Computer-Assisted Analysis and Study of the Structure of Mixed-Content Miscellanies[1]

David J. Birnbaum
Department of Slavic Languages and Literatures
University of Pittsburgh
djbpitt+@pitt.edu

Draft of 2003-09-14 7:27 PM

Abstract: The present report describes and illustrates computational approaches to resolving two difficult research tasks connected with the study of medieval Slavic mixed-content miscellanies: 1) identifying the manuscripts in a corpus that are most like other manuscripts in terms of their contents and 2) visualizing graphically the relationship between the contents of two manuscripts. All core technology is XML-based (XML, XSLT, SVG).

Introduction

The Study of Mixed-Content Miscellanies

A *mixed-content miscellany* is a manuscript book that consists of an arbitrary set of texts (*articles*)[2] selected and arranged without the application of any particular organizational principle, that is, without a common genre, function, etc.[3] For example, The Loveč (Lovčanski) Miscellany (of

---

[1] The author is grateful to Andrej Bojadžiev, Dave Dubin, Sibelan E. S. Forrester, M. A. Johnson, Patrick Juola, Scott Malec, Predrag Matejić, Anisava Miltenova, Dave Mundie, Hugh Olmsted, Wendell Piez, Diljana Radoslavova, Miranda Remnek, Allen Renear, Robert Romanchuk, Bruce Rosenstock, Elizabeth Shaw, and Cynthia Vakareliyska for discussion, comments, and suggestions concerning some of the issues addressed here. Portions of this research have been or will be presented at the Thirty-Eighth International Congress on Medieval Studies (Kalamazoo, MI, US, May 2003), the 2003 Summer Research Laboratory on Russia and Eastern Europe and the Graduate School of Library and Information Science Electronic Publishing Research Group (University of Illinois at Urbana-Champaign, IL, US, June 2003), the Medieval Slavic Summer Institute at the Ohio State University Research Center for Medieval Slavic Studies (Columbus, OH, US, July 2003), Extreme Markup 2003 (Montreal, QUE, CA, August 2003), and the Thirteenth International Congress of Slavists (Ljubljana, Slovenia, August 2003).

[2] The constituent components of a miscellany are sometimes called *texts, works,* or *chapters.*

[3] A *fixed-content miscellany* consists of articles with a stable structure, such as church books (Oktoix, Triodion, Trebnik) or books organized according to the church calendar. "Fixed" thus does not necessarily mean unvarying, but it does mean relatively con-

King Ivan Alexander), pre-1331, N. 13.3.17 from the Library of the Russian Academy of Sciences, Saint Petersburg, consists of the following articles (Gagova 1995):

"Skitski" Patericon, excerpt[4]
Vita of St. Benedict, excerpt
Narrative from the Books
Sermon for the Assumption
Interpretations on the Holy Trinity and the Christian Faith, excerpts
Vita of St. Nikita
Vita of St. Mark of Athens
Revelation of St. Methodius of Patara
Nomocanon, excerpt[5]
Patericon, excerpts
Acts of our Lord Jesus Christ
Thunder-Book[6]
Kalendologion[7]

While these texts are all suitable as edifying or instructional readings for monks, they belong to a variety of genres (however one understands that term), and despite their common educational function, they are structured differently from one another, they would have been employed for different practical purposes within a monastic community, and they have no common properties that might make them a natural class of texts that one might expect to recur regularly in the same order in different manuscripts.

The contents and arrangement of one mixed-content miscellany often coincide partially with the contents and arrangement of another, and given the absence of any clear selectional or organizational principle governing the makeup of the books, one might explain conspicuous correspondences as evidence of shared textual transmission. For example, a scribe who sets about creating a mixed-content miscellany might open an exist-

---

strained and stable. In practice, the mixed/fixed dichotomy is a partial simplification, with texts varying in the stability of their tradition.

[4] A *patericon* is a collection of stories about monks. The "skitski," or "skete" patericon, an important seventh-century collection of stories about Egyptian monastic figures, was popular in Slavonic translation.

[5] A *nomocanon* is a guide to ecclesiastical law and procedure.

[6] A *thunder book* (*brontologion*) is an omen book dealing with "thunder in terms of signs of the zodiac and the age of the moon when it is heard." (Mathiesen 1995: 167)

[7] A *kalendologion* is an omen book about "the day of the week on which Christmas falls." (Mathiesen 1995: 167)

ing mixed-content miscellany, copy a favorite text, and then continue copying the text after that, and the one after that, as long as they engage his interest. At some point he might then return his source book to its shelf, select another, and copy a set of texts from it. This method of copying is consistent with the available evidence, where two manuscripts often show a set of the same articles in the same relative order, although not necessarily in the same absolute locations.

To take a hypothetical example, mixed-content miscellany manuscript X might contain articles A, B, *C, D, E,* F, and G and mixed-content miscellany manuscript Y might contain articles *C, D, E,* H, I, and J, so that the third through fifth articles of manuscript X correspond to the first three articles of manuscript Y (corresponding articles are italicized). On the basis of this correspondence, one might reconstruct a scenario in which the scribe of one of these manuscripts had access to a manuscript that was structured the same way as the other, and may have copied a series of articles from it in order.[8]

The conclusions one can draw from this type of hypothetical textual relationship are subject to certain constraints:

1. It is not necessarily the case that the scribe of one of these manuscripts would have had direct access to the other physical manuscript. For example, the scribe of manuscript Y may have used a manuscript Z that is no longer extant, which could have been a) a copy (child) of X, b) an ancestor from which X was copied, or c) a sibling of X (that is, X and Z may both have been copied from the same ancestor).
2. The hypothesis is not directional, which is to say that by itself it is not capable of determining whether the scribe of X had access to a manuscript similar to Y or vice versa.[9]

---

[8] See Miltenova 1986, Miltenova 1986a, Miltenova 1987 and, for an application of the same methodology to a different text type, Miltenova 2001.

[9] One mode of textual production for miscellanies with a strong organizational principle and moderately stable tradition, such as different redactions and arrangements of works that are united by having all been authored by Maksim Grek, involves copying all of the works from one manuscript source in order and then turning to a second source and copying in order the works that were not already present in the first source. In this case, the scribe turns to the second source manuscript precisely because it overlaps substantially with the first, seeking to produce a composite copy that unites the contents of both sources. In such cases, the absolute position of the correspondences may frequently disclose both the directionality and the sequence of copying. (Olmsted 1994: 115)

3. The hypothesis is independent of the age of the manuscripts, which is to say that even in cases where X is older than Y and the two are related, it is not the case that Y must be based on X. X could instead have been based on an even older non-extant ancestor of Y.

What the preceding hypothesis addresses, then, is not the relationship between physical manuscripts, but the relationship between the textual organizations witnessed by the manuscripts. When we hypothesize a relationship between two manuscripts, what we are proposing is not direct physical copying, but partially shared textual transmission.

It is, of course, possible for the contents of mixed-content miscellany manuscripts to overlap by chance, but it is reasonable to hypothesize that the likelihood of chance decreases in inverse proportion to the length of the sequence of shared articles. That is, a correspondence of a single text constitutes very weak evidence for shared textual transmission, while a very long correspondence is unlikely to have arisen by chance.[10] In general, the significance of a correspondence as evidence of shared textual transmission increases at a greater-than-arithmetic (linear) rate, which is to say that, for example, a six-article correspondence between two manuscripts constitutes much stronger evidence of shared transmission than two three-article correspondences between the same two manuscripts. Additionally, the absolute position of the correspondences is irrelevant, which is to say that, for example, a three-article correspondence is equally significant whether the articles occur in the same absolute position in both manuscripts (e.g., first three, third through fifth, etc.) or in different absolute positions (as in hypothetical manuscripts X and Y, above).[11]

Finally, it is important to note that the hypothetical textual relationships described above are based on comparing not the full text of two manu-

---

[10] The assumption that correspondence implies shared transmission is credible precisely because the manuscripts in questions are *mixed-content* miscellanies, that is, manuscripts whose contents are not organized according to any independently identifiable principle. Books with fixed content can be expected to resemble one another even when scribes independently seek to create the same type of book, but there is no obvious reason why specific mixed-content miscellanies, which observe no organizational principle whatsoever, should happen by chance to wind up with similar contents and structures. Furthermore, as Olmsted (1994) demonstrates, shared transmission is often unmistakeable even in cases involving miscellanies whose contents are subject to significant constraints.

[11] As was noted previously although absolute placement may not be important for discovering the existence of textual relationships, Olmsted (1994) has demonstrated that it may be relevant for determining directionality, an issue not addressed in the present report.

scripts, but lists of their contents (according to titles determined and assigned by researchers, since the same work may bear different titles in different manuscripts [Krushelnitskaya 2003]). While full-text comparisons, which have long been used for stemmatic purposes in humanities computing, provide valuable measurements of relationship and distance, the strategy described here, based on descriptive "tables of contents," differs from full-text comparison in at least two ways: 1) it is applicable in situations where one has available only descriptions, and not full transcriptions, of manuscripts; and 2) it is not affected by editorial intervention during copying.[12]

Problems of Studying Mixed-Content Miscellanies

The study of mixed-content miscellanies raises questions about locating and identifying related manuscripts, and also about visualizing in an accessible way the textual correspondences among manuscripts. These two types of problems are described in greater detail below.

Locating and Identifying Related Manuscripts

A scholar armed with a corpus of mixed-content miscellanies might wish to look for evidence of textual transmission by asking two types of questions:

1. Which manuscripts in the corpus are most like manuscript X?
2. Which manuscripts in the corpus are most like which other manuscripts in the corpus?

As long as the corpus is sufficiently small, both types of questions are easily answered merely with the application of memory and paperwork. However, given the very large number of mixed-content manuscripts in existence, one can increase the chances of discovering important correspondences by working with very large corpora, and also by working collectively, in which case not every researcher will be able to become familiar with the contents of every manuscript in the corpus. Because this

---

[12] The relative immunity to editorial intervention is also a liability, since the reliance on tables of contents, rather than full text, also increases the likelihood of false matches. For example, one may find a correspondence in titles that turns out to mask two established redactions so different that transmission from one manuscript to the other is impossible. Ideally, different redactions will have been assigned different titles by researchers during cataloging, but in many cases scholars will not have had direct access to the manuscripts, and will have needed to work with old catalogs and collection descriptions that do not implement the necessary distinctions. Note also the prevalence, exemplified in Gagova 1995, of titles qualified with the term "excerpt," which is incapable of distinguishing among different excerpts from the same work.

sort of large-scale comparative work would not be manageable in memory or on paper, it invites computational assistance.

The solutions to the two types of questions posed above are different, although related. To answer question #1, one compares manuscript X to every other manuscript in the corpus, a task of linear (O(n)) complexity. To answer question #2, one must compare every manuscript in the corpus to every other manuscript in the corpus, a task of quadratic (O(n^2)) complexity. Heuristics can simplify the specific comparison routines,[13] but it is nonetheless necessary to consider separately each pair of manuscripts. Sample numbers of required comparisons are:

Number of Manuscripts in the Corpus (n)	*Number of Comparisons to Find Manuscripts Like a Specific Manuscript X*	*Number of Comparisons to Find Manuscripts Like (Any) Other Manuscripts*
n	$n - 1$	$n(n - 1) / 2$
5	4	10
10	9	45
20	19	190
50	49	1225
100	99	4950
300	299	44850
500	499	124750

Specific comparison algorithms and their computational implementations are described below (“Locating and Identifying Related Manuscripts”).

[13] For example, as described below, once one looks for correspondences, there are processing advantages to searching for longer series before shorter ones. Nonetheless, as a heuristic one might first check for one-article correspondences as a way of eliminating the need to look for longer correspondences in cases where there is no correspondence at all. Consider also that if manuscripts A and B have identical contents and one compares manuscript A to manuscript C, it is not also necessary to compare manuscript B to manuscript C, since the degree of similarity between B and C will be identical to the degree of similarity between A and C.

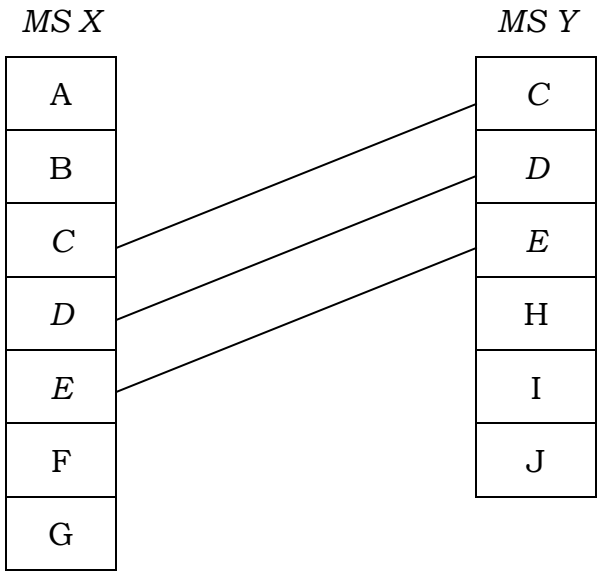
Visualizing Relationships between Manuscripts

Relationships among miscellanies traditionally have been described either in prose (see the discussion in Olmsted 1994: 110–12) or in tables (as in, e.g., Miltenova 1986a: 125, Miltenova 1987: 23, and Miltenova 2001: 105). As Olmsted (1994: 112) notes concerning prose descriptions, "mere strings of numbers can be somewhat opaque," and although Miltenova's tables allow her to collate data from several manuscripts, it is not easy to see the relationships among manuscripts within a grid of numbers unless one highlights the correspondences graphically.

Olmsted (1977, 1994) addresses directly the problem of visualizing correspondences among the contents of manuscripts, and proposes as a solution the use what he terms a *plectogram,*[14] which he describes as follows:

> Two manuscripts to be compared are represented by two columns, left and right, with one row, or line, for each composition [...] Between these two columns, lines are drawn connecting identical works [...] . (Olmsted 1994: 112)

A plectogram of our hypothetical example above would look as follows:



As Olmsted (1994) emphasizes, "plectograms [...] are presented as working aids showing the degree of existing relationship, not as proposed genealogical models." (125) Their virtue is that "the device lets us see at a glance the contours of the relationship. It allows us to give focus and pri-

[14] "We shall refer to this sort of representation as a **plectogram** (cf. Greek πλέκω 'weave, braid', πλεκτός'woven, braided'). (Olmsted 1994: 112)

ority to information that otherwise would be unwieldy; it allows us to set manageably small tasks for further work in strategic sampling of textual and other relationships." (131) In other words, plectograms do not prove the existence of relationships between either physical manuscripts or the texts that they witness, but they do enable the researcher to visualize the correspondences as part of the process of determining whether such genealogical relationships can be supported.

Computational Approaches and Solutions

Goals and Guidelines

Philological Goals and Guidelines

One might envision the two tasks described above (locating and identifying related manuscripts, visualizing relationships between manuscripts) as early stages in the process of studying mixed-content miscellanies. A more complete description of that process might be:

1. Produce tables of contents for each manuscript in a corpus of mixed-content miscellanies.
2. Using a computer program that conducts pairwise comparisons across the entire corpus of manuscripts, measure and record the degree of similarity between each pair of manuscripts.
3. Either:
   a. Locate and identify those manuscripts that are most similar in content and structure to a particular manuscript used for comparison; or
   b. Locate and identify those manuscripts that are most similar in content and structure to one another within the corpus.

   In either case, hypothesize that these are the manuscripts most likely to be related to one another through textual transmission, at least in the case of mixed-content miscellanies. Where the scope of a corpus makes manual pairwise comparison impractical, or even impossible, the use of computer technology will enable researchers to concentrate their attention on pairs of manuscripts that are flagged by the program as more likely than others to be related to one another.
4. Construct plectograms of pairs of manuscripts that are likely to be related to each other as a way of visualizing the textual correspondences.
5. Examine the results of the preceding steps to confirm or refute the hypothetical relationships.

6. Having identified related manuscripts, conduct whatever additional philological research one might wish to undertake.

It is important to note that the use of computer technology in steps 2–3 is not a black box that produces a definitive report of instances of textual transmission in the corpus. Computer technology in steps 2–3 is a heuristic that saves the researcher from having to conduct intensive and time-consuming manual pairwise comparisons of all manuscripts in the corpus, including those that are very unlikely to be related. Computer technology in step 4 offers a technique that, as was noted above, enables the researcher to see at a glance patterns that are more salient and more easily apprehended when viewed graphically than when described in words and numbers.

Technological Goals and Guidelines

Olmsted's plectograms "were generated on a Macintosh computer by a program written in the graphic spreadsheet application WINGZ by Informatix" (1994: 113). One goal of the present paper has been to operate entirely within open and accessible standards, producing at every stage files that are legible in plain-text editors, without relying on commercial or otherwise proprietary document architectures or programs. This strategy protects researchers from the risk of having their results become inaccessible when a particular software company goes out of business.

With this goal in mind, the contents of each manuscript are encoded in a modified version of the Text Encoding Initiative (TEI)[15] Extensible Markup Language (XML)[16] Document Type Definitions (DTDs).[17] Taking advantage of one of the general strengths of markup technology, these encoded descriptions are multi-purpose. In fact, they were created initially as input to transformations that would produce electronic and paper catalogs of manuscript descriptions, but although electronic comparison and graphic visualization were not explicit goals of that initial encoding, the unmodified XML files are suitable for these new purposes without manual editing. XML data was accessed with Extensible Stylesheet Language Transformations (XSLT),[18] and subsequent processing was conducted with either XSLT or arbitrary scripting languages

---

[15] See http://www.tei-c.org/.

[16] See http://www.w3.org/XML/.

[17] See http://clover.slavic.pitt.edu/~repertorium/.

[18] See http://www.w3.org/TR/xslt. XSLT transformations were implemented with the freeware Saxon processor, about which see http://saxon.sourceforge.net/.

(Spitbol[19] was used for prototyping because of its legibility and strong string-, list-, and pattern-processing abilities, but the programming in question is not complex, and any general-purpose scripting language could have been employed in its place). Intermediary textual output was formatted as XML or plain text, depending on whether subsequent processing was to use XSLT or Spitbol, respectively, and final textual output was formatted as XML for viewing in any XML browser (such as Microsoft Internet Explorer). Plectograms were formatted as Scalable Vector Graphics (SVG),[20] an XML tag set for graphic imaging. XML, XSLT, and SVG are all published World Wide Web Consortium (W3C)[21] standards and are supported by a wide range of software tools and products, many of which are distributed at no cost.

Algorithms and Implementation

Locating and Identifying Related Manuscripts

For reasons described above ("The Study of Mixed-Content Miscellanies"), the algorithm for determining which pairs of manuscripts should be considered more closely related than others was developed according to the following assumptions:

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).
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.[22]

The input files for this report were a set of 104 (modified) TEI-based XML-encoded manuscript descriptions. The total number of articles in all 104 manuscripts is 1539, with 751 different article titles. The number of

---

[19] See http://www.snobol4.com/.

[20] See http://www.w3.org/TR/SVG11/.

[21] See http://www.w3.org/.

[22] An alternative metric would incorporate the length of the manuscripts, thus calculating something comparable to the percentage of material that is shared by the two, rather than the raw number of articles. Because the length of mixed-content miscellany manuscripts is variable, the percentage of articles that correspond seemed less significant than the literal length of the match in terms of nmber of artcles.

articles per manuscript ranges from a low of 1 to a high of 66, with a mean of approximately 14.8.

The procedure used to locate and identify correspondences was as follows:

1. Extract plain-text lists of all articles from each manuscript in the corpus using an XSLT script.
2. Merge the resulting manuscript-specific article lists into a single list without repetition and output a version of this list that assigns a unique index value to each article. This was accomplished by using the Unix *cat* command to combine the lists produced in the previous step and then piping the output through *sort* and *uniq* processes to produce a unified list. A Spitbol script then read in the list and output each line with an associated unique index number (four-place hex numbers, which allow for 2^{16} [65536] different article titles, were employed).
3. Using the individual article lists produced in step #1 and the index numbers assigned to each article in step #2, produce a coded representation of the contents of each manuscript for subsequent comparison. It would have been possible to use XSLT to compare the actual article titles, rather than the intermediary index numbers, but the index numbers are simpler to process because they are of uniform length.
4. Generate a list of all pairs of manuscripts to compare. Each pair needs to be compared only once (that is, if one compares X to Y, it is not necessary to compare Y to X separately).
5. Compare the lists of index numbers for each pair of manuscripts according to the comparison algorithm described below and generate a non-negative integer value for each pair of manuscripts representing the extent of the similarity.

The comparison algorithm for determining similarity was as follows:

1. Identify the longer and shorter manuscripts and call the shorter N and the longer M.
2. For manuscript N with n articles, try to find a sequence of all n articles in manuscript M, then a sequence of $n - 1$ articles, etc., ending in single-article matches.
3. Include submatches, which is to say that a three-article match also involves two embedded two-article matches and three embedded one-article matches. The inclusion of submatches ensures that long matches will be weighted more heavily than sets of short matches

(e.g., a six-article match will be weighted more heavily than two three-article matches). Neither the absolute location of the matching content within each manuscript nor the absolute number of articles in each manuscript affects the weighting.

4. The measure of similarity between two manuscripts is the sum of all matches.

The number of comparisons required for a pair of manuscripts is:

$$\sum_{i=1}^{n} i(m-n+i)$$

where n = the number of articles in the first manuscript of the pair, m = the number of articles in the second, and $1 \le n \le m$.[23] Sample numbers of comparisons are:

Articles in Manuscript X (n)	*Articles in Manuscript Y (m)*	*Formula*	*Number of Comparisons*
5	5	$\sum_{i=1}^{5} i(i)$	55
5	10	$\sum_{i=1}^{5} i(5+i)$	130
5	20	$\sum_{i=1}^{5} i(15+i)$	280
5	30	$\sum_{i=1}^{5} i(25+i)$	430
5	50	$\sum_{i=1}^{5} i(45+i)$	730

A corpus of 300 manuscripts with 5 articles per manuscript would require (300 * 299 / 2) = 44850 manuscript pairs * 55 article sequences = 2466750 comparisons. A corpus of 300 manuscripts with 10 articles per manuscript would require the same (300 * 299 / 2) = 44850 manuscript pairs * 385 article sequences = 17267250 comparisons.

Sample values of matches are: [24]

---

[23] The algorithm is more efficient if the pair is arranged so that the shorter manuscript is first, since that strategy avoids, for examples, looking to match a seven-article sequence from the first manuscript within a second manuscript that contains only six articles.

Length of Match(es)	Formula	Weight
n	$\sum_{i=1}^{n} i(n+1-i)$	$\sum_{i=1}^{n} i(n+1-i)$
1	$\sum_{i=1}^{1} i(1+1-i)$	1
3	$\sum_{i=1}^{3} i(3+1-i)$	10
6	$\sum_{i=1}^{6} i(6+1-i)$	56
3 + 3	$\sum_{i=1}^{3} i(3+1-i) + \sum_{i=1}^{3} i(3+1-i)$	20
3 + 6	$\sum_{i=1}^{3} i(3+1-i) + \sum_{i=1}^{6} i(6+1-i)$	66

A pairwise comparison of all manuscripts in the sample corpus of 104 manuscripts identified the following as the ten weightiest matches:

[24] A more precise weighting strategy, i. e., one that accurately reflected the relative likelihood of textual transmission in cases of one six-article correspondence vs two three-article correspondences, could be determined by verifying empirically the effect of match length on the likelihood of a genuine textual relationship.

Weight	*Manuscript X*	*Manuscript Y*
9149	AM100MCB	AM433
5520	AM100MCB	AM149NBW
4560	AM149NBW	AM433
3655	AM17DUJC	AM309
1330	DR57CMS	DR944CMS
548	DR54CMS	DR971
220	DR248	DR940CMS
181	DR248	DR972
157	DR944CMS	DR972
156	DR57CMS	DR972

Visual inspection of the tables of contents of the manuscript pairs that the programs identified as similar supports the hypothesis that the programs are capable of ranking pairs of manuscripts according to the length of the ordered correspondences in their contents, at least in a general or relative way. While there is a strong and obvious likelihood that corresponding contents suggest shared textual transmission, at least in the case of mixed-content miscellanies, this hypothesis must ultimately be tested by examining the actual manuscripts, and not merely lists of their contents.

The likelihood that this type of correspondence can be attributed reliably to shared textual transmission varies according to the type of text. For example:

1. This type of comparison is most useful for mixed-content miscellanies because the tremendous variability of their makeup, owing to the complete lack of compositional constraints, increases the likelihood that correspondence implies a direct relationship. On the other hand, in the case of manuscripts with a very stable tradition, such as the Trebnik manuscripts added to the corpus for control purposes, one expects to find a high degree of correspondence simply because the

makeup of a Trebnik admits relatively little variation. [25] This is true of several of the pairs of (relatively fixed-content) miscellanies listed in the table above.

2. Very strong correspondences may help locate potential "twins" within a large corpus.
3. Except in the case of short manuscripts, low weights identify pairs that are relatively less likely to merit further attention.

The list of manuscripts most likely to be related textually to a specific other manuscript can be identified by using an XSLT transformation to extract from the master list of comparisons those comparison elements that refer to the specific target manuscript. For example, such a transformation can identify quickly the thirty-two manuscripts in the corpus that have a positive weight when compared to AM326 ("Adžarski sbornik N 326 NBKM").

Visualizing Relationships between Manuscripts

Scalable Vector Graphics (SVG) is an XML tag set designed for describing complex graphic objects and rendering them within an SVG viewer.[26] While SVG documents look like graphic images in an SVG viewer or browser, they are actually XML documents, which means that they consist of plain-text content plus XML markup and can be viewed and edited like any plain-text or XML document. Furthermore, because both the input files for this project and the eventual SVG-encoded plectograms are XML documents, it is easy to transform data from the original files into plectograms with XSLT transformations, and the use of XSLT to produce the SVG output ensures that that output will be well-formed XML. Because XML, XSLT, and SVG are all published, non-commercial standards, and because the files at all times are plain-text documents with markup that can be viewed in plain-text editors, this technology avoids the risk of making the data crucially dependent on a particular vendor's proprietary file format. Furthermore, the ability to generate the plecto-

---

[25] A Trebnik is a book of rites, rituals, and prayers for specific occasions, similar to the Euchologion, but without the eucharistic liturgies. See Radoslavova 2000 for a detailed study of Trebink manuscripts, including the two contrasted in the plectogram. Of the ten weightiest pairs, listed above, the files beginning with "AM" represent mixed-content miscellanies and those beginning with "DR" represent Trebnik manuscripts.

[26] The most popular SVG viewers currently available are the Adobe SVG Viewer, which works as a web-browser plug-in (http://www.adobe.com/svg/main.html), and the stand-alone SVG Browser included as part of the Apache Batik SVG Toolkit (http://xml.apache.org/batik/). Both are supported on multiple operating systems and distributed without charge.

grams from XML files that were originally produced for a different purpose exploits the multipurposing strength of structured text in general and XML in particular.

To generate a plectogram, an XSLT transformation is applied to two XML manuscript-description files (supplied as command-line parameters). Because the article titles are often so lengthy that they would not fit legibly into a plectogram, the index numbers generated earlier for use in identifying similarities in the contents of manuscripts are used in the plectograms in place of the corresponding article titles. The XSLT stylesheet that produces the plectograms retrieves these index values automatically by looking up the index number associated with each article title in a separate XML file (a by-product of the earlier comparison routines) that maps each title to its associated index number. One strength of XML-related technology is that the XSLT transformation was able to extract the information needed directly from the manuscript descriptions and then manipulate it to generate the plectogram, without the need for any intermediary files.

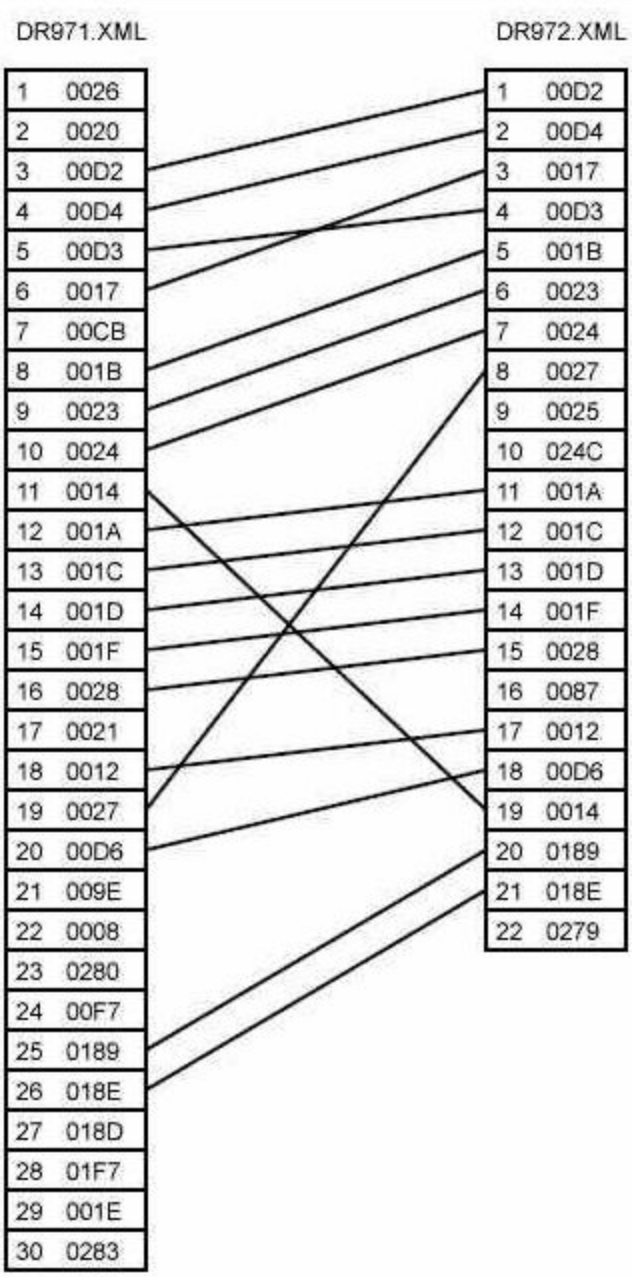
The following plectogram illustrates the correspondences between DR971 (Trebnik of Dimit″r Joanovič N 971 NBKM) and DR972 (Trebnik of Daskal Filip N 972 NBKM), which have a correspondence weight of 59.[27]

---

[27] The SVG images in this report were converted from JPG images through a screen-capture utility, a process that is not capable of retaining SVG animation. The original SVG file is animated, and responds to mouse-over events in two ways:

1. When the mouse cursor enters a cell, the four-digit index number in that cell turns red, as does the identical index number when it occurs in any cell in any column. Additionally, all connecting lines associated with the selected index number (in any column) also turn red.
2. When the mouse cursor enters a cell, the Bulgarian-language title of the article is displayed in red above the plectogram. Index numbers were used in place of full article titles because the latter are often too long for legible diagrammatic rendering, but their animated display in response to mouse-over events ensures that they remains easily accessible, and the user is not required to look them up separately.

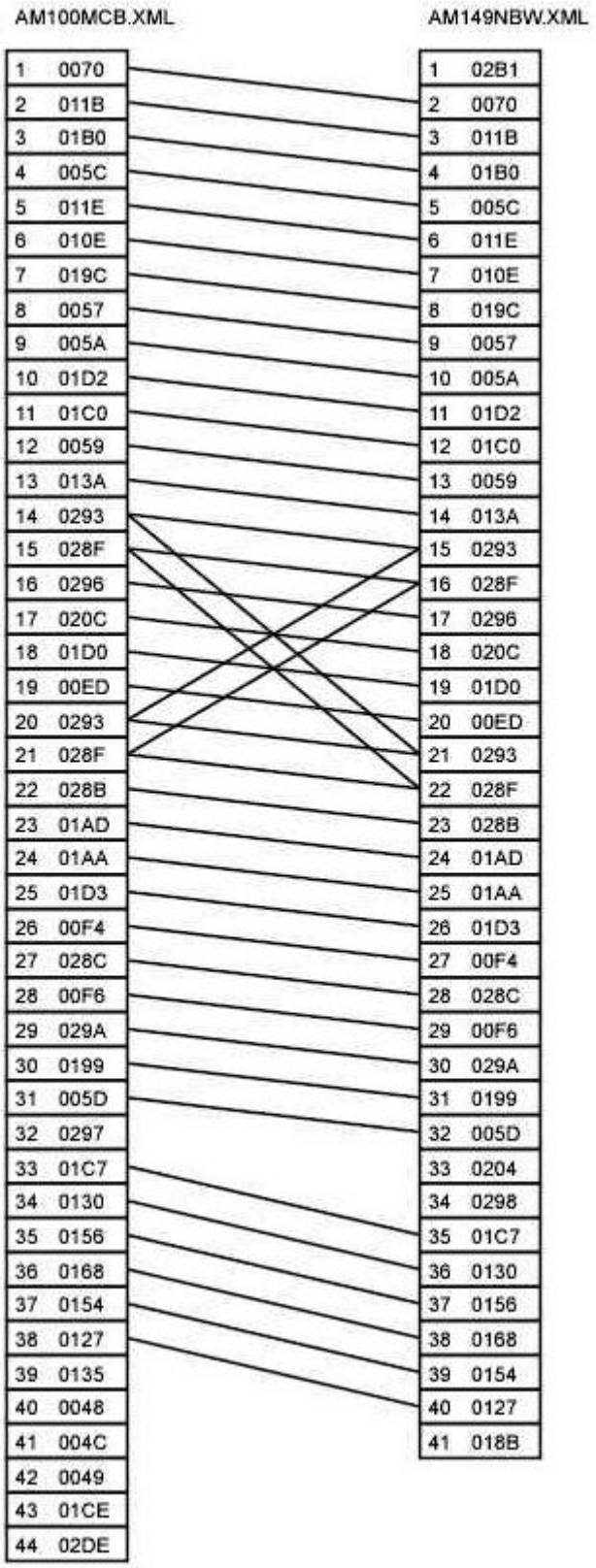
The current version of the Adobe SVG viewer (3.0) supports animation. The current version of the Batik viewer (1.1.1) does not.



In this example, the two textual traditions overlap to such an extent that there is clearly some sort of relationship, but the crossing lines suggest that the relationship is probably indirect, that is, that neither manuscript was a particularly close source for the other. Because these are both Trebnik manuscripts, reflecting a relatively stable “fixed-content” tradition, the correspondences are most easily attributable to the stability of that tradition, rather than to any specific relationship between these two particular manuscripts.

In some cases plectogram visualization can draw a researcher’s attention to possible imprecisions in the identification or encoding of the manu-

script contents. For example, the following plectogram compares AM100MCB (Daniilov Miscellany N 100 MSPC Belgrad) and AM149NBW (Vienna Apocryphal Miscellany N 149), which have a correspondence weight of 5520.



This plectogram draws the researcher's attention to the following details:

1. Two sets of two texts occur twice in both manuscripts: 0253 and 024F are items 14 and 15, respectively, and also items 20 and 21, respectively in AM100MCB, and the same articles are items 15 and 16, respectively, and 21 and 22, respectively in AM149NBW. The overwhelming pattern of parallel lines suggests a very close textual relationship, making it most likely that the actual transmission-related correspondences are between 14, 15, 20, and 21 in AM100MCB and 15, 16, 21, and 22, respectively in AM149NBW. The fact that there is repetition not of one, but of two articles in sequence constitutes an additional argument in support of a direct relationship between the manuscripts.

   The repetition described above suggests that the protograph of this branch of the tradition was compiled from two sources, each of which separately contained the two articles in question. There is no natural alternative explanation for why a scribe would have copied the same materials twice.

2. The only gap in the parallel structures involves item 32 (0257) in AM100MCB and items 33 (01DC) and 34 (0258) in AM149NBW. When we consult the master list of articles, we find that 0257 is "Questions of St. John the Theologian to Abraham" and 0258 is "Questions of St. John the Theologian to Abraham, fragment." This may not be an error, especially in light of the insertion of an adjacent additional text (item 33 [01DC]) in AM149NBW, and it may show either an abridgement in AM149NBW during copying from a source like AM100MCB or the use of a source that was already abridged. It is also possible that the designation by researchers of the article in one manuscript as fragmentary and the other as complete is erroneous, which should be verified.

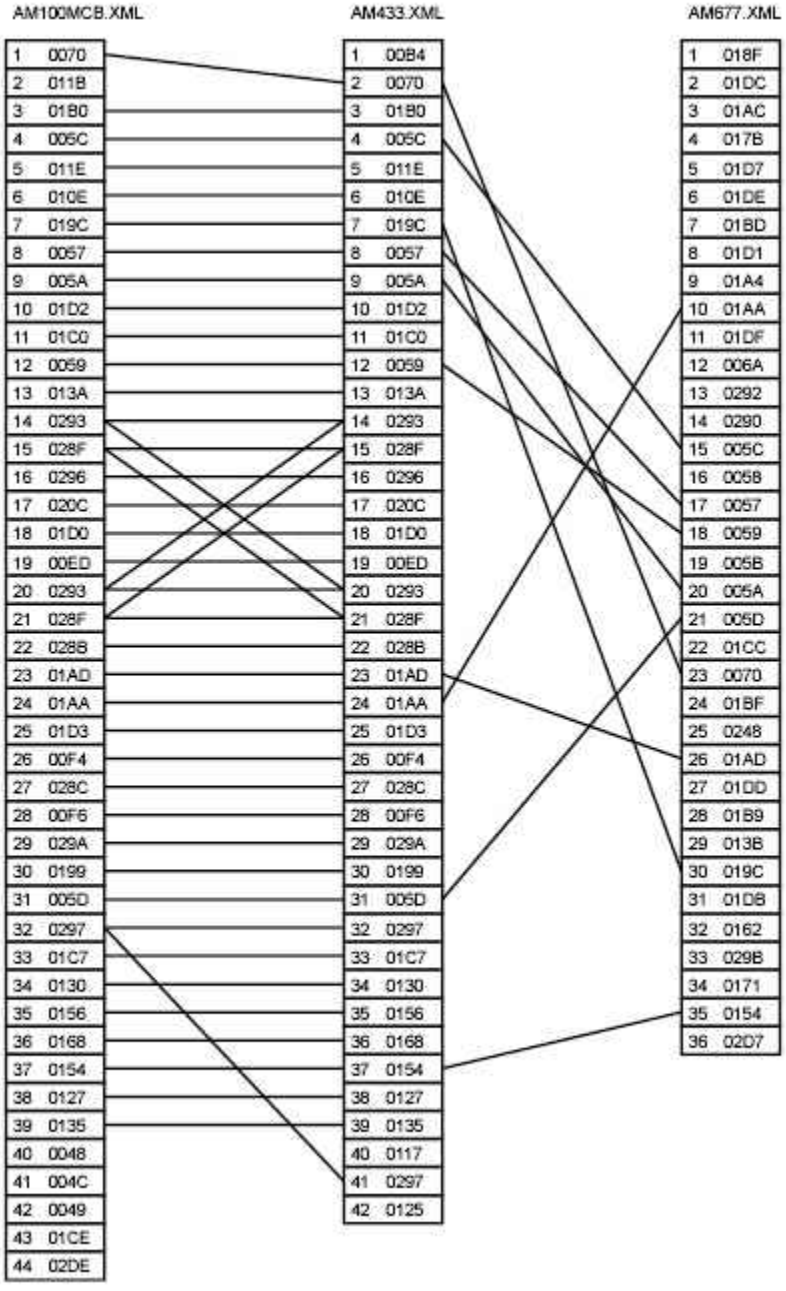
In the example above, then, the plectogram not only shows the close relationship between the two manuscripts, but it also draws the researcher's attention in an obvious and accessible way to a place where there appears not to be a complete correspondence in contents, an interruption in the pattern that calls for special attention.

What the Plectograms Reveal

The descriptive strengths of plectography can be illustrated by examining plectograms of assorted mixed- and fixed-content miscellanies.

Two Closely-Related Mixed-Content Miscellanies and an Outsider

Consider the following plectogram of three mixed-content miscellanies, AM100MCB (Daniilov Miscellany N 100 MSPC Belgrad), AM433 (Panagjurski Miscellany N 433 NBKM), and AM677 (Tikveški Miscellany N 677 NBKM):

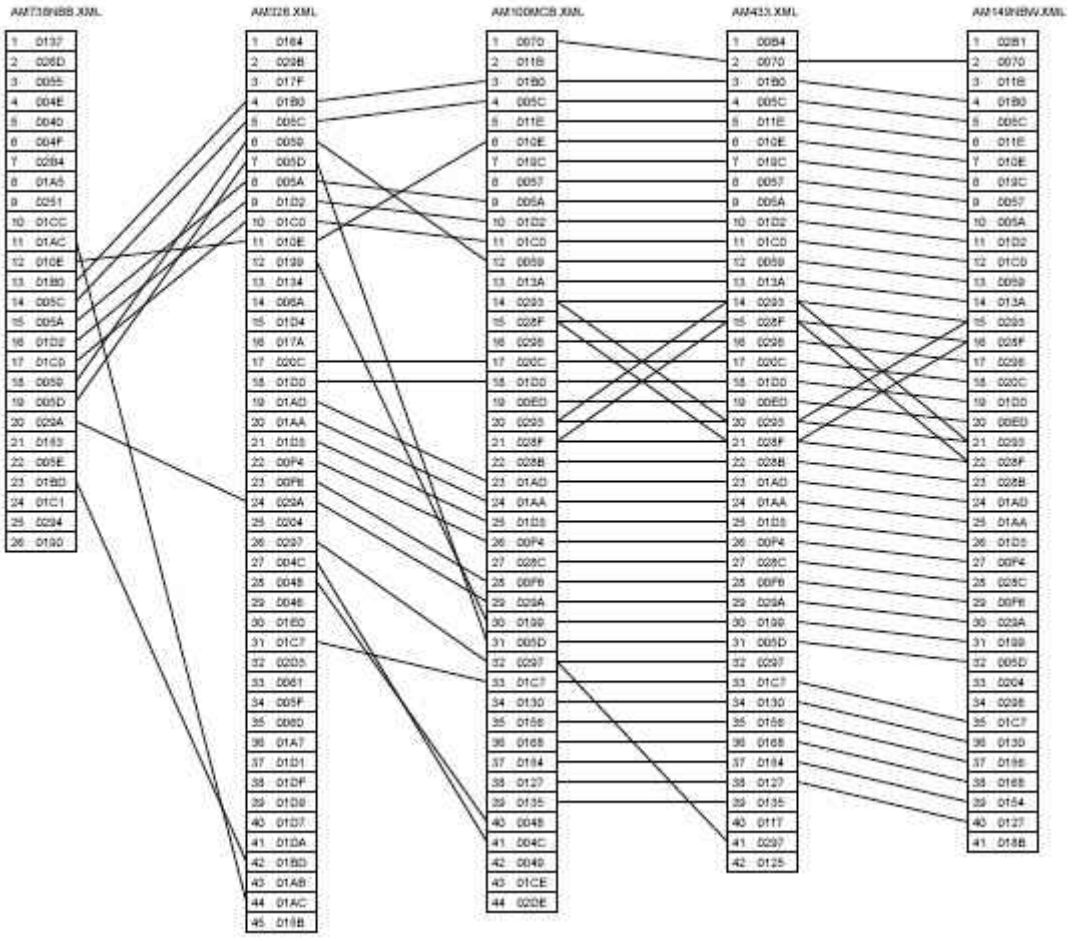


It is clear from the the number of connecting lines between AM100MCB and AM433, the fact the lines fall substantially in an uninterrupted par-

allel pattern, and the presence of the same series of two articles twice (and in the same place) in both manuscripts that the two are very closely related textually. But this plectogram shows just as clearly that although AM677 shares many articles with the other two manuscripts, any shared textual transmission must be very indirect, since although there are many connecting lines, no two lines are parallel and adjacent to each other.

Five Mixed-Content Miscellanies

The following plectogram represents five manuscripts that Anisava Miltenova (personal communication) attributes to a common archetype: AM738NBB (Belgrade Miscellany N 738), AM326 (Adžar Miscellany N 326), AM100MCB (Daniilov Miscellany N 100 MSPC Belgrade), AM433 (Panagjurski Miscellany N 433 NBKM), and AM149NBW (Vienna Apocryphal Miscellany N 149).

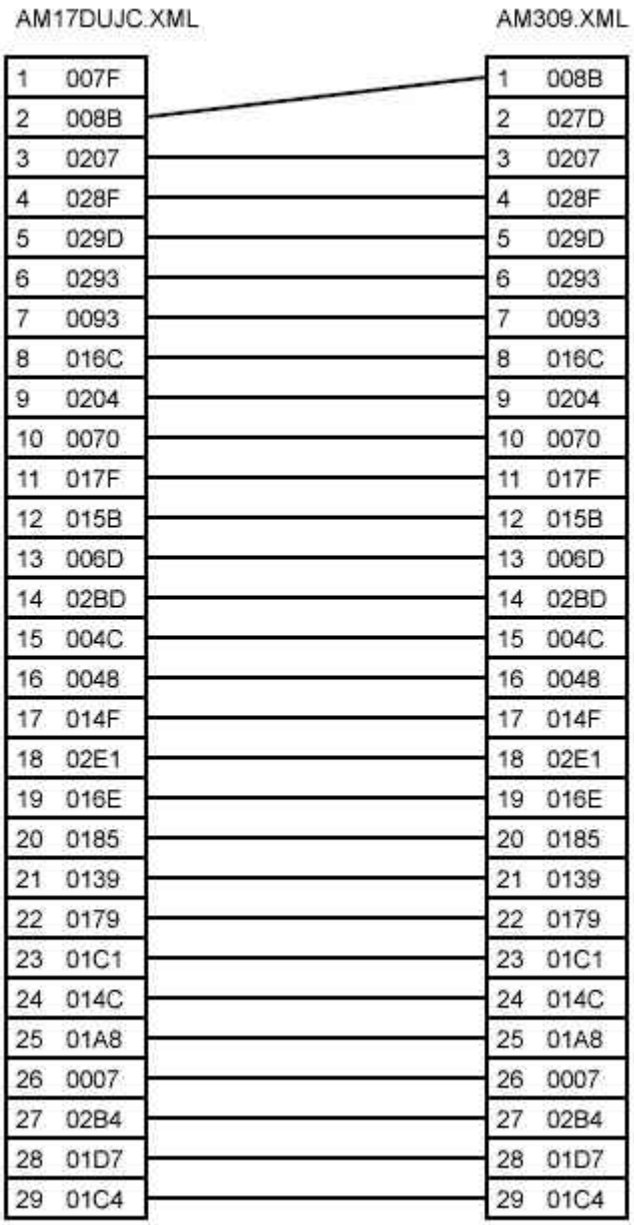


This pattern clearly shows that the three rightmost manuscripts (AM100MCB, AM433, AM149NBW) are very closely related textually, AM326 (second from left) is similar to the last three in composition, al-

though not to the same extent as they are to one another, and AM738NBB (leftmost) bears structural similarities to AM326 that are not all shared with the other three.

Twin Manuscripts

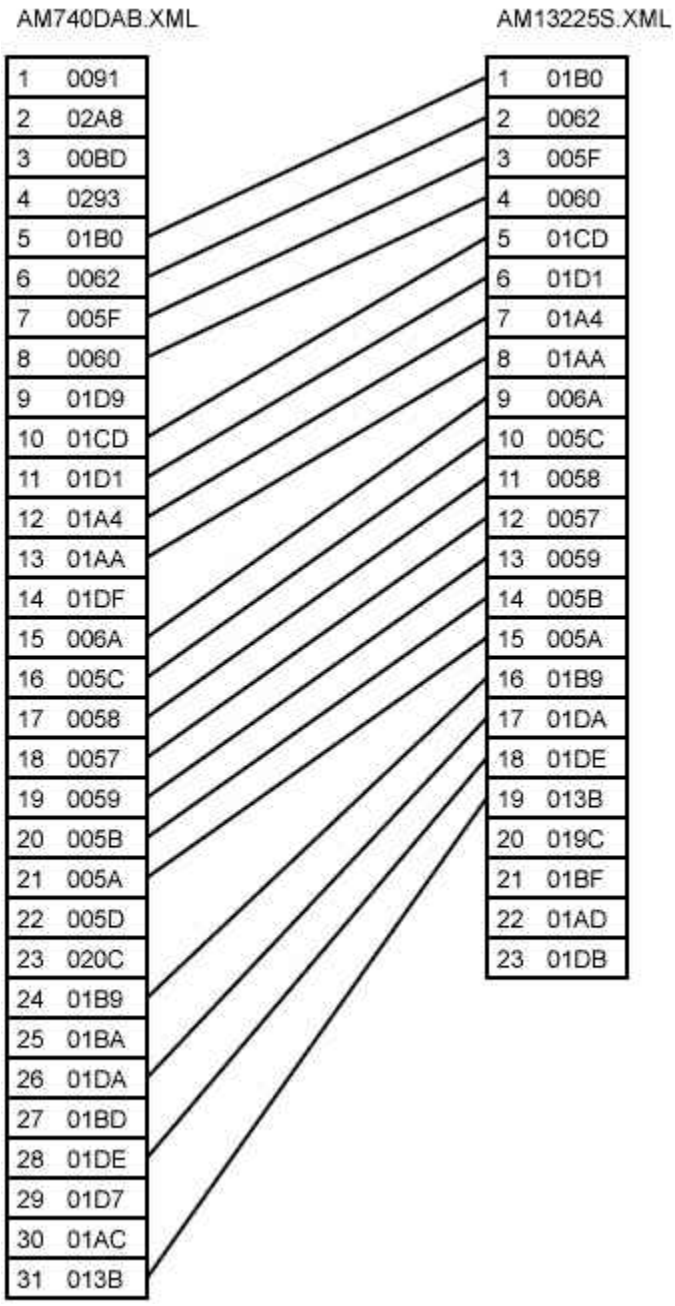
The following plectogram represents “twin” mixed-content miscellanies, i.e., manuscripts that reflect not simply shared textual transmission, but almost perfect textual alignment. The two miscellanies are AM17DUJC (Miscellany N 17 from the collection of the Ivan Dujčev Center for Slavonic-Byzantine Studies) and AM309 (Beljakovec Miscellany N 309 NBKM). According to Anisava Miltenova (personal communication), AM17DUJC is copied from AM309.



Focusing

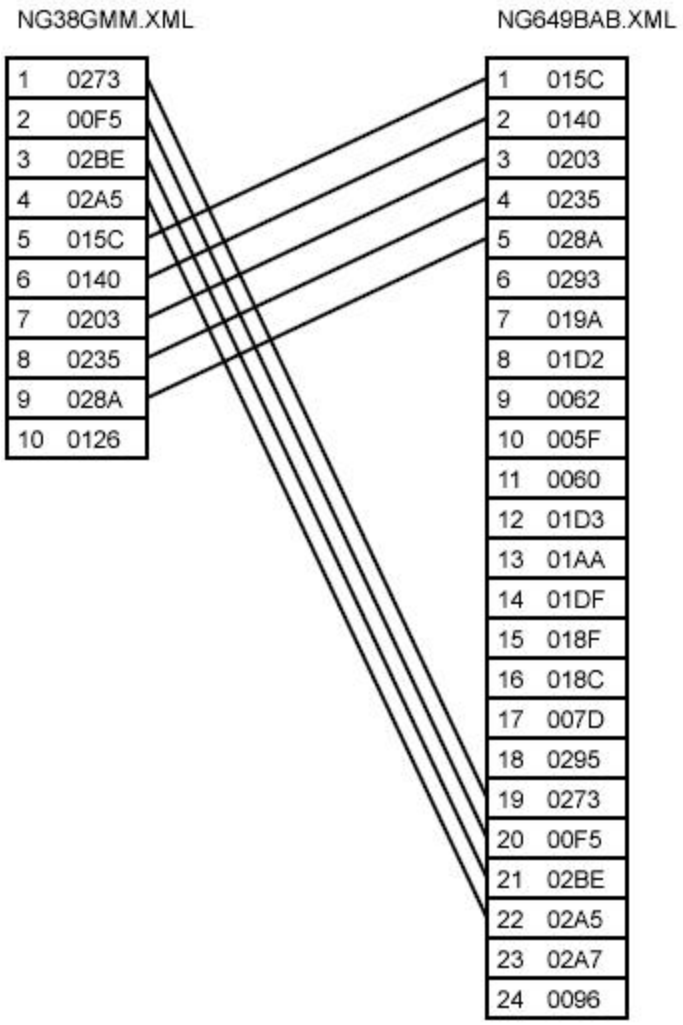
The following plectogram illustrates a variation on a pattern that Olmsted (1994: 113ff) identified as “focusing.” In this plectogram, which compares the contents of AM740DAB (Bucharest State Public Record Office Mixcellany N 740) and AM13225S (Jacimirskij Miscellany N 13.2.25 BAN St Petersburg), almost all articles in AM13225S correspond to articles in AM740DAB in the same relative order. There are no overlapping lines, but several articles in AM740DAB that do not have correspondences in AM13225S are interspersed with those that do. There are two logical interpretations of this pattern:

1. The AM13225S tradition was created when a scribe read through a manuscript of the AM740DAB tradition in order and selected the texts to include.
2. The AM740DAB tradition is an expansion of the AM13225S tradition. In this case the AM740DAB type of compilation would have been produced by copying a text of the AM13225S tradition and inserting additional articles where the scribe considered them necessary or appropriate. Anisava Miltenova, who has worked with both manuscripts *de visu*, has concluded that although the two share a protograph, AM13225S is based on an intermediate antigraph that may have been abridged or defective.



Rearrangement

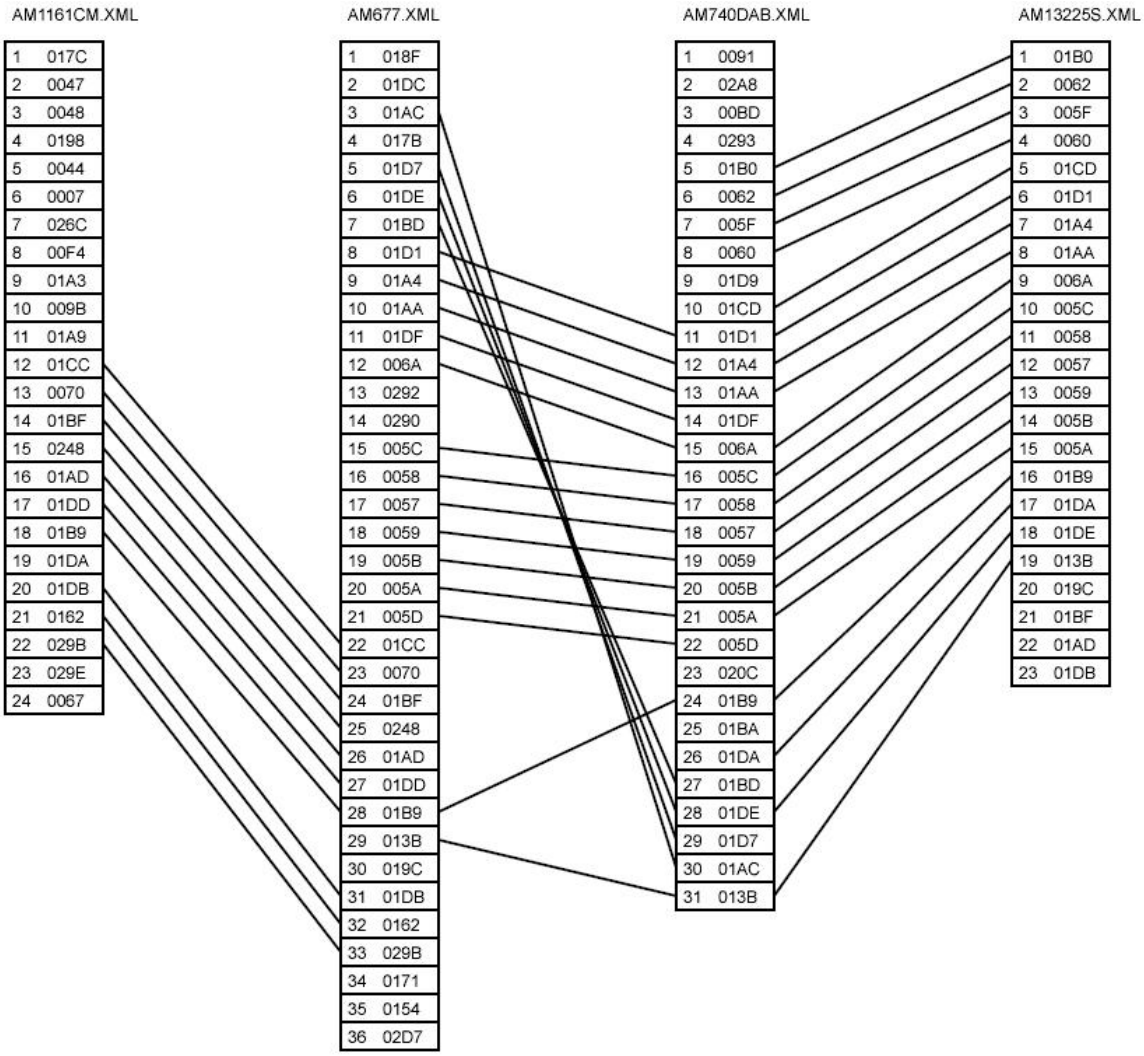
A comparison of NG38GMM (Synodal Miscellany N 38, GIM, Moscow) and NG649BAB (Tulča Miscellany N 649, Romanian Academy of Sciences) reveals a chiastic pattern of correspondences:



This particular pattern results from the rearrangement and rebinding of the contents of NG649BAB after it was originally compiled, and its presence might be considered suggestive of precisely this type of modification.

Compilation

The following plectogram illustrates the textual correspondences among four mixed-content miscellanies: AM1161CM (Apocryphal Miscellany N 1161, Church Historical-Archival Institute), AM677 (Tikveški Miscellany N 677 NBKM), AM740DAB (Bucharest State Public Record Office Mixcellany N 740), and AM13225S (Jacimirskij Miscellany N 13.2.25 BAN St Petersburg).



The relationship between the last two of these is discussed above (Focusing), but the plectogrammatic representation of all four manuscripts suggests two additional hypotheses:

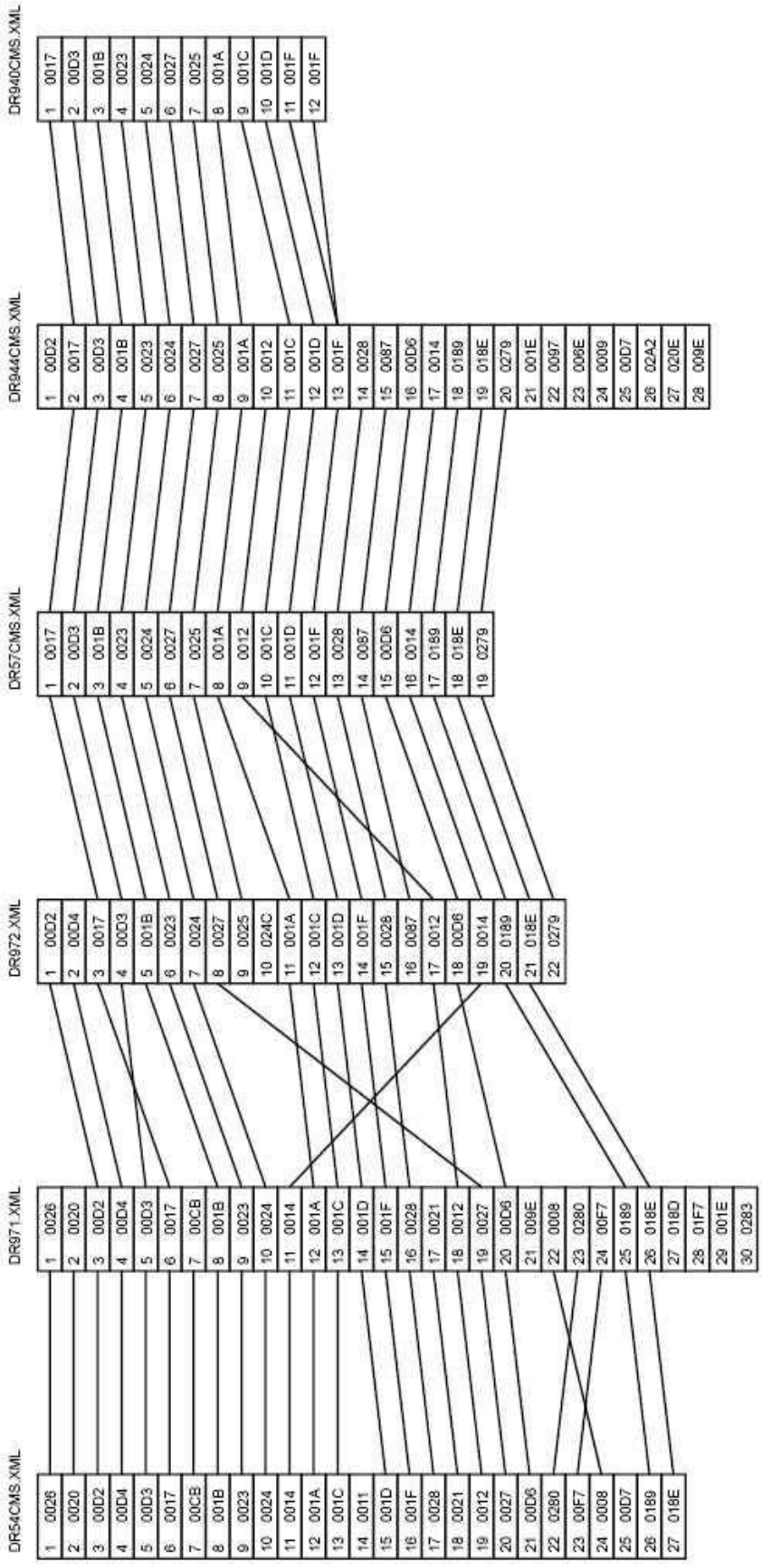
1. AM677 (second from left) is compiled from sources similar to AM1161CM (leftmost), or, perhaps, the protograph of the second half of AM1161CM, on the one hand, and AM740DAB (second from right), on the other. In support of this suggestion, note that most the first twenty-one articles in AM677 have correspondences in AM740DAB, while most of the articles numbered 22 and above have correspondences in AM1161CM. None of the first twenty-one articles in AM677 has a correspondence in AM1161CM, and although two articles numbered 22 and above in AM677 do have correspondences in AM740DAB, the absence of any systematic pattern suggests that these may be accidental.

2. AM677 and AM740DAB share an ancestor, but the two have become disarranged with respect to each other. This hypothesis is compatible with the general chiastic pattern, which could have arisen by moving a set of folios from one part of a manuscript to another.[28]

Six Fixed-Content Miscellanies (Trebniki)

The following plectogram illustrates the textual correspondences within a set of six Trebniki (fixed-content miscellanies): DR54CMS (Trebnik N 54 CIAI), DR971 (Trebnik of Dimit″r Joanovič N 971 NBKM), DR972 (Trebnik of Daskal Filip N 972 NBKM), DR57CMS (Trebnik N 57), DR944CMS (Trebnik N 944 CIAI), and DR940CMS (Trebnik N 940 CIAI):

[28] The completely reversed order of articles #3 and 5–7 of AM677, on the one hand, and 27–30 of AM740DAB, on the other, is poses a more complicated problem. Anisava Miltenova (personal communication) suggests that AM740DAB represents the original order, and that AM677 reflects a deliberate rearrangement of the texts, which were expanded by the addition of article #4, the preface by St. John Chysostom concerning evil women.


DR940CMS.XML
1 0017
2 00D3
3 001B
4 0023
5 0024
6 0027
7 0025
8 001A
9 001C
10 001D
11 001F
12 001F
DR944CMS.XML
1 00D2
2 0017
3 00D3
4 001B
5 0023
6 0024
7 0027
8 0025
9 001A
10 0012
11 001C
12 001D
13 001F
14 0028
15 0087
16 00D6
17 0014
18 0189
19 018E
20 0279
21 001E
22 0097
23 006E
24 0009
25 00D7
26 02A2
27 020E
28 009E
DR957CMS.XML
1 0017
2 00D3
3 001B
4 0023
5 0024
6 0027
7 0025
8 001A
9 0012
10 001C
11 001D
12 001F
13 0028
14 0087
15 00D6
16 0014
17 0189
18 018E
19 0279
DR972.XML
1 00D2
2 00D4
3 0017
4 00D3
5 001B
6 0023
7 0024
8 0027
9 0025
10 024C
11 001A
12 001C
13 001D
14 001F
15 0028
16 0087
17 0012
18 00D6
19 0014
20 0189
21 018E
22 0279
DR971.XML
1 0026
2 0020
3 00D2
4 00D4
5 00D3
6 0017
7 00CB
8 001B
9 0023
10 0024
11 0014
12 001A
13 001C
14 001D
15 001F
16 0028
17 0021
18 0012
19 0027
20 00D6
21 009E
22 0008
23 0280
24 00F7
25 0189
26 018E
27 018D
28 01F7
29 001E
30 0283
DR54CMS.XML
1 0026
2 0020
3 00D2
4 00D4
5 00D3
6 0017
7 00CB
8 001B
9 0023
10 0024
11 0014
12 001A
13 001C
14 0011
15 001D
16 001F
17 0028
18 0021
19 0012
20 0027
21 00D6
22 0280
23 00F7
24 0008
25 00D7
26 0189
27 018E

This plectogram clearly shows the close textual relationship between the two leftmost manuscripts, on the one hand, and among the four rightmost, on the other. Whether the correspondences between the second and third manuscripts, which show several small sets of parallel lines, reflects shared textual transmission or coincidence can be determined by examining whether the clusters of parallel lines traditionally represent a stable part of the tradition (that is, whether those articles may tend to cluster for a specific content-dependent or functional reason, even in manuscripts that are not very closely related).

Enhancements

The programs developed for this report are robust and fully functional, but because they employ command-line utilities, they are not suitable for use by computationally-naïve researchers. Possible eventual enhancements include:

1. Better integration of components. For example, users might wish to move by hypertext links among a list of all correspondences within the corpus sorted by weight, a list of correspondences involving a particular manuscript, tables of contents or other data about particular manuscripts, plectograms comparing particular manuscripts, etc.
2. The integration mentioned above might be supported with a point-and-click graphical interface. For example, the cells in the SVG plectograms currently respond to mouse-over events by changing color and displaying the article title. This interface might be enhanced by having the cells also respond to various click events; for example, clicking on a cell might open a list of manuscripts in the corpus that contain the article represented by the index number in that cell.
3. The present implementation runs on a local platform and accesses existing (static) files, but might be enhanced using client-server technology to generate custom reports. For example, a user might select two manuscripts from a checklist as input for a plectogram and then submit a request to a server that would generate the SVG on the fly and return the results to the user.
4. Comparisons in the present report are undertaken on the level of articles, a term used in the XML description files to indicate the component texts of a miscellany. Some of these component texts may themselves be subdivided into components, so that the makeup of, for example, an *erotapokriseis*[29] or *fiziolog*[30] text within a mixed-content

---

[29] An *erotapokriseis* is a set of questions and answers.

miscellany may itself be analyzed using the same comparative technology.

5. The weighting programs currently operate on pairs of manuscripts, generating a value intended to represent the degree of similarity in the structure of two manuscripts. An alternative approach might evaluate the entire corpus at once using clustering software, thus grouping all members of the corpus simultaneously according to similarity.
6. All Cyrillic text, both modern and medieval, has been recorded in ASCII transliteration, and should be converted to Cyrillic.

Conclusion

This report has demonstrated how standards-based computer technology can be used to support the traditional philological tasks of identifying and studying manuscripts with overlapping content that might reflect a common textual tradition. The two computer applications described here, one of which identifies likely candidates for study and the other of which produces a graphic representation of the correspondences in the contents of a pair of manuscripts, are useful tools for simplifying philological research. Most significantly, the traditional nature of the philological tasks in question demonstrates that humanities computing is not merely a way of doing new things (such as producing electronic editions). It is also, and no less importantly, a way of accomplishing basic and traditional philological research on a scale that would have been impossible (or so impractical as to be virtually impossible) without the aid of computers.

---

[30] A *fiziolog* is a set of descriptions of real and mythological animals.

Works Cited

Gagova, Nina. 1995. NG13317S.XML. Available at http://clover.slavic.pitt.edu/~repertorium/.

Krushelnitskaya, Ekaterina V. 2003. "Description of Medieval Slavic Sources in the Electronic Catalog of the Department of Manuscripts in the National Library of Russia: Goals, Research Problems, and Prospects." In: David J. Birnbaum, Anisava Miltenova, and Sarah Slevinski, eds. *Computational Approaches to the Study of Early and Modern Slavic Languages and Texts.* (= Proceedings of the "Electronic Description and Edition of Slavic Sources" conference, 24–26 September 2002, Pomorie, Bulgaria.) Sofia: Boyan Penev Publishing House, Bulgarian Academy of Sciences. In press. Preprint available at http://clover.slavic.pitt.edu/~repertorium/.

Mathiesen, Robert. 1995. "Magic in Slavia Orthodoxa: The Written Tradition." In: Henry Maguire, ed. *Byzantine Magic.* Washington, DC: Dumbarton Oaks Research Library and Collection. 155–77. Available on line at: http://www.doaks.org/ByzMagic/magic08.pdf.

Miltenova, Anisava. 1986. "K″m metodikata na uzučavane na sbornicite s″s smeseno s″d″ržanie v starite južnoslavjanski literaturi." In: M. Colucci, G. Dell'Agata, and H. Goldblatt, eds. *Studia Slavica Mediaevalia et Humanistica Riccardo Picchio Dicata.* Rome: Edizione Dell-Ateneo. 517–26.

Miltenova, Anisava. 1986a. "Sbornik s″s smeseno s″d″ržanie, delo na etropolskija knižovnik Jeromonax Daniil." *Starob″lgarska literatura* 9. 114–25.

Miltenova, Anisava. 1987. "Apokrifnijat sbornik ot manastira Savina XIV v. v sravnenie s drugi podobni južnoslavjanski r″kopisi." *Arxeografski prilozi* 9. 7–30.

Miltenova, Anisava. 2001. "'Ustroenie na [svetite] slova' v starob″lgarskata literatura." *Starob″lgarska literatura* 32. 99–110.

Olmsted, Hugh. 1977. "Studies in the Early Manuscript Tradition of Maksim Grek's Collected Works." Unpublished doctoral dissertation, Harvard University.

Olmsted, Hugh. 1994. "Modeling the Genealogy of Maksim Grek's Collection Types: The 'Plectogram' as Visual Aid in Reconstruction." In: Michael S. Flier and Daniel Rowland, eds. *Medieval Russian Culture. Volume II.* (= C*alifornia Slavic Studies* XIX.) Berkeley: University of California Press. 107–33.

Radoslavova, Dilyana. 2000. "Problems Concerning the Computer Description and Processing of Slavic Euchological Manuscripts. (A typology of Bulgarian *trebnici* of the 17th Century, Preliminary results.) In:

Anisava Miltenova and David J. Birnbaum, eds. *Medieval Slavic Manuscripts and SGML. Problems and Perspectives.* Sofia: Professor Marin Drinov Academic Publishing House. 170–204.

Appendix A: Corpus Contents[31]

Filename	*Manuscript Title (Bulgarian)*	*Number of Articles*
AA1322	Samokovski sbornik N 1322 NBKM	5
AA1348	Vechen Calendar 1348 NBKM	19
AA308	Apokrifen sbornik 308	27
AA325	Apokrifen sbornik 325 NBKM	17
AA36NBB	Prisrenski Sbornik, Rs 36NB Belgrad	46
AA53NBB	Belgradski apokrifen sbornik Rs 53	23
AA698	Gabrovski sbornik N 698 NBKM	4
AA724	Damaskin N 724 NBKM	5
AA761	Pouchenija za zhenite N 761 NBKM	10
AA771	Sbornik N 771 NBKM	17
AM1-102O	Sbornik ot zhitija	5
AM1-103O	Oktoih	3
AM1-108O	Sbornik N 1/108 ot Odeskata biblioteka	10
AM100MCB	Daniilov sbornik N 100 MSPC Belgrad	44
AM109PAT	Atonski sbornik ot manastira "Sv. Pavel" N 109	8

[31] The 104 XML files used in this project were produced between the mid-1990s and the present by Adelina Anguševa, Desislava Atanasova, Dimitrinka Dimitrova, Margaret Dimitrova, Nina Gagova, Anisava Miltenova, Maja Petrova, Diljana Radoslavova, Ana Stojkova, and Elena Tomova, all of the Institute of Literature of the Bulgarian Academy of Sciences. As of July 2003, these scholars and others associated with the Repertorium Project (http://clover.slavic.pitt.edu/~repertorium/) have encoded descriptions of approximately three hundred manuscripts. For control purposes, the present study includes mixed-content miscellanies, fixed-content miscellanies, and other types of manuscripts.

Filename	*Manuscript Title (Bulgarian)*	*Number of Articles*
AM1161CM	Apokrifen sbornik ot C7rkovnija Institut N 1161	24
AM11ODES	Sbornik N 11 ot sbirkata na V. Grigorovich v Odeskata biblioteka	19
AM13225S	Sbornik na Jacimirski N 13.2.25 BAN Sankt Peterburg	23
AM13410S	Bajchov sbornik N 13.4.10 BAN St. Peterburg	17
AM13613S	Sbornik na Jacimirski N 13.6.13 BAN St. Peterburg	26
AM149NBW	Vienski apokrifen sbornik N 149	39
AM17DUJC	Sbornik N 17 ot Cent7ra za slavjano-vizantijski prouchvanija	28
AM241HLU	Hludov sbornik N 241, GIM, Moskva	23
AM29SAV	Savinski sbornik N 29	38
AM305NBB	Belgradski fragment ot apokrifen sbornik N 305 ot NB Belgrad	4
AM309	Beljakovski sbornik N 309 NBKM	28
AM326	Adzharski sbornik N 326 Nbkm	43
AM38NBB	Belgradski apokrifen sbornik N 38	12
AM39A	Atinski fragment ot apokrifen sbornik N 39	5
AM413BDL	Sbornik ot Bodleian Library N 413	19
AM433	Panagjurski sbornik N 433 NBKM	42
AM52NIK	Nikolashki sbornik N 52	67
AM677	Tikveshki sbornik N 677 NBKM	36

Filename	*Manuscript Title (Bulgarian)*	*Number of Articles*
AM738NBB	Belgradski sbornik N 738	26
AM740DAB	Bukurewtki sbornik N 740 ot D7rzhavnija arxiv	30
AM76NBW	Vienski sbornik N 76	10
AM828NBB	Pribilov sbornik N 828 NB - Belgrad	36
AM82NIK	Nikolashki sbornik N 82	15
AS1052	Sbornik N 1052 NBKM	15
AS1053	Sbornik NBKM 1053	19
AS1055	Pouchitelno sborniche	7
AS22PANT	Sbornik ot manastira Pantelejmon N 22	15
AS26054W	Sbornik N I 26054 ot Universitetskata biblioteka - Viena	23
AS447BAB	Sbornik RAN N 447	4
AS681	Sbornik N 681 NBKM	20
AS685	Koprivwtenski sbornik N 685 NBKM	53
AS9D15MP	Sbornik IX D 15 ot Naroidnija muzej v Praga	9
AS9H10MP	Trebnik IX H 10 ot Naroidnija muzej v Praga	8
DA01	Slepchenski apostol	2
DA02	Slepchenski apostol	1
DA03	Slepchenski apostol	1
DA04	Slepchenski apostol	1
DA05	Slepchenski apostol	1

Filename	*Manuscript Title (Bulgarian)*	*Number of Articles*
DA06	Slepchenski apostol	1
DA51	Vraneshnichki Apostol	1
DA53	Apostol izboren	2
DA82	Shafarikov apostol	2
DA87	Apostol	2
DD1118	Nomokanon N 1118 NBKM	6
DD1170	Nomokanon N 1170 NBKM	1
DD295	Nomokanon N 295 NBKM	1
DD296	Nomokanon N 296 NBKM	3
DD310	Zhitie na Vasilij Novi N 310 NBKM	1
DD312	Zhitie na Vasilij Novi N 312 NBKM	2
DD333	Samokovski sbornik N 333 NBKM	15
DD437	Kotlensko sborniche N 437 NBKM	9
DR124CMS	Trebnik	8
DR149NBP	Trebnik	21
DR248	Trebnik N 248 NBKM	17
DR280CMS	Trebnik N 280 CIAI	14
DR54CMS	Trebnik N 54 CIAI	28
DR57CMS	Trebnik	20
DR621	Trebnik N 621 NBKM	29
DR940CMS	Trebnik N 940 CIAI	12

Filename	*Manuscript Title (Bulgarian)*	*Number of Articles*
DR944CMS	Trebnik N 944 CIAI	28
DR971	Trebnik na Dimit7r Joanovich N 971 NBKM	30
DR972	Trebnik na daskal Filip N 972 NBKM	22
ELENA1	Nikiforov sbornik 1	6
ELENA10	Sbornik s7s sluzhbi i zhitija za sv. Ivan Rilski na Xristaki Pavlovich i trebnik	11
ELENA2	Nikiforov sbornik 2 (Sluzhebnik s7s sluzhba i zhitija na sv. Ivan Rilski)	12
ELENA3	Sbornik ot sluzhbi i zhitija na sv. Ivan Rilski	6
ELENA4	Sbornik s7s sluzhbi i zhitija na sv. Ivan Rilski	3
ELENA5	Sbornik s7s sluzhba i zhitija na sv. Ivan Rilski	4
ELENA6	Sbornik ot kanoni, sluzhbi i zhitija na sv. Ivan Rilski	8
ELENA7	Sbornik ot sluzhbi i zhitija na sv. Ivan Rilski	10
ELENA8	Sbornik ot sluzhbi i zhitija na sv. Ivan Rilski	11
ELENA9	Sbornik s7s sluzhbi i zhitija na b7lgarski svetci (konvoljut)	5
MD1143	Chetirievangelie N 1143 NBKM	4
MD13518S	Kotlenski damaskin N 13.5.18 RAN Peterburg	23
MD838	Vlashki psaltir N 838 NBKM	1
MD845	Chetirievangelie N 845 NBKM	1
MD846	Evangelie N 846 NBKM	2
MD923	Oktoix N 923 NBKM	3

Filename	*Manuscript Title (Bulgarian)*	*Number of Articles*
MP408G	Bdinski Zbornik N 408 ot Universiteta v Gent	14
NG13317S	Lovchanski (Ivan-Aleksandrov) sbornik ot predi 1331, N 13.3.17 ot BAN Sankt Peterburg	13
NG17PANT	Xronika na Georgi Amartol N 17 ot manastira "Sv. Pantelejmon" na Aton	2
NG2BAN	Sofijski psaltir (Pesnivec) na Ivan Aleksand7r N 2 BAN	8
NG2VAT	Xronika na Konstantin Manasij N 2 ot Bibliotekata na Vatikana	2
NG320BAB	Xronika na Georgi Amartol N 320 ot Bibliotekata na Rum7nskata Akademija	3
NG376PBS	Lavrentiev sbornik N F.I.376 GPB Sankt Peterburg	15
NG38GMM	Sinodalen sbornik N 38 GIM, Moskva	10
NG434HIL	Xilendarski sbornik N 434, manastira "Hilendar" na Aton	4
NG649BAB	Tulchanski sbornik N 649 ot bibliotekata na RAN	22
NG97BAB	T7lkuvanie na evangelieto ot Matej ot Teofilakt Oxridski N 97 ot bibliotekata na RAN	5

Appendix B: Edit Distance

Similarity in two organized structures (as different as written texts, on the one hand, and DNA sequences, on the other) is traditionally measured by *edit distance,* which may be defined as the smallest number of changes (insertions, deletions, substitutions) required to transform one structure into the other. According to this model, the similarity of two structures to each other is in inverse proportion to their edit distance, which is to say that if fewer changes are required to transform one into the other, they are more similar.

The process of identifying an algorithm for evaluating the structural similarity of the contents of two mixed-content miscellanies revealed that edit distance for the purpose of comparing these sorts of virtual tables of contents might be different from edit distance as defined for comparing full text strings. The reason for this difference is that the process of editing full text is different from the hypothesized strategy for compiling mixed-content miscellanies.

Consider the following strings of symbols:

Manuscript	**Articles or Words**										**Note**
#1	**A**	**B**	**C**	D	**E**	**F**	**G**	*H*	*I*	*J*	
(Hypothetical)	K	**A**	**B**	**C**	D	**E**	**F**	**G**	L	M	(A–G)
#2	K	**A**	**B**	**C**	N	**E**	**F**	**G**	L	M	3 + 3: A–C, E–G
#3	O	*H*	*I*	*J*	P	**A**	**B**	**C**	Q	R	3 + 3: A–C, H–J

Assume first that we are evaluating the edit distance between #1 and #2, on the one hand, and #1 and #3, on the other, and that the letters under "Articles or Words" represent words in continuous text. Both #2 and #3 share two three-item sequences with #1; #1 and #2 share **ABC** and **EFG** (bolded) and #1 and #3 share **ABC** (bolded) and *HIJ* (italic). Despite the superficially comparable surface differences between #1 and #2, on the one hand, and #1 and #3, on the other, the edit distances are different because one can move from #1 to #2 by passing through the "hypothetical" manuscript, continuing A–G and then changing D in the middle during the transition from the hypothetical manuscript to #2, but no such process is available for the transition from #1 to #3.

If, on the other hand, the letters under "Articles or Words" represent articles in a miscellany, this distinction becomes less significant. The hypothetical strategy discussed in the body of this report for compiling a mis-

cellany does not include changing a single article in the middle of a series, and while this is not impossible, it is difficult to imagine that scribes would frequently have had occasion to copy three works from a source, omit the fourth, inserting a different work from a different source in its place, and then return to the original to copy the next three works.