A Toolbox for Manuscript Analysis

Melanie Gau,[1] Maria Vill,[2] Florian Kleber,[2] Markus Diem,[2] Heinz Miklas,[1] and Robert Sablatnig[2]

[1] Institute of Slavic Studies, University of Vienna. Austria.

[2] Institute for Computer Aided Automation, PRIP, Vienna University of Technology. Austria.

Abstract

Manuscript analysis has long been solely the domain of scientists in the humanities who had to cope with their complex tasks without the aid of specialized tools and result management facilities. In a conjoint philological and computational approach, we have created a toolbox for semi-automated manuscript analysis. Using high resolution digital images, it combines tools for the codicological (and in the case of documents, diplomatic) and paleographical investigation of manuscripts with sophisticated state of the art image processing methods. The results of the individual tools interact with one another and can be stored as images, metadata, and database entries for statistical processing. This interaction produces valuable results for the field of graphemics, the psycholinguistic comparison of human versus machine analysis, computer aided script recognition, and the reconstruction of damaged source material.

Keywords: *codicology, paleography, detail and character analysis, database application*

1 INTRODUCTION

In 1975 a hidden collection of manuscripts (mss.; singular, ms.) was discovered in St. Catherine's Monastery on Mt. Sinai in Egypt. Among them six mss. written in Glagolitic script were found, comprising two highly important fragments, the so-called Glagolitic *Missale Sinaiticum* (Sin. Slav. 5/N) and the new part of the *Euchologium Sinaiticum* (Sin. Slav. 1/N), both dating from the eleventh century and belonging to the Old Church Slavonic canon. Our Austrian Science Foundation project, "Critical Edition of the New Sinaitic Glagolitic Euchology (Sacramentary) Fragments with the Aid of Modern Technologies," centers on the decipherment and reconstruction of these two most exciting finds.

Due to the extremely bad state of the fragments, especially of the Missal, an interdisciplinary working group was set up, consisting of a philological team at the University of Vienna and two technical teams at the Vienna University of Technology and the Vienna Academy of Fine Arts.

Beyond the actual engagement with the mss., this arrangement offers the unique opportunity to combine philological knowledge with image processing methods and develop new routines for ms. research. Digital high resolution (565 dpi) multispectral images, acquired on an expedition to Mt. Sinai in 2007, were registered (aligned onto one another)[1] and provide the basis for our work.

Standard procedures for codicological and paleographical ms. analysis have long been based on a series of laborious manual processes. Electronic data processing opened the field to a new medium, the computer. Its facilities for batch processing and bundling of pertinent procedures simplify these processes. Still, commercial image processing software usually requires complex handling, is often expensive to purchase, and the combination of many different software applications can make holistic results difficult. A specifically created toolbox, on the other hand, combines the relevant procedures and provides interacting complex research methods, while still ensuring clear and straightforward usability for the philologist.[2]

Apart from routine paleographical procedures, this toolbox also contributes to research in the automated recognition and completion of characters and produces additional empirical results for the philological branch of graphetics (script analysis).

We intend the toolbox to be an extensible, modular system that can be expanded to include new requirements as they arise.

[1] Markus Diem and Robert Sablatnig, "Registration of Ancient Manuscript Images Using Local Descriptors," paper presented at the 14th International Conference on Virtual Systems and MultiMedia (VSMM—Dedicated to Cultural Heritage), Limassol, 2008.

[2] Heinz Miklas, "Analysis of Traditional Written Sources with the Aid of Modern Technologies," paper presented at the Conference on Electronic Visualisation and the Arts (EVA), Moscow, 2004.

2 THE TOOLBOX

So far, the toolbox (see figure 1) is based on a MathWorks Matlab (version 7.5.0) Graphical User Interface (GUI), as most of its functions were individually programmed in Matlab in the first place. It includes general functionalities like loading an image, loading the next image in the file, zooming in and out, and panning the image in the display window.

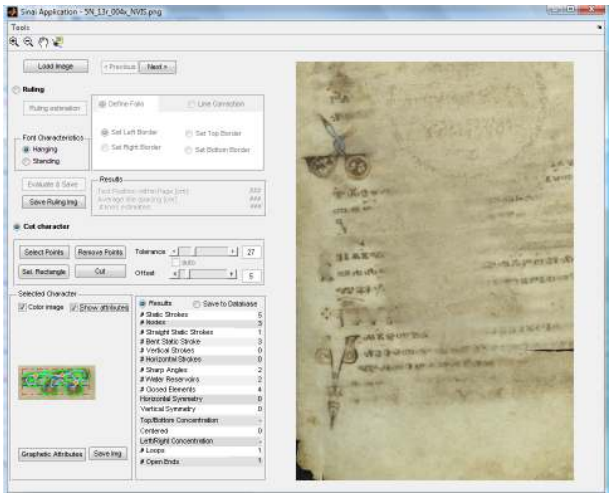

Figure 1. *The toolbox.*

Feature 1: Automatic Page Layout and Text Line Structure Analysis

The layout of ms. pages reveals information about its scribe and spatio-temporal origin. Therefore, layout description not only gives the first clues to the origin of historical mss., but also serves as the basis for the reconstruction of lost textual data. Knowing the layout, both the number of lines and the number of characters in an average line can be reconstructed and make it possible to estimate the missing text.

The challenge of automated layout analysis mainly lies in the corrupted condition of mss., which may be due to text pollution, blur of the background, faded ink, washed out text, degradations of the parchment, and skewed writing material.[1] Since the page segmentation (ruling, margins) on parchment was scratched into the writing material by the scribe, it can appear almost as irregular as the writing itself.

Based on the orientation of the text, the ruling is automatically detected (*line structure analysis*) and the lines are displayed directly on the loaded image (see fig. 2). For faded pages, the line structure is extrapolated according to *a priori* information on the ruling scheme.[2] The developed algorithm has a preprocessing stage, which comprises a skew estimation,[3] an adaptive image binarization, and a noise removal. Subsequently, the text components (words, characters, et cetera) are segmented and finally grouped to extract the text lines,[4] thus providing the basic information for the calculation of the ruling scheme.

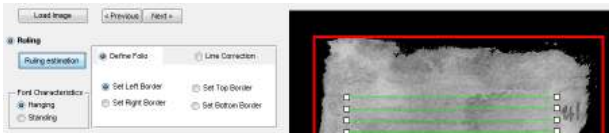

Figure 2. *Ruling estimation and automatic text line extraction.*

If necessary, as in the case of damaged mss., manual corrections and a final fine tuning are possible. The folio frame can be adjusted at the left, right, top, and bottom border of the loaded folio, and the calculated ruling scheme can be corrected by moving, removing, or adding individual lines, by skew correction, and adjustment to the left or right line border. The layout detector also accounts for font position characteristics like hanging (e.g., early Glagolitic script variants) and standing scripts.

The meta information that is calculated from the ruling is shown in the GUI and includes a) the text position within the page (cm); b) the average line spacing (cm); and c) the number of lines estimated. It can be saved in the metadata of the image. So far, the *a priori* information is encoded directly in the program code, but as the number of analyzed layouts grows we plan to implement an option to configure it directly in the GUI.

Feature 2: Cut Character

Another paleographic routine is collating sample alphabets of all scribes (hands) of a ms. for character analysis. This tedious procedure has commonly been

[1]Florian Kleber et al., "Ruling Estimation for Degraded Ancient Documents Based on Text Line Extraction," paper presented at the Conference on Electronic Visualisation and the Arts (EVA) Digital Cultural Heritage—Essential for Tourism, Vienna, 2008.

[2]Heinz Miklas, "Zur editorischen Vorbereitung des sog. Missale Sinaiticum (Sin. Slav. 5/N)," in *Glagolitica. Zum Ursprung der slavischen Schriftkultur*, ed. Heinz Miklas, *Schriften der Balkan-Kommission, Philologische Abteilung / Österreichische Akademie der Wissenschaften, Philosophisch-Historische Klasse* (Vienna: Verl. d. Österr. Akad. d. Wiss., 2000) 117–129.

[3]Florian Kleber and Robert Sablatnig, "A Skew Detection Technique Suitable for Degraded Ancient Manuscripts," paper presented at the 36th Conference on Computer Applications and Quantitative Methods in Archaeology (CAA): On the Road to Reconstructing the Past, Budapest, 2008 (forthcoming).

[4]Florian Kleber et al., "The Sinaitic Glagolitic Sacramentary Fragments," paper presented at the Conference on Electronic Visualisation and the Arts [EVA], Berlin, 2008 [forthcoming]; Florian Kleber et al., "Ancient Document Analysis Based on Text Line Extraction," paper presented at the 19th International Conference on Pattern Recognition [ICPR], Tampa, FL, 2008.

executed with image editing software offering a wide range of features; thus, the process of cutting out characters a) becomes more complicated than necessary and b) requires physical skills like a steady hand and a good eye focus in order not to cut off too much or include unnecessary noise.

The simple but efficient *Cut Character* application combines a variety of image editing steps to cut out elements from a ms. page. Depending on contrast and shape of the glyph, there are two options to select the character (see fig. 3). The user selects the object by clicking once or several times directly into it. The optimal contrast between ink and background can be determined by adjusting the *Tolerance* slider. Each set point can be removed and reset again. In practice, a tolerance of about 10–20% (of an 8-bit grey value image) performs well for all characters in one text section or even a whole page. The slider *Offset* adjusts the border buffering as the number of background pixels that will be cut out together with the glyph. Multi-part objects, for example split characters like the Latin "i" or objects with extremely low contrast between glyph and background, can be encased in an adaptable rectangle.

Figure 3. *Cut character function: points* (left) *and rectangle* (right).

With the command *Cut*, the content of both types of selection, points or rectangle, will be copied into the *Selected Character* display.

Presently we are working on an automatic elimination of background and noise around the selected objects to represent them more clearly. Furthermore, together with each object we will save as metadata its a) coordinates (size), b) position on the page, and c) relation to the text line. Thus, sample alphabets will truthfully reflect the character's position in the ms. layout. This is a major improvement on the current situation, as a simple sequence of glyphs does not necessarily account for the proper line position (especially hanging scripts) and actual size.

Feature 3: Calculation of Character Features

In the box *Selected Character* there are three display options. 1) The *Color Image* option shows the cut out object in its original form, which can be saved without further calculations. 2) The default display shows a binarized (black and white) image of the selected object. The binarization algorithm takes into account the tolerance value of the cut character function. The option *auto* for rectangle selection calculates a tolerance based on the *Otsu algorithm*[1] for the binary image. Optimizing the threshold improves both the binary image and the following calculation steps in quality. 3) The third display option, *Show Attributes*, relates to the calculation of distinctive character features, *Graphetic Attributes*, which we describe below.

While the linguistic analysis of spoken language has long disposed of a feature catalogue for the system of phonology,[2] we are still at the beginning of developing and, what is more, evaluating a standard catalogue for the description of writing systems. Today, the field of script analysis comprises not only paleography, but also graphemics, graphology, forensics, pedagogy, etc.[3]

The inventory for the description of character-based scripts according to linguistic and computational aspects we are using has been developed by H. Miklas since the late 1980s. The formal part of the catalogue[4] consists of two superordinate levels of graphetic character attributes, *static* and *dynamic*. The former characterizes the actual shape of the letter, that is the state *as it is*, whereas the latter focuses on its production and consecutiveness, that is *how the character was realized*.

At this time, only static features of the *Graphetic Attributes* function have been considered, here selecting those that could be implemented with computational methods.[5]

1. Number of static strokes per character
2. Number of nodes per character
3. Number of straight static strokes per character[6]
4. Number of bent static strokes per character
5. Number of vertical static strokes per character

[1]Nobuyuki Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics* 9 (1) (1979).

[2]The description of (distinctive) speech sounds.

[3]Cf., for example, Peter Rück, *Methoden Der Schrift-tbeschreibung* (Stuttgart: Thorbecke, 1999): Introduction.

[4]Heinz Miklas, "Geschriebene Sprachen im Vergleich: Graphematische Modellbildung und slavische Sprach-typologie" (Freiburg i. Br.: 1992 (unpublished)).

[5]See Maria Vill and Robert Sablatnig, "Static Stroke Decomposition of Glagolitic Characters," paper presented at the Conference on Electronic Visualisation and the Arts [EVA] Digital Cultural Heritage–Essential for Tourism, Vienna, 2008.

[6]"Straight and bent strokes are discriminated by means of the formfactor of a skeleton branch. According to our investigation a stroke is to be considered as straight, if the formfactor is larger than 4. It is to be considered as bent, if the formfactor is smaller than 4" (Vill and Sablatnig, "Static Stroke Decomposition of Glagolitic Characters").

6. Number of horizontal static strokes per character
7. Number of loops[1]
8. Number of open ends
9. Number of closed elements per character
10. concentration left[2]
11. concentration right
12. concentration top
13. concentration bottom
14. concentration center[3]

The algorithms for the automatic feature calculation were developed on samples from a professional calligrapher. Automated stroke detection consists of the following procedures: thresholding the image, smoothing the contours with snakes, thinning it to a skeleton, and dissecting it into analyzable segments (numbers 1–9): *nodes* and *strokes*. *Nodes* are defined as crossings of at least three line segments (see fig. 4). Each segment constitutes a single element[4] and is defined as *stroke*. Since these strokes can be classified by their individual features, like straight (number 3) or horizontal (number 6), and strokes with the according features are countable, they give first empirical information on a character.[5]

Figure 4. *Glagolitic character "b" with a static stroke partitioning.*[6]

The *character feature classification* is a prerequisite step for an automatic recognition of those features that will eventually discriminate each character from the others. This prepares the ground for Optical Character Recognition (OCR) and automatic script reconstruction of incomplete, damaged characters.[7]

The results of the automatic feature calculation (see fig. 6) are rendered graphically on the binary image in the *Show Attributes* view as skeleton, nodes, and outline feature graph onto the binary image (see also fig. 5c) and listed in detail in the *Results* box.

The *Save Img* Button saves all three versions of the cut out character: the image in its original form, the binarized image, and the binarized image with the calculated features as a graph representation (fig.5).

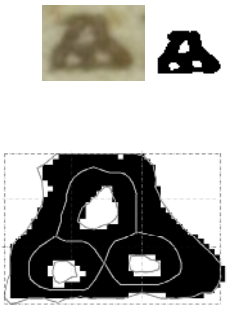

Figure 5. *Character display as a) color image (original), b) default (binarized), and c) show attributes (binarized with feature graph representation) view.*

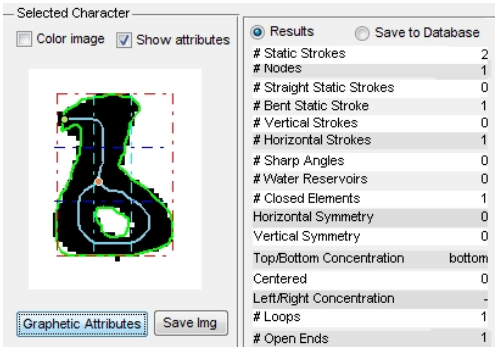

Figure 6. *Calculation of character features.*

Via *Save to Database* these attributes are automatically transferred to a database for graphetic script description.

Feature 4: Character Database

In order to grasp the specific characteristics of the script of a ms. for hand and variants analysis (or for an entire script or subsystem of writing in general), it is necessary to build up a comprehensive character corpus of all hands appearing in the ms. If this corpus is to be used for computational purposes, it must contain a

[1]Features 7 and 8 were not included in the static stroke feature catalogue by H. Miklas, but proved effective for computational analysis. On stroke ending analysis, see Maria Vill and Robert Sablatnig, "Automated Stroke Ending Analysis for Drawing Tool Classification," paper presented at the 19th International Conference on Pattern Recognition (ICPR), Tampa, FL, 2008 and Maria Vill, *Automated Ending Analysis of Drawn Strokes* (Master's thesis, Vienna University of Technology, 2008).

[2]The concentration attributes are derived from the binary image. Here the image is divided into a 3 x 3 grid. If a row or column contains a minimum amount of white pixels compared to the other rows/columns, it is defined as top/bottom or left/right concentrated (cf. Vill and Sablatnig [p. 88 n5]).

[3]A character is defined as centered, if the minimum amount of pixels is contained both in the middle row and the middle column of the grid.

[4]Exception: The crossing of a line segment and a loop (number 7) also constitutes a node (see figure 6), but only two strokes.

[5]See Vill and Sablatnig (p. 88n5).

[6]Ibid.

[7]Heinz Miklas and Melanie Gau, "New Technical Methods for the Study of Damaged Manuscripts," paper presented at the Conference on *Sovremennye informacionnye technologii i pis'mennoe nasledie: ot drevnich tekstov k ėlektronnym bibliotekam*, Kazan', 2008.

representative number of samples of each character in the best possible resolution, that is, the original resolution of the digital image of the ms.
For our tests, we have stored a corpus of characters with about 10 samples per character of each of the three hands in the *Missale Sinaiticum* and a single character set each from the *Euchologium Sinaiticum* and other available Glagolitic mss.

The database for graphetic script description is set up in Microsoft Office Access 2003. So far, there is a selection of English, Russian, and German as operating languages. The database is divided in two main frames. The *Feature Frame* gives general characteristics of the character (for example, *Name*, *Sample Number*, *Folio* number, *Line* number, whether it was processed manually or computationally), as well as all available images (see fig. 7). It also displays our set of character features of the two subcategories *static* (see feature 3) and *dynamic* for each character in the database. So far we have focused mainly on the static features, that is, the actual visual impression of the character.

Each feature is based on the strictly binary distinction *exists* or *does not exist*, but as some features apply to single strokes or segments of one and the same character, they can appear in one object several times. The actual number of occurrences can be distinctive for a character and can be recorded in the database.

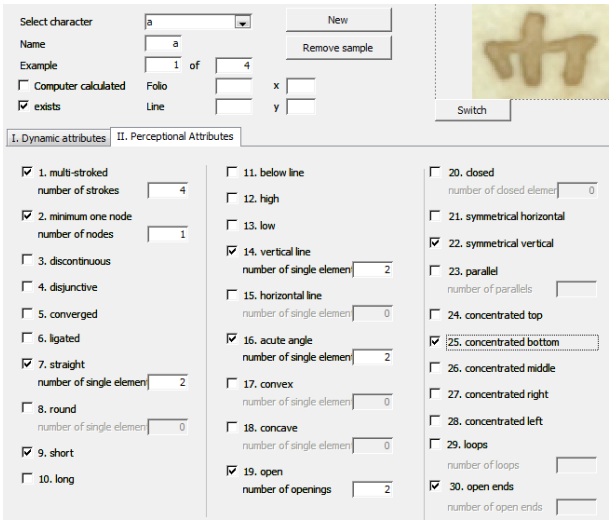

Figure 7. *Feature Frame.*

For automatic primitive analysis via the *Character Feature* calculation of the toolbox (see feature 3), the information can be imported directly into the database. For manual evaluation the user can fill in the character features directly in the GUI and change them at any time. Manual evaluation can also be filled into special Microsoft Word 2003 Excel style sheets to be then imported into the database.

This dual input facility also enables comparison of computer generated and human classifications of the same characters. This is of essential importance for research in the psychology of graphetic perception, because the human mind tends to interpret and recognize patterns subconsciously. In practice this helps us to identify even unusual forms of a character, but may obstruct objective evaluation. The five *concentration* attributes (see feature 3, 10–14 and figure 7, numbers 24–28) especially pose this problem, because here the human perception of the concentration/density of a character shape is biased by a natural conception of its center that may deviate from the empirical center as identified by the computer. Here, the necessity of computational calculations becomes most evident.[1]

The second frame (*Control Frame*) allows setting the infrastructure for new scripts, documents, and hands, and provides selection functions for already stored scripts.

Figure 8. *Control Frame*.

Furthermore, it permits running a number of fundamental queries, statistics, and print out options:

a) *Which characters have the following features...?*–Extracts all characters with the indicated characteristics from a preselected set of mss./hands in one or more scripts.
b) *Differences between hands...*–Enables the direct comparison of two hands. This feature is used for the detailed study of single scripts.
c) *What attributes were used?*–For both manual and computational purposes it is important to know which character features contain relevant information about a script. Thus, unused attributes can be discarded from further automatic script recognition processes.

The print-out options include an overview sheet of the attributes of a certain character, a complete character

[1]Cf. Vill and Sablatnig, "Static Stroke Decomposition of Glagolitic Characters" (p. 88n5).

inventory of a single script, codex or hand, and an overview of the whole content of the database.

Separate Feature: Color Contrast Tool

Concomitant with the other developments, further algorithms have been developed to enhance the readability of latent texts via contrast enhancement. Since this step is crucial for the decipherment of textual material in as bad a condition as one of our mss., this facility was also integrated into our set of tools. Since an immediate visual feedback of color changes in a ms. image is not feasible with a Matlab setup, a standalone tool was developed in Java.

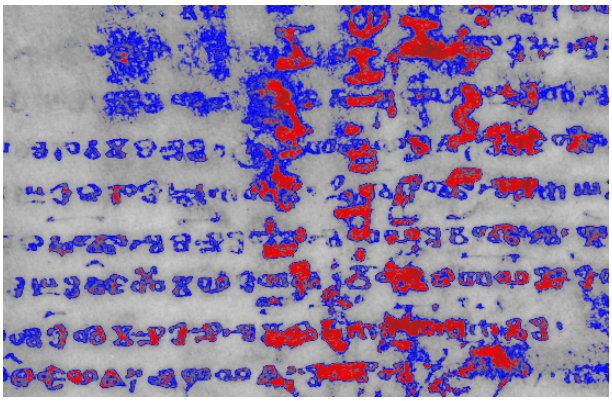

Figure 9. *Color contrasting two scripts of a ms.*

The application allows selecting a specific color of, for example, an (underlying) palimpsest text, the ruling, or other low contrast elements of the ms. By directly clicking into the image the color is selected and can be transformed into a more contrastable false color from a color palette.

3 Conclusion and Outlook

The Toolbox for Manuscript Analysis is a major facilitation of ms. investigation and provides the philologist with a powerful set of new and sophisticated research instruments. It allows the post-processing and reassessment of all related data. Ultimately, it is possible to measure and evaluate codicological, paleographic, and especially graphetic (graphemic) information empirically. Some of its achievements have already proven helpful for further computational script analyses and OCR-development. Nonetheless, there are still certain drawbacks. The toolbox came into being as an aggregation of separately (mostly in Matlab) developed algorithms that were then united. Due to limitations of the Matlab environment regarding modularity, scalability, interaction, and performance, the toolbox and all separate and new tools will have to be joined in a more flexible Java framework.

At the moment, new tools are being developed for a) the automatic extraction of multicolored and exceptionally sized objects, for example (large) initials; b) the automatic layout analysis of more complex layouts, for example multi-column layouts; c) the positioning and combination of fragments; d) the expansion of metadata and their storage in a separate file; and e) amendments to the database output, for example statistical evaluations and graphical results. We also plan to test the tools on a wider range of scripts. Before sharing the toolbox with a larger public the documentation will be extended to a fully fledged manual and sample procedures will be added to the descriptions. Thus we will ensure that even complicated features with a wider range of parameters can be fully exploited.

Acknowledgements

This project is supported by the Austrian Science Foundation (www.fwf.ac.at) under grant P19608-G12.

Bibliography

Diem, Markus, and Robert Sablatnig. “Registration of Ancient Manuscript Images Using Local Descriptors,” paper presented at the 14th International Conference on Virtual Systems and MultiMedia (VSMM)—Dedicated to Cultural Heritage, Limassol, 2008, 188–192.

Kleber, Florian, Martin Lettner, Maria Vill, and Robert Sablatnig. “The Sinaitic Glagolitic Sacramentary Fragments,” paper presented at the Conference on Electronic Visualisation and the Arts (EVA), Berlin, 2008 (forthcoming).

Kleber, Florian, and Robert Sablatnig. “A Skew Detection Technique Suitable for Degraded Ancient Manuscripts,” paper presented at the 36th Conference on Computer Applications and Quantitative Methods in Archaeology (CAA): On the Road to Reconstructing the Past, Budapest, 2008 (forthcoming).

Kleber, Florian, Robert Sablatnig, Melanie Gau, and Heinz Miklas. “Ancient Document Analysis Based on Text Line Extraction,” paper presented at the 19th International Conference on Pattern Recognition (ICPR), Tampa, FL, 2008 (CD publication).

Kleber, Florian, Robert Sablatnig, Melanie Gau, and Heinz Miklas. “Ruling Estimation for Degraded Ancient Documents Based on Text Line Extraction,” paper presented at the Conference on Electronic Visualisation and the Arts (EVA) Digital Cultural Heritage—Essential for Tourism, Vienna, 2008, 79–86.

Miklas, Heinz. "Analysis of Traditional Written Sources with the Aid of Modern Technologies," paper presented at the Conference on Electronic Visualisation and the Arts (EVA), Moscow, 2004: http: //conf.cpic.ru/eva2004/ rus/reports/report_211.html.

Miklas, Heinz. "Geschriebene Sprachen im Vergleich: Graphematische Modellbildung und slavische Sprachtypologie." Freiburg i. Br. (1992): unpublished.

Miklas, Heinz. "Zur editorischen Vorbereitung des sog. Missale Sinaiticum (Sin. Slav. 5/N)," in *Glagolitica. Zum Ursprung der slavischen Schriftkultur*, edited by Heinz Miklas, 117–129. Vienna: Verlag d. Österr. Akad. d. Wiss., 2000.

Miklas, Heinz, and Melanie Gau. "New Technical Methods for the Study of Damaged Manuscripts," paper presented at the Conference on *Sovremennye informacionnye technologii i pis'mennoe nasledie: ot drevnich tekstov k ėlektronnym bibliotekam*, Kazan', 2008, 177–180.

Otsu, Nobuyuki. "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics* 9 (1) (1979): 62–66.

Rück, Peter. *Methoden der Schriftbeschreibung*. Stuttgart: Thorbecke, 1999.

Vill, Maria. *Automated Ending Analysis of Drawn Strokes*, Master's thesis, Vienna University of Technology, 2008.

Vill, Maria, and Robert Sablatnig. "Automated Stroke Ending Analysis for Drawing Tool Classification," paper presented at the 19th International Conference on Pattern Recognition (ICPR), Tampa, FL, 2008 (CD publication).

Vill, Maria, and Robert Sablatnig. "Static Stroke Decomposition of Glagolitic Characters," paper presented at the Conference on Electronic Visualisation and the Arts (EVA) Digital Cultural Heritage—Essential for Tourism, Vienna, 2008, 95–102.