Report on the Sanskrit Text Archive Conference Austin, Texas, October 28–29, 1988.

D. Wujastyk

November 30, 1988

Indologists in Europe and America have long been aware of the advantages that their colleagues in classics have been reaping because of the existence of the Thesaurus Linguae Graecae and other digital archives of classical literature.

We have not only envied the classicists their ability to generate concordances and to study metre and syntax efficiently. But we have been acutely aware of the special benefits that a large Sanskrit text archive would confer simply as a rapid access library. Sanskrit literature, which is many times larger in volume than Greek and Latin put together, presents such problems of sheer magnitude that no one can write anything about it without knowing full well that what is said is partial. This is not always just because no scholar can read more than a fraction of the surviving material, but also because only a very few libraries in the world own more than a fraction of that material.

Many Sanskritists are highly computer literate, and indeed several successful computer-based research projects have been completed. It was probably only a matter of time before someone proposed a TLG for Sanskrit.

On October 28th and 29th this year, a meeting was held at the University of Texas at Austin to explore a project to set up a substantial text archive of Sanskrit language materials. A brief note in advance of the meeting was sent to HUMANIST (Humanist Mailing List, Vol. 2, No. 239. Saturday, 15 Oct. 1988.), and this is a report on how the meeting went.

The impulse for the project has come from Richard Lariviere, Professor of Sanskrit at Austin and Director of the Center for Asian Studies, and from Stephen Phillips, Professor of Philosophy at the same university. The participants of this first pilot meeting included the following mixture of Sanskritists, linguists, classicists and computer scientists:

Subhas C. Biswas (Director of Central Secretariat Library, Ministry of Human Resource Development, New Delhi)
Theodore Brunner (TLG, UC Irvine),
George Cardona (Linguistics, U. Penn.),
Colin Foote (LA),
John Freeburn (Nyack, NY),

Michael Gagarin (Classics, UT Austin),
Edwin Gerow (Sanskrit, Reed College, Portland),
Robert P. Goldman (Sanskrit, UC Berkeley),
Wilhelm Halbfass (Sanskrit, U. Penn.),
Julie Hiebert (Sanskrit, UT Austin),
Susan Hockey (Oxford U. Computing Service),
Daniel Ingalls, Jr. (formerly Xerox PARC, SmallTalk, etc.),
Daniel Ingalls, Sr. (Sanskrit Prof. Emeritus, Harvard),
Robert D. King (Dean, College of Liberal Arts, UT Austin),
Richard W. Lariviere (Convener; Sanskrit, UT Austin)
Win Lehmann (Director, Linguistics Research Center, UT Austin),
Tony Meadow (Bear River Assocs., Berkeley, CA),
Barbara S. Miller (Sanskrit, Columbia),
James Nye (Bibliographer for South Asia, Regenstein Library, Chicago U.),
Herman van Olphen (Oriental and African Languages and Literatures, UT Austin),
Stephen Phillips (Philosophy, U. Texas),
Sheldon I. Pollock (Sanskrit, U. Iowa),
Edgar Polomé (Oriental and African Languages and Literatures, UT Austin),
K. Kunjunni Raja (Adyar Library, Madras),
Raja Rao (Author, UT Austin),
Mythili Rao (Tata Inst., Bombay),
Ludo Rocher (Sanskrit, U. Penn.),
Dale Steinhauser (Maharshi Vedic U.),
Gary Tubb (Sanskrit, Brown U.)
Om Vikas (Centre for Advanced Study of Electronics, New Delhi)
Michael Witzel (Sanskrit, Harvard),
Dominik Wujastyk (Sanskrit, Wellcome Institute, London)

The meeting opened with a talk by Theodore Brunner about the history and background of the TLG project, with special attention to its organization. Susan Hockey followed this with a strong exhortation to the planners of the new project to use the Standard Generalized Markup Language as the basis for their coding scheme. She explained what SGML was, and also outlined the aims of the new ACH/ALLC/ACL Text Encoding Initiative.

After a break for lunch, Daniel Ingalls Sr., a grand old man of American Indology, and the former teacher of many of those present, outlined his own efforts to study parts of the *Mahābhārata* from a metrical point of view, using computer methods. His son Daniel Ingalls Jr., a programmer of renown, described a SmallTalk program he had written at his father's instigation some years ago, which performed character recognition on a bitmap image of a Sanskrit page. This program has subsequently been translated into Object Pascal by Dale Steinhauser, and a working version of this was demonstrated on a Macintosh.

James Nye discussed his work on parsing and encoding Monier-Williams, one of the main Sanskrit-English dictionaries, with a view to providing the text in machine readable form. He also raised some interesting points concerning

the sociology of knowledge and its relation to the media of transmission, noting that the large scale migration of Sanskrit literature to a new medium was bound to have repercussions beyond the merely technical.

On the second day, George Cardona demonstrated a mouth-watering application of HyperCard on the Mac to the problem of multiple Sanskrit commentaries. He showed the Sanskrit grammatical text of Pāṇini, well known for its complexity and multi-layered commentarial tradition, with commentaries appearing in screen windows with correct and full contextual reference, examples, and many other features. The demonstration only covered part of the text and commentaries, but was already enough to be a useful teaching aid. Cardona also gave a detailed specification for a possible future project to create a linguistic database of Sanskrit with built in grammatical and lexical systems, a sort of Sanskrit expert system, based on Pāṇini's grammar.

Sheldon Pollock then initiated a discussion about the possibility of founding a new association of Sanskritists, with a new journal, to fill a perceived gap in the professional institutions available to Sanskritists in the USA. The discussion turned towards the suitability of the American Oriental Society, which already provides such services, and could provide more if some of its dormant committees were brought up to full function. Since several of the people in the room were either on the main AOS committees, or were shortly to be so, it was felt that things might begin look up in this area.

A concluding open discussion was chaired by Richard Lariviere, with much fruitful discussion of matters both general and particular.

Amongst the most central issues were whether current OCR technology (including the system that had been demonstrated to us) was adequate, or whether the data input should be done by professional offshore typing agencies, as is done for the TLG. A consensus for the latter position emerged.

A related issue concerned the level of manual markup necessary before the texts were sent for data input. Obvious features such as headings, speakers, prose, verse, and so on would naturally be tagged. But what about sandhi? This phonetic feature of continuous speech is explicit in all Indian orthographies. I.e., whereas we might write "dogs and cats", a Sanskrit speaker would write "dogzen kats". There is also an orthographic feature in most Indian alphabets whereby some adjacent characters join together graphically. Undoing sandhi and conjunct characters automatically appear not to be trivial tasks. Furthermore, there are many texts in which Sanskrit authors make deliberate play on the ambiguities that can be introduced by sandhi ("'Gladly', the cross-eyed bear"). Should the archive contain "dogzen kats" as printed in an imaginary edition, or "dogs and cats", "dogz-en kats", "dogs-and-cats", or what? All agreed that the added burden of coding for sandhi would be a colossal task, and the recent work of Verboom ("Towards a Sanskrit Word Parser", *LLC* 3.1 (1988), 40–44) was circulated by way of a suggestion that it might not be impossible to build an automatic sandhi parser. But the issue was not fully decided.

A third major issue discussed was whether a Sanskrit project should build up its text archive in the manner of the TLG, i.e., through a single, expensive, centralized and carefully controlled data input scheme, or in the manner

of the Oxford Text Archive, which accepts texts from all quarters, with the addition of an attempt to impose standards and quality control "after the fact" of data input, i.e., a sort of cottage industry approach to data capture. My view is that the latter approach might appear to hold the seductive promise of quicker, cheaper initial results, and could perhaps be workable to a limited extent, through the application of such tools as SGML parsers to validate and standardize formats, but that in the long run, the formation of a serious, professional textual databank of the magnitude envisaged absolutely requires the kind of monolithic approach that has been proven to work for the TLG.

Naturally there was discussion of financial and organizational matters. Ted Brunner underlined the effectiveness of the committee structure that has evolved to run the TLG project, and it was generally felt that this was a very suitable model. The AOS could probably provide an advisory committee much as the APA had done for the TLG. There was quite a lot of murmuring about how it would be easy to get the project off the ground if there were a million dollar gift of the kind that had been given to Ted Brunner at the beginning of the TLG. But Brunner explicitly countered this by noting that the importance of such a generous initial endowment could be overestimated. It carried with it the danger of complacency, and in the overall scheme of a project of the TLG's magnitude, a million dollars was spent rather quickly. While such a pump-priming fund would be a marvellous help, it would be equally important to focus sharply on getting longer term fund-raising and financing mechanisms in place at the outset.

After two hours of discussion, Stephen Phillips proposed a twelve point document summarizing the majority views of those present, and after some further discussion, a vote was taken and an abbreviated version of the document was accepted as the final view of the meeting.

What the Conference decided–in outline–was this: that a Sanskrit text archive was a very good idea; that the texts in it should be encoded according to an SGML based scheme; that the American Oriental Society should be approached to provide an Advisory Board to the project; that the project to create such an archive should be based at UT Austin; that Richard Lariviere was the person best suited to have this particular millstone hung around his neck (and, incidentally, the only volunteer); that the project should try to affiliate itself with an appropriate Indian cultural body; that a pilot project of encoding a single text should be undertaken at UT in the first instance, during the next year, in order to assess the costs and difficulties involved (the *Mahābhārata* or the *Rāmāyaṇa* were the most popular candidate texts).

We dispersed to the four corners of the world with bright hopes that in not too many years a magnificent new tool for Sanskrit studies may become available.