Information Technologies and Innovation in Sanskrit-Based Indian Studies

Abstracts

Dominik Wujastyk (Universität Wien)

Friday, 9:10 – 9:55

Text, structure and embedded meaning

Since computers came into widespread use amongst university academics in the early 1980s, scholars have been feverishly entering text into computers and creating databases of textual material. What general principles can be applied to these activities? What is important, and what is ephemeral? How can we ensure that the texts we create, both our own writings and transcriptions of classical texts, remain useable in the middle to long term future? What is the meaning of "added value" in relation to e-text creation? What is the difference between a database and a tagged and structured text file? This presentation will explore the importance of the Text Encoding Initiative in relation to these questions, and reflect upon how we may make the best use of public standards in planning our e-activities.

Oliver Hellwig (Universität Heidelberg)

Friday, 9:55 – 10:40

Improving the automatic tokenization and lexical analysis of Sanskrit texts

Indology nowadays disposes of large collections of digital Sanskrit texts that offer new insights into the linguistic, intellectual and religious history of ancient India. To use such collections effectively for research purposes, scholars need to have access to the linguistic structures of these texts and especially to their lexical and grammatical information. The presentation describes which methods from Computational Linguistics can be applied to extract and store such information in a semi-automatic way. The main part of the presentation is dedicated to the questions of how the accuracy of such automatic analysis methods can be improved and which accuracy rates we can expect for unsupervised linguistic analysis at the moment. The presentation is concluded with a sketch of how a complete linguistic analysis system for a text collection such as GRETIL could be designed.

Anand Mishra (Universität Heidelberg)

Friday, 11:10 – 11:55

Integrating traditional analyses in computational processing of Sanskrit texts

The *Aṣṭādhyāyī* of Pāṇini is a device to produce linguistic expressions using its constituent elements. It prescribes a set of fundamental components that constitute the language, characterizes them using a number of attributes, and specifies rules to form the final linguistic expressions. In the first part of my presentation, I will briefly outline my modelling of the Pāṇinian system of Sanskrit grammar and demonstrate its implementation on a computer. Further, I would discuss the employment of these algorithms for processing Sanskrit texts.

In the second part of my presentation, I would like to brief about the process of annotating Sanskrit texts (belonging to the *Puṣṭi Sampradāya*) with a view to comprehend the inherent structure of the text and the traditional approach of organizing its content (e.g. an *adhikaraṇa* consisting of *viṣaya*, *saṃśaya*, *pūrvapakṣa*, *uttarapakṣa*, *saṃgati*). Apart from that, I will talk about the extension of the tag-set to represent the later reception of a text, and the processes for understanding and interpreting it.

Jan Westerhoff (Durham University)

Friday, 14:30 – 15:15

Śāstravid: A new tool for the study of Indian philosophy

In my talk I will give a brief outline of a project funded by the European Research Council which started in October 2010. Its aim is to transform the way Indian philosophical texts are currently studied. In order to do this it will provide a philosophical analysis of a set of central works from the Indian tradition, a set well known for its demanding content and the conceptual complexity of the arguments it contains. This analysis will incorporate a set of cutting-edge methodological principles, the most important of which is the intricate interlinking of conceptual analysis and its textual basis. These principles will be encoded in a web-based electronic tool which will be developed during the course of the project. This tool, called Śāstravid, incorporates an example of the new research paradigm and at the same time facilitates further academic research based on the same approach. Its aim is to provide a key that unlocks the contents of Indian philosophical texts by bringing together the information contained in commentarial works, both ancient and modern, in such a way that it is easily accessible from the text itself.

Apart from structuring the works, providing commentarial background and linking texts to other texts, the present project develops a radically new way of analyzing the text's conceptual contents. This analysis is linked directly to the texts themselves, which makes it easy to switch between philological and philosophical modes of research. This linkage between conceptual and textual analysis embodies a radically new way of thinking about Indian philosophical texts that is located at the very frontier of the discipline. This kind of conceptual analysis with direct links to its textual basis has never been applied to the study of philosophical texts before and constitutes a highly original approach to the study of these materials. It pushes the study of Indian philosophical works beyond the domain of mere textual scholarship into the emerging field of research studying Indian philosophy as philosophy that is characterized by a close engagement with the concepts and arguments present in these texts without sacrificing philological accuracy. The system's electronic nature ensures that the new approach developed here is dynamic and can constantly reflect the current state of research. It is modular rather than monolithic and can be enlarged and enhanced in a step-by-step manner. It is flexible since it does not superimpose pre-fabricated structures onto texts but derives them from the texts themselves. It is multi-dimensional as it allows a simultaneous engagement with the philological and philosophical aspects of a text.

The project will facilitate cooperation between scholars working on Indian philosophical texts, widen access to these materials, encourage their interdisciplinary study, and transform the way research on them is carried out.

Birgit Kellner (Universität Heidelberg)

Friday, 15:15 – 16:00

The Indian Logic Knowledge Base: Towards the development of terminological resources in Sanskrit knowledge systems

General dictionaries of Sanskrit do not reflect the semantic nuances of technical terminology developed in Sanskrit knowledge systems. These dictionaries for the most part date back to the 19th century, but Indological research made much progress since their publication. This progress is often hard to see. Research results are often published in journals that are difficult to obtain; important observations are at times contained in lengthy footnotes to critical editions or translations of Sanskrit texts, which makes it hard especially for students and non-specialists to keep track of the state of the art.

Collaboratively created online resources can be used for keeping a common repository of up-to-date knowledge on technical terminology. This idea was the driving force behind the creation of the Indian Logic Knowledge Base (ILKB), developed since 2004 and intended as an information resource for the terminology of ancient Indian theories of reasoning. Special attention was given to presenting definitions, explanations and exemplifications of terms on the basis of primary sources.

In this paper, I present the ILKB, discuss its history and the practical and technological problems that were encountered in its implementation; I will also outline future directions of development.

Olga Serbaeva Saraogi (Universität Zürich)

Friday, 16.30 – 17:15

The place and role of the Śaiva Tantric texts in early mediaeval Indian literature with a particular accent on the Purāṇas: A reassessment based on the computer-assisted statistical analysis of textual parallels and indexes

The aim of the proposed project is to ascertain the position of a selected corpus of Śaiva tantric texts within the larger body of early mediaeval Indian literature with a particular accent upon the Śaiva Purāṇas. The way forward is to find, analyze and present all textual parallels within and between purāṇic and tantric corpora, and to combine this evidence with that obtained from other forms of textual analysis in order to establish with precision the position of the Śaiva Tantras within early mediaeval Indian literature as a whole.

Research will proceed in three stages: (1) A specially created program will be used to locate, analyze and present in an accessible form all the textual parallels existing: in the selected corpus of Śaiva tantric texts, in the selected corpus of the purāṇic texts, and between the two corpora. The similarities and differences between the texts of the two corpora will be defined through the use of statistical analysis, on the basis of which a preliminary relative chronology for both will be established. (2) This chronological structure will be further refined through the analysis of the occurrence of selected concepts in each of the two corpora. All the material discovered in the course of stages 1 and 2 will contribute to (3) the establishment of the place and role of Śaiva tantric texts in the development of early mediaeval Indian literature through demonstrating the interrelations of the purāṇic and the tantric corpora.

The project will offer a powerful tool for Sanskrit textual analysis to Indology in general, and the parallels found between the tantric and non-tantric corpora uncovered by it are bound to change scholarly representations of textual interrelations in early mediaeval Indian literature.

Sven Sellmer (Adam Mickiewicz University, Poznań)

Friday, 17:15 – 18:00

IT-based methods in metrical studies of the *Mahābhārata*

This paper aims at showing some possibilities and limitations of IT-supported methods in metrical studies of the *Mahābhārata*. It consists of three parts:

1. presentation of some specially designed PHP programs that make possible the extraction of metrical (and cognate) data, like verse and caesura patterns, from electronic versions of epic texts and their preparation for further analysis;
2. application of several statistical methods to the collected data and preliminary evaluation of the results;
3. reflections on future possibilities of IT-supported methods in metrical studies and on the limitations of such an approach; it appears that the most fruitful strategy will be based on a combination of IT tools, statistical methods and, last but not least, a thorough analysis of single verses, based on a better understanding of the mechanisms of (oral and written) versification.

Himal Trikha (Universität Wien)

Saturday, 9:00 – 9:45

A Study of the Manuscripts of the Woolner Collection

The A. C. Woolner Collection at the Punjab University Library, Lahore, Pakistan, is a fine collection of about 8000 Sanskrit manuscripts on a wide range of topics. A recently concluded FWF (Austrian Science Fund) project aimed at bringing the content of the philosophical manuscripts in the collection to wider notice and accessibility, using a blend of modern digital technology and traditional philological skills.

See further http://www.istb.univie.ac.at/swc.

Philipp A. Maas (Universität Wien)

Saturday, 9:45 – 10:30

On solving the problem of textual contamination by means of computer-aided stemmatics

As is well known, the study of the interrelationships of manuscripts (stemmatics) as formulated by Paul Maas is based on the premise that every manuscript is a copy of a single exemplar. Since in the written transmission of Sanskrit works contamination (i.e. the mixture of two or more text versions in the process of producing a new copy) is widely spread, many editors regard the study of text genealogy as virtually impossible in the case of Sanskrit texts. In evolutionary biology, which has to cope with similar methodological problems, special methods have been developed in order to create hypotheses concerning the genealogy of biological species even from partly contradictory data. "Cladistics" is one of these methods, which by means of its underlying theoretical assumptions qualifies for an application in stemmatics. In this presentation I would like to show that a complementary approach to text genealogy combining the application of cladistic software with a text-critical discussion of variant readings, may lead to plausible stemmatic hypotheses even in the case of contaminated transmissions.

Informal Presentations

Saturday, 11:00

Olga Serbaeva Saraogi (Universität Zürich)

Brief description of the computer program SANSKRIT_PARALLELS

The present program runs on FoxPro9. Its main aim is to find parallels between Sanskrit Śaiva tantric texts written in *ślokas*. “Sanskrit_Parallels“ enables analysis on a basis other than word separation because the chosen texts have the two following features: they do have an important quantity of parallels between them, and they are metrical. Two different algorithms of comparison are used.

The current version of the program has produced preliminary results for about 30 Śaiva tantric texts. These texts, consisting altogether of around 108,000 lines, with the minimum of correspondence selected at 50%, contain 59,000 parallels. The results of comparison are incorporated into a table used for the construction of "trees of relations" showing the closeness of groups of texts to each other.

This database of parallels constitutes the basis of the analytical part of the future research project, which consists in establishing precise relations between the texts and the reconstruction of a new relative chronology.

Jonas Soiné (Universität Tübingen)

Word sense disambiguation in the context of Sanskrit

The M.A. thesis *Wordbedeutungsdisambiguierung im Kontext des Sanskrit* deals with the subject of automatic word sense disambiguation. Since this field of scientific research has not yet been applied to digitized Sanskrit data, it seemed reasonable to give an introductory approach to this task.

The thesis is composed of three parts. In the first step, an overview of the underlying linguistic terminology is given. This comprises concepts like ambiguity, polysemy, homonymy or hyponymy, and the phenomenon of semantic change in Sanskrit. In the second part, the formal NLP methods for the task of word sense disambiguation (WSD) are introduced, and with the sense-tagging system San-SemAn occurrences of the lexemes *jana* and *śārdūla* from the Rāmāyaṇa. Part of the Digital Corpus of Sanskrit is sense-tagged by two human annotators. In the last part, this tagging exercise is evaluated with an observed ITA of 91.2%. Then the Decision Lists algorithm of Yarowsky is applied to the lexical sample task of WSD for these two words. A measured accuracy of approximately 97% shows convincingly that future research in this area is worth looking at. The presentation sketches the central ideas of WSD and describes how such a system can be integrated into Indological research.

Keywords: WSD, NLP, CL, AI, Decision Lists, computational Sanskrit

Markus Schüpbach (Universität Zürich)

Determining and comparing concepts in the Mokṣadharmaparvan – requirements of a database-assisted inventory

Ernst Prets (Österreichische Akademie der Wissenschaften, Wien)

A web-based database of Nyāya text fragments

The early school of Nyāya, one of the six orthodox Hindu philosophies, is mainly represented by the four main preserved commentaries and sub-commentaries on the school's founding text, the *Nyāyasūtra*, which is ascribed to the sage Akṣapāda (? 2nd century CE) and was probably finalized in its classical form by anonymous redactors around 400 CE. Early Nyāya literature includes a number of works that we only know of through references found in later texts. The information on these works is often limited to the names of the authors and/or those of their works, but sometimes includes quotations or paraphrases from them.

An analysis of such quotations and doxographies is highly important, since relevant developmental steps of the system's ideas are often documented only indirectly by the four major commentaries. Some basic ideas seem to be related to authors whose works are lost and are often referred to in the works of authors of opposing schools and systems.

The first step of the project "Fragments of Indian Philosophy" has therefore been to assemble fragments of these early Naiyāyikas' works, providing the first overall view of all preserved fragments and doxographies. The creation of an internet-supported database to collect and analyze fragments of early Nyāya texts has been an essential component of this project. In my presentation, I will introduce the database with a practical demonstration of its functions, discussing both its present applications and how these could be broadened and developed further. I will also talk about some of the technical challenges we were confronted with in building the relational database, and what lessons can be learnt from them for future projects involving similar databases.

The core of the interactive database is a relational database based on the open-source software "PostgreSql" (FN: http://www.postgresql.org/). It runs on a dedicated Linux-server maintained by the Department for Information Technology Services (ITS) of the Austrian Academy of Sciences (FN: http://www.oeaw.ac.at/arz). All entries stored in the database are encoded in UTF-8, thus allowing to store and to represent all special characters needed. The whole database is dumped daily and saved to the central backup system in order to avoid any loss of data.

Patrick Mc Allister (Österreichische Akademie der Wissenschaften, Wien)

A tool for collaborating on e-texts

In this talk a common tool used by programmers will be introduced, and a suggestion made as to how it might be useful for joint work of nonprogrammers on e-texts (such as exist in GRETIL, e.g.). The tool I'll introduce is a distributed version control system called "git". It is a system that controls different versions of a number of files (usually text files), and allows for simultaneous work on the same set of files, tracking each contributor's efforts, and indicating conflicting changes. I'll propose that such a system might be a very useful addition to currently existing text repositories, and, coupled with the TEI guidelines, could enable anyone using those repositories to also contribute to them in a structured and simple manner.