Computational Database of Mahābhārata for Assisting Research/Analysis

Dr. P. Ramanujan

Mahābhārata is well-known to exist in various versions. Under a project funded by Central Secretariat Library, Dept. of Culture, New Delhi, the Indian Heritage Group developed a database of Mahābhārata as per BORI edition, but included the additional texts from other versions as given therein. The present paper seeks to demonstrate the database with its navigational and retrieval features.

Facilities to preserve study and publish information contained in manuscripts (palm leaf, paper, etc.) with the help of advanced computing tools and technologies are welcome in order to unearth the treasures hidden in them for betterment of mankind. C-DAC has developed a comprehensive Manuscript Processing Software ***Pānḍu-Lipi Samśodhaka*** for the purpose of the critical edition of Sanskrit texts. Having evolved PC-ISCII standards for proper representation of all Sanskrit and Vedic character set in computers, developed an exhaustive knowledgebase of Vedas, Vedāṅgas and Upāṅgas and application programs for the fourteen Vidyāsthānas and tools and utilities like Editor, index, search, concordance etc., we now undertake to extend these to the deciphering of manuscripts in scripts like Grantha, Nandinagari, Telugu, Malayalam etc., of Sanskrit/Vedic texts, many of which are not yet published. We also help collation of various version/variant forms of texts from different sources for critical editions of such rare, unpublished works in these domains.

The ***Pānḍu-lipi Samśodhaka*** is useful in the collation, search, view, print etc. This promises to facilitate the study of the text and variations in the Mahābhārata, notwithstanding its regional and contextual factors. We describe the salient features of this further.

Features

The functional modules of the system cover acquisition, formatting, inputting, indexing, creating database, searching, locating, printing, collation and publishing. The range of texts covered include Śāstric texts, ṚgVeda, Kriṣna Yajurveda, Samaveda, Lakṣaṇa Granthas, texts in Tamil, a combination of Tamil and Sanskrit, called Maṇipravāla, etc., in a variety of scripts. The sample includes about fifty works, one hundred and twenty manuscripts, six scripts and many domains. There are about 3500 leaves (pages) as images to accompany the PC-ISCII texts. Two of these texts, viz. Ṣhaḍvimśati Sūtra and Yohi Bhāṣya, are chosen, for illustration and possible publication of a critical edition with a Sanskrit commentary.

Description of the modules:

Acquisition

a) One typically starts with consulting catalogues, indices, lists, reports, etc., of Manuscript collection of desired texts through a number of sources Bibliographic survey.
b) Select the ones feasible to obtain from the list (shortlisting). Provide for balanced representation of various regions, script and versions (i.e. with commentaries, with accents etc.)
c) Acquire copies xeroxed/scanned /microfilmed
d) Convert/export to a single (uniform) format i.e. jpg in the current case. Factors like clarity, condition of original, resolution of scanning and size of the image files, all influence the choice of common format used.

Formatting

The inputs come in various forms, when raw, i.e. from the institutions/library collections. We may have 3.5 or even 10 folios per scanned image. Here the two sides of the folios are in different files and the job of sequencing the image as per text and separation of folios is involved. Numbering them serially according to the text is done. An important task here, in the case of manuscript bundles containing different texts, is separation of the texts and folios belonging to multiple texts. They must be present in all the works concerned. Usually, libraries offer separation, if catalogued already.

Inputting

We strongly recommend the entry of the data contained in the manuscripts for the purpose of study, word-split, index, search (phrases), editing and collation. This, of course, requires domain experts who are difficult to get. However, the IHG offer expertise in this endeavour. We also have another possible source for data entry, which is loading text, if the work in the manuscript is one of available digital texts from our repository. (A list of about 250 texts from all Vidyāsthānas is available. C-DAC Indian Heritage Portal would make this available on the web soon).

Adding commentaries, translations, hyperlinks, annotations for collation etc, are the factors necessitating data-entry. Also transliteration, training in rare scripts etc, is enabled. However, efforts may be launched to develop efficient OCR or

speech recognition systems of high quality simultaneously and when these mature, we can minimize data entry needed.

Editing

This step involves aligning the data entered, with the original manuscript, line by line and page by page. This also can be done in an edit box/window below (or adjacent to) the image of the manuscript or entered through Vedic Editor and inserted into database. The pages and line boundaries are as before. Adding information for retrieval, hyperlinks etc., can also be done. Multilingual texts, currently require LEAP-like software for data entry and use in RTF format in the system for further processing. Here ISCII-ISFOC conversions are employed. Currently Vedic texts of Sāmaveda Gāna require use of only Grantha script and transliteration is not available. Śrautam and Guruparamparā Prabhāva etc. are multilingual samples. These are typed in LEAP and processed through rtf. controls.

Creating database

The PC-ISCII text files (*.pci) created by data entry or loading data are to be converted into database format. This is either Microsoft Access or Microsoft FoxPro format covering various fields for facilitating information retrieval. There is a utility that converts from aci./pci. format to db. format. Databases of works, institutions, manuscripts, books etc., are also created and linked in the application list of abbreviations. Scheme of data for reference in these texts etc. are also created as tables.

Searching

This is the crux of the system and helps in providing word or phrase level search (with and without accent-markers) across the database, text-wise, and lists the manuscript reference numbers where the search string occurs. In future, this can be extended across texts if need be (this feature is there in our Vedic Editor, wherein a string occurring in any of the 250+ texts are listed as a concordance). Choice of script, facility to transliterate, and seeing the results in the same manner of alignment as the manuscript are the useful aspects.

Locating

This refers to locating the search string in the image of the particular page of the manuscript where it occurs, including the line number and location in it. We see the string ‘highlighted’ in the text window by choosing ‘find’ in the page and

physically looking in the corresponding line and 'location' in it on the image above by selecting view in 'search' mode. The text window is provided with line numbers to facilitate this manual locating in the image.

Printing

Provision to print the texts in database, search results etc., in any script of choice or script of the original etc. so that further reference or insertion into documents can be enabled. Report generation kind of printing needs can also be addressed. List of texts, institutions, reference details etc. can be printed.

Collation

From the search function, we can organise the readings of different texts (like 'file compare') across the manuscripts combined with report generators. A scheme for annotating can be devised to assist here. Work will follow to enrich features here.

Publishing

Publication through Desk-top-publishing can be done by exporting to some DTP software and adding embellishments as desired.

The Mahābhārata Database Project was funded by Central Secretariat Library, Department of Culture, Government of India. It began as a Pilot Project for preparing the Database of Śānti Parvan with 18,000 Ślokas (verses) in the year 2000. After the successful completion of the pilot phase, the total project was undertaken, which was completed in the year 2001. The source material for the text to be followed was the B.O.R.I. Edition.

The work got over with many value-added features over and above the original project specifications. Some of these were:

"The complete Ślokas of Mahābhārata would be converted into machine readable form using GIST Card with appropriate database system. The main stress will be on the retrieval of information.
The retrieval will be by –
1. Any word of Śloka,
2. Any number of Śloka,
3. Any part of the Śloka,
4. Complicated samāsas,
5. Event of ashīrvād/shāp and the actual happening with links of chapter, parva, Śloka number etc. and

6. Reference to the context, who said to whom and when in which Śloka with links of chapter, parvan, Śloka number etc.

The following index files will be prepared:
1. Name index
2. Chapter index
3. Parvan index
4. Geographical index
5. Family tree index
6. Compound words with sandhi-viccheda and hyperlinks listing
7. Character index
8. Authority files of variant personal names and variant geographical names"

Against these specifications, the funding and duration, the achievements are much higher than expectations. Particularly on compound word dissolution, tagging and analysis. This is, in fact, the single most significant feature of the entire project. A Technical Advisory Committee comprising eminent Sanskritists (chaired by Prof. Ramkaran Sharma) have monitored the progress, and come out in high praise of the effort at Sanskrit analysis through computers.

Regarding compliance, items 1 to 4 and 6 (partly) are in retrieval and all but geographical/family/character names are achieved. To prepare the material required for these items (including samāsa analysis), a panel of experts was selected country-wide after organizing a workshop on the scheme and methodology. The scheme of mark-up is quite elaborate and envisages around 60 distinct tags to be used appropriately for helping analysis. This feature is by far the most exhaustive tagging scheme for the Epic, and hence should rank as one of the significant achievements of the project. The scheme is appended at the end.

Salient Features of the Mahābhārata Database

The database can be browsed for any desired parvan (and Śarga) as text (optionally as word-split and marked-up or *tagged* form as well) with choice of scripts from among Assamese, Bengali, Devanagari, Gujarati, Kannada, Malayalam, Oriya, Punjabi, Roman (with diacritics), Tamil and Telugu [Hindi, Konkani, Marathi, Nepali and Sanskrit share Devanagari Script]. Additional details provided are Sarga name and Anthar-parvan name. Prose form of the text are wrapped around to new line after about 50 characters. On-line help is provided in all screens.

In the retrieval mode, multiple ways of choosing/searching the desired information are provided like parvan, sarga, Śloka number, word (split form), part of a Śloka (phrase search includes blanks and multiple words also), by sarga name, speaker name, topic name (prakaraṇa/ vishaya), *prātipadika* or nominal stem search, be it the initial, middle or final member of a compound word and Boolean search to cover these with logical operators like AND, OR, NOT etc.

All these are provided with keyboard short-cut (hot keys) and icons with descriptions. Details like parvan name, sarga name, speaker name, anthar-parvan name are available. Script change, marked-up form view, help and back (exit) are standard features. In the search by number option, on selecting the parvan, admissible limits of sarga numbers and thereupon, those of Śloka numbers are displayed for valid values to be entered.

In the 'search word' option, on entering few initial characters (even one), all admissible words beginning with the typed characters are listed and choosing anyone thereupon, the Śloka numbers are displayed with the ref. no. scheme for desired values to be selected. The selected Śloka with all other details is displayed as before.

In the 'phrase search' option, any particular parvan(s) or all can be chosen and the phrase can be typed. All occurrences, with statistics and details of information are displayed. On choosing any desired number, the particular Śloka is shown. Sarga, speaker and topic names are also similarly selected. In pratipadika search, the desired stem as beginning, middle or end are additionally singly or severally selectable, and with statistics, detailed display of the Ślokas containing those compound words are shown.

In the 'Index mode', word, name, sarga, Śloka, samāsa and speaker indices are provided. Śloka index covers the entire Mahābhārata and samāsa index has two-tier selection for first-level and subsequent detailed types. Help files are also accessible from any screen by pressing F1 key or Help button. There is also a demo of the program as a guided tour included in the CD-ROM.

The Samāsa Mark-up Scheme Used for the Mahābhārata Database Project

The proposed scheme for mark-up is given here for discussion and adoption. It may be noted that name of the compound, its notation (tag) to be used for mark-up and examples are included in the scheme. Marked-up examples are also included for many cases for purposes of illustration.

Multiple word compounds are also shown to clarify the scheme. The bracketing

in the tags also has distinction between words where compounding proceeds sequentially (i.e, from left to right) and where there is a change in its direction. These are also suitably illustrated. Nested brackets could be used later with our programs.

*Multiple mark-up*s for words denoting different possibilities/interpretations or allegories etc. are also encouraged to be indicated by using ? and placing alternate tag(s) in curly brackets. e.g, **word**>_*Tag1*/{?*Tag2*...}.

By the same token, *multiple splits* of words also (especially in classical literature) may be indicated by copying the text under question and splitting alternately but in *curly bracket*s. Preference among them is left to the experts.

We have suggested certain specific features like not marking *samāsānta taddhita* suffixes etc. in the end where possible (with exceptions as in +t option in Dvigu, for example) and also included *lyabanta avyaya*s as samāsa etc. In all cases, a compound word has at least a pair of *angle bracket*s (< & >). The *underscore* character (_) will be the delimiter for tags. *Hyphen*s separate stems.

The marked-up files were processed for tags and statistically analysed. The following interesting summary of the break-up of compounds is obtained:

Total number of slokas (in main text) – 73815 [There are about 28,000 slokas in the appendices].

Total no. of words - 7,00,000

Total no. of compound words -

Total no. of compounds - 1,35,827

Main samāsa type	Sub Type	Count
Avyayibhava		**1808**
	A1	1701
	A2	31
	A3	76
Tatpurusha		**66458**
	T1	67

	T2	529
	T3	7569
	T4	620
	T5	595
	T6	*32107*
	U	8944
	Tds	383
	Tds+t	154
	Tdt	4
	Tdu	6
	Tg	7244
	Tp	8096
	Tm	134
	Tb	6
Karmadharaya		**18838**
	K1	*16468*
	K2	47
	K3	106
	K4	198
	K5	1595
	Km	424
Bahuvrīhi-samanādhikaraṇa		**22663**
	Bs2	33
	Bs3	1208
	Bs4	13
	Bs5	264
	Bs6	*16442*
	Bs7	272

	Bsd	4
	Bsp	2
	Bsmn	951
	?Bsg	0
	Bsmg	220
	Bss	40
	BsU	161
Bahuvrihi-vyadhikaraṇa		
	Bv	204
	BvS	2151
	Bvs	237
	BvU	364
	Bb	97
Dvandva		**8425**
Itaretara	Di	7893
Samahara	Ds	453
Itaretara - viruddha	DiV	79
Ekasesha		9
S - kevala		805
! (Aluk)		83
Ambiguous/Unknown		16738
Total		**135827**

Of these, we see that the three types, viz, T6, Bs6 and K1 cover nearly half of all compounds. Hence, devising a rule-based program for generating the vigraha-vakyas will be attempted.