Content-aware web browsing and visualization tools for cantillation and chant research

Steven R. Ness, George Tzanetakis
Department of Computer Science
University of Victoria, BC, Canada.
sness@sness.net,gtzan@cs.uvic.ca

Dániel Péter Biró
School of Music
University of Victoria, BC, Canada
dpbiro@uvic.ca

Abstract

Chant and cantillation research is particularly interesting as it explores the transition from oral to written transmission of music. The goal of this work to create web-based computational tools that can assist the study of how diverse recitation traditions, having their origin in primarily non-notated melodies, later became codified. One of the authors is a musicologist and music theorist who has guided the system design and development by providing manual annotations and participating in the design process. We describe novel content-based visualization and analysis algorithms that can be used for problem-seeking exploration of audio recordings of chant and recitations.

1 Introduction

In recent years there has been increasing research activity in the areas of multimedia learning and information retrieval. Most of it has been in traditional specific domains, such as sports video [5], news video [4] and natural images. There is broad interest in these domains and in most cases there are clearly defined objectives such as identifying highlights in sports videos, explosions in news video or sunsets in natural images. Our focus in this paper is a niche domain that shares the challenge of effectively accessing large amounts of data but has specific characteristics that preclude the use of existing multimedia tools.

Although there is much related work little of it is directly relevant to our particular application. Work on melodic similarity is typically based on symbolic representations [3] and therefore not applicable. Even in the cases where audio recordings are used [2] there are no interactive visualizations which limits their use by expert musicologists. An earlier version of the web-based system we describe that did not have support for content-based similarity retrieval of pitch contours was presented in Ness et. al [10].

The goal of this project is to develop tools to study chants from various traditions around the world including Hungarian *siratok* (laments)[7], Torah cantillation[14], tenth century St. Gallen plainchant[6], and Koran recitation[9]. These diverse traditions share the common theme of having an origin in primarily non-notated melodies which then later became codified. The evolution and spread of differences in the oral traditions of these different chants are a current topic of research in Ethnomusicology [13].

It has proved difficult to study these changes using traditional methods and it was decided that a combined approach, using field recordings marked up by experts, mathematical models for analyzing the pitch content, automatic alignment for pitch contour similarity and a flexible GUI, would help figure out what questions needed to be asked. Unlike traditional multimedia data where most users can be used as annotators, in our case annotation requires trained experts. This is a problem seeking domain where there are no clearly defined objectives and formulating problems is as important as solving them. We believe that despite these challenges it is possible to develop semi-automatic tools that can assist researchers in formulating questions regarding how symbols are used in chant and recitation.

Web-based software has been helping connect communities of researchers since its inception. Recently, advances in software and in computer power have dramatically widened its possible applications to include a wide variety of multimedia content. These advances have been primarily in the business community, and the tools developed are just starting to be used by academics. We have been working on applying these technologies to ongoing collaborative projects [10]. We leverage several new technologies including *Flash*, *haXe*, *AJAX* and *Ruby on Rails*, to rapidly develop web-based tools. Rapid prototyping and iterative development have been key elements of our collaborative strategy. Although our number of users is limited compared to other areas of multimedia analysis and retrieval, this is to some degree compensated by their passion and willingness to work closely with us in developing these tools.

2 Chant research

Our work in developing tools for chant research is a collaboration with Dr. Daniel Biro, a professor in the School of Music at the University of Victoria. He has been collecting and studying recordings of chant with specific focus on how music transmission based on oral transmission and ritual was gradually changed to one based on writing and music notation. The examples studied come from improvised, partially notated, and gesture-based [8] notational chant traditions: Hungarian siratok (laments) [1], Torah cantillation [15] [2], tenth century St. Gallen plainchant [11] [3], and Koran recitation [4]. This work falls under the more general area of Computational Ethnomusicology [13].

Although Dr. Biro has been studying these recordings for some time and has considerable computer expertise for a professor in music, the design and development of our tools has been challenging. This is partly due to difficulties in communication and terminology as well as the fact that the work is exploratory in nature and there are no easily defined objectives. The tool has been developed through extensive interactions with Dr. Biro with frequent frustration on both sides. At the same time, a wonderful thing about expert users like Dr. Biro is that they are willing to spend considerable time preparing and annotating data as well as testing the system and user interface which is not the case in more traditional broad application domains.

3 Analysis and Browsing

3.1 Melodic Contour Analysis

Our tool takes in a (digitized) monophonic or heterophonic recording and produces a series of successively more refined and abstract representations of the segments it contains as well as the corresponding melodic contours. More specifically the following analysis stages are performed:

- Hand Labeling of Audio Segments
- First Order Markov Model of Sign Sequences
- F0 Estimation
- F0 Pruning
- Scale Derivation: Kernel Density Estimation
- Quantization in Pitch
- Scale-Degree Histogram
- Histogram-Based Contour Abstraction
- Dynamic Time Warping for Contour Similarity
- Plotting and Recombining the Segments

The recordings are manually segmented and annotated by the expert. Even though we considered the possibility of creating an automatic segmentation tool, it was decided that the task was too subjective and critical to automate. Each segment is annotated with a word/symbol that is related to the corresponding text or performance symbols (for example cantillation marks) used during the recitation.

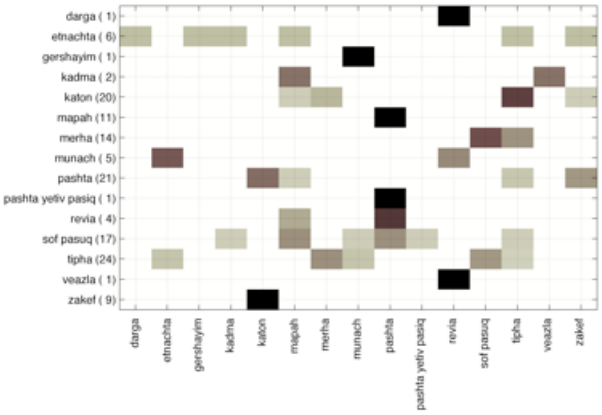


Figure 1. Syntagmatic analysis with a 1st-order Markov model of Torah trope signs for Shir Ha Shirim ("Song of Songs").

In order to study the transitions between signs/symbols we calculate a first order Markov model of the sign sequence for each recording. We were asked to perform this type of syntagmatic analysis by Dr. Biro. Although it is completely straightforward to perform automatically using the annotation, it would be hard, if not impossible, to calculate manually. Figure 3.1 shows an example transition matrix. For a given trope sign (a row) it shows how many total times does it appear in the example (numeral after row label), and in what fraction of those appearances is it followed by each of the other trope signs. The darkness of each cell corresponds to the fraction of times that the trope sign in the given row is followed by the trope sign in the given column. (NB: Cell shading is relative to the total number of occurrences of the trope sign in the row, so, e.g., the black square saying that "darga" always precedes "revia" represents 1/1, while the black square saying that "zakef" always precedes "katon" represents 9/9.) This type of analysis can help identify the syntactic role of different signs.

After the segments have been identified, the fundamental frequency ("F0" in this case equivalent to pitch) and signal energy (related to loudness) are calculated for each segment

[1] Archived Examples from Hungarian Academy of Science (1968-1973)

[2] Archived Examples from Hungary and Morocco from the Feher Music Center at the Bet Hatfatsut, Tel Aviv, Israel

[3] Godehard Joppich and Singphoniker: Gregorian Chant from St. Gallen (Gorgmarienhtte: CPO 999267-2, 1994)

[4] Examples from Indonesia and Egypt: in Approaching the Koran (Ashland: White Cloud, 1999)

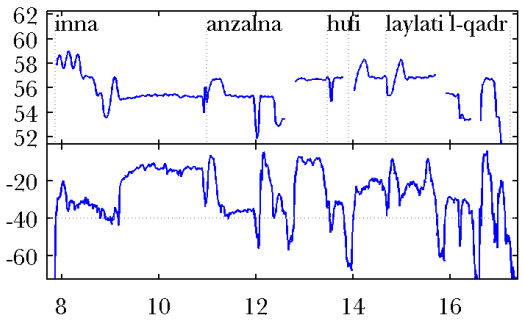


Figure 2. F0 contour

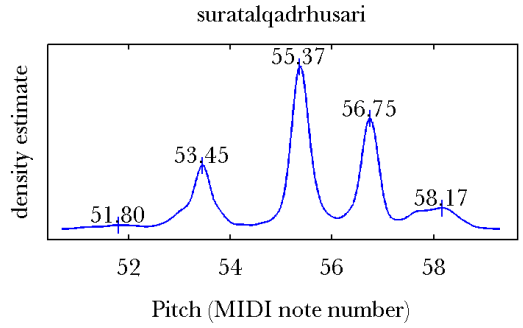


Figure 3. Recording-specific scale derivation

as functions of time. We use the SWIPEP fundamental frequency estimator [1] with all default parameters except for hand-tuned upper and lower frequency bounds for each example. For signal energy we simply take the sum of squares in 10-ms rectangular windows.

The next step is to identify pauses between phrases, so as to eliminate the meaningless and wildly varying F0 estimates during these noisy regions. We define an energy threshold, generally 40 decibels below each recording's maximum. If the signal energy stays below this threshold for at least 100 ms then the quiet region is treated as silence and its F0 estimates are ignored. Figure 3.1 shows an excerpt of the F0 and energy curves for an excerpt from the Koran sura ("section") Al-Qadr ("destiny") recited by the renowned Sheikh Mahmud Khalil al-Husari from Egypt.

Following the pitch contour extraction is pitch quantization, which is the discretization of the continuous pitch contour into discrete notes of a scale. Rather than externally imposing a particular set of pitches, such as an equal-tempered chromatic (the piano keys) or diatonic scale, we have developed a novel method for extracting a scale from an F0 envelope that is continuous (or at least very densely sampled) in both time and pitch. Our method is inspired by Krumhansl's time-on-pitch histograms adding up the total amount of time spent on each pitch [8]. We demand a pitch resolution of one cent [5], so we cannot use a simple histogram. Instead we use a statistical technique known as non-parametric kernel density estimation, with a Gaussian kernel [6]. More specifically a Gaussian (with standard deviation of 33 cents) is centered on each sample of the frequency estimate and the Gaussians of all the samples are added to form the kernel density estimate. The resulting curve is our density estimate; like a histogram, it can be interpreted as the relative probability of each pitch appearing at any given point in time. Figure 3.1 shows this method's density estimate given the F0 curve from Figure 3.1.

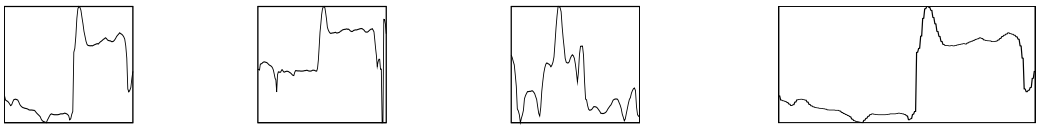
We interpret each peak in the density estimate as a note of the scale. We restrict the minimum interval between scale pitches (currently 80 cents by default) by choosing only the higher peak when there are two or more very close peaks. This method's free parameter is the standard deviation of the Gaussian kernel, which provides an adjustable level of smoothness to our density estimate; we have obtained good results with a standard deviation of 30 cents.

Once we have determined the scale, pitch quantization is the trivial task of converting each F0 estimate to the nearest note of the scale. In our opinion these derived scales are more true to the actual nature of pitch-contour relationships within oral/aural and semi-notated musical traditions. Instead of viewing these pitches to be deviations of pre-existing "normalized" scales our method defines a more differentiated scale from the outset. With our approach the scale tones do not require "normalization" and thereby exist in an autonomous microtonal environment defined solely on statistical occurrence of pitch within a temporal unfolding of the given melodic context. Once the pitch contour is quantized into the recording-specific scale calculated using Kernel density estimation, we can calculate how many times a particular scale degree appears during an excerpt. The resulting data is a scale-degree histogram which is used create simplified abstract visual contour representations.

The basic idea is to only use the most salient discrete scale degrees (the histogram bins with the highest magnitude) as significant points to simplify the representation of the contour. By adjusting the number of prominent scale degrees used to represent the simplified representation the researchers can view/listen to the melodic contour at different levels of abstraction and detail. Figure 3.1 shows an original continuous contour, the quantized representation using the recording-specific derived scale and the abstracted representation using only the 3 most prominent scale degrees.

[5] One cent is 1/100 of a semitone, corresponding to a frequency difference of about 0.06%

[6] Thinking statistically, our scale is related to a distribution given the relative probability of each possible pitch. We can think of each F0 estimate (i.e each sampled value of the F0 envelope) as a sample drawn from this unknown distribution so our problem becomes one of estimation the unknown distribution given the samples



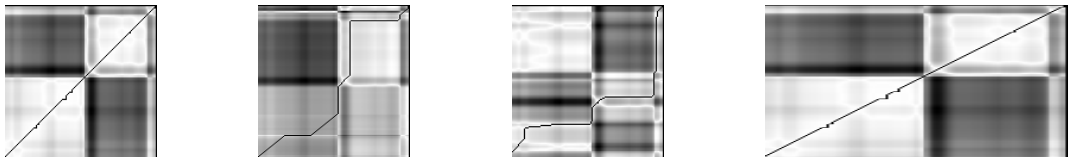
(a) F0 Contour of 11 Pashta

(b) F0 Contour of 42 Pashta

(c) F0 Contour of 18 Sof Pasuq

(d) F0 Contour of 11 Pashta Doubled

Figure 5. F0 contours of 4 different gestures from a Torah recitation from Hungary. The first two show different versions of the pashta gesture and the third shows the gesture for sof pasuq. The last is a version of the first pashta gesture stretched by two.



(a) DTW of 11 Pashta vs 11 Pashta

(b) DTW of 11 Pashta vs 42 Pashta

(c) DTW of 11 Pashta vs 18 Sof Pasuq

(d) DTW of 11 Pashta vs 11 Pashta Doubled

Figure 6. Similarity Matrices of the above four gestures compared with the first pashta gesture. Superimposed on the figures is the DTW curve showing the alignment between the signs.

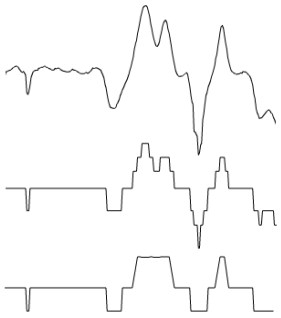


Figure 4. Melodic contours at different levels of abstraction (top: original, middle: quantized, bottom: simplified using 3 most prominent scale degrees

3.2 Dynamic Time Warping for Contour Similarity Calculation

One of the main aspects in the studying of signs in the context of chant and recitation is to what extent they convey gesture information that is invariant with respect to the underlying text. To study this question it was necessary to develop a method to compare the pitch contours of different realizations of the same sign. Dynamic Time Warping (DTW) is a technique by which the similarity between two different time sequences can be measured. It allows a computer to find an optimal match between two sequences by performing a non-linear warping of one sequence to the other. The technique of dynamic programming is used for efficient implementation. An example of DTW in Music Information Retrieval is comparing the tempo variations between two different performances of a symphony. The DTW algorithm would identify the parts of the two symphonies that were played at the same tempo as a diagonal line, with the line varying above and below the diagonal due to tempo variations.

First the similarity matrix between the two pitch contours is calculated. The DTW algorithm finds the optimal alignment of the two sequences and calculates the cost for that alignment. When the contours are similar the alignment cost will be small compared to when the contours are dissimilar. The matching process is pitch shift invariant and allows variations and tempo stretching. That way for any particular sign (pitch contour) we can sort the signs (pitch contours) by similarity.

To illustrate the technique we use the gestures of two separate annotated recordings of a section of the Torah. One of these was recorded in Morocco, and the other was recorded in Hungary. Figures 5(a), 5(b), 5(c) and 5(d) show the F0 contour of the sections of the audio file from a Torah recording from Hungary. Figure 5(a) shows a pashta

Gesture (Hungary)	Average Precision (Hungary)	Gesture (Morocco)	Average Precision (Morocco)
tipha	0.662	katon	0.453
pashta	0.647	mapah	0.347
mapah	0.641	tipha	0.303
katon	0.604	sofpasuq	0.285
etnachta	0.601	pashta	0.242
sofpasuq	0.591	merha	0.251
merha	0.537	etnachta	0.150
revia	0.372	zakef	0.125
zakef	0.201	revia	0.091
kadma	0.200	kadma	0.043

Table 1. Average precision for different signs

sign, Figure 5(b) shows another pashta sign from further along in the audio file. Figure 5(c) shows a sof pasuq gesture and Figure 5(d) shows the first pashta gesture, but with the sample stretched by a factor of two.

The figures 6(a),6(b) , 6(c) and 6(d) show Similarity Matrices and the alignment paths computed using DTW for these four gestures compared to the first. White areas are highly similar and black areas have low similarity. In Figure 6(a) the first pashta gesture is compared to itself. The DTW curve is overlayed in black and is basically a straight diagonal line from one corner to the opposite corner showing direct alignment. Figure 6(d) shows a similar behavior, except that the slope of the line is shallower. Figure 6(b) shows the comparison of one pashta gesture to another. This path had a DTW cost of 23.8442. Figure 6(c) shows an alignment between the pashta gesture and a sofpasuq gesture. One can see that the line is not only not diagonal, but that the line is often on dark areas which results in high alignment cost.

Table 1 shows the average precision for particular signs for two recordings of the same excerpt from the Torah - one from Hungary and one from Morocco. Each recordings contains approximately 130 realizations of each sign with a total of 12 unique signs. Two pitch contours are considered relevant to each other if they are annotated by the same sign. For each "query" contour we return a list of results which are the pitch contours sorted by the alignment cost of the DTW. Average precision emphasizes returning more relevant contours earlier. It is the average of precisions computed after truncating the list of returned results after each of the relevant documents in turn. Unlike traditional retrieval systems where the mean average precision can be used to characterize the overall system performance in our cases we are more interested in the individual difference in precision among different signs. These differences show which signs have well-defined gestural characteristics and which signs are not interpreted consistently. Ultimately the numbers are only meaningful after careful interpretation by an expert. For example based on Table 1 one can infer that the performer in the Hungarian version had more consistent interpretations of the signs than the performer in the Moroccan version.

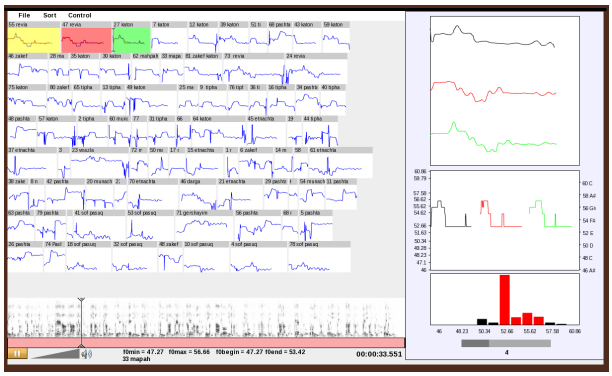


Figure 7. Web-based *Flash* interface to allow users to listen to audio, and to enable interactive querying of gesture contour diagrams.

3.3 Cantillion interface

We have developed a browsing interface that allows researchers to organize and analyze chant segments in a variety of ways (`http://cantillation.sness.net`). Each recording is manually segmented into the appropriate units for each chant type (such as trope sign, neumes, semantic units, or words). The pitch contours of these segments can be viewed at different levels of detail and smoothness using a histogram-based method. The segments can also be rearranged in a variety of ways both manually and automatically. The audio analysis (pitch extraction and dynamic time warping) are performed using the Marsyas audio processing framework [7] [12].

The interface (Figure 7) has four main sections: a sound player, a main window to display the pitch contours, a control window, and a histogram window. The sound player window displays a spectrogram representation of the sound file with shuttle controls to let the user choose the current playback position in the sound file. The main window shows all the pitch contours for the song as icons that can be repositioned automatically based on a variety of sorting criteria, or alternatively can be manually positioned by the user. The name of each segment (from the initial segmentation step) appears above its F0 contour.

When an icon in the main F0 display window is clicked, the histogram window shows a histogram of the distribution of quantized pitches in the selected sign. Below this

[7] `http://marsyas.sourceforge.net`

histogram is a slider to choose how many of the largest histogram bins will be used to generate a simplified contour representation of the F0 curve. In the limiting case of selecting all histogram bins, the reduced curve is exactly the quantized F0 curve. At lower values, only the histogram bins with the most items are used to draw the reduced curve, which has the effect of reducing the impact of outliers and providing a smoother "abstract" contour. Shift-clicking selects multiple signs; in this case the histogram window includes the data from all the selected signs. We often select all segments with the same word, trope sign, or neume; this causes the simplified contour representation to be calculated using the sum of all the pitches found in that particular sign, enhancing the quality of the simplified contour representation. Figure 7 shows a screenshot of the browsing interface.

In the current work we implemented a mode that allows the researcher to sort the samples based on the Dynamic Time Warping cost from one sample to the other. The interface allows the user to select an arbitrary gesture from the interface, and then perform a sorting of all other gestures to it. In the example shown in Figure 7 the user has chosen a "revia", and has sorted all the other gestures based on their DTW-based alignment distance from this first revia. One can see that the gesture closest to this revia is another revia gesture from a different section of the audio file.

4 Summary and discussion

By combining the expert knowledge of our scientific collaborators with new multimedia web-based tools in an agile development strategy, we have been able to ask new questions that had previously been out of reach. Chant research is a challenging domain where problem seeking is important. Participatory design together with content-aware visualizations and analysis tools can help researchers interact with large collections of annotated audio recordings of chant in interesting new ways. The integration of all the different components in a single web-based interface is critical for an effective system. Given the subjective interpretive nature of musicological research each algorithm in isolation would be of little use. This necessitates the development of the system as a whole and makes evaluation harder. Ultimately we only have few experts users (one in our case) and the only feedback we can receive is through them. By including them in the design we having been able to create a system that our expert finds useful and is willing to spend significant time interacting with it.

There are many directions for future work. We are planning to explore the histogram-based contour simplification in conjunction with the dynamic time warping alignment process to identify what is the "optimal" simplification of the pitch contours. More careful study of the results by musicologists is also required. Making the system available on the web can help collaborative approaches and reduce the learning curve required for usage. We also hope to make the annotation process part of the web interface and enable uploading of recordings from researchers around the world.

5 Acknowledgments

We would like to thank Matt Wright for initial work and Emiru Tsunoo for the implementation of dynamic time warping as well as the Social Sciences and Humanities Research Council (SSHRC) of Canada for financial support.

References

[1] A. Camacho. *A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music*. PhD thesis, University of Florida, 2007.

[2] B. Duggan, B. O' Shea, and P. Cunningham. A system for automatically annotating traditional irish music field recordings. In *Int. Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2008.

[3] P. Hanna and P. Ferraro. Polyphonic music retrieval by local edition of quotiented sequences. In *Int. Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2007.

[4] A. Hauptman and et al. Informedia at trec 2003 : Analyzing and searching broadcast news video. In *Proc. of (VIDEO) TREC 2003*, Gaithersburg, MD, 2003.

[5] A. Hauptman and M. Witbrock. *Informedia: News-on-demand Multimedia Information Acquisition and Retrieval*. MIT Press, Cambridge, Mass, 1997.

[6] T. Karp. *Aspects of Orality and Formularity in Gregorian Chant*. Northwestern University Press, Evanston, 1998.

[7] Z. Kodaly. *Folk Music of Hungary*. Corvina Press, Budapest, 1960.

[8] C. L. Krumhansl. *Cognitive Foundations of Musical Pitch*. Oxford University Press, Oxford, 1990.

[9] K. Nelson. *The Art of Reciting the Koran*. University of Texas Press, Austin, 1985.

[10] S. Ness, M. Wright, L. Martins, and Tzanetakis.G. Chants and Orcas: Semi-automatic tools for Audio Annotation and Analysis in Niche Domains. In *Proc. ACM Multimedia*, Vancouver, Canada, 2008.

[11] L. Treitler. The early history of music writing in the west. *Journal of the American Musicological Society*, 35, 1982.

[12] G. Tzanetakis. *Marsyas-0.2: A case study in implementing music information retrieval systems*, chapter 2, pages 31–49. Intelligent Music Information Systems: Tools and Methodologies. Information Science Reference, 2008. Shen, Shepherd, Cui, Liu (eds).

[13] G. Tzanetakis, K. A, W. Schloss, and M. Wright. Computational ethnomusicology. *Journal of Interdisciplinary Music Studies*, 1(2), 2007.

[14] G. Wigoder and et al. *Masora, The Encyclopedia of Judaism*. MacMillan Publishing Company, New York, 1989.

[15] H. Zimmermann. *Untersuchungen zur Musikauffassung des rabbinischen Judentums*. Peter Lang, Bern, 2000.