The origin of the sayings has been debated. Previously, it was generally assumed that they originated as only orally transmitted sayings around the first monastic communities in Egypt, possibly already in the late 4th and/or during the 5th century, that they early on were associated with certain fathers, and then were written down, probably in Palestine, to collect the memories of the first fathers after the devastation of Scetis. This view has been questioned by recent research that has highlighted the didactic functions of these texts in accordance with ancient rhetorical education (Larsen, 2008, 2013; Rapp, 2010; Rubenson, 2013). In addition, it has been shown (Rubenson, 1995, pp. 145–62; Faraggiana di Sarzana, 1997) that some of the earliest collections partially consist of sayings that were extracted from earlier written sources and other literary works, such as narratives and lives of monastic figures (*Vita Arsenii*, *Epistulae Antonii*, the works of Cassian, etc.). Therefore, the collections were only partially orally transmitted, if at all, in the first stages. Scholars have distinguished two main types of organizations; the sayings can be arranged alphabetically according to the names of the desert fathers, or systematically according to a range of themes concerning Christian virtues and vices. However, collections also attest a variety of other types of organization, including mixtures of the two types mentioned, as well as collections without any kind of organization. Another question, which has been the focus of scholarly debate, is the origin of and relationship between the earliest collections. It is generally assumed that a Greek alphabetically organized collection was first created, and from it a collection organized according to certain themes was formed (Chitty, 1974; Faraggiana di Sarzana, 1997). It has, however, also been suggested that other unorganized collections must have preceded or been developed independently from the known Greek ones (Faraggiana di Sarzana, 1997; Dahlman, 2013).

Even though scholars in general agree that the sayings were first written down in Greek, the extant manuscripts in Greek are not particularly old. There are older manuscripts both in Syriac and in Latin. The sayings were probably first written down during the second half of the 5th century, and already from the early 6th century we have around ten manuscripts preserved in Syriac, and another five from before 1000 AD (Holmberg, 2013). In Latin, the earliest manuscripts are dated to around 650 AD. On the other hand, the oldest preserved Greek manuscripts date from the 9th century (except for a few early fragments). There are early text witnesses also in Coptic. Apart from these languages, the collections were also translated early on into Christian Palestinian Aramaic, Sogdian, Armenian, Georgian, Arabic (and further, to Ethiopic via Arabic), and, after 870 AD, into Old Church Slavonic (Veder, 2012, p. 31). From the 12th and 13th centuries onwards, collections of sayings were translated into vernacular languages all over Europe, including Old Norse. It is thus a vast textual tradition represented in many manuscripts and languages. In the largest text traditions in Greek, Latin, and Slavonic,[2] there are many hundreds of manuscripts containing collections of AP, since after entering a new cultural milieu they became sources to new compilations.

3 The Purpose of the Study

The purpose of this study is to present and discuss methods that can be used to analyze collections of sayings in order to identify similarities and dissimilarities both when it comes to which sayings they contain and the order of the sayings in collections. By exploring this, relationships between text traditions also across the languages in this very rich and multifaceted material could be revealed that have not been known previously.

The manuscripts can contain compilations of different types of collections; one manuscript can contain several collections. The compilations of the collections into manuscripts are mostly of a later date, and mark the ambitions of scribes from the 9th century onwards: they assembled the type of material associated with one particular genre in these compilatory manuscripts; this happened not only in the (former) Roman Empire, but also all over the Christian world. Thus, monastic compilations may contain collections of sayings embracing a large number of sayings deriving from many different sources. In this study, the sequences and contents of the individual collections, rather than the manuscripts they are part of, will be in focus. Even with the more common collections that are systematically or alphabetically organized, there are always some differences in the set of sayings in the specific collection in one manuscript witness. As a matter of fact, the sequence of sayings in a certain collection, that is their occurrence and the order in which they appear in one manuscript, is seldom totally identical to the order in another manuscript.

This study confines itself to the following linguistic traditions represented in the database Apophthegmata Patrum Database (APDB)/*Monastica*: Arabic, Coptic, Greek, Latin, Old Norse, Slavonic, and Syriac. In the database, each small narrative, designated as a text segment, has been given a unique ID, making comparisons possible. In this study, the collections within the manuscripts are analyzed individually; thus, different parts of the same manuscript are treated separately. We include the datasets of the sets of sayings and their order contained in the different parts, which in this case are the relevant collections that the individual manuscript contains. Only systematically organized collections are analyzed in this study, along with a few that have a 'mixed' type of organization, disregarding the collections that are clearly alphabetically organized with one exception (more on this below). Manuscripts containing systematic collections have been selected, since this type is present in the majority of these languages. Besides, for an extended comparison, the mixed types of collections found in Syriac, Arabic, and Old Norse text witnesses are included. In addition, a Greek manuscript, Vat_gr_2592, containing an old alphabetic-anonymous collection is used as a point of reference (for a table of the collections used in the study, see Appendix A). This collection is known to be an early witness, that is, it is thought to represent an early stage in the complex textual transmission of collections (Faraggiana di Sarzana, 1997). It contains two parts: the first one, 'A', is an alphabetically organized collection, and the second one, 'B', contains sayings that are 'anonymous', that is, normally not attributed to a certain monk.

4 Distance Measures

The range of quantitative and particularly statistical tools that can ultimately be brought to bear on the problem of inferring how different processes have shaped the evolution of fixed-content miscellanies such as the monastic sayings is vast. Developing and refining these methods constitute a substantial and long-running research program. Here we focus exclusively on the problem of developing suitable distance measures, as a kind of initial 'beachhead' from which more sophisticated methods may be developed. All the methods discussed in this article are implemented in the accompanying *seqsim* python library (Tresoldi *et al.*, 2021).

Having a precisely defined notion of 'distance' between items in a dataset—in other words some measure of how much meaningful difference separates two items—enables the use of a broad suite of quantitative methods for both visualization and analysis, which can collectively be termed distance methods. This toolkit facilitates such things as: visualization, e.g. multi-dimensional scaling (e.g. Cox and Cox, 2008) can produce 2D and 3D plots displaying the 'shape' of a dataset in an intuitively accessible way, while NeighbourNets (Bryant and Moulton, 2004) can represent distance relationships between items in the dataset in a way that preserves possible 'conflicting' signal; clustering, e.g. algorithms such as k-medoids (Kaufman and Rousseeuw, 1987) or hierarchical agglomerative clustering (e.g. Zhao and Karypis, 2005) can be used to sort the items in a dataset into nonoverlapping groups in a way, which minimizes the distance between items in the same group while simultaneously maximizing the distance between items in different groups (see Birnbaum, 2016 for an application of hierarchical agglomerative clustering to stemmatological data); and the construction of phylogenies, e.g. NeighborJoining (Saitou and Nei, 1987) can be used to build ‘family trees’ representing hypotheses on how items in a dataset are related to one another through a process of descent with modification from an unobserved common ancestor.

In a broader evolutionary analytic context, distance methods are often distinguished from the so-called model-based methods, which incorporate explicit models of the historical processes of change which act on evolving entities, rather than simply considering the (dis)similarity of the final results of those processes. Generally speaking, model-based methods are better able to exploit all of the information which may be in a dataset, while distance-based methods tend to be considerably less computationally demanding. This makes distance methods especially well suited to exploratory analysis early in a project. More advanced model-based approaches can then be used to address specific research questions, for example taking advantage of their ability to reconstruct, or test, hypotheses about past states of observed entities.

For a sample of witnesses of collections in selected manuscripts, denoted M, a distance measure is a systematic means of assigning to any pair of manuscript witnesses, which we would call m_1 and $m_2 \in M$, a non-negative number denoted $d(m_1, m_2)$ (mathematically, d is a *function*, d: $M \times M \rightarrow \mathbb{R}^+$), such that the value of this number—the distance between the two witnesses—captures some important notion of what the two have in common, or what they do not. For any given kind of dataset, many different distance measures may be defined. Cross-examining such analyses can prove to be fruitful. There is no ‘best’ or ‘one true’ sense of distance between witnesses of a certain text. Rather, different distance measures are best suited to different research problems. This is analogous to how the distance between cities, even if their geographic locations are fixed, must be considered differently depending upon whether we are trying to estimate the travel time of either an aircraft that can fly directly between two points, a train that is constrained to follow fixed tracks along a relatively flat route, or a piece of mail whose passage is also influenced by non-physical traits like customs agreements between countries.

While researchers have considerable freedom in specifying a distance measure for a dataset, there are certain ‘common sense’ properties that are generally desirable for distance measures to satisfy. Some of these properties may be a strict requirement for the applicability of certain distance methods. For example, the distance between any text witness and itself should be usually equal to zero, $d(m, m) = 0$. For many distance methods, it will be required that the distance measure used is *symmetric*, i.e. $d(m_1, m_2) = d(m_2, m_1)$, but symmetry is not a necessary property of distance metrics in general. The difference between symmetric and asymmetric distance measures can be illustrated by considering two approaches to measuring the distance between a town on the top of a mountain and one at the bottom of the mountain. The distance in kilometers is symmetric—it makes no difference whether one measures from the top to the bottom or vice versa. But the effort required by a cyclist to move between the towns is asymmetric: rolling downhill requires considerably less effort than climbing uphill.

A very closely related concept to distance is similarity, which is essentially the ‘opposite’ idea: a mapping from pairs of datapoints to numeric values such that low values indicate substantial differences between the points, unlike low distance values. Often, similarity measures can be transformed into equivalent distance measures, and vice versa. This is possible if a similarity measure s has some maximum possible value $s_{\max}$, in which case it can be transformed into a distance measure d via $d(m_1, m_2) = s_{\max} - s(m_1, m_2)$. The most straightforward example of such a transformation would be a similarity measure which simply counts the number of features for which two datapoints have equal values. This can be transformed into a distance measure which counts the number of features for which two points differ.

Here we consider distance measures which are suitable for quantifying the differences and similarities between collections consisting of an ordered sequence of items (in this case, the smallest comparable ‘story’, defined as the segments of the sayings with unique IDs) drawn from a finite set of options, such as the monastic sayings. Whereas traditional textual scholarship normally considers primarily different readings witnessed in a text tradition, in this type of text genre, the actual sets of text segments, their appearance, and order of appearance constitute the first and foremost form of text variation on the ‘macro level’. We wish to study the structures of these sets of narratives—narratives that are not in themselves, as texts, so varying as are their appearance in the first place. Both their appearance and their order can be analyzed by using different distance measurements.

5 Measuring Distance between 'Bags of Stories'

A very simple distance measure which is applicable in this context, and which can serve as a useful base case for comparison, is the Jaccard distance. This measure disregards the linear order of items and instead considers each collection purely as a 'bag of stories'—each collection is a set of text segments, which either contains or does not contain any given item, with no notion of order or any other internal structure. The Jaccard distance between two such manuscripts is $1 - I/U$, where I is the number of items that are present (at any position) in *both* manuscripts, and U is the number of items that are present (again, at any position) in *either* manuscript.[3] When two manuscripts contain precisely the same items, regardless of order, $I = U$ and the distance is 0, while if the two manuscripts have no items in common then $I = 0$ and the distance is 1.

Put in the context of the collections of sayings, this method can be used to see what is not so clear to the eye when viewing the long lists of sayings organized in different ways in the different types of collections. Many sayings are common for different 'types' of organization, e.g. alphabetically or thematically arranged—exactly how common, and how they cluster, can be easily recognized by using the Jaccard distance measuring: for example, hitherto unknown relations between systematically, alphabetically organized collections of sayings, and also relations to collections that seem to be arranged according to a mixture of types, can be identified. Both distance in textual variation and the distance measured in the appearance of the sayings, that is, the 'bag of stories' that the individual collections of sayings contain, could be a way of even more firmly establishing the links between such parts in this huge text tradition, also across language barriers, that include collections organized in various ways.

Below, a comparison of different collections in the various languages by using the Jaccard distance is given as an example.[4] The results visible in the graph—which uses nonmetric multidimensional scaling to arrange the witnesses in a 2D display such that witnesses with a low Jaccard distance appear closer together than those with a high Jaccard distance—are quite clear in that we see a main cluster that contains most of the selected collections in the manuscripts in the different languages. The visualization also reveals that some of the collections, which previously have been tentatively defined as 'mixed' collections, indeed represent selections that seem to be a mixture that has a concentration of the 'bags of stories' present in the other types of collections: they appear in the center of the graph along with the other collections.[5] Furthermore, concerning the alphabetically organized collection of sayings in Vat_gr_2592, that is, Part A, we get a confirmation that the text segments contained in this collection, even though they are arranged alphabetically, are more or less the exact same ones as those in the thematically organized collections that are in the center of the cluster, since this collection is to be found very near the most central part of the main cluster. The second collection in this manuscript, Part B, however, is peripheral in this context, which means that this collection do not contain the same set of text segments than do most of the other collections. This is also the case with some other collections. All in all, twelve of the selected collections place themselves in the periphery. Not only are they far from the main cluster; they are also not particularly close to one another, with one exception, StPeterb_BAN_Belokr_2_F and Beog_NBS_Dec_93_B: the latter two obviously contain more or less identical 'bags' of stories. The four examples of collections in Arabic witnesses in the Jaccard graph have been spread in an interesting way indicating that there are more than small differences in between what text segments are included in these collections. The witnesses Strasb_4225_A and Mil_Ambr_L120sup_A seem to contain more or less exactly the same texts as do the collections in different languages that are contained in the main cluster. The text witness Par_ar_276_B is clearly not very similar to them.

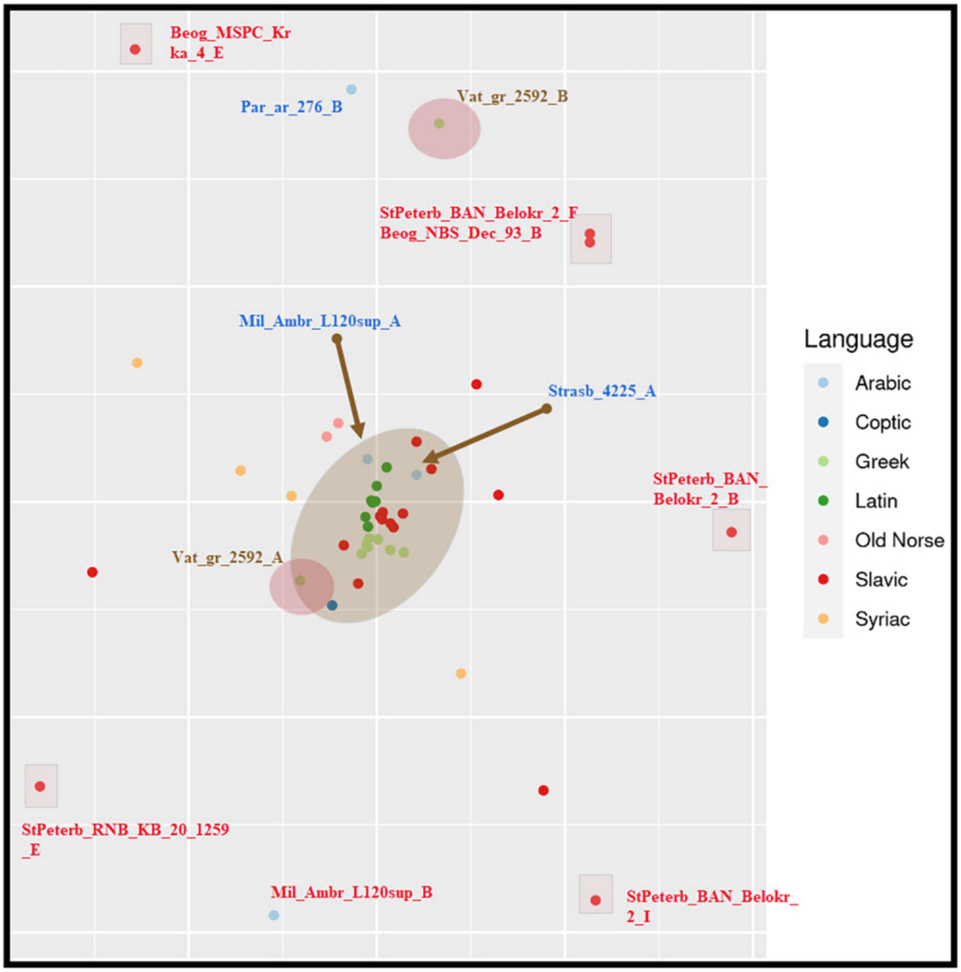

Fig. 1A MDS visualization of selected collections in manuscripts using Jaccard distance[6]

6 Measuring Distance between Sequences of Narratives

While using the Jaccard distance to analyze the collections of sayings as unordered 'bags of stories' has provided some insight, it is also clear that the collections are in fact sequentially organized and that discarding information about the order in which stories appear removes historical evidence of shared textual transmission from the data. We therefore now turn our attention to distance measures, which take sequential structure into account.

Such a measure designed specifically for analyzing relationships between manuscripts containing 'mixed-content miscellanies' was proposed by David J. Birnbaum (2003). More precisely, Birnbaum proposed a similarity measure, where the similarity between two manuscripts is equal to the sum of the lengths of all the common subsequences between them, such that increasing the number of common subsequences, or increasing the length of any individual common subsequence, both result in increased similarity. Since the maximum number of subsequences two sequences can share is a maximum value for this similarity measure, it can be converted to a distance, which lies between 0 and 1 by subtracting the similarity from its maximum value and dividing the result by this maximum value. The division step is necessary because the maximum value of the similarity measure is dependent on the length of the shortest manuscript. Therefore, without normalizing the distance measures between different pairs of manuscripts, they are not directly comparable.

This measure was designed in accordance with the following assumptions, quoted from Birnbaum's article regarding detecting common transmission in mixed-content miscellanies (here 'matches' refers to subsequences; Birnbaum, 2003, p. 24):

(1) Long matches are more highly-valued than sets of short matches, e.g. a six-article correspondence constitutes much stronger evidence of shared transmission than two three-article correspondences (see below for clarification).
(2) Matching articles must be adjacent and in the same relative sequence in both manuscripts.
(3) Absolute position in the manuscripts is irrelevant for identifying or weighting relationships.
(4) The total number of articles in the manuscripts is irrelevant for identifying or weighting relationships.

To clarify the first requirement, suppose we have three manuscripts. Manuscript 1 contains text segments labeled A, B, C, D, E, F, G, H, and I, manuscript 2 contains sequences A, B, C, D, E, F, X, Y, Z, and manuscript 3 contains segments A, B, C, U, V, W, G, H, I. Manuscripts 1 and 2 share a subsequence of length 6 (A, B, C, D, E, F, occurring at the start of each). Manuscripts 1 and 3 share two subsequences each of length 3 (A, B, C at the start of each and G, H, I at the end). Birnbaum's assertion is that manuscripts 1 and 2 should be considered much more likely to be related to one another than manuscripts 1 and 3, even though both pairs have six segments in common (i.e. all pairs are equidistant under the Jaccard distance). This assumption is driven by a concern with chance resemblance. Short common subsequences can be expected to appear simply by chance in sufficiently long sequences assembled independently at random, with the expected frequency of such subsequences decreasing rapidly with their length. In order to avoid 'false positives', i.e. unrelated manuscripts receiving a high similarity value due to chance similarities, it does indeed make sense to place greater emphasis on individual long common subsequences.

However, multiple (relatively) short subsequences do not necessarily indicate resemblance due entirely to chance. Would-be long subsequences between manuscripts with shared transmission can ultimately result in multiple shorter subsequences as a result of either physical damage to the manuscript or accidental omission during copying, either of which could remove a single folio page worth of stories (in this case the smallest identifiable parts of the sayings, that is, the text segments with unique IDs) somewhere in a long common sequence. The loss of just a single story in the middle of a long common subsequence results in a potentially very large increase in Birnbaum distance between two manuscripts, despite being a comparatively minor and common event. Thus, we see that the measure has a high risk of 'false negatives', i.e. closely related manuscripts receiving a similarity value, which is lower than less closely related manuscripts due to the exaggerated influence of minor, not uncommon damage.

Multiple short common subsequences due to chance and those due to the disruption of long subsequences via historic processes are not indistinguishable. If what would have been a common subsequence of length 10 becomes one of length 5 and one of length 4 due to the loss of a single story, the two shorter subsequences will be immediately adjacent to one another. In contrast, if two common subsequences of lengths 5 and 4 occur due to random chance, they are overwhelmingly more likely to be non-adjacent. Thus, we see that Birnbaum's third assumption, that absolute position within a manuscript is not relevant in weighing the relative importance of different subsequences, is not consistent with the realities of the relevant transmission process, at least for the present subject of study.

As mentioned before, according to this method, a six-article correspondence between two manuscripts constitutes much stronger evidence of a shared transmission than two three-article correspondences between the same text witnesses. That would mean that the sequence of text segments which we could name A, B, C, D, E, F, and G in two collections would indicate a closer relation than the sequence of only five of them, A, B, C, E, F, and G in one of the collections and the full sequence of six in the other. This probably

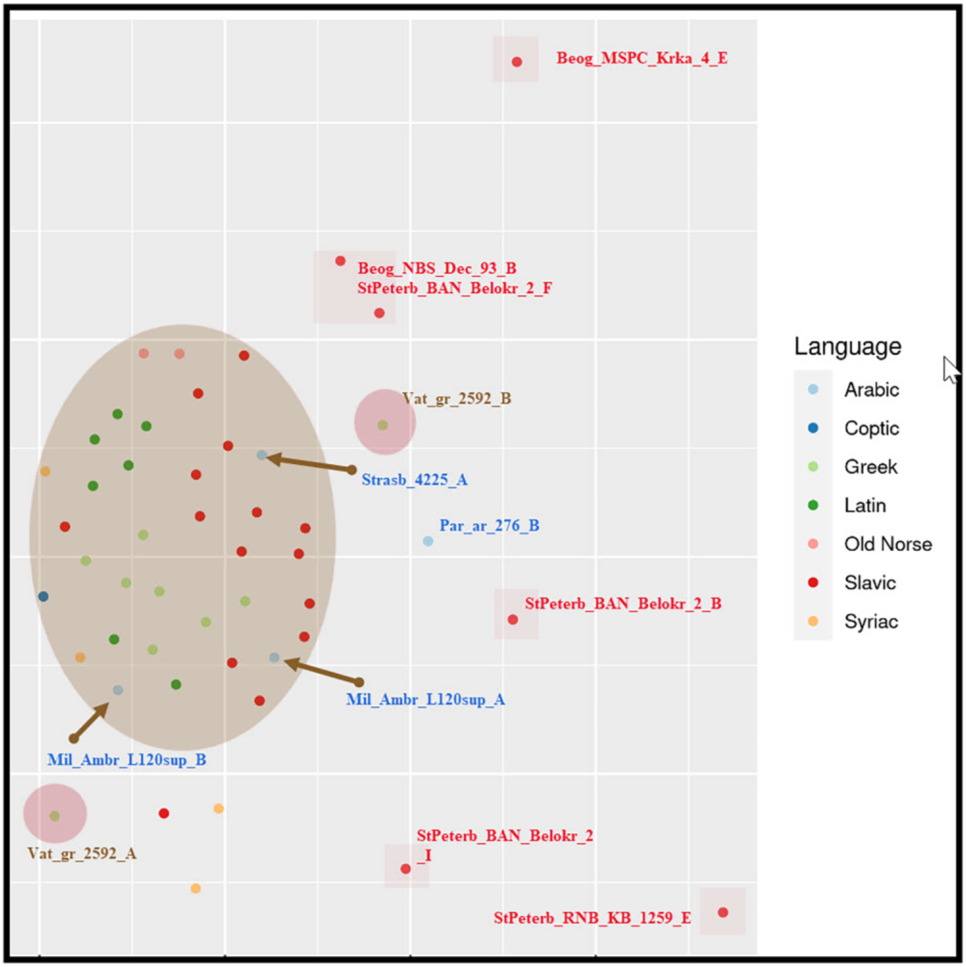

Fig. 2A MDS visualization of selected collections in manuscripts using Birnbaum distance

serves other text genres better; in the case of the sayings of the desert fathers and mothers, however, minor differences in the sequences might just be a result of a mistake, an omission in the copying process, resulting in a missing saying in a 'normal' sequence. Therefore, this type of difference is of minor importance.

When comparing the outcome of the same dataset plotted with the Birnbaum algorithm, we can see that the graph shows a less centered cluster compared with the Jaccard measurement graph. Now, the actual order of the sayings matter, which it did not when using the Jaccard 'bag of stories' method. Even so, some of the manuscripts containing collections that were rather far away from the main cluster in the Jaccard graph are still more peripheral. The four selected Arabic manuscript collections, however, distribute differently: three out of four are contained in the main cluster, which means that their sequence of the specific sayings is similar to the 'core' of the collections. The collection Mil_Ambr_L120sup_B that was more peripheral in the previous graph, when only the same contents mattered, apparently has a structure that is quite similar to many of the other collections in the main cluster, despite the fact that it contains less shared sayings (which was seen in the Jaccard graph). The Greek and Latin manuscripts appear a bit differently compared with the Jaccard graph, however. Both collections (Parts A and B) of the Greek manuscript Vat_gr_2592 are now outside of the main cluster, even though they are not very peripheral. This makes sense since this manuscript contains collections that do not have the same type of organization when it comes to the order of the sayings, even though, as we have seen in the Jaccard graph, the alphabetical collection (Part A) in fact shares a great number of sayings with the core of the selected material for comparison.

Within the textual traditions of several of the selected collections in Greek and Latin, we know that there are smaller but still notable differences. Scholars have tentatively suggested that manuscripts belong to different groups of families or to stages in the textual transmission, primarily based on how similar the contents of the collections are and on how similar the sequences of sayings seem to be. This has, however, not been done systematically. Recently, further studies on textual variation *per se* have started up, that is, studies on the text variants in the sayings themselves. According to preliminary results from these studies (see Dahlman *et al.*, forthcoming), some of the groups that have been suggested previously by scholars can be confirmed, at least to some extent, both concerning the Greek and the Latin material. In the Birnbaum graph, the Greek and Latin manuscripts are rather evenly distributed, even though two Latin manuscripts, Mun_SB_Clm_18 093 and Par_lat_5387_A, are placed on the other side of the center of the cluster, in which the Greek and Latin manuscripts otherwise appear evenly distributed and rather close to one another. The distribution of the Greek and Latin manuscripts does not give any clarification of the relations between the groups of manuscripts that have been identified. Thus, it seems as if this type of model is not optimal to use for this type of material, even though it certainly paints a rough picture illustrating the main differences between the sequences of sayings in the present collections.

7 Levenshtein Distance and Generalized Edit Distances

Levenshtein distance is a distance measure for sequential information which is widely used in bioinformatics (where it is used to analyze DNA, RNA, and protein sequence data) and computer science (where it is used for example in spell-checking software to find dictionary words close to unrecognized words). In our context, given two sequences of stories, the Levenshtein distance between them is the minimum number of insertions, deletions, or substitutions of individual stories required to transform one sequence into the other. The Levenshtein distance measure is perhaps the best known representative of a broader family of distance measures, called edit distance measures. In general, these measures define the distance between two ordered sequences by counting the minimum number of required editing operations to transform one sequence into the other. Different edit distances are defined by different sets of available operations. For example, while the ordinary Levenshtein distance permitted the insertion, deletion, or substitution of individual items, the more complicated Damerau–Levenshtein distance allows these same operations but also the additional operation of transposing two adjacent items (i.e. swapping their order). This distance measure is especially applicable in spell-checking contexts, where swapping the order of two letters is a common mistake when typing rapidly. Under ordinary Levenshtein distance, such an error would contribute two edits to the distance, making it 'as bad' an error as typing the wrong letter twice within a word. In addition to allowing different editing operations, different edit distances can also assign weights to operations, such that individual instances of one kind of operation make a lesser or greater contribution to the distance than individual instances of another operation.

The transposition operation included in the Damerau–Levenshtein distance may seem at first blush to be a valuable addition in a stemmatological context, with swapping the order of two consecutive stories seeming a plausible copying error for a scribe to make. However, analysis of the AP data suggests that this is in fact a very rare occurrence. Nevertheless, this serves to illustrate an essential point: the generalized edit distance framework permits scholars to define a set of edit operations and corresponding weights that are a good match to their study domain. By their very nature, edit distance measures encourage explicit consideration of the processes of change, which operate on study subjects; in the case of stemmatology, of the cultural evolutionary processes which shape manuscript traditions as they are translated and otherwise adapted to new environments and new applications.

Because Levenshtein distance relies upon explicit processes of change that can be applied to sequences, it avoids the shortcomings of Birnbaum's distance measure with regard to historical processes acting on manuscripts. For example, physical damage to a manuscript resulting in the loss of a small number of sayings will not contribute much to the Levenshtein distance between the damaged manuscript and an undamaged original from which it was copied—this corresponds to only a single edit operation. At the same time, multiple short common subsequences distributed throughout the manuscripts due to chance resemblance are not likely to unduly reduce the distance between unrelated manuscripts, as the large number of operations required to explain the intervening non-common sequences will dominate the overall distance.

When looking at Fig. 3A, presenting the dataset according to the Levenshtein plot, similar to the Jaccard plot, the dataset is more centered in one main cluster, even though it is shaped differently. The same collections that are peripheral in the other graphs are still peripheral, and the ones that are rather

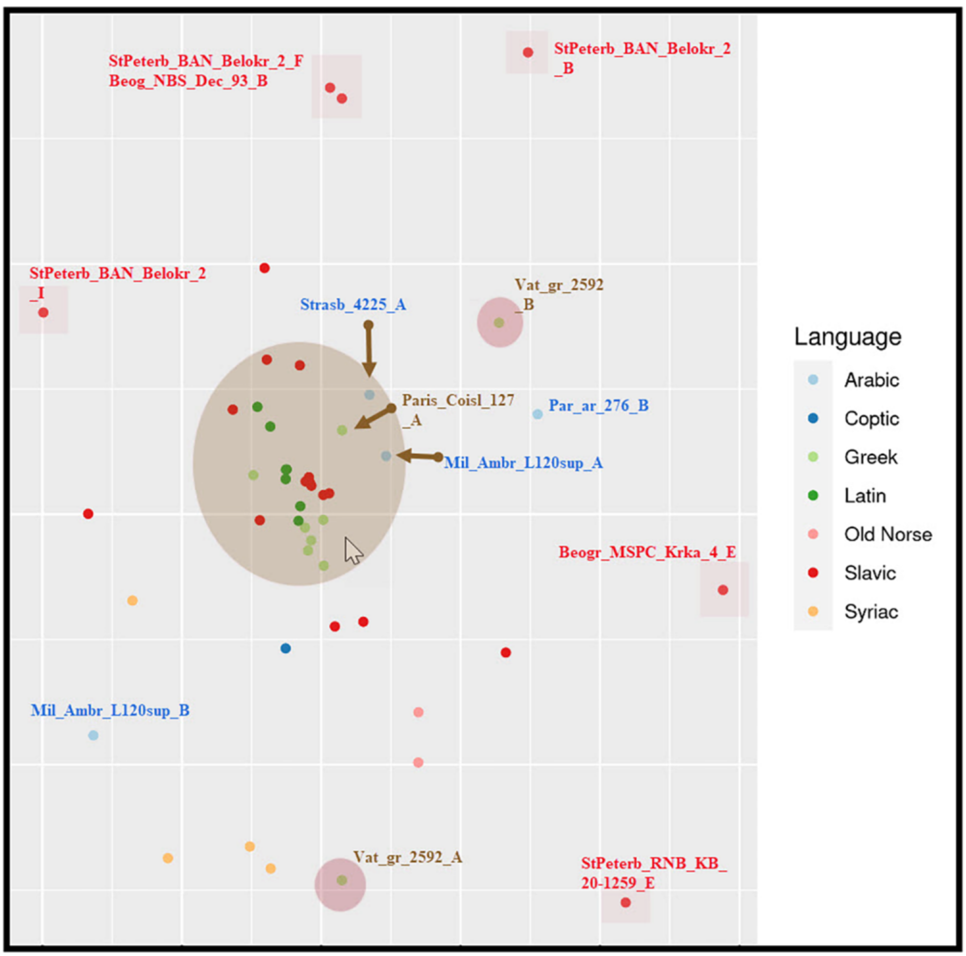

Fig. 3A MDS visualization of selected collections in manuscripts using Levenshtein distance close to one another in the periphery are the same, with a few notable exceptions. The difference with the Jaccard plot can be seen when looking specifically at the Greek and Latin witnesses mentioned before that we already know are quite close both in their contents and in the sequence of the sayings. They are now placed more or less on a line in the middle of the graph. However, they are so close to one another that it is still not possible to see any further differences in between them, except for a few collections. First of all, again, the Greek manuscript Vat_gr_2592 stands out, both concerning the alphabetical collection (Part A) and the anonymous collection (Part B). They are not contained in the main cluster now either. Moreover, the Greek collection Par_Coisl_127_A is further away from the other Greek and Latin collections, but on the other hand, it is close in the sequence of its sayings with some collections in Slavonic and Arabic.

This method accordingly gives a good deal of interesting information that makes it possible to investigate the relations further. Even so, there is room for improvement. In the remainder of this section, we develop one proposal for a stemmatology-specific edit distance measure, with special attention on the nature of the manuscripts containing the *Apophthegmata Patrum* collections.

8 A Proposal for a Customized Distance Solution: The FCM Levenshtein Distance Method

In the manuscripts containing the collections of sayings an omission of a set of sayings that would fit a folio page is fairly common. Such an omission does not reflect a 'true' difference either, since it may simply be the result of the loss of a folio page in the manuscript or in its model manuscripts. However, in the next manuscript generation, this omission of a full sequence could have been incorporated in the copying process: in this way, such omissions can in fact signal relations between manuscript witnesses, even entire families of witnesses (Göransson, 2019). The absolute location of matching content within each manuscript is certainly an important factor to consider when analyzing the relations, but the effect of the lacunes should therefore also be considered in the process. Lacunes, that is, unintentional or accidental omissions, can appear anywhere in manuscripts: in the beginning, middle, or at the end of a manuscript. Sometimes the lacune is large; at other times, it only consists of a folio or two in a manuscript that has been torn out at a later stage. Since the sets of sayings in some collections, in particular the systematically organized ones, is fairly stable, is it often possible to state exactly the sayings that have disappeared. In these instances, they can be added to the data sets, so as to eliminate that type of data disturbance (Göransson, 2019). However, the approach has to be carefully considered, since the omissions caused by lacunes sometimes could have had an impact on the text transmission as well, as mentioned above; we have seen this happen sometimes. Therefore, whenever different methods are tested, it should be remembered that manuscripts that have not been affected by lacunes, or parts of manuscripts that have not been affected, could be compared on the one hand, and the full datasets, including the possible data disturbance, on the other, so as to be able to compare the results and see if it makes a difference or not.[7] In this pilot study, however, a more refined analysis distinguishing between the types of lacunes has not been made.

Here we propose a novel edit distance measure designed specifically for use with datasets like the Apophthegmata Patrum data and with particular attention to the lacunes. Our measure is a variant of the Levenshtein edit distance, in that insertion and substitution of individual text segments in a manuscript are permitted and count as a single operation (we used Levenshtein and not Damerau–Levenshtein distance as a starting point because an investigation of our data suggested that the transposition of two adjacent segments is a very rare occurrence). However, the particulars of the deletion operation have been considerably altered. We refer to this measure as the Fixed Content Miscellanies Levenshtein distance, or FCM-Levenshtein distance.

The omission of an entire leaf's worth of sayings during the copying process, or the loss of a whole leaf due to physical damage to the manuscript, are not uncommon events and as such they should not make a heavy contribution to the distance between two manuscripts. However, the Levenshtein distance models such changes as a large number of independent deletions of individual sayings, and as such they do contribute considerable distance. Therefore, the first change is that we permit the deletion of up to ten consecutive text segments in the collection, that is, segments of sayings, in a manuscript as a single operation. This number has been chosen since the length of the text segments that constitutes the sayings is rather stable in the collections of sayings, regardless of language. Even if some sayings are very long and others only contain one row, normally the sayings are similar in size. Given the large number of text segments in the material studied, an estimation of the average length of a text segment in the database is a help when calculating the number of text segments that would have been lost in case a full folio page was omitted. The number ten has been calculated based on a sample study of average count of words on a folio page, that is, one leaf with front and back side (see further below). An estimation of the average number of text segments in one folio page extant in this manuscript material was made of a smaller part of the material in a few Latin, Greek, and Arabic manuscripts. From this investigation, we concluded that ten consecutive segments missing is a normal number for a missing folio page. This allows the distance measure to better reflect the physical embodiment of the manuscripts. Furthermore, based on the pilot investigation undertaken, we can conclude that the text density in manuscripts written in different languages is not very different; therefore, it is possible to make this type of rough estimation.

In the following, a more detailed explanation is given of how the average number of text segments based on an average word length on a folio page was made. The assumption that the number of words on a folio page is not so different regardless of language and date of the manuscript was based on the fact that the number of lines in a manuscript is normally fixed because of the fact that the lines of the parchment quires were first ruled before writing started. This means that all the folio pages in the same quire had the same ruling. Since the markings of ruling were made quire by quire and the quires were prepared in the same way, a typical medieval manuscript has a fixed set of lines on each manuscript page; this only varies with a couple of lines between the quires. This fact makes it safer to assume that an average of number of words on a folio page is similar in the entire manuscript. Moreover, even if the number of lines in medieval manuscripts vary to some extent, the average number of words on a folio page can be checked in manuscripts in different languages, written during different centuries. If the average of words on a folio page is similar, it is then possible to estimate the amount of text loss when a folio page has been omitted for different reasons, regardless of language or date of the manuscript.

For the purpose of this study, a comparison of words per folio page was made as an average of a number of words in ten full folio pages in selected manuscripts in four languages relevant in the present study.[8] If a folio page in the selected sequence includes extensive blank space it was not counted. We selected manuscripts that have been transcribed and the text inserted into the database, which facilitates the word count. Three manuscripts each in Slavonic, Latin, and Greek, and two in Arabic (only two have been transcribed so far) were included.[9] The manuscripts are dated from the 7th century up to the 16th century. The average word count is given in Appendix C. The results reveal that there are no visible differences in words per folio page, or text density, depending on the language, nor depending on the date of the manuscript. Instead, there are individual differences between manuscripts in the same language. A couple of manuscripts have lower average of words per folio (henceforth wpf) page compared with the rest of the chosen manuscripts (one in Greek: 255 wpf, another in Latin: 225 wpf). The remaining manuscripts have an average of 300–400 wpf. The overall average wpf including all the manuscripts is 351; in the individual languages 332 in Greek, 342 in Latin, 396 in Slavonic, 325 in Arabic.

The average word count per folio page is thus the basis for a calculation of average number of text segments on a folio page. The calculation of average number of text segments was based on the average number of words in each text segment calculated from the extensive chapter IV of the so-called Pelagius and John collection, that is, a systematic collection in the Latin manuscript Brux_BR_9850-52, with a general word count of 385 per folio page. This chapter contains eighty-eight text segments in this manuscript. The division of total number of words in all the folio pages included in the word count (nineteen full folio pages) by eighty-eight text segments gave the number 39.4 words per average text segment. Three hundred and eighty-five words per average folio page in this manuscript divided by 39.4 words per text segment gives 9.8 text segments per folio page as an average. Hence, the average number of text segments of a folio page was concluded to be ten.

Before making more nuanced analyses using this method for other material, the same procedure as described above is needed for the material selected for analysis. Calculations should be made to fix the average number of text segments based on the average number of words per folio page that would be a representative number in a missing folio page in the specific case.

The second change implemented with the FCM-Levenshtein distance method is that we assign a lower weight to deletion operations if the entire sequence of consecutive sayings being deleted lies entirely within the first 10% of the manuscript sequence or the final 10% of the manuscript sequence. Again, this brings the distance measure into closer agreement with physical reality for this particular material. As can be seen in the online platform *Monastica*, presenting the structures of the manuscripts, sometimes there are rather large lacunes in the beginning of the manuscripts; also towards the end of the manuscripts this is often the case. Damage or loss of folio pages or entire quires at the beginning and end of the manuscript is more probable than to several folios in the middle of the manuscript, which are shielded by the outermost pages. We have estimated that an average would count for 10%. Thus, we make 'outer 10% deletions' carry half the weight of 'inner 80% deletions'. While these modifications were inferred considering the AP material specifically, we believe that the same method could be applied to many similar genres, since the codicological history of many manuscripts is no different from the ones discussed in this article. However, the character of the manuscripts should first carefully be evaluated; the percentage of the sequences that should be assigned lower weight according to this kind of lacunes is probably different in different types of source material.

Interestingly enough, the collections in the graph produced through this customized Levenshtein method (Fig. 4A) do not agglomerate in one, main cluster. Instead, the distribution of the different collections compared is more diffuse. This might indicate that the previous clusters are more based on accidental features. Even though the collections that are most peripheral in the graphs (Figs 1A, 2A, and 3A) are still distributed at the more peripheral places, thanks to the modifications inferred to the method, we can now see distinctions, e.g. in between the Greek and Latin collections more clearly. Three Greek collections that share quite a few textual variants, Athens_500_A, Athos_Prot_86_A, and Par_gr_914_A, are close to one another in this graph. However, also Par_gr_2474_A, is close to the latter. The results displayed in the graph are similar to the recent conclusions drawn from studies of textual variation, identifying the Greek Athens_500_A, Athos_Prot_86_A, and Par_gr_914_A in one group of witnesses (see Dahlman *et al.*, forthcoming). However, the Latin witnesses, which according to another forthcoming study based on textual variation seem to belong to at least two main branches (Göransson, forthcoming), are distributed less clearly in the graph. Two witnesses that belong to the second main group, according to this study, are marked with arrows; the other Latin witnesses in the graph all belong to the first main group (for the labels, see Fig. 4B in Appendix B).

Furthermore, a Greek collection, Par_Coisl_282, seems to be close to a group of Slavonic collections: Mosc_Gim_Cudov_318A_C, StPeterb_BAN_Belokr_2_D, Mosc_GIM_Sin_3_C, and StPeterb_RNB_Pog_267_A; the relationship between the two latter manuscripts has been pointed out by Åkerman Sarkisian (2020). Another Greek collection, Par_Coisl_127_A,

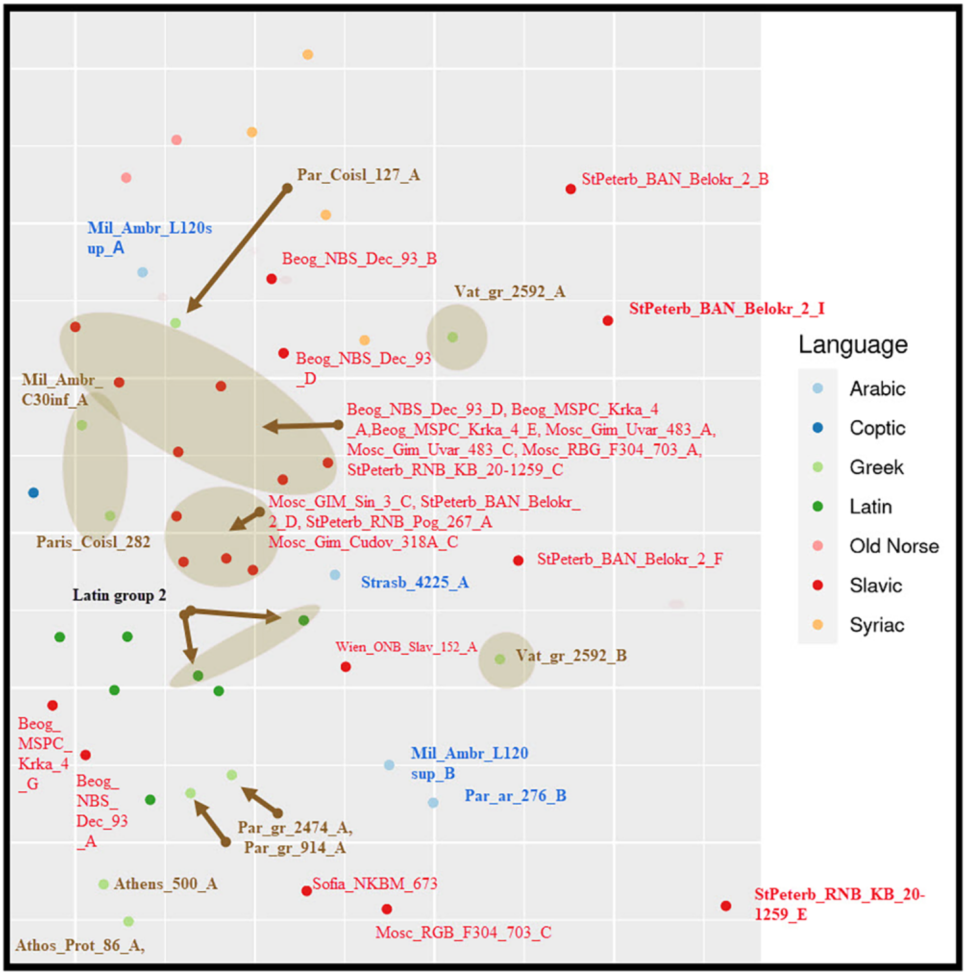

Fig. 4A. Visualization of selected collections in manuscripts using FCM-Levenshtein method

is close to the Arabic Mil_Ambr_L120sup_A. Interestingly, it could also be noted that several collections witnessed as parts of Slavonic manuscripts group around the same Greek collection. These Slavonic collections are Beog_NBS_Dec_93_B and D, Beog_MSPC_Krka_4_A, Mosc_Gim_Uvar_483_A and C. Another Greek collection close to them is Mil_Ambr_C30inf_A.

It thus seems, as if the last of the four models we have discussed here is the one that gives us the most nuanced picture of the relations between the collections compared. Through this graph, we can see relations that have not been identified hitherto, and it will be possible to explore the relations further based on this result. At the same time, however, the results produced in all three previous models are also relevant.

9 Conclusion

The text traditions in manuscripts containing ‘mixed-content miscellanies’, and in this case ‘fixed-content miscellanies’, are complex in that they are heterogeneous and fluid. In this article, different methods for exploring the text traditions based on quantitative data have been explored. Instead of focusing only on the texts themselves, scholars can take aspects concerning the selection of material and the organization of it into account when it comes to these fixed-content miscellanies, which constitute a substantial part of the legacy we have in medieval manuscripts, regardless of language. Analyses based on quantitative methods, and in particular on distance measurements of the sequences of the texts in collections and anthologies, can reveal relations that cannot be detected by manual analysis. Indications that are found can then be further investigated, contextualized, and checked against qualitative and quantitative analysis of the textual variation in between the identified correlating collections in manuscripts and groups of manuscripts.

In this study, the same material containing sets of data in the individual collections in the selected manuscripts representing the collections of sayings in Greek, Latin, Slavonic, Arabic, Coptic, Syriac, and Old Norse languages have been compared by using four different distance measures. The results tell us that it is indeed rewarding to use different methods to compare data, not the least in order to see which type of method is best suited to promoting research on a huge and quite complex set of material such as the collections of the sayings of the desert fathers and mothers represent. Since the methods first investigated, the Jaccard, Birnbaum, and Levenshtein distance measuring methods, did not seem to present graphs that were nuanced enough to get a richer picture of the relations between the collections, a modification of the Levenshtein method has been proposed in the present article. By analyzing the results from the four methods taken together, relations between the thematically organized and the 'mixed' collections have been made clearer than before. The clusters that are visible in the graphs are rather consistent in the four different methods; they distribute differently, but the most peripheral ones remain the same when analyzed using the four different methods, thus confirming that they are all relevant. Groups of collections witnessed in the parts of the manuscripts that are close to one another in all the three methods measuring sequences of sayings have been identified: they also include collections in different languages.

Furthermore, the results help in corroborating studies of the actual texts in the sayings, and the textual variation in them across the collections witnessed in manuscripts, which is an important factor to consider when studying the development of these text traditions. Through the analyses, relations between collections that have previously been suggested based only on text variants could be confirmed. We also saw how the Greek collection Vat_gr_2592_A, which is believed to represent an early type of alphabetic collection, indeed relates to the collections organized systematically, thus both confirming its importance and giving a more nuanced image of the relations. The distribution of the Arabic collections in relation to the collections in other languages also revealed relations that have not previously been known.

We believe that the methods presented in this article concerning collections of 'fixed-content miscellanies' can be useful for studies of texts similar to the present one, that is, text collections that can be characterized as 'mixed-content miscellanies' as defined by David J. Birnbaum. The modification of the Levenshtein method proposed here, the Fixed-Content-Miscellanies Levenshtein method (or 'FCM-Levenshtein method'), incorporates expert, real-world knowledge of the domain of study in a flexible way, which can also be adapted to other mixed-content miscellanies-type systems. All code is published in the accompanying python package *seqsim* (Tresoldi *et al.*, 2021).

Funding

The research presented in this article is part of the 'Cultural evolution of texts' project (MXM19-1087:1) headed by Michael Dunn at Uppsala University, Sweden. It is supported by funding from Stiftelsen Riksbankens Jubileumsfond.

Appendix A

Table of collections used in the study

Manuscriptid_part	Language	Type	Source
Athens_500_A	Greek	S	Manuscript
Athos_Prot_86_A	Greek	S	Manuscript
Beog_MSPC_Krka_4_A	Slavonic	S	Manuscript
Beog_MSPC_Krka_4_E	Slavonic	S	Manuscript
Beog_MSPC_Krka_4_G	Slavonic	S	Manuscript
Beog_NBS_Dec_93_A	Slavonic	S	Manuscript
Beog_NBS_Dec_93_B	Slavonic	S	Manuscript
Beog_NBS_Dec_93_D	Slavonic	S	Manuscript
Brux_BR_8216-18_C	Latin	S	Manuscript
Brux_BR_9850-52_A	Latin	S	Manuscript
Cologn_DB_165_A	Latin	S	Manuscript
Dayr_alAbyad_MONB-EG	Coptic	S	Manuscript
HML-Klemming	Old Norse	M	Edition
HMS-Unger	Old Norse	M	Edition
Lond_Add_12173_E	Syriac	M	Manuscript
Lond_Add_14626_B	Syriac	M	Manuscript
Lond_Add_17176_B	Syriac	M	Manuscript
Mil_Ambr_C30inf_A	Greek	S	Manuscript
Mil_Ambr_L120sup_A	Arabic	M	Manuscript
Mil_Ambr_L120sup_B	Arabic	M	Manuscript
Mosc_GIM_Cudov_318A_C	Slavonic	S	Manuscript
Mosc_GIM_Sin_3_C	Slavonic	S	Manuscript
Mosc_GIM_Uvar_483_A	Slavonic	S	Manuscript
Mosc_GIM_Uvar_483_C	Slavonic	S	Manuscript
Mosc_RGB_F304_703_A	Slavonic	S	Manuscript
Mosc_RGB_F304_703_C	Slavonic	S	Manuscript
Mun_SB_Clm_18093	Latin	S	Manuscript
Par_ar_276_B	Arabic	M	Manuscript
Par_Coisl_127_A	Greek	S	Manuscript
Par_Coisl_282	Greek	S	Manuscript
Par_gr_2474_A	Greek	S	Manuscript
Par_gr_914_A	Greek	S	Manuscript
Par_lat_13756_A	Latin	S	Manuscript
Par_lat_5387_A	Latin	S	Manuscript
Sin_syr_46	Syriac	M	Manuscript
Sofia_NBKM_673	Slavonic	S	Manuscript
StPeterb_BAN_Belokr_2_B	Slavonic	S	Manuscript
StPeterb_BAN_Belokr_2_D	Slavonic	S	Manuscript
StPeterb_BAN_Belokr_2_F	Slavonic	S	Manuscript
StPeterb_BAN_Belokr_2_I	Slavonic	S	Manuscript
StPeterb_RNB_KB_20-1259_C	Slavonic	S	Manuscript
StPeterb_RNB_KB_20-1259_E	Slavonic	S	Manuscript
StPeterb_RNB_Pog_267_A	Slavonic	S	Manuscript
Strasb_4225_A	Arabic	M	Manuscript
Vat_gr_2592_A	Greek	A	Manuscript
Vat_gr_2592_B	Greek	N	Manuscript
Vat_lat_600_F	Latin	S	Manuscript
Wien_ONB_Slav_152_A	Slavonic	S	Manuscript

Type: A = Alphabetical, M = Mixed, N = Anonymous, S = Systematical.

Appendix B

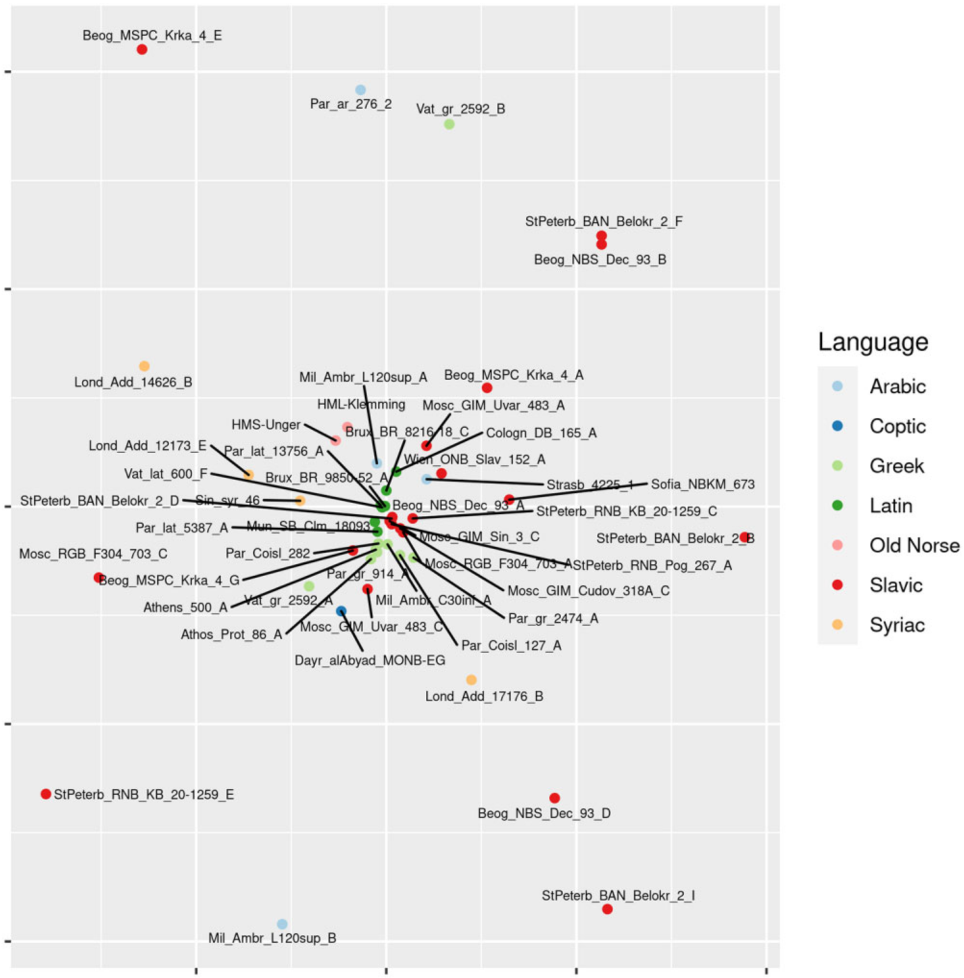

Fig. 1B Jaccard labeled

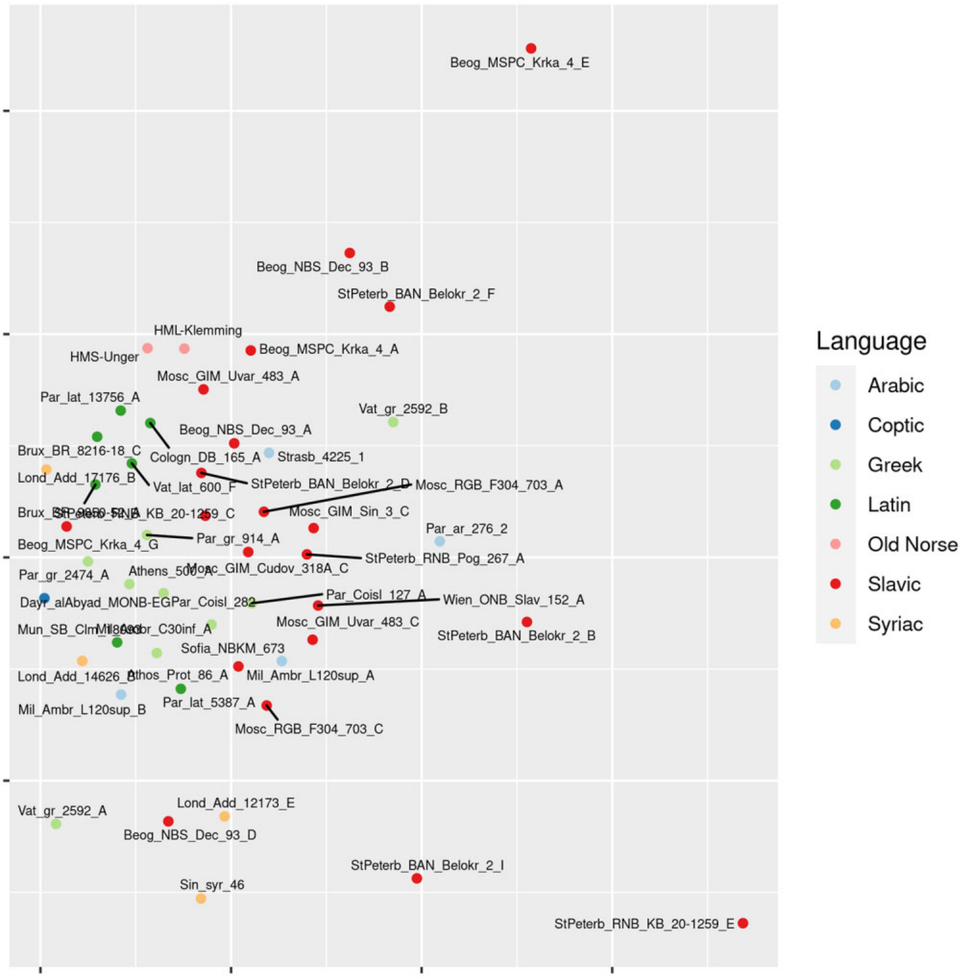

Fig. 2B Birnbaum labeled

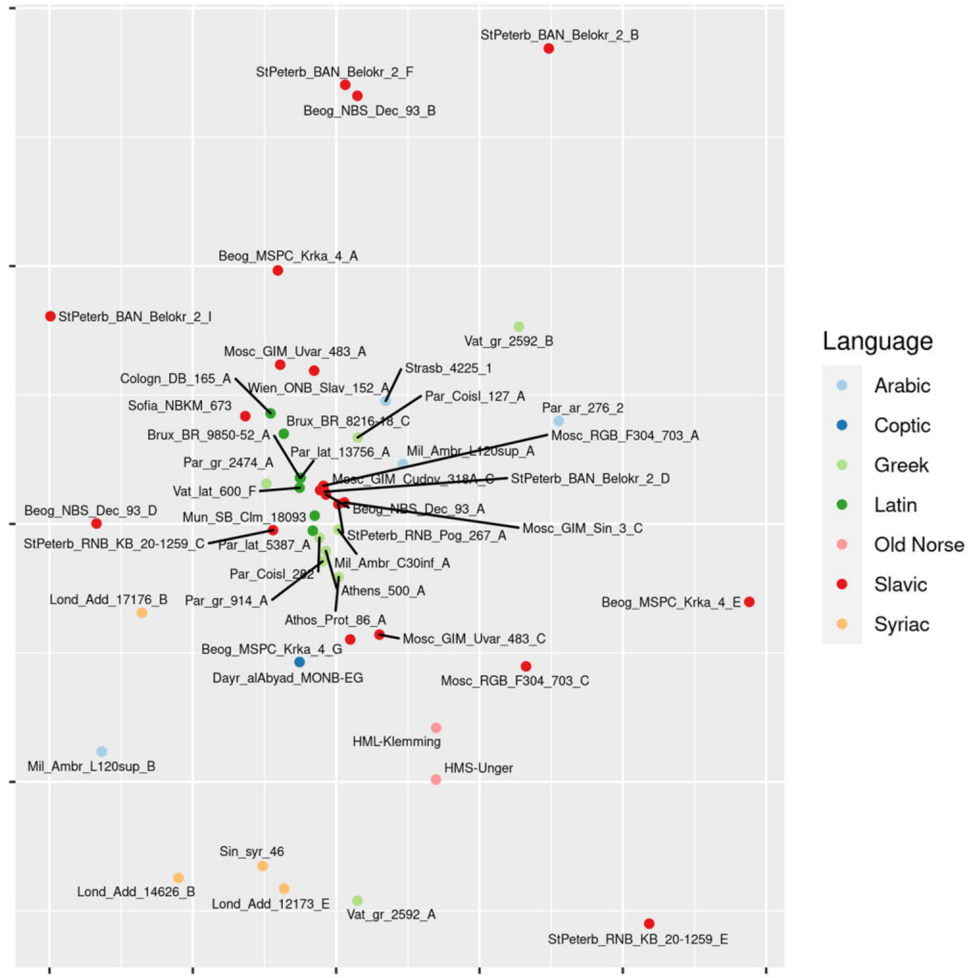

Fig. 3B Normalized Levenshtein labeled

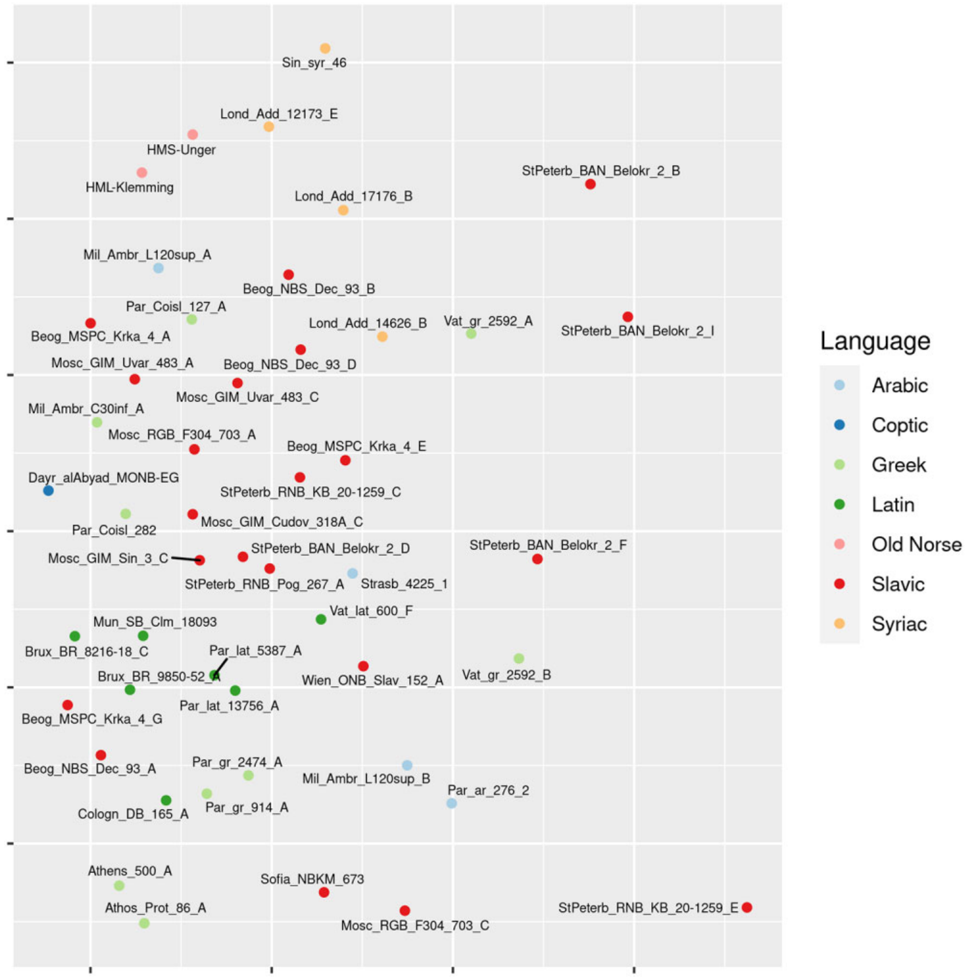

Fig. 4B FCM-Levenshtein labeled

Appendix C

Word count per folio page in selected manuscripts

Manuscript	Par_gr_2474 17r-26v	Lund_UB_54 2r-4v, 6rv	Athos_Prot_86 24r-33v	Brux_BR_8216-18 76r-85v	Brux_BR_9850-52 16r-25v	Cologn_DB_165, 34r-43v	StPet_BAN_Belokr_2 93r-102v	Beog_MSPC_Krka_4, 2r-12v (not 7 lacunose)	Beog_NBS_Dec_93 13r-22v	Strasb_4225 62r-71v	Vat_ar_71 180r-189v
Date	s. 13 AD	s. 11 AD	9th c. AD	819 AD	695-711 AD	675-725 AD	s. 16 AD	s. 14 AD	s. 12-13 AD	s. 10 AD	885 AD
	Greek average 332 wpf			Latin average 342 wpf			Slavonic average 396 wpf			Arabic average 325 wpf	
Words per folio page (wpf)	254	433	309	404	407	212	402	395	410	232	318
	278	432	323	426	367	220	389	342	415	352	326
	254	447	318	426	373	206	398	375	368	315	342
	259	437	290	475	394	208	397	371	432	307	376
	250		313	398	438	228	397	374	394	311	350
	242		310	428	357	221	382	385	397	310	353
	252		294	420	410	230	463	396	382	358	339
	247		314	415	386	225	410	405	412	306	365
	269		301	412	363	246	412	396	387	289	330
	248		290	396	324	236	461	382	366	307	313
Average	255	437	306	420	382	223	411	382	396	309	341
Overall average											351

Fig. 5 Words per folio page in selected Greek, Latin, Slavonic, and Arabic manuscripts

References

Åkerman Sarkisian, K. (2020). The *Apophthegmata Patrum* in the Slavonic context: a case study of textual doublets. In Ashbrook Harvey, S., Rydell Johnsén, H., Westergren, A., and Arentzen, T. (eds), *Wisdom on the Move: Monastic Sayings and Stories in Multicultural Conversation*. Leiden: Brill, pp. 119–46.

Birnbaum, D. (2003). Computer-assisted analysis and study of the structure of mixed content miscellanies. *Scripta & e-Scripta*, **1**: 15–64.

Birnbaum, D. (2016). *Repertorium of Old Bulgarian Literature and Letters*. http://repertorium.obdurodon.org/dendrogram-colored.xhtml (accessed 27 April 2022).

Bryant, D. and Moulton, V. (2004). Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, **21**(2): 255–65.

Chitty, D. J. (1974). The books of the old men. *Eastern Churches Review*, **6**: 15–21.

Cox, M. A. A. and Cox, T. F. (2008). Multidimensional scaling. In Chen C., Härdle, W., and Unwin A. (eds), *Handbook of Data Visualization*. Berlin, Heidelberg: Springer, pp. 315–47.

Dahlman, B. (2013). The Collectio Scorialensis Parva: an alphabetical collection of old apophthegmatic and hagiographic material. In Rubenson, S. (ed.), *Early Monasticism and Classical Paideia*, Studia Patristica LV: 3. Leuven: Peeters, pp. 23–33.

Dahlman, B., Göransson, E., and Åkerman Sarkisian, K. (forthcoming). A crosslinguistic approach to the study of the *Apophthegmata Patrum*. A case study of textual variation in the Greek, Latin and Slavonic traditions.

Faraggiana di Sarzana, C. (1997). *Apophthegmata Patrum*: Some crucial points of their textual transmission and the problem of a critical edition. In Livingstone, E. (ed.), *Historica, Theologica et Philosophica, Critica et Philologica*, Studia Patristica 29. Leuven: Peeters, pp. 455–67.

Göransson, E. (2019). The identification, definition and typology of manuscript lacunae. In Miltenova, A., Baranov, V., Miklas, H., Hawkins, K., Fuchsbauer, J. (eds), *Digital and Analytical Approaches to the Written Heritage. Proceedings of the 7th international conference El'Manuscript Textual Heritage and Information Technologies*, **2018**. Sofia: Gutenberg, pp. 79–96.

Göransson, E. (forthcoming). More than one translation behind the so-called Pelagius and Johannes translation of the *Apophthegmata Patrum*?

Holmberg, B. (2013). The Syriac collection of *Apophthegmata Patrum* in MS. Sin. Syr. 46. In Rubenson, S. (ed.), *Early Monasticism and Classical Paideia*, Studia Patristica LV:3. Leuven: Peeters, pp. 35–58.

Kaufman, L. and Rousseeuw, P. J. (1987). Clustering by means of medoids. In Dodge, Y. (ed.), *Statistical Data Analysis Based on the L1 Norm and Related Methods*. Amsterdam: Elsevier, pp. 405–16.

Larsen, L. (2008). The *Apophthegmata Patrum*: Rustic rumination or rhetorical recitation. *Meddelanden från Collegium Patristicum Lundense*, **22**: 21–30.

Larsen, L. (2013). 'On Learning a New Alphabet': The sayings of the desert fathers and the monostichs of Menander. In Rubenson, S. (ed.), *Early Monasticism and Classical Paideia*, Studia Patristica LV:3. Leuven: Peeters, pp. 59–77.

Miltenova, A. (1986a). Към методиката на изучаване на сборниците със смесено съдържание в старите южнославянски литератури. In Colucci, M., Dell'Agata, G., and Goldblatt, H. (eds), *Studia Slavica Mediaevalia et Humanistica: Riccardo Picchio Dicata*. Rome: Edizioni dell' Ateneo, pp. 517–26.

Miltenova, A. (1986b). Сборник със смесено съдържание, дело на етрополския книжовник йеромонах Даниил. С*таробългарска литература*, **19**: 114–25.

Miltenova, A. (1987). Апокрифният сборник от манастира Савина XIV в. в сравнение с други подобни южнославянски ръкописи. *Археографски прилози*, **9**: 7–30.

Miltenova, A. (2001). 'Устроение на [светите] слова' в старобългарската литература. С*таробългарска литература*, **32**: 99–110.

Rapp, C. (2010). The origins of hagiography and the literature of early monasticism: purpose and genre between tradition and innovation. In Kelly, C., Flower, R., and Stuart Williams, M. (eds), *Unclassical Traditions*, Vol. **I**: *Alternatives to the Classical Past in Late Antiquity*. Cambridge: Cambridge Philological Society, pp. 119–30.

Rubenson, S. (1995). *The Letters of St. Antony: Monasticism and the Making of a Saint*. Minneapolis: Fortress Press.

Rubenson, S. (2013). The formation and re-formations of the sayings of the desert fathers. In Rubenson, S. (ed.), *Early Monasticism and Classical Paideia*, Studia Patristica LV:3. Leuven: Peeters, pp. 5–22.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**(4): 406–25.

Tresoldi, T., Maurits, L., and Dunn, M. (2021). Seqsim, a library for computing measures of distance and similarity for sequences of hashable data types. Version 0.3.2. Uppsala: Uppsala universitet. https://github.com/evotext/seqsim (accessed 27 April 2022).

Veder, W. (2012). *The Scete Paterikon – Patericon Sceticum – Skitskiĭ paterik*, Vol. **1**: *Pegasus Oost-Europese Studies* 12. Amsterdam: Pegasus.

Zhao, Y. and Karypis, G. (2005). Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, **10**: 141–68.

Notes

1 The modified Levenshtein model has been developed by Luke Maurits and Michael Dunn. The code and the library including the underlying data for this article have been published by Tiago Tresoldi, Luke Maurits, and Michael Dunn in 'seqsim, a library for computing measures of distance and similarity for sequences of hashable data types'. Version 0.3.2. Uppsala: Uppsala universitet, 2021, available at: https://github.com/evotext/seqsim. The data from analyses of the contents in different manuscripts have been provided by Samuel Rubenson, Britt Dahlman, Karine Åkerman Sarkisian, and Elisabet Göransson. The manuscript data are part of a relational MySQL database (the *Apopththegmata Patrum DataBase* (APDB)) developed by Kenneth Berg. The database output is available on a web-based research platform, *Monastica*—a dynamic library and research tool, https://monastica.ht.lu.se/, with a new improved educational site at https://edu.monastica.ht.lu.se/. The construction of APDB and *Monastica* has been part of projects led by Samuel Rubenson.

2 For the sake of simplicity and for the purposes of this article, we will use the term Slavonic without taking any stance on either periodization of linguistic evolution or the differentiation of redactions, referring to this Byzantine legacy transmission to the Slavic lands.

3 Note that Jaccard distance, $1–I/U$, is derived in just the manner we described earlier from a similarity measure, the *Jaccard index* I/U, which has a maximum possible value 1.

4 Two collections in single-manuscript-based editions have been included as well; see Appendix A for a full table including metadata on language, source type, and collection for the selected collections. In the *Monastica* platform available online at https://edu.monastica.ht.lu.se/ the collections are normally marked as "Parts" (usually labeled "A", "B", "C", and so on) in the hierarchical structure of a source.

5 Cf. the list of types of collections in Appendix A, and a detailed graph with labels of all the manuscripts in each graph given in Appendix B.

6 Labeled versions of all the figures are to be found in Appendix B (Figs 1B, 2B, 3B, and 4B).

7 In the database APDB/*Monastica*, scholars mark the different types of omissions with specific tags in order to facilitate such refined analyses. There, "Lacuna" (L) stands for a physical loss, and "Missing" (M) stands for a loss of sayings depending on scribal mistakes or because of a lacuna in the Vorlage. In the manuscript Par_gr_2474, for example, there is both a Lacuna in the beginning of the manuscript, and a section marked Missing for the chapters VIII to half of XV (thus not due to a physical loss).

8 An exception was one of the Greek manuscripts, Lund_UB_54, which is fragmentary; in this case, only four full folio pages could be counted.

9 See Appendix C for a detailed list of manuscript labels, dating, and individual and average word count per folio page.