PROPOSED WORKSHOP ON
DIGITISING INDIC TEXTS IN EDITABLE FORMAT
TO BE JOINTLY ORGANISED BY
THE ASSOCIATION FOR LITERARY AND LINGUISTIC COMPUTING, UK
AND
JADAVPUR UNIVERSITY, KOLKATA, INDIA

AIMS AND SCOPE:

The Workshop will address the digitization of texts in editable format with special reference to Indic languages, particularly Bengali. The matter is of importance from two angles:

(a) technical: solving the special problems of digitizing Indic scripts.
(b) scholarly, archival and bibliographical: creating literary archives and fostering an editorial culture.

(a) Technical:

The special problems of many Indic alphabets, including Bengali, are as follows:

a) Only consonants are written in full, with the accompanying vowel sounds indicated by tagged-on vowel markers. Although these vowel sounds phonetically follow the consonant, they are sometimes written before it, and at other times above or below. This makes font creation and screen visualization more difficult. It also makes certain functions of text analysis – e.g., phonetic analysis, or collation of texts with variant spellings – specially problematic, as the visual sequence presented on screen does not match the phonetic sequence followed during keying-in and hence registered in the processing unit.

b) The Bengali alphabet has about 50 letters, without majuscule/miniscule variation. (The 'about' is significant: there is some debate as to what constitutes a letter.) There are also a huge number of conjunct letters (2, 3 or even 4 conjunct consonants plus a vowel), besides a range of vowel tags (in several forms for each vowel, depending on the consonant it is attached to) and some 'half-letters' (consonants without vowels). All this vastly increases the number of glyphs, to a total of 450-500 items in the average 'typecase'. Moreover, the forms of these conjuncts vary from font to font in print. Despite a trend towards simplification, most of these conjunct letters will be with us for a long time to come. And needless to say, they will always have to be processed in the case of extant texts.

c) This makes it a great challenge both
   i) to generate these conjuncts in fonts for electronic use: though many Bengali fonts have been generated, they nearly always have a measure of glitches or instability; and
   ii) to develop an OCR programme that can 'read' these conjuncts in extant print fonts. Again, much work has been done, but with an accuracy of 95%+ only in certain text situations. It is often no more than 85%.

The proposed workshop would offer a rare chance for specialists from abroad to learn of these problems and the ways Indian experts are tackling them, and in turn to suggest new approaches and solutions based on their experience with Roman or other alphabets.

The issues to be taken up could be:

1. improving and stabilising Unicode fonts in Bengali and related Indian languages;
2. improving and extending the OCR programmes developed so far, to create texts suitable for all kinds of textual and phonetic analysis;
3. improving an experimental collation programme already developed.

(b) Scholarly, archival and bibliographical

Bengali has an extensive literature, whose 'modern' phase began in the early 19th century. It was the first Asian language to come into extensive contact with Western literature and thought. From the 19th century, it produced a great range of educational, social, religious and even scientific/technological works, as well as the first body of modern creative literature in any Indian language. This corpus forms the textual basis of what is often called the Bengal Renaissance, reaching its climax in the work of Rabindranath Tagore (1861-1941). Despite dramatic changes over the last half-century, Bengali literature and culture can still be said to live in the aftermath of the Bengal Renaissance.

Bengali was also (except for a brief, soon closed chapter in western India) the first Asian language to earlier material is seminal for Indian – indeed, world – printing history. It shows a specially interesting amalgam of Western techniques developed over 350-400 years with innovations specific to the local script and conditions of production.

Bengal is famous for its vibrant literary culture, with a rich body of creative works and their scholarly interpretation. But so far as textual scholarship and editorial attention is concerned, this creative and critical activity is taking place in a near-vacuum. Hardly a score of Bengali texts are available in critical editions as international scholarship understands the term. Original 19th-century works are often hard to come by, surviving only in one or two copies, often badly preserved and deteriorating in the hot and humid climate.

There is thus a great need for

1. Ensuring the sheer physical record of this rich body of works in digital form.
2. Generating an editorial culture by producing electronic texts in editable format.

Combining the technical and scholarly imperatives, the Workshop would help us work towards the goal of ensuring that texts in Indic languages can be processed in every way considered standard for the Roman alphabet – search/concordance, collation, OCR, phonetic analysis etc. – and hence made accessible for all kinds of editorial and scholarly activity.

We propose Bengali as the language to focus on, as (a) the language of the region where Jadavpur University is located; and (b) a language where a good deal of work has already been done in the above respects, ensuring an informed and interactive milieu. Today, efficient word-processing programmes exist for all major Indian languages. Some work has been done on OCR programmes in Devanagari (Hindi) script, tha being the country's official language, and in Brahmi-based scripts including Bengali. But it seems fair to say that the other advanced functions noted above, essential for textual study and editorial processing, have hardly been pursued elsewhere in India. In Bengali, at least a foundation has been laid which can be consolidated in the Workshop.

Of course we shall invite experts in other languages, as most of the issues are germane to their work as well. And we need hardly say that the ALLC experts need not have knowledge of Bengali or other Indic scripts. Rather, their contribution will be valuable precisely by virtue of drawing on a broader field of research and experience.

LOCAL PARTICIPANTS;

We are planning for intensive participation by a core group of 15-20 local members. Another 15-20 persons – a few senior members, but chiefly young project and research staff – may attend to absorb the culture of electronic texts. These 'auditors' will be welcome to take active part, but they are unlikely to do so often. We hope, nonetheless, that they will feel encouraged to interact with the experts outside the workshop and in the future. In particular, young staff working on a single limited aspect of electronic texts will benefit greatly from this broader experience.

Among the established scholars and workers in the field who, we hope, will attend the Workshop are the following. This list is neither confirmed nor complete.

a) Professor Kalyan Kumar Datta and Professor Samar Bhattacharya, School of Education Technology, Jadavpur University: members of the 'Vidyasagar' group that developed the first Bengali electronic fonts.

b) Professor Mita Nasipuri, Dr Anirban Ray Chaudhuri, and other members of the Department of Computer Science and Engineering, Jadavpur University, associated with CMATER. an active OCR development centre attached to their Department.

c) Professor Bidyut Baran Chaudhuri, Indian Statistical Institute, Kolkata, who developed the first viable OCR programme in Bengali.

d) Professor Ashok Mukhopadhyay, sometime Professor of Printing Engineering, Jadavpur University and CEO of the University Press attached to Visva-Bharati, the university founded by Rabindranath Tagore and till recently custodian of his works.

e) Professor Gautam Sengupta, Professor of Linguistics, University of Hyderabad: a noted applied linguist with much work on Bengali fonts and electronic texts.

f) Professor Palash Baran Pal and Professor Somendra Mohan Bhattacharya, Saha Institute of Nuclear Physics, Kolkata: physicists who have also worked extensively on Bengali fonts, word-processing programmes and online text databases.

'Ankur' group, who are working with Bengali electronic fonts and texts.

h) Members of the School of Cultural Texts and Records, Jadavpur University: literary and humanistic scholars with expertise in electronic texts: e.g., Professor Subha Chakraborty Dasgupta, Professor Amlan Das Gupta, Dr Amitava Das, Dr Samantak Das.

Among younger delegates and 'auditors', we would specially welcome the young project staff and ancillary workers attached to the School of Cultural Texts and Records, Jadavpur University, and various relevant units of the Faculty of Engineering and Technology.

N.B.: Jadavpur University is planning a two-day conference on electronic texts to precede the Workshop. Here, the audience will be bigger and less specialized. The ALLC need not be directly involved in organizing or funding this conference, but we hope its delegates will agree to come two days earlier to attend it and offer papers.

ALLC PARTICIPANTS: To be provided by ALLC

PROGRAMME SCHEDULE (to be finalised after consultation with the ALLC):

We propose a two-day workshop, on 9-10 February 2007 (Friday and Saturday). There will be two sessions each day. We would propose the following division of topics:

1. Improvement of Unicode font generation in Bengali, for manual keying-in of texts in editable format.
2. Degraded document processing and core OCR engine.
3. Advanced OCR system with enhanced graphical user interface for document digitization in editable format retaining original layout.
4. Further development of the collation programme for Indic scripts already available with the School of Cultural Texts and Records.

WORKSHOP OUTCOMES AND DISSEMINATION:

As remarked above, despite the rich textual heritage in Bengali, there is relatively little textual an editorial awareness. Today, electronic resources can enable us to achieve this awareness, and thus leap-frog into an advanced editorial culture, in a relatively short span of time. Once we have a 'bank' of digital texts in editable format, we can proceed to electronic editing and other advanced processing of texts.

This calls for the close interaction of textual and documentary scholars in Indic languages with experts in electronic texts and literary computing. As yet, there is little such contact. We hope the Workshop will help to bridge the gap. It will bring technological and moral support to the people actually working on electronic texts in Bengali and other Indic languages, and help them find their place in an international context. At the same time, it will foster a more informed level of operational skill among general textual and literary scholars. Delegates from both these categories will form the first group of beneficiaries.

There will be a 'spread effect' extending to other Indian languages, where the problems are often the same. Experts in electronic texts in these language areas would constitute a second tier of beneficiaries.

There would also be 'spread effect' of another sort: raising textual awareness, and exercising that awareness through electronic texts, among all students and archivists of Indic languages, literature and history – in fact, any discipline requiring textual documentation. This raising of consciousness would confer an unquantifiable benefit at a third level.

COSTS:

Jadavpur University would meet all local expenses of the workshop, and expenses relating to Indian delegates. It could, at its own cost, put up the ALLC delegates at the Jadavpur University Guest House.

We would request the ALLC to sponsor the travel of their delegates, and their stay if elsewhere than at the Jadavpur University Guest House. We would also be grateful if the ALLC were to provide Jadavpur University with appropriate software not readily available in India, or too costly for local levels of funding.

Note on Jadavpur University:

Jadavpur University began as a technological institution. But it is unique among Indian universities in that, over the last 20-30 years, it has built up one of India's most successful Arts Faculties, including four departments of language and literature: Bengali, Comparative Literature, English and Sanskrit. It is arguably here. The people who developed them are still around, and will take part in the Workshop. Notable work on Bengali and Devanagari OCR is currently going on in the Department of Computer Science and Engineering: they have developed 'Anulikhan', the first OCR system in any Indic script with editable format and original layout retention.

The School of Cultural Texts and Records (comprising technologists as well as literary scholars and historians) conducts a range of textual projects using electronic resources including an experimental collation software – the first in any Indic language. The School already has a major digital archive of Bengali literary manuscripts in non-editable form, various bibliographical databases including the first Short-Title Catalogue in any Indian language, and a large music archive in digitized form. There is interaction between members of the Arts Faculty and the Faculty of Engineering and Technology in matters of textual computing. We hope the Workshop will enhance this.

Technical/Personnel Requirements:

No special requirements. Jadavpur has most if not all facilities likely to be required. The workshop would be held in an air-conditioned seminar room with round-table seating, and projection facilities using a laptop (loaded with Windows XP) and LCD projector. There will be personnel to set up and handle all this equipment. Of course, if a delegate wishes to demonstrate any special software, he/she will have to bring it along.

We are not thinking of providing hands-on computer access to individual participants. But if necessary, we could arrange such a venue with up to 10 computers, each accessed by 2-3 participants.