

MAGAZIN FÜR DIGITALE EDITIONSWISSENSCHAFTEN

*Herausgegeben vom Interdisziplinären Zentrum
für Editionswissenschaften
der Friedrich-Alexander-Universität Erlangen-Nürnberg*

Vorstand:

BORIS DREYER
KLAUS MEYER-WEGENER
SONJA GLAUCH

Board:

FLORIAN KRAGL
KLAUS MEYER-WEGENER
WOLFGANG WÜST

4 / 2017

FAU University Press
Magazin für digitale Editionswissenschaften
ISSN 2364-0855

Herausgeber:
Interdisziplinäres Zentrum für Editionswissenschaften
Prof. Dr. Boris Dreyer (Sprecher)
Universität Erlangen-Nürnberg
Department Geschichte
Alte Geschichte
Kochstr. 4, Postfach 8
D-91054 Erlangen

EDITORIAL

Das »Magazin für digitale Editionswissenschaften« versteht sich als ein offenes Forum zur Vorstellung von »best practises« für Online-Editionen. In den fünf- bis zehnsseitigen Darstellungen sollen anhand konkreter Beispiele aus aktuell bearbeiteten Projekten insbesondere die »technische Seite« von Online-Editionen dargelegt werden: von der digitalen Codierung bis hin zu Visualisierungsstrategien, von theoretischen Erwägungen bis hin zu pragmatischen Überlegungen. Im Zentrum stehen fachspezifische Ansprüche, Standards und Methoden bei der Editionsarbeit sowie die verwendeten digitalen Werkzeuge und Präsentationsformen.

Die Mitglieder des Interdisziplinären Zentrums für Editionswissenschaften der Friedrich-Alexander-Universität Erlangen-Nürnberg, die das Magazin tragen, wollen mit diesem Forum einen Beitrag dazu leisten, dass Kriterien für Online-Editionen in der Öffentlichkeit zur Diskussion gestellt und die Anwendung derselben Editionswerkzeuge in verschiedenen Projekten vergleichend gegenübergestellt werden können. Auf diese Weise erfolgt eine Systematisierung der genutzten Editionsmittel und es wird möglich, gemeinsame Standards auf dem immer breiter werdenden Schnittfeld zwischen Philologie und Informationstechnik zu entwickeln.

Die Herausgeber

INHALT

STEFAN FIEL UND ROBERT SABLATNIG

Writer Identification on Historical Glagolitic Documents

7

FABIAN SCHWABE

Im stillen Kämmerlein.

Zur digitalen Leseausgabe der

›Hrafnkels saga freysgoða‹

21

WRITER IDENTIFICATION ON HISTORICAL GLAGOLITIC DOCUMENTS

STEFAN FIEL UND ROBERT SABLATNIG

ABSTRACT

This work aims at automatically identifying scribes of historical Slavonic manuscripts. The quality of the ancient documents is partially degraded by faded-out ink or varying background. The writer identification method used is based on image features, which are described with Scale Invariant Feature Transform (SIFT) features. A visual vocabulary is used for the description of handwriting characteristics, whereby the features are clustered using a Gaussian Mixture Model and employing the Fisher kernel. The writer identification approach is originally designed for grayscale images of modern handwritings. But contrary to modern documents, the historical manuscripts are partially corrupted by background clutter and water stains. As a result, SIFT features are also found on the background. Since the method shows also good results on binarized images of modern handwritings, the approach was additionally applied on binarized images of the ancient writings. Experiments show that this preprocessing step leads to a significant performance increase: The identification rate on binarized images is 98.9%, compared to an identification rate of 87.6% gained on grayscale images.

Keywords: Writer identification, historical documents

INTRODUCTION

This work deals with the automated identification of writers of ancient manuscripts. Currently, paleographers are performing this task mainly manually in order to localize, date or authenticate historical writings.^{1, 2} A large number of historical documents has been digitized in the past decade and has been made accessible to a growing number

- 1 Itay Bar-Yosef, Isaac Beckman, Klara Kedem and Itshac Dinstein: Binarization, character extraction, and writer identification of historical hebrew calligraphy documents, in: IJDAR 9 (2007), p. 89–99.
- 2 Marius Bulacu and Lambert Schomaker: Automatic handwriting identification on medieval documents, in: 14th International Conference on Image Analysis and Processing (ICIAP 2007), Modena 2007, p. 279–284.

of users – for example see Wolf et al.³ or Easton et alii⁴. By automating the task of writer identification, it can be applied to a vast amount of historical documents and thus become a valuable tool for paleographers.

The historical manuscripts investigated in this work originate from the 10th to 11th centuries and are written in Glagolitic, the oldest Slavic script.⁵ Writings belonging to five different manuscripts have been examined: Three books have been imaged at Mt. Sinai⁶ and the remaining writings were photographed in libraries in Austria and Italy. Scholars found that the investigated manuscript leaves were written by seven scribes, whereby two manuscripts were written by several different hands, while another manuscript, which today is stored in two parts in different libraries, was written by one single scribe. In Figure 1 three examples of the manuscripts investigated are shown.

Recently, we proposed a writer identification method⁷ that has been designed for Latin texts. The aim of the current work is to evaluate the applicability of this approach to the Glagolitic writings examined. To the best of our knowledge this is the first time that a writer identification approach is applied to Glagolitic handwritings. Compared to modern handwritings, the scribe identification task is complicated by the circumstance that the writings are partially in bad condition, since characters are faded-out and the documents are degraded by background clutter.

While the majority of the writer identification approaches deals with modern handwritings,⁸ recently several approaches have been applied to historical writings: Bensafia et al.⁹ use graphemes as features for the identification task. The approach is applied on a database consisting of modern handwritings, as well as on a data set, which contains

3 Lior Wolf, Lior Litwak, Nachum Dershowitz, Roni Shweta and Yaacov Choueka: Active clustering of document fragments using information derived from both images and catalogs, in: 2011 International Conference on Computer Vision (ICCV), Barcelona 2011, p. 1661–1667.

4 Roger L. Easton, William A. Christens-Barry and Keith T. Knox: Spectral Image Processing and Analysis of the Archimedes Palimpsest, in: 19th European Signal Processing Conference (EUSIPCO 2011), Barcelona 2011, p. 1440–1444.

5 Heinz Miklas: Zur editorischen Vorbereitung des sog. Missale Sinaiticum (Sin. slav. 5/N), in: *Glagolitica – Zum Ursprung der slavischen Schriftkultur*, ed. by Miklas, H., Sadovski, V., and Richter, S., Wien 2000 (OAW, Phil.–hist. Kl., Schriften der Balkan-Kommission, Philologische Abt. 41), p. XV–XVI, 117–129.

6 Martin Lettner, Markus Diem, Robert Sablatnig and Heinz Miklas: Registration and Enhancing of Multispectral Manuscript Images, in: 16th European Signal Processing Conference (EUSIPCO08), Lausanne 2008, p. 1–5.

7 Stefan Fiel and Robert Sablatnig: Writer Identification and Writer Retrieval using the Fisher Vector on Visual Vocabularies, in: 2013 International Conference on Document Analysis and Recognition, Washington, DC 2013, p. 545–549.

8 Bulacu and Schomaker [n. 2], p. 279–284.

9 Ameur Bensafia, Thierry Paquet and Laurent Heutte: Information retrieval based writer identification, in: Seventh International Conference on Document Analysis and Recognition (ICDAR 2003), 2003, p. 946–950.



Figure 1. Examples taken from the dataset. (From left to right) Euchologium Sinaiticum folio 22 recto (Writer 3), Psalterium Demetrii Sinaitici folio 10 recto (Writer 2), Codex Clozianus folio 8 verso (Writer 7). Portions taken from the same images, from top to bottom: Writer 3, Writer 2, Writer 7.

documents written by 39 scribes and originating from the 19th century. The results obtained on the historical database are significantly lower compared to the results gained on modern documents, which can be attributed to the presence of noise and slant amongst others.

Bulacu and Schomaker¹⁰ propose a writer identification system that combines textual and allographic features. The approach is tested on 70 medieval English documents that were written by 10 different scribes. The authors show that the combination of the feature groups leads to an increased performance.

Brink et al.¹¹ suggest to use the width of the ink trace along with directionality measurements as features for writer identification purposes. The approach is evaluated on two historical handwriting datasets of English and Dutch documents. Similar to the approaches mentioned above the dataset consists of manually cropped manuscript images.

Yosef et alii¹² note that historical documents are generally in a poor condition, which impedes a proper binarization. Therefore, the authors propose a multi-stage binarization approach that is especially designed for ancient manuscripts. Afterwards, several selected letters are automatically found and used for the writer identification, which is based on style analysis. For the classification task, K-nearest neighbors and Linear Bayes classifier are compared, whereby the latter mentioned is better suited for classification.

While the approaches mentioned above are applied on binarized document images, Bres et alii¹³ perform writer identification on grayscale images. Therefore, the Hermite transformation is used for denoising and identification of handwritings. The system has

¹⁰ Bulacu and Schomaker [n. 2], p. 279–284.

¹¹ Axel Brink, Jorrit Smit, Marius Bulacu and Lamert Schomaker: Writer identification using directional ink-trace width measurements, in: Pattern Recognition 45 (2012), p. 162–171.

¹² Yosef, Beckman, Kedem, Dinstein [n. 1], p. 89–99.

¹³ Stéphane Bres, Véronique Eglín and Catherine Volpillac-Augér: Evaluation of Handwriting Similarities Using Hermite Transform, in: Tenth International Workshop on Frontiers in Handwriting Recognition, ed. by Guy Lorette, La Baule 2006. <http://www.suvisoft.com> Université de Rennes 1.

been tested on 1438 historical documents stemming from 189 different writers, whereby the writings are written in different languages and alphabets.

Recently, Wolf et alii¹⁴ proposed an unsupervised approach that is related to writer identification: In order to detect historical documents that have been taken from the same book, but have been dispersed, a semi-automatic clustering method is suggested. The approach is based on graphical models and image similarities are determined by using a Bag of Words approach. By using their technique about 1000 new connections between so far unrelated manuscript folios could be found and verified by scholars. Although the authors note that their approach is not entirely suited for writer identification, this result proves the valuable support that image processing techniques can provide for the paleographer.

On modern handwritings Li and Ding¹⁵ proposed Grid Microstructure Features for Writer Identification. First the edges of the handwriting are extracted and on each pixel of the border the neighborhood is described. A feature vector is generated with the probability distribution of the different pixel pairs. This method won the »ICDAR 2011 Writer Identification Contest«.¹⁶ Jain and Doermann¹⁷ are using K-adjacent segment features in a Bag of Words framework for writer identification. The relationship between sets of neighboring edges in an image is represented.

The current work is structured as follows. In Section 2 the methodology is introduced and in Section 3 the performance of the system is evaluated. Finally, in Section 4 a conclusion is drawn.

METHODOLOGY

As preprocessing step the document images are cropped to regions containing solely text. It should be noted that the manuscripts have been imaged in different places and their conditions are varying considerably. Thus, the background is heterogeneous – for example on several images there is a color chart in the background – or on other pages only a minor region of the parchment contains legible text, whereas the remaining characters are faded out. In order to remove the background, but also parchment regions

14 Wolf, Litwak, Dershowitz, Shweka, Choueka [n. 3], p. 1661–1667.

15 Xin Li and Xiaoqing Ding: Writer Identification of Chinese Handwriting Using Grid Microstructure Feature, in: *Advances in Biometrics*, ed. by Massimo Tistarelli and Mark S. Nixon, Berlin / Heidelberg 2009 (Lecture Notes in Computer Science 5558), p. 1230–1239.

16 Georgios Louloudis, Nikolaos Stamatopoulos and Basilis Gatos: ICDAR 2011 Writer Identification Contest, in: 2011 International Conference on Document Analysis and Recognition, Beijing 2011, p. 1475–1479.

17 Rajiv Jain and David Doermann: Offline Writer Identification Using K-Adjacent Segments, in: 2011 International Conference on Document Analysis and Recognition, Beijing 2011, p. 769–773.

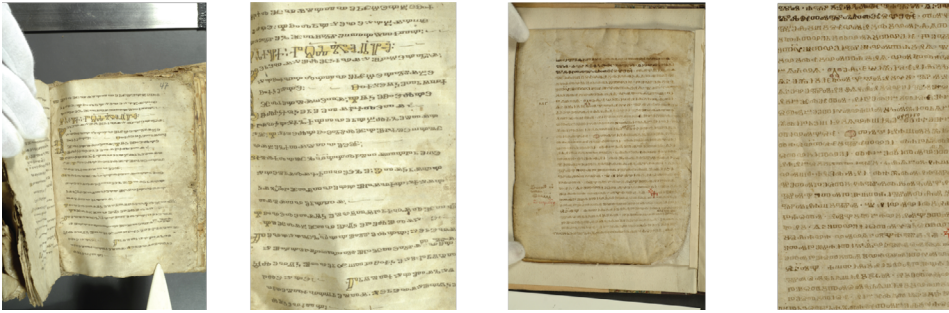


Figure 2. Input and result images of the cropping procedure. (From left to right) Psalterium Demetrii Sinaitici folio 47 recto (Writer 1). Corresponding result image. Codex Marianus folio 2 verso (Writer 6). Corresponding result image.

containing no characters or text with a considerably low contrast, the following simple approach is used for the extraction of the main text block:

Text lines are found by using a text line detection method similar to the one proposed by Yosef et alii:¹⁸ First, Local Projection Profiles (LPP) are applied to the image considered. Afterwards, the LPP image is filtered with a Gaussian column kernel and zero crossings of its first derivative are found. In a final step, a non-extremum suppression is performed in order to remove false positives. The zero crossings found are located at local minima and maxima and the minima encode the text lines. Afterwards, text lines with a length smaller than a predefined threshold are rejected and the bounding box containing the residual text lines is used for the cropping of the text region. Two examples for the cropping procedure are given in Figure 2. It can be seen that the main text block of both manuscripts is successfully extracted, but it has to be mentioned that only rectangular regions are extracted from the images. In order to reject for example decorative elements – like initials – a more sophisticated layout analysis technique would be more appropriate.

For the task of writer identification the method by Fiel and Sablatnig¹⁹ is used. This method is designed to work without binarization, on pages with uniform background, modern handwriting, and if there is not only handwriting on the page, with a segmentation of areas which contain handwriting. In contrast, the manuscripts examined are corrupted by background clutter, faded-out ink, and water stains. Since experiments²⁰ have shown that when applying the method on binarized images it can keep up with other state of the art methods, it is also applied to binarized images of the dataset. This method is described shortly afterwards.

18 Itay Bar-Yosef, Nate Hagbi, Klara Kedem and Itshac Dinstein: Line segmentation for degraded handwritten historical documents, in: Proceedings of the 2009 10th International Conference on Document Analysis and Recognition (ICDAR 2009), Washington, DC 2009, p. 1161–1165.

19 Fiel and Sablatnig [n. 7], p. 545–549 (2013).

20 Fiel and Sablatnig [n. 7], p. 545–549 (2013).

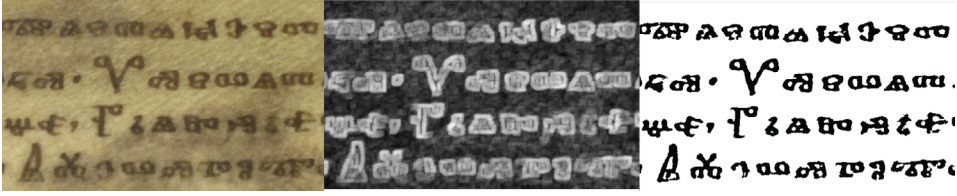


Figure 3. Binarization of a manuscript portion. (Left) Input image. (Middle) Contrast image. (Right) Binarization result.

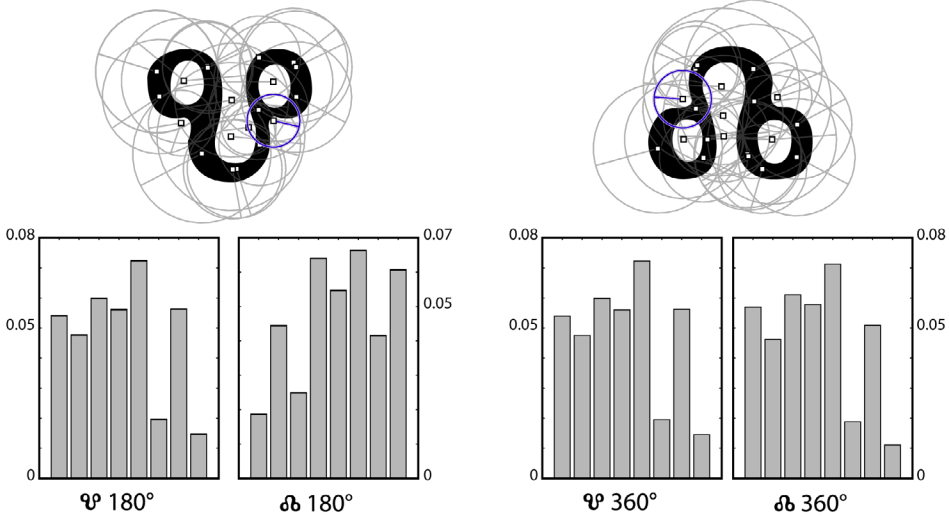


Figure 4. Two Glagolitic characters v and d and their corresponding SIFT features. The second row are the generated histograms of the two marked features (colored blue). In the bottom row on the right the SIFT features are calculated rotational invariant, thus they both generate a similar histogram. On the left side the features are calculated rotational dependent, which makes the two characters distinguishable. By courtesy of Diem and Sablatnig.²¹

The Glagolitic writings are suffering from the aforementioned degradations, and thus the images are binarized by employing a binarization approach suited for historical documents that has been proposed by Su et alii²² The algorithm starts with a calculation of a contrast image, whereby the pixels in this contrast image encode normalized intensity differences between the maximum and minimum gray levels within a local neighborhood.

21 Markus Diem and Robert Sablatnig: Recognizing Characters of Ancient Manuscripts, in: Proceedings of IS&T SPIE 7531 (2010), p. 7531.

22 Bolan Su, Shijian Lu and Chew Lim Tan: Binarization of historical document images using the local maximum and minimum, in: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, Boston, MA, 2010, p. 159–166.

Afterwards, high contrast pixels, which are located at stroke boundaries, are found by applying a global Otsu²³ threshold. Those high contrast pixels are used in the final binarization step: A pixel is classified as a foreground pixel, if the following two requirements are met: First, the pixel should be in the near of a predefined number of high contrast pixels. Second, the pixel intensity must be smaller or equal than the mean intensity of the high contrast pixels in a local neighborhood window. An example for the binarization of a Glagolitic writing is given in Figure 3. It can be seen that the majority of the characters is successfully segmented, although the parchment portion is corrupted by clutter and the foreground to background contrast is varying.

The method for writer identification is based on the Fisher Kernels, introduced by Perronnin and Dance²⁴ and improved by Perronnin et alii,²⁵ which are calculated on Visual Vocabularies. The first step is the application of the Scale Invariant Feature Transform (SIFT) from Lowe.²⁶ These features have been modified by mirroring the angle of the keypoint, if the angle is larger than 180 degrees, proposed by Diem and Sablatnig²⁷ for character recognition of Glagolitic characters. Figure 4 shows two Glagolitic characters and their SIFT features. It can be seen that the interest points are located in the middle of the circles and at the corners. Also the down sampled histograms, rotation invariant on the right side and rotationally dependent on the left side, of the highlighted SIFT features are shown. When the features are calculated rotation invariant, they are not distinguishable. In contrast, when the features are calculated rotational dependent they generate different histograms, allowing the distinction whether the features are located at the upper or the lower profile of the writing. It has been shown that the upper and lower profile of a writing is a discriminative feature for writer identification. Since the background contains a lot of noise the contrast threshold for the SIFT features has been set to a higher level and also the edge threshold is lowered to reduce the number of edge-like features – compared to the parameters used in Fiel and Sablatnig.²⁸

After the calculation of the SIFT features the visual vocabulary is generated. This is done on a separate training set to ensure the independence of the writers in the evaluation set. For performance reasons a Principal Component Analysis (PCA) is applied on

23 Nobuyuki Otsu: A threshold selection method from grey-level histograms, in: IEEE Transactions on Systems, Man and Cybernetics 9 (1979), p. 62–66.

24 Florent Perronnin and Christopher Dance: Fisher Kernels on Visual Vocabularies for Image Categorization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007 (CVPR 2007), Minneapolis, MN, 2007, p. 1–8.

25 Florent Perronnin, Jorge Sanchez and Thomas Mensink: Improving the fisher kernel for large-scale image classification, in: Proceedings of the 11th European Conference on Computer Vision: Part IV (ECCV'10), Berlin/Heidelberg 2010, p. 143–156.

26 David G. Lowe: Distinctive Image Features from Scale-Invariant Keypoints, in: International Journal of Computer Vision 60 (2004), p. 91–110.

27 Diem and Sablatnig [n. 21], p. 7531.

28 Fiel and Sablatnig [n. 7], p. 545–549.

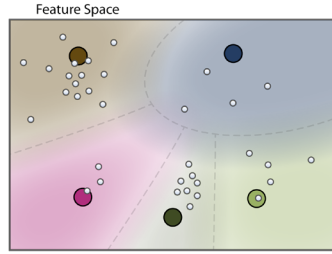


Figure 5. Schematic partitioning of the feature space with k-means (dashed lines) and with GMMs (colors). Since GMMs do not have strict borders the features space can be described more precisely.

the features of the training set to reduce the dimensionality from 128 to 64. The visual words are represented by means of Gaussian Mixture Models (GMM). The amount of the GMMs has to be set in advance. Experiments have shown that for this task the best number of GMM is 100. It is assumed that the generation process of all SIFT features in all images of the training set can be modeled by a probability density function.²⁹ The parameters of this density function can be estimated by an Expectation-Maximization (EM) algorithm. The EM algorithm estimates iteratively the three parameters of the different Gaussians. The advantage of using GMM for visual vocabularies instead of k-means, which is normally used for Bag of Words, is that the feature space can be described more precisely (see Figure 5). When using k-means for clustering strict borders are introduced and the distance of the feature to the cluster center and influences from other cluster centers are not taken into account. Since the GMM does not have any strict borders it overcomes these problems. After estimating the parameters for the GMM the feature vector for each image can be generated. This is done by calculating the SIFT features $X = \{x_t, t = 1 \dots T\}$ for one image and then applying the PCA with the parameters calculated for the training set. Afterwards the Fisher Kernel is applied. It is computed by³⁰

$$G_k^X = \frac{1}{\sqrt{w_k}} \sum_{t=1}^T P(k|x_t) \left(\frac{x_t - \mu_k}{\sigma_k} \right)$$

where G_k^X is the feature vector for one specific distribution k . w_k are the weights of the k -th distribution, μ and σ are the means respectively the variation of the particular distribution. The results for all distributions are then concatenated, resulting in a feature vector of the image with ND -dimensions, where N is the number of distributions and D the dimension of the SIFT features (for this task $N = 100$ and $D = 64$ since we apply the PCA). As proposed by Perronnin et alii the vector is additionally raised by the power of 0.8. They also showed that the cosine distance is a natural measure of similarity for the Fisher Vector.

²⁹ Perronnin, Sanchez and Mensink, T [n. 25], p. 143–156.

³⁰ Perronnin, Sanchez and Mensink, T [n. 25], p. 143–156.

EVALUATION

The dataset for the evaluations contains 361 images with Glagolitic writing on it. Seven different writers have been identified on this pages.³¹ Table 3 shows the distribution of the writers in the dataset - along with the associated manuscripts. It can be seen that by far the most documents in the dataset were written by Writer 1 (207) whereas Writer 6 has only 3 pages. With 58 documents Writer 5 has the second most documents in the dataset, followed by Writer 3 with 40 and Writer 7 with 24. Writer 2 has 22 documents. The documents of Writer 5 are from two different manuscripts and the documents of Writer 7 are stored at two different locations. Additionally, the approximated average character height (in pixel) and the corresponding standard deviation (std) are provided in Table 3, in order to enable a comparison between the character resolutions of each writer. The character height per writer is approximated by averaging the median height of the segmented characters within a binarized page. It should be noted that this approximation is dependent on the performance of the binarization method applied. It can be seen that character reso-

Table 1. Distribution fo writers in the dataset.

Writer Id:	1	2	3	4	5	6	7
Psalterium Demetrii Sinaitici:	207	22	40				
Euchologium Sinaiticum:					7		
Missale Sinaiticum:				7	51		
Codex Marianus:						3	
Codex Clozianus:							24
Total:	207	22	40	7	58	3	24
Average Character height:	27.7	31.5	32.3	34.4	28.8	29.3	21.4
Std. of character height:	2.5	5.0	3.8	9.5	4.2	1.5	3.4

lution of the pages belonging to Writer 7 is considerably lower compared to the remaining writers. Additionally, the character height is partially varying within a page due to decorative elements – such as initials – and warping effects. Exemplar pages with varying character heights can be seen in Figure 1. For generating the visual vocabulary 8 document fragments of unknown writers are used.

For the evaluation of the writer identification the method of the ICDAR 2011 Writer Identification Contest³² has been used. Each document in the database is taken as reference document and the distance to all other documents are calculated. Then two criteria,

³¹ Miklas, [n. 5], p. XV–XVI, 117–129.

³² Louloudis, Stamatopoulos and Gatos [n. 16], p. 1475–1479.

namely the soft and the hard criterion, are evaluated. For the soft criterion the first N most similar documents are observed. If one of these documents is from the same writer as the reference document then it is considered as a correct hit. The result is the percentage of the correct hits for all documents. For the hard criterion all N documents have to be from the same writer. The soft criterion is evaluated with N equals 1, 2, 5, and 10. For the hard criterion N is 2, 3, and 4. It has to be noted, that since Writer 6 has only 3 documents in the dataset the hard criterion with N equals 4 cannot be satisfied. These evaluation methods have been carried out on the cropped dataset and also on the binarized cropped dataset.

Table 2. Evaluation of the soft criterion (in %)

	Top 1	Top 2	Top 5	Top 10
grayscale dataset	87.6	89.8	92.5	93.6
binarized dataset	98.9	98.9	99.7	100.0

Table 2 shows the evaluation of the soft criterion on the dataset. It can be seen that the method performs better on the binarized dataset. The maximal difference is 11.3% for the Top 1, which means that the writer of 40 additional documents is identified correctly when using binarized data. Since the images have a non-uniform background, SIFT features are not only located on the writing itself, but also on the background, holes in the document, and border of water stains. Also because only a simple cropping method is used parts of the book cover can occur in the images and SIFT features are also calculated there. These features influence the performance on the non binarized dataset. In contrast, by binarizing a document image parts of the image which are not text are deleted.

Table 3. Evaluation of the hard criterion (in %)

	Top 2	Top 3	Top 4
grayscale dataset	80.9	78.1	75.3
binarized dataset	98.1	95.3	93.6

Table 3 shows the evaluation of the hard criterion. Again, the method performs better on the binarized dataset. The maximal difference is 18.3% for N equals 7. Like when using the soft criterion features, which are not located on the writing itself influence the performance on the non binarized dataset. This time the influence is greater, since all 4 most similar documents have to be from the same writer. Another possible effect is that documents with water stains are more likely to be considered as similar because of features which are located on the boarder of these stains.

Figure 6 shows plots of the different distances for all documents to two reference documents. The distances are calculated on the binarized dataset. For the left plot the reference document is page number 10 and for the right plot page number 353. The vertical

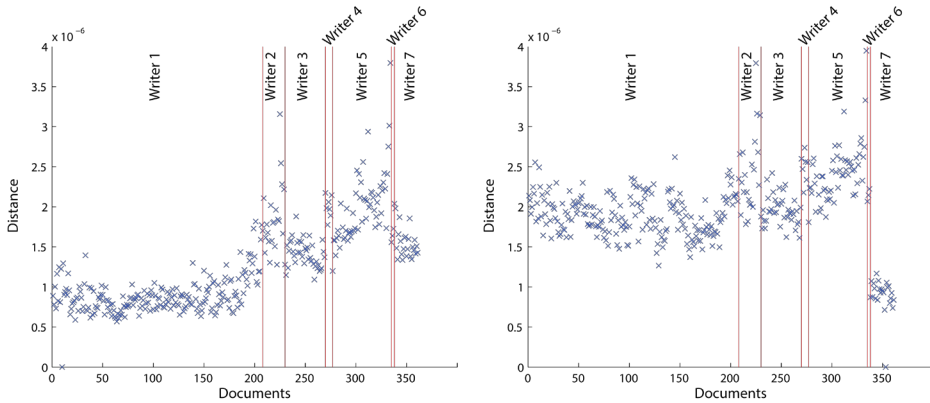


Figure 6. Distances of all images to two reference documents for the evaluation on the binarized dataset (left: page number 10, right: page number 353). The vertical red lines indicate a change of the writer.

red lines indicate a change of the writer. The reference document has a distance of 0 to itself. For all other documents the distance is calculated using the cosine distance. It can be seen that documents written by the same writers as the particular reference document have a smaller distance than documents from other writers. There are also some outliers in the distances of other writers, which are pages where the binarization does not give exact results and parts of characters are lost or the border of the characters cannot be determined exactly.

Table 4 shows the mean precision per writer with varying N . For all documents of one writer the percentage of the documents in the first N in the ranking is calculated. For Writer 4 and Writer 6 the evaluation was skipped if N was higher than the number of their documents in the dataset.

Table 4. Mean precision of each writer with varying N . Cells which have no text are skipped because the corresponding writers do not have enough documents in the dataset for evaluation.

Top- N	Writer Id						
	1	2	3	4	5	6	7
1	100	90.1	100	100	96.6	100	100
2	100	88.6	97.5	100	96.6	100	100
3	100	90.1	96.7	100	93.7		100
5	99.9	83.4	97	100	93.1		100
10	99.9	77.7	94.3		91		100
15	99.9	61.2	91.5		86.8		100
20	99.8	51.4	89.4		82.1		98.8

The documents of Writer 1 have a big influence on the result of the evaluation because, as stated above, the largest contingent of documents was written by this scribe. Thus, a further evaluation has been carried out. This time the documents of Writer 1 are not ta-

ken as reference document, but the corresponding images remain in the dataset, so these documents can still occur in the most similar pages of the other writers.

Table 5 and Table 6 are showing the results of the evaluation on the databases where the documents of Writer 1 are not considered as reference document. Compared to the results with Writer 1 every evaluation has worse results. On the binarized dataset the performance drops 2.5% for the soft evaluation, for the hard evaluation the maximal drop is 7.9%. As expected, the fact that Writer 1 has by far the most documents in the dataset has an influence on the results, but at least for the evaluation on the binarized dataset the influence is tolerable. When regarding the evaluation on the grayscale dataset the difference is significant. For the soft Top 1 evaluation the performance drops by 16.7% compared to the soft evaluation on the grayscale dataset with Writer 1, but compared to the results on the binarized dataset the difference is 19.9% worse. For the hard evaluation this gap increases. That means on the grayscale images the method has performed better on the images from Writer 1 than on the other images. The reason for this is, that the manuscript where the images of Writer 1 originate from has, compared to the other manuscripts, the most uniform background and also the ink is not faded out or blurred. Due to the binarization these factors do not influence the results on the binarized dataset, here the background is removed and the characters are clearly visible.

Table 5. Evaluation of the soft criterion (in %) without the pages of Writer 1 as reference documents

	Top 1	Top 2	Top 5	Top 10
grayscale dataset	70.8	76.0	82.5	85.1
binarized dataset	97.4	97.4	99.4	100.0

Table 6. Evaluation of the hard criterion (in %) without the pages of Writer 1 as reference documents.

	Top 2	Top 3	Top 4
grayscale dataset	55.2	48.7	42.2
binarized dataset	95.5	89.0	85.7

One example for the correct identification of manuscripts written by the same scribe is given in Figure 7. The two portions shown – respectively the corresponding folios – are both belonging to a manuscript named Codex Clozianus, but the folios are stored in different libraries – namely libraries in Innsbruck, Austria and Trent, Italy. It can be seen that the writing on folio 8 recto has a higher contrast to the remaining background than the text on folio 3 verso. Additionally, the latter mentioned folio is corrupted by background stains. These circumstances lead to a poor binarization result. Nevertheless, the method is capable of determining that both folios were written by the same scribe, as can be seen in Figure 6 (right): In this plot, folio 8 recto of the Codex Clozianus is used as reference

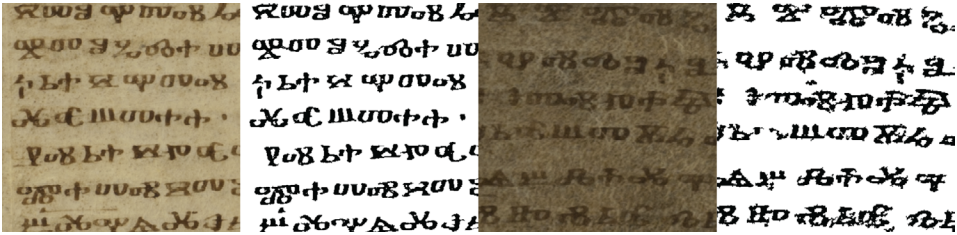


Figure 7. Two portions of Codex Clozianus. (From left to right) Folio 8 recto, stored at the City Museum in Trent, Italy. Corresponding binarization result. Folio 3 verso, stored at the Ferdinandeum museum in Innsbruck, Austria. Corresponding binarization result.

document and the distances to the remaining folios of the same codex – including folio 3 verso – are smaller than the distances to the other writings in the dataset.

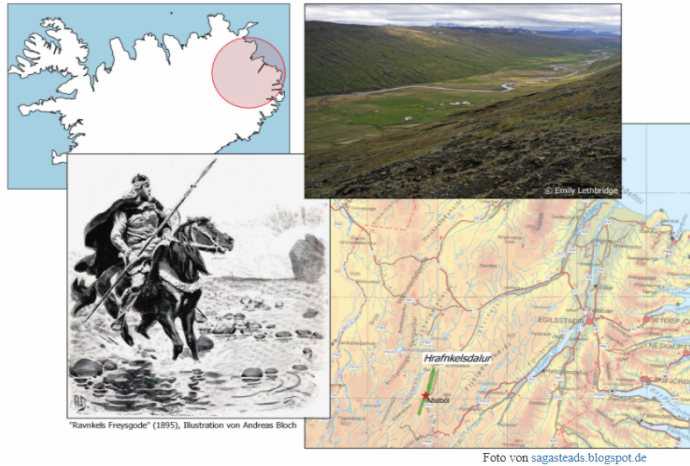
CONCLUSION

This paper presented the first application of writer identification on Glagolitic documents. This is done by using the Fisher Vector on visual vocabularies. First SIFT features are calculated on the document image and with help of GMM, which has been generated in advance on a training set, the Fisher Vector can be generated. The documents are then ordered by their similarity to a reference document and the evaluations are carried out using the nearest neighbors. As input image a dataset with 361 images with Glagolitic writing is used, which were preprocessed by cropping the area which contains text. On the complete dataset the best performance, with 98.9% correct first nearest neighbor, was achieved by applying a binarization to the image. Since the method used has been designed to avoid a binarization step, the evaluation also has been carried out on the unprocessed images, which resulted in significant worse identification rates. The reason for this is that SIFT features are generated on the non-uniform background, water stains, and holes in the document. Also, since only a simple cropping method is used, features have been calculated on the book cover on some images. For work to come, the detection of the text region has to be improved, to calculate features only on the written text.

IM STILLEN KÄMMERLEIN. ZUR DIGITALEN LESEAusGABE DER ›HRAFNKELS SAGA FREYSGOÐA‹

FABIAN SCHWABE

HRAFNKELS SAGA FREYSGOÐA



UNIVERSITÄT
TÜBINGEN
Zum eScience-Center



Kontakt: Fabian Schwabe
fabian.schwabe[at]uni-tuebingen[dot]de



Download der Saga
als XML-Datei

Abb. 1: Landingpage der Leseausgabe
(<http://ecenter.uni-tuebingen.de/hrafnkels-saga/start.html>)

Edieren und digitales Edieren

Wissenschaftliche Editionen erscheinen mittlerweile als digitale Editionen regelmäßig online. Gedruckte Editionen hingegen sind heutzutage häufig nur noch eine Zugabe zu einer Online-Edition¹ bzw. das Ergebnis kleinerer Editionsprojekte², bei denen

1 Bspw. Alfred-Escher-Briefedition: <https://www.briefedition.alfred-escher.ch/>. Parallel dazu wurden die Briefe bisher in sechs gedruckten Bänden publiziert: <https://www.alfred-escher.ch/publikationen/> [zuletzt besucht am 26.05.2017].

2 Bspw. Gesine Mierke (Hg.): Die Crecentia-Erzählung aus der »Leipziger Kleinepikhandschrift Ms 1279, in: Saxofodina. Fundstücke zur Literatur- und Kulturgeschichte Sachsens in Mittelalter und Früher Neuzeit 1, Chemnitz 2013.

während der Bearbeitung der technische Aufwand einer digitalen Edition vermutlich gescheut wurde, und man sich auf die philologische Arbeit konzentrierte.

In der Editionsarbeit lässt sich also ein Paradigmenwechsel erkennen. Während gedruckte Editionen sehr häufig nur einen Herausgeber haben, entstehen digitale Editionen fast ausschließlich in Zusammenarbeit mit Spezialisten für die einzelnen Arbeitsbereiche, ins besondere für die technische Infrastruktur. Diese Arbeitsteilung ist sinnvoll und hat natürlich ihre Vorteile, bringt aber zugleich mit sich, dass kleinere und weniger prominente Texte nur sehr selten als digitale Edition bearbeitet werden. Gerade wenn der von Patrick Sahle als ›digitale Dekanonisierung³ bezeichnete Prozess, der der Forschung neue zuverlässige Grundlagentexte zur Verfügung stellt, seine volle Wirkung entfalten soll, ist es nötig, die Arbeit an digitalen Editionen neben den großen Projekten, die natürlich einen wichtigen Vorbildcharakter für die Forschung haben können, im Allgemeinen voranzutreiben.

In den letzten Jahren entstanden unter anderen auch mehrere digitale Editionen von einem oder mehreren Texten, die ihren Fokus nicht auf die Wiedergabe von einzelnen Handschriften oder die Zugänglichkeit zu neuem Material legten, sondern die Vernetzung von verschiedenen Forschungsergebnisse bzw. die Integration von zusätzlichem (linguistischen) Material zum Ziel hatten.⁴ Diese Editionen lassen sich sehr gut für die Vermittlung von Wissen einsetzen, selbst wenn dies nicht unbedingt das angestrebte Ziel war. Zudem demonstrieren sie den enormen Vorteil, den eine annotierte digitale Lesefassung gegenüber einem herkömmlichen Druck als »studentische Leseausgabe« hat.

Die digitale Editionsphilologie kann durch die heutigen Webtechnologien und mit einfachen Arbeitsprozessen nach kurzer Einarbeitungszeit auch von Philologen ohne dezidierten Informatikhintergrund bewältigt werden. In diesem Zusammenhang sollte man jedoch nicht Herman Brinkmans Mahnung über die Zweckgebundenheit einer Edition vergessen:

»Die wissenschaftliche Edition von heute ist kein Selbstzweck, kein Denkmal für die ältere Literatur oder für das sprachliche Erbe einer Nation. An erster Stelle ist sie ein Arbeitsinstrument für die Wissenschaft. Die Edition soll dem Forscher dienstbar sein.«⁵

3 Patrick Sahle: Digitale Editionstechniken, in: Digitale Arbeitstechniken für die Geistes- und Kulturwissenschaften, hg. von Martin Gasteiner und Peter Haber, Wien 2009, S. 231–249 [insbesondere S. 231–33].

4 Als Beispiele möchte ich folgende drei Websites mit digitalen Editionen bzw. Leseausgaben nennen, die trotz unterschiedlichen Fokus viele Ähnlichkeiten aufweisen:

1. Sammlung von Texten verschiedener Sprachen: Perseus Digital Library (seit 1995), gehostet von der Tufts University (<http://www.perseus.tufts.edu/hopper/collections> [zuletzt besucht am 26.05.2017])
2. die gotischen Bibel (2004), gehostet von der Universiteit Antwerpen (<http://www.wulfla.be/gothic/browse> [zuletzt besucht am 26.05.2017])
3. altenglische Lesetexte zu Peter Bakers Introduction to Old English, gehostet von der University of Virginia (<http://www.oldenglishaerobics.net/anthology> [zuletzt besucht am 26.05.2017])

5 Hermann Brinkman: Dilemmata der Dienstbarkeit. Editionen mittelniederländischer Texte im Spannungsfeld zwischen Literatur- und Sprachgeschichte, in: Edition und Sprachgeschichte. Baseler Fachtagung 2.–4. März

Übertragen auf die Leseausgabe eines Textes hieße dies, dass es sich um ein Arbeits- oder Lerninstrument für die wissenschaftliche Ausbildung handelt, und die Leseausgabe dem Studierenden bzw. Lernenden dienstbar sein soll. Diesem Ansatz wurde in der Arbeit an der hier vorzustellenden Leseausgabe der ›Hrafnkels saga Freysgoða‹ gefolgt.

Die Saga und ihre Überlieferung

Die ›Hrafnkels saga Freysgoða‹ (dt. ›Die Saga vom Freyspriester Hrafnkell‹) ist »eine kurze, literar[isch] aber wohl die bedeutendste Isländersaga«⁶, die in der zweiten Hälfte des 13. Jahrhunderts entstanden und deren Handlung im 10. Jahrhundert in Ostland angesiedelt ist.

Die titelgebende Hauptperson ist der mächtige und arrogante Gode Hrafnkell, der seinen Totschlag des Schafhirten Einar nicht büßen will, weil jener unerlaubt sein Pferd geritten ist, das er dem Gott Freyr geweiht hat, und nur von ihm selbst geritten werden durfte. Hrafnkell fühlt sich lediglich seinem Gelöbnis, nicht aber dem isländischen Recht verpflichtet. Nachdem Einars armer Vater ein sehr großzügiges Kompensationsangebot abgelehnt hat, überredet er seinen gesetzeskundigen, aber gesellschaftlich unbedeutenden Neffen Samr, auf dem Althing einen Prozess gegen Hrafnkell anzustrengen. Dank der Hilfe eines westisländischen Brüderpaars gelingt es Samr, den Goden zu ächten, in dessen Folge Hrafnkell seinen Besitz und das Godentum einbüßt, aber von Samr das Leben geschenkt bekommt. Samr übernimmt darauf den Hof und das Amt, während Hrafnkell sich in einem anderen Tal ein neues Leben aufbauen muss, was ihm schnell gelingt. Aus Vergeltung tötet er schließlich Samrs Bruder, überfällt Samr, der sich daraufhin Hrafnkell unterwerfen muss, und übernimmt wieder seinen Hof und das Godentum. Die Ausgangssituation der Saga ist wiederhergestellt.

Im Online-Katalog handrit.is, in dem die Arnamagnäanische Sammlung und weitere isländische Handschriftensammlungen erfasst sind, werden insgesamt 42 Handschriften aufgelistet⁷, die die Saga in Teilen oder komplett wiedergeben. Hinzu kommen mindestens drei weitere Handschriften, die heute in Uppsala, Stockholm bzw. London aufbewahrt sind.⁸ Die Saga ist mit – Ausnahme eines einblättrigen Pergamentfragments (AM 162 I fol., um 1500) – nur in Papierhandschriften aus dem 17. Jahrhundert und später überliefert. Man unterscheidet in der Überlieferung mehrere Redaktionen, die – wenn auch mit unterschiedlicher Nähe – alle auf AM 162 I fol. zurückgeführt werden können.

2005 (Beihefte zu Editio 26), hg. von Michael Stolz, Robert Schöller und Gabriel Viehhauser, Tübingen 2007, S. 185.

6 Rudolf Simek und Hermann Pálsson: Lexikon der altnordischen Literatur, Stuttgart 2007, S. 189.

7 <https://handrit.is/is/search/results/FSmkVP> [zuletzt besucht am 26.05.2017].

8 Jón Helgason (Hg.): Hrafnkels saga Freysgoða (Nordisk filologi A, 2), København 1964, S. VIII.

Der Hauptzeuge der häufig als A bezeichneten Redaktion ist AM 156 fol. Weitere indirekte Abschriften des Pergamentfragments sind AM 158 fol. und AM 443 4to sowie AM 496 4to (unvollständiger Text), die Unterschiede zu AM 156 fol. und Gemeinsamkeiten untereinander aufweisen (B-Redaktion). Eine vierte und von AM 162 I fol. recht unabhängige Handschrift ist AM 551 c 4to (C-Redaktion). Ihr schlechter Zustand mit vielen Textlücken kann durch AM 451 4to (Anfang des 18. Jh.) aufgefüllt werden.⁹ Die bedeutende Saga wurde mehrmals ediert¹⁰ und in mehrere moderne Sprachen übersetzt¹¹.

Die Idee zur Leseausgabe

Die Arbeit an der Leseausgabe zur ›Hrafnkels saga Freysgoða‹ entstand als eine spontane Idee des Kursleiters eines Lektürekurses der Tübinger Skandinavistik, zu dem der Autor dieses Artikels als Gast eingeladen wurde. Für diesen Kurs – wie für jeden anderen Lektürekurs alter Sprachen auch – bereiteten sich die Teilnehmer auf die Sitzungen vor, indem sie den aktuellen Textabschnitt hinsichtlich Morphologie, Syntax und Lexik Wort für Wort durcharbeiten, um den Abschnitt in der kommenden Sitzung vor Ort nach Möglichkeit *ad hoc* übersetzen und auf Fragen des Kursleiters hinsichtlich morphologischer und syntaktischer Phänomene antworten zu können. Für jedes Wort müssen die grammatikalische Bestimmung der vorliegenden Wortform und das Lemma zur Ermittlung der Wortbedeutung in einem Wörterbuch bestimmt werden. Diese zum Teil zeitraubende Arbeit der Teilnehmer sollte nach Systematisierung und Vereinheitlichung der Angaben bzw. ermittelten Informationen weiter nutzbar gemacht werden.

Die zentrale Idee war also, die Teilnehmer des Lektürekurses für die inhaltliche Bearbeitung des Sagatextes zu gewinnen. Die grammatikalische Bestimmung der Wortformen durch die Teilnehmer stand dabei im Vordergrund. Am Ende zeigte sich jedoch, dass die Vorbereitung der Teilnehmer bei weitem nicht so einheitlich und genau war, wie es für das Vorhaben nötig gewesen wäre. Die grammatikalisch exakte Bestimmung von Wortformen bürdete allen Teilnehmer mehr Arbeit auf, als sie bisher in ihrer Vorbereitung

9 Jón Helgason [Anm. 8], S. V–VII.

10 Die wichtigsten wissenschaftlichen Ausgaben sind: Peder Thorsen und Konráð Gíslason (Hgg.): *Sagan af Hrafnkeli Freysgoða*, København 1839. Jakob Jakobsen (Hg.): *Austfirðinga sögur* (Samfund til udgivelse af gammel nordisk litteratur (= STUAGNL) 29), København 1903. Jón Helgason (Hg.): *Hrafnkels saga Freysgoða* (Nordisk filologi A, 2), Reykjavík 1950. Jón Jóhannesson (Hg.): *Austfirðinga sögur* (Íslensk fornrit XI), Reykjavík 1950. Walter Baetke (Hg.): *Hrafnkels saga freysgoða*. Mit Einleitung, Anmerkungen und Glossar (Altnordische Textbibliothek 1), Halle a. Saale 1952. Darüber hinaus gibt es zahlreiche weitere Editionen, die zumeist unkommentiert sind und den Charakter einer Leseausgabe haben.

11 Gängige deutsche Übersetzungen sind: Die Geschichte vom Freyspriester Hrafnkel, in: *Sieben Geschichten von den Ostland-Familien* (Thule. Altnordische Dichtung und Prosa 12), hg. und übs. von Gustav Neckel, Jena 1913, S. 75–101. Die Hrafnkels Saga, in: *Sagas aus Ostisland. Die Hrafnkels Saga und andere Geschichten von Macht und Fehde* (Saga Bibliothek der altnordischen Literatur) hg. und übs. von Dirk Huth, München 1999, S. 71–101. Andreas Vollmer (Üb.): Die Saga von Hrafnkell Freysgoði, in: *Die Isländersagas in 4 Bänden mit einem Begleitband*, 4. Bd., hg. von Klaus Bödl, Andreas Vollmer, Julia Zernack, Frankfurt am Main 2011, S. 177–212.

gemacht hatten. Darüber hinaus erforderte die Mehrarbeit einen sicheren Umgang mit grammatikalischen Kategorien und der Kodierung für die Speicherung der Information. Dies führte nach und nach zum Ausstieg einzelner Teilnehmer aus der projektrelevanten Mehrarbeit. Am Ende wurden nur die Wortformenbestimmung und Lemmatisierung der ersten beiden Kapitel der insgesamt 20 Kapitel von Studenten vorgenommen.

Weil aber das Konzept der Leseausgabe weit fortgeschritten war und viele Arbeitsprozesse zur HTML-Fassung des Sagatextes schon automatisiert waren, wurde das Projekt nicht verworfen. Aus der ursprünglichen angesetzten Gruppenarbeit mit studentischer Beteiligung wurde schließlich ein klassisches geisteswissenschaftliches Ein-Personen-Projekt im ›stillen Kämmerlein‹.

Vom digitalen Text zur digitalen Leseausgabe

Für die Umsetzung der digitalen Leseausgabe konnte auf mehrere Vorarbeiten Dritter zurückgegriffen werden. Da es sich nicht um die Anfertigung einer Edition auf Basis handschriftlicher Textzeugen handelte, konnte die Frage nach der Textgrundlage pragmatisch beantwortet werden. Der Sagatext liegt im Web in mehreren Fassungen (Altisländisch, Neuisländisch) und in Übersetzungen vor. Als Ausgangstext wurde der von Guðni Jónsson 1945 herausgegebene altisländische Text von heimskringla.no gewählt.¹²



Abb. 2: Auflistung der externen Quellen der digitalen Leseausgabe

12 Guðni Jónsson (Hg.): Hrafnkels saga Freysgoða (Íslendinga sögur 8), Reykjavík 1945. Die Edition ist selbst eine unkritische Leseausgabe. Der digitale Text der Edition ist auf http://heimskringla.no/wiki/Hrafnkels_saga_Freysgo%C3%B0a [zuletzt besucht am 26.05.2017].

Der Text wurde zunächst in ein selbst definiertes und zu 100% TEI-konformes XML-Format umgewandelt, so dass er schließlich als Vertikaltext – also ein Wort pro Zeile – mit wenigen, die Struktur des Textes betreffenden Metadaten vorlag. Die zeitlich aufwendigste Arbeit war die grammatikalische Bestimmung der insgesamt 9230 Wortformen und die Angabe des Lemmaansatzes nach dem Kopenhagener Wörterbuchprojekt ›Ordbog over det norrøne prosasprog / Old Norse Prose (ONP). Diese Bestimmung ermöglicht die Verknüpfung der einzelnen Wortformen des Textes mit den externen digitalen Quellen, den beiden Wörterbüchern von Cleasby/Vigfusson¹³ und Fritzner¹⁴ sowie der Grammatik von Noreen¹⁵.

Die Normalisierung des Ausgangstextes orientiert sich am späteren Altnordischen aus der Mitte des 14. Jahrhunderts und verwendet die im Neuisländischen gebräuchlichen Buchstaben. Die Leseausgabe hingegen soll der normalisierten Orthographie des ONP folgen¹⁶, weil sich dessen Normale hoffentlich als quasi verbindliche Standardisierung durchsetzen und die verschiedenen älteren Ansätze ablösen wird.

Hallfreðr kom út ok reisti bú dt. Schönhaar

Þat var á dögum Haralds konungs **ins härfağra**, Halfdanarsonar ins svarta, Guðrøðarsonar veiðikonungs, Halfdanarsonar ins milda ok ins matarilla, Eysteinnssonar frets, Ólafssonar trételgiu sviakonungs, at sá maðr kom skipi sinu til Íslands í Breiðdal, er Hallfreðr hét. Þat er fyrir neðan Fljótsdalsherað. Þar var á skipi kona hans ok sonr, er Hrafnkell hét. Hann var þá fimmtán vetra gamall, mannvænn ok gerviligr.

Hallfreðr setti bú saman. Um vetrinn andaðisk útlend ambátt, er Arnþrúðr hét, ok því heitir þat síðan á Arnþrúðarstöðum. En um várit fórdi Hallfreðr bú sitt norðr yfir heiði ok gerði bú þar, sem heitir í Geitdal.

Ok eina nátt dreymdi hann, at maðr kom at honum ok mælti: „Þar liggir þú, Hallfreðr, ok heldr óvarliga. Fór þú á burt bú þitt ok vestr yfir Lagarfljót. Þar er heill þin qll.“

Eftir þat vaknar hann ok fórir bú sitt út yfir Rangá í Tungu, þar sem síðan heitir á Hallfreðarstöðum, ok bjó þar til elli. En honum varð þar eftir góltr ok hafr, ok inn sama dag, sem Hallfreðr var í brott, hljóp skriða á húsin, ok týndusk þar þessir gripir, ok því heitir þat síðan í Geitdal.

Abb. 3: Das erste Kapitel der Leseausgabe

- 13 Richard Cleasby und Gudbrand Vigfusson: An Icelandic-English Dictionary, Oxford 1874. Digital aufgearbeitet von Sean Crist, http://lexicon.ff.cuni.cz/texts/oi_cleasbyvigfusson_about.html [zuletzt besucht am 26.05.2017].
- 14 Johan Fritzner: Ordbog over det gamle norske sprog, 4 Bde, Kristiania 1883-96. Digital aufgearbeitet von Eining for digital dokumentasjon (EDD), <http://www.edd.uio.no/per1/search/search.cgi?appid=86&tabid=1275> [zuletzt besucht am 26.05.2017].
- 15 Adolf Noreen: Altisländische und altnorwegische Grammatik. Laut- und Flexionslehre unter Berücksichtigung des Urnordischen (Sammlung kurzer Grammatiken germanischer Dialekte A, 4), Halle a. Saale 1923. Als HTML-Version umgearbeitet von Andrea de Leeuw van Weenen, <http://www.arnastofnun.is/solofile/1016380> [zuletzt besucht am 26.05.2017].
- 16 http://onp.ku.dk/english/onp_access/orthography/ [zuletzt besucht am 26.05.2017].

Mittels einer XSL-Transformation wurde das XML-Dokument in mehrere, voneinander abhängige HTML-Dokumente (Kapitel) und HTML-Codeschnipsel (Wortannotationen) umgewandelt, die mit Hilfe von JavaScript-Funktionen miteinander verbunden sind. Das Layout der Website fußt auf dem CSS-JavaScript-Framework Bootstrap¹⁷ und eigenen Stilangaben.

Für die fertige Leseausgabe wurde ganz bewusst ein schlichtes Layout gewählt, das trotz Annotationen und mehreren Ebenen in der Edition zunächst ein druckähnliches Äußeres haben soll. Bis auf eine dezente Hintergrundfarbe von einzelnen Wortgruppen, die durch die Hoverfunktion auf Übersetzungshinweise als HTML-Tooltip aufmerksam macht, und die seitenweise Darstellung der einzelnen Kapitel unterscheidet sich die Präsentation des Sagatextes nicht von einem gedruckten Werk.

Alle weiteren Annotationen, die zum Textverständnis beitragen können, werden als Popup präsentiert. Der Leser muss hierzu auf das betreffende Wort klicken. Er erhält zunächst die Wörterbuchform, Informationen zur Wortklasse (mit weiterer Kategorisierung bei Substantiven und Verben) und die genaue grammatikalische Bestimmung der Wortform. Mit einem weiteren Klick im Popup-Fenster kann er die das Lemma oder die Grammatik betreffende Inhalte der externen Quellen abrufen. Als blitzschnelle Lesehilfe,

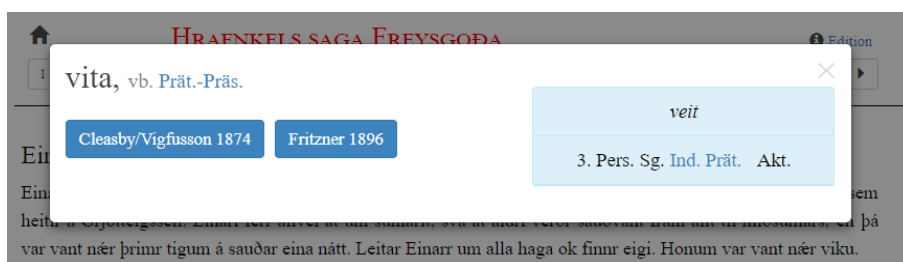


Abb. 4: Popup-Fenster zur Wortform *veit*

wenn man die Bedeutung eines Wortes nicht weiß, ist diese Aufteilung der Information natürlich nicht sonderlich bequem, aber dies ist auch nicht die Intention dieser Aufteilung.

Die Leseausgabe versteht sich vielmehr als ein Hilfsmittel, um Altnordisch zu erlernen, und nicht um altnordische Texte schnell durchzulesen. Dem Leser soll die Möglichkeit geboten werden, den betreffenden Wörterbuchartikel genau zu lesen, um die geeignetste Übersetzung ermitteln und so einen eigenen differenzierten (passiven) Wortschatz aufbauen zu können. Das Nachschlagen in der externen Grammatik hat ebenfalls den Zweck, das Wissen des Lesers zu festigen und zu erweitern. Die punktuelle Einbindung

¹⁷ <http://getbootstrap.com> [zuletzt besucht am 26.05.2017].

der externen Quellen zielt also darauf ab, diese zu konsultieren und zu einer auf den Satz abgestimmten Übersetzung zu gelangen.

Der Arbeitsprozess

Wie oben angedeutet, unterteilte sich die Arbeit an der digitalen Leseausgabe in mehrere Teilarbeitsprozesse (*workflows*), auf die – insbesondere, wenn sie vollständig oder teilweise automatisiert werden konnten – nun näher eingegangen wird.

Die Umwandlung des Ausgangstextes in einen Vertikaltext mit XML-Kodierung erfolgte unter Verwendung Regulärer Ausdrücke (*regular expressions*), durch die die Textstruktur mit nur wenigen manuellen Anpassungen überführt wurde. Für die XML-Kodierung des Sagatextes (Textstruktur, grammatikalische Annotation) wurden lediglich sieben Elemente und vier Attribute der TEI im Zieldokument im Bereich `<text>` / `<body>` benötigt.

Element	Beschreibung
<code><div></code>	Kapitel
<code><head></code>	Überschrift
<code><p></code>	Absatz
<code><q></code>	Wörtliche Rede
<code><seg></code>	Wortgruppe
<code><w></code>	Wort
<code><pc></code>	Satzzeichen

Attribut	Beschreibung
<code>@lemma</code>	Form des Wörterbucheintrags
<code>@ana</code>	grammatikalische Bestimmung
<code>@type</code>	Kategorisierung
<code>@n</code>	Übersetzung einer Wortgruppe

Die Überführung des altisländischen Ausgangstextes in die gewünschte normalisierte Orthographie ließ sich nur teilweise automatisiert vornehmen, da die Zeichen und Zeichenkombinationen nicht immer uneindeutig waren. Die Normalisierung des Textes nach den Vorgaben des ONP konnte erst während der grammatikalischen Bestimmung abgeschlossen werden und blieb in erster Linie Handarbeit, die nur minimal durch den Computer unterstützt wurde.

1. æ => æ oder ó
2. ö => o oder ø
3. -st und -zt in Verbindungen => -sk bzw. -zk
4. á|ó|ú+l+f|g|l|m|p|s => a|o|u+l+f|g|l|m|p|s
5. kleinere Unterschiede in der Lemmaansetzung
6. kleinere Unterschiede in flektierten Formen

Obwohl die XML-Struktur nach der automatisierten Umwandlung bereits nahezu fehlerfrei vorlag, wurde in einer Schemadatei der Aufbau des XML-Dokumentes definiert. Anders als bei den Kodierungsvorschlägen der TEI oder von The Medieval Nordic Text Archive¹⁸ (Menota), einem weiteren XML-Standard, auf den zurückgegriffen wurde, wurde ein sehr restriktives XML-Schema gewählt, um die Richtigkeit von Hand vergebener Werte bestmöglich kontrollieren zu können. Neben einfacheren Beschränkungen (`<xs:restriction>`) wie etwa der Festlegung von akzeptierten Satzzeichen und der Definition von Werten durch `<xs:enumeration>` wurden mit `<xs:pattern>` und Regulären Ausdrücken zwei eigene Datentypen für die formale Überprüfung der ONP-Normalen und der grammatikalischen Analyse definiert.

Mit dem Datentyp `normaliseONP` können die Werte in `<w>` und `@lemma` auf die Richtigkeit der verwendeten Zeichen überprüft werden. Dies gibt natürlich keine Gewissheit, ob die eingegebenen Werte tatsächlich eine korrekte altnordische Wortform angeben. Für eine solche Überprüfung bräuchte man eine Liste mit allen korrekten Wortformen. Der Aufwand, eine nur annähernd brauchbare Liste zu erzeugen, ohne direkt nutzbare Vorarbeiten Dritter, wäre zu groß.

```
<xs:simpleType name="normaliseONP">
  <xs:restriction base="xs:string">
    <xs:pattern value="[A-ZÆÍÓÚÝÞǼǾǿǿ]*(-)?([A-ZÆÍÓÚÝÞǼǾǿǿ])*[a-zæíóúýþǽǿǿ]*"/>
  </xs:restriction>
</xs:simpleType>

<xs:simpleType name="msaMenota">
  <xs:restriction base="xs:string">
    <xs:pattern
      value="(x(UA|IM|IT|CC|CS|VP|RP))|(x(N[CP](i(A|I|O|U|IA|IO|JA|WA|JO|WO|AN|ON|JN|IN|R|ND|RT|UA)) g[MFNU]
      n[SPU] c[NGDAU] s[IDU] |A V r[PCSU] g[MFNU] n[SPU] c[NGDAU] s[IDU] |PE p[123U] ( g[MF])? n[SDPU] c[NGDA]
      (eE)? |PE c[GDA] |D[PD] |P[QI] ) g[MFNU] n[SPU] c[NGDAU] |AT g[MFNU] n[SPU] c[NGDAU] |N[AO] ( g[MFNU] n[SPU]
      c[NGDAU])? |VB i(S([1-6])? |W([1-4])? |RD|PP|U) f((I( tPT)? |F t(P[ST]|U) p([1-3]|U) n[SPU] m(IN|IP|SU|U)) |
      (P tP([ST]|U) g[MFNU] n[SPU] c[NGDAU] s[DI])) v[AR] |A V r[PCSU] |A P y[GDAU]))"/>
    </xs:pattern>
  </xs:restriction>
</xs:simpleType>
```

Abb. 5: Definition der Datentypen `normaliseONP` und `msaMenota`

18 <http://menota.org> [zuletzt besucht am 26.05.2017].

Der Datentyp `msaMenota` für die grammatikalische Bestimmung der Wortformen in `@ana` wurde dicht an das Kodierungssystem `Menota` angelehnt. Die XML-Kodierung von `Menota` ist eine Erweiterung der `TEI` mit Hinblick auf eine sprachwissenschaftlich orientierte Kodierung für mittelalterliche Handschriften. Für die zulässigen Werte der `Menota`-eigenen Spezifizierung `@msa:ana` (*morpho-syntactical analysis*) existieren konkrete Vorgaben¹⁹, die wiederum auf die Verwendung im Zusammenhang mit Handschriften abgestimmt sind. Für einen normalisierten Text mit grammatikalischen Informationen zu Lemma und Wortform, die von außen an den Text herangetragen werden, musste das Kodierungssystem erweitert werden. Gerade die Verknüpfung mit einer Grammatik erforderte eine Bestimmung, die die morphologische Vielfalt der Verben und Substantive im Altnordischen adäquat einfängt. Für Verben unterscheidet `Menota` zwischen starken und schwachen Verben sowie Präteritopräsenzien und ehemals reduplizierenden Verben. In

```
<text>
  <body>
    <head>
      <w lemma="Hrafn-kell" ana="xNP iA gM nS cG sI" type="person">Hrafnkells</w>
      <w lemma="saga" ana="xNC iON gF nS cN sI">saga</w>
      <w lemma="freys-goði" ana="xNP iAN gM nS cG sI">Freysgoða</w>
    </head>
    <div type="chapter" n="1">
      <head>
        <w lemma="Hall-freðr" ana="xNP iUA gM nS cN sI" type="person">Hallfreðr</w>
        <w lemma="koma" ana="xVB iS4 fF tPT p3 nS mIN vA">kom</w>
        <w lemma="út" ana="xAV rP">út</w>
        <w lemma="ok" ana="xCC">ok</w>
        <w lemma="reisa" ana="xVB iW3 fF tPT p3 nS mIN vA">reisti</w>
        <w lemma="bú" ana="xNC iA gN nS cA sI">bú</w>
      </head>
      <p>

        <w lemma="sá" ana="xDD gN nS cN">Þat</w>
        <w lemma="vera" ana="xVB iS5 fF tPT p3 nS mIN vA">var</w>
        <w lemma="á" ana="xAP yD">á</w>
        <w lemma="dagr" ana="xNC iA gM nP cD sI">dögum</w>
        <w lemma="Har-aldr" ana="xNP iA gM nS cG sI" type="person">Haralds</w>
        <w lemma="konungr" ana="xNC iA gM nS cG sI">konungs</w>
        <seg type="specTrans" n="Schönhaar">
          <w lemma="inn" ana="xAT gM nS cG">ins</w>
          <w lemma="hárfagr" ana="xAJ rP gM nS cG sD">hárfagra</w>
        </seg>
        <pc>,</pc>
        <w lemma="Halfdanarson" ana="xNP iU gM nS cG sI" type="patronymic">Halfdanarsonar</w>
        <w lemma="inn" ana="xAT gM nS cG">ins</w>
        <w lemma="svartr" ana="xAJ rP gM nS cG sD">svarta</w>
        <pc>,</pc>
        <w lemma="Guðrøðarson" ana="xNP iU gM nS cG sI" type="patronymic">Guðrøðarsonar</w>
        <w lemma="veiði-konungr" ana="xNC iA gM nS cG sI">veiðikonungs</w>
        <pc>,</pc>
```

Abb. 6: Anfang des ersten Kapitels in XML-Kodierung

19 http://menota.org/HB2_ch8.xml [zuletzt besucht am 26.05.2017].

Grammatiken werden die Verben in verschiedene Klassen bzw. Ablautreihen weiter unterteilt. Gleiches gilt für die Substantive, die in sogenannte Stämme unterteilt werden. Mit msaMenota kann die formal korrekte Angabe zu einer Wortform überprüft werden, von denen die meisten händisch vorgenommen wurden. Die Struktur der XML-Kodierung wird in Abb. 6 deutlich.

Die unauffällige Einbettung der externen Inhalte in die Popup-Fenster hängt vor allem davon ab, ob die Anbieter der jeweiligen Seite an eine Möglichkeit des Referenzierens ihrer Inhalte gedacht haben. Zwar konnte die Verlinkung auf die Einträge der beiden Wörterbücher und der Grammatik durch eine Anweisung in der XSL-Transformation automatisch mit einem einfachen Direktlink gesetzt werden, doch waren dazu zwei verschiedene Verfahren nötig.

Im Falle von Noreens Grammatik handelt es sich um ein reines HTML-Dokument, in dem alle Abschnitte, Paragraphen, Anmerkungen und Fußnoten mit dokumentinternen IDs versehen sind, nicht zuletzt wegen der zahlreichen Querverweise innerhalb des Werks. Es genügte also ein einfaches Mapping der im XML-Dokument kodierten grammatikalischen Informationen, um auf die entsprechende ID im HTML-Dokument von Noreens Grammatik zu verweisen. Zur Webadresse <http://www.arnastofnun.is/solofile/1016380> gesellt sich die ID, z. B. zum Abschnitt ›Flexionslehre‹ (`diviflex`), so dass sich daraus die Adresse <http://www.arnastofnun.is/solofile/1016380#diviflex> ergibt.

Die Adressen der beiden Wörterbücher mussten auf andere Weise ermittelt werden. Hier handelt es sich um Datenbankabfragen mittels einer Scriptsprache. Die anzusteuern Website ist ein durch Nutzereingabe erzeugter sogenannter Deep Link oder eine dynamisch erzeugte und für den Nutzer nicht sichtbare Adresse. Nutzt man z. B. die Portalseite zu Fritzners Wörterbuch²⁰ kann man dieses zwar durchsuchen und mit einem Perl-Script Anfragen an die Datenbank schicken, aber das Ergebnis ist eine dynamische, nicht zitier- bzw. referenzierbare URL. Nach Anfrage bei dem Seitenbetreiber wurde der Aufbau eines direkten Links auf einen bestimmten Eintrag mitgeteilt, in dem das Lemma-wort in einer eigenen, bisweilen von der normalisierten Orthographie des Wörterbuchs abweichenden Kodierung eingetragen werden muss. Dies bedeutete für Verwendung der Wörterbucheinträge in der Leseausgabe, einen automatisierten Workflow zu schaffen, der die ONP-Normale in die entsprechende Notation für die Generierung der Webadresse umwandelt. Der Großteil der Lemma wurde mit einer kurzen Verkettung der XSL-Funktionen `replace()` und `translate()` in die gewünschte Form überführt. Dies ließ sich für die Einträge in Fritzners Wörterbuch leicht bewerkstelligen, weil die zu generierende Webadresse moderne skandinavische Buchstaben, darunter auch *ð*, *þ*, *æ*, *ö* und *ø*, und Vokale mit Akuten (*á*, *é* etc.) akzeptiert, so dass die ONP-Normale nur geringfügig

20 <http://www.edd.uio.no/perl/search/search.cgi?appid=86&tabid=1275> [zuletzt besucht am 26.05.2017].

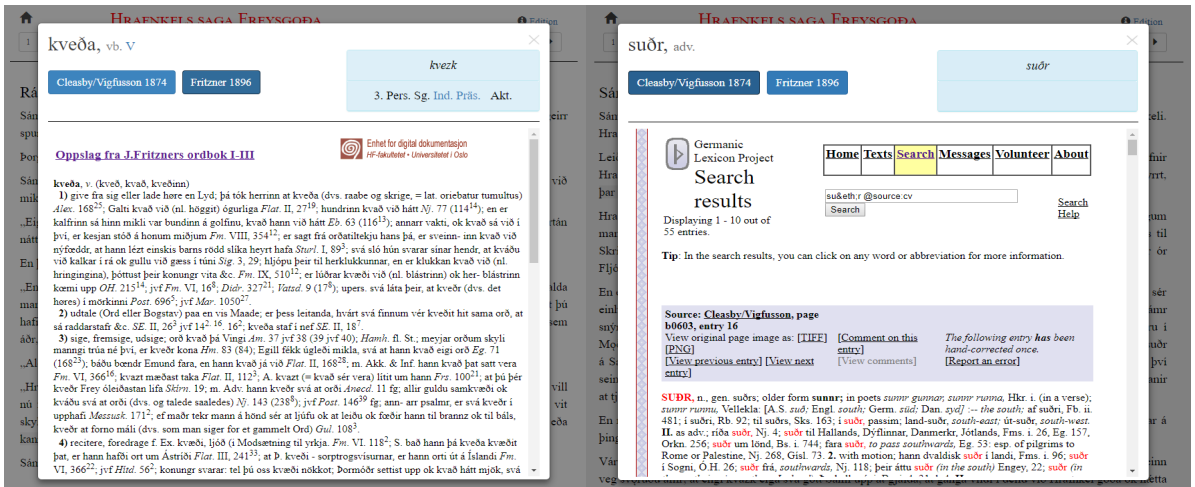


Abb. 7: Einbindung der externen Wörterbücher von Fritzner (li) und Cleasby/Vigfusson (re)

abgeändert werden musste. Für wenige Lemmata war die automatische Umwandlung nicht möglich; für diese wurde auf ein einfaches händisch angelegtes Mapping per Liste zurückgegriffen.

Für das Wörterbuch von Cleasby/Vigfusson wurde auf das Germanic Lexicon Project (GLP) zurückgegriffen, das die Volltextdigitalisierung mehrere Nachschlagewerke vorantreibt. Im Prinzip wird derselbe Workflow benutzt, um den gewünschten Wörterbucheintrag zu erhalten. Jedoch musste dieser an die Eigenheiten der Website angepasst werden, wie z. Bsp. zu aller erst die Einschränkung der Suche der Website auf das Wörterbuch von Cleasby/Vigfusson. Allerdings wird die Suche auf den gesamten Text und nicht auf Wörterbucheinträge ausgeführt und liefert somit auch weitere Ergebnisse neben dem gesuchten Eintrag. Im Gegensatz zum Portal zu Fritzners Wörterbuch erscheint das Ergebnis einer Anfrage an die Website als für den Nutzer sichtbarer, dynamisch kreierter Deep Link. An der URL erkennt man leicht, dass die Kodierung des Suchbegriffs für die zu generierende Webadresse leider sehr umständlich ist, weil alle Buchstaben, die nicht im Englischen vorkommen, mit Codes wiedergegeben werden. Um diesen Arbeitsschritt ebenfalls vollkommen automatisieren zu können, musste ein aufwendigeres Mapping durchgeführt werden, während dessen alle Zeichen (z. B.: á => %E1 und æ => a%2Be) in die richtige Form für den betreffenden Link umgewandelt wurden. Außerdem wurde erneut händisch eine Liste angelegt, um die wenigen nicht durch den Workflow umwandelbaren Wörterbucheinträge bzw. Lemmata in die richtige Form zu überführen.

Ausblick

Die Arbeit an der digitalen Leseausgabe der ›Hrafnkels saga Freysgoða‹ führte zu zwei Ergebnissen: die Leseausgabe selbst und die automatisierten Arbeitsprozesse (*workflows*). Die erarbeiteten Workflows und genutzten Standards bieten großes Potential für eine schnelle Erweiterung des Textmaterials hin zu einer Sammlung digitaler Leseausgaben mit ähnlichem oder gar gleichem Funktionsumfang. Die enge Anlehnung an der Kodierung nach Menota eröffnen die Möglichkeit, auf die Texte in Menotas Textarchiv²¹ zurückzugreifen, die vollständig normalisiert und lemmatisiert sind. Hierfür bedarf es freilich eines weiteren Arbeitsschritts, der Menotas Kodierung in eine XML-Kodierung überführt, damit die Transformation der XML-kodierten Information in eine funktionale digitale Leseausgabe automatisiert gemacht werden kann.

Ein Text, der nach den Vorgaben von Menota oder eben auch TEI kodiert wurde, ist aufgrund von zum Teil sehr unspezifischen Vorgaben ohne eigene Aufbereitung nur schwer nutzbar. Um dieses Ziel zu erreichen, müssten die Kodierungsvorgaben der beiden Initiativen sehr viel strikter (im Schema) formuliert sein. Momentan streben weder Menota noch die TEI dieses Ziel an. Die Kodierungsvorschläge der TEI sind für ein solches Ziel viel zu global und unspezifisch angesetzt, und die Vorgaben von Menota geben zu viele Möglichkeiten, die Elemente und Attribute zu nutzen bzw. konkrete Werte für Attribute zu vergeben. Beide Ansätze stellen damit lediglich eine Verständlichkeit der Kodierung von Informationen für Dritte sicher, die den Austausch von Forschungsdaten (*interchange*), aber nur sehr selten deren direkte Nachnutzung (*interoperability*) ermöglicht.²²

Die Anwendung der Workflows ist aber auch im Zusammenhang mit der universitären Lehre denkbar. Mit der Wiederaufnahme der ursprünglichen Idee und unter Berücksichtigung der Gründe für das letztliche Scheitern des Ansatzes im ersten Versuch wäre es durchaus möglich, Lektürekurse derart zu konzipieren, dass die Teilnehmer als ein Teil ihrer persönlichen Kursarbeit einzelne Textstellen kodieren. Aus einer solchen im Unterricht integrierten Gruppenarbeit kann ausreichend viel Material entstehen, so dass ein neuer Text vollständig oder fast vollständig lemmatisiert und als XML kodiert vorliegt, auf den die XSL-Transformation zur Erstellung einer digitalen Leseausgabe anwendbar ist. Unabdingbar ist allerdings ein ausreichendes Verständnis der Bearbeiter für die Lemmatisierung und Normalisierung der Wortformen und Lemmata.

21 <http://clarino.uib.no/menota/page> [zuletzt besucht am 26.05.2017].

22 Zu dieser Problematik siehe bspw. Desmond Schmidt: Towards an Interoperable Digital Scholarly Edition, Journal of the Text Encoding Initiative [Online], Issue 7, November 2014, online since 01 January 2014. URL: <http://jtei.revues.org/979>; DOI : [10.4000/jtei.979](https://doi.org/10.4000/jtei.979) [zuletzt besucht am 26.05.2017].

Darüber hinaus stellt eine fertige digitale Leseausgabe natürlich auch selbst Unterrichtsmaterial dar, mit dem die Teilnehmer eines Lektürekurses wesentlich mehr Textmenge bewältigen können als mit konventionellen Hilfsmitteln. Hier soll nicht unerwähnt bleiben, dass eine digitale Leseausgabe nicht das Erlernen der Grammatik und des Vokabulars einer Sprache, geschweige denn das Verständnis der Feinheiten einer Sprache ersetzen kann, sondern im Zuge dessen lediglich als Hilfsmittel fungiert. Ob die digitale Leseausgabe der ›Hrafnkels saga Freysgoða‹ in jetziger Gestalt für studentische Teilnehmer des akademischen Unterrichts geeignet ist, muss die Erfahrung der kommenden Jahre zeigen.

Nicht zuletzt lässt sich der etablierte Arbeitsablauf selbstredend von anderen Bearbeitern für die Erstellung weiterer Leseausgaben verwenden, die sich dann lediglich auf die inhaltlichen Aspekte der XML-Kodierung konzentrieren müssen. Ob und inwiefern eine automatisierte Lemmatisierung ohne großen Aufwand unter Rückgriff auf das bereits lemmatisierte Wortmaterial erreicht werden kann, kann ich zwar momentan nicht beantworten, aber dies ist technisch durchaus möglich, wie isländische Computerlinguisten bereits demonstriert haben.²³

23 Eiríkur Rögnvaldsson & Sigrún Helgadóttir: Morphological Tagging of Old Norse Texts and Its Use in Studying Syntactic Variation and Change, in: Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series, hg. von Caroline Sporleder, Antal P.J. van den Bosch und Kalliopi A. Zervanou, Berlin 2011, S. 63–76. [online: https://notendur.hi.is/~eirikur/Tagging_ON.pdf, zuletzt besucht am 26.05.2017].
In neuisländischer Textgestalt liegen als ›The Saga Corpus‹ mehrere Isländersagas, Königssagas und zeitgenössische Sagas automatisch lemmatisiert vor: <http://www.malfong.is/index.php?lang=en&pg=fornritin> [zuletzt besucht am 26.05.2017].