Report on the Cologne Sanskrit Dictionary Project

Paper read at the 10th International Sanskrit Conference,
Bangalore, January 3-9, 1997
Author Dieter B. Kapp and Thomas Malten
Institute of Indology and Tamil Studies (IITS)
University of Cologne
Pohligstr. 1
50969 Köln
Germany
Email: ami01@uni-koeln.de

Abstract

The Cologne Digital Sanskrit Lexicon (CDSL) project undertakes to digitize and merge the major bilingual Sanskrit dictionaries compiled in the 19th century. Its aim is to provide a basic lexical corpus to provide an easy access to all available meanings of Sanskrit words and to allow the creation of a number of computer programs that will help to analyze Sanskrit texts.

In the first stage Monier-William's Sanskrit-English dictionary (MW) has been digitized to be followed at a second stage by three other dictionaries (Cap, PW2 and Sch). All these will be structured and unified to allow access to the meanings as developed by the different lexicographers.

As a final goal it is hoped that a step can be taken towards an integrated Sanskrit word catalogue which codifies the distribution of lexical units in Sanskrit text corpora by linking them to the existing descriptions in dictionaries by a numeric system which functions as a placeholder for a word sense which can be expanded or changed. Last but not least connecting Sanskrit with Tamil vocabulary is envisaged. To this end the major Tamil dictionaries have already been converted into digital form.

The main object of this paper is to describe in some detail how the printed MW has been encoded without changing its rather complicated structure and without losing any of the information contained in it. 7-bit encoding has been used for the transliteration of DevanAgari to make it directly readable for humans as well as making it accessible to general text processing tools.

{Assistance and advice given by the following persons is gratefully acknowledged: P. Schreiner, Zürich, J. Smith, Cambridge, J. Nye, Chicago, M. Koerbl and J. Tümmers, Cologne, G. Rajasundaram, Annamalainagar, R. Bunker, Fairfield (Iowa). The project has been financially supported by the vice-chancellor of Cologne University.}

Abbreviations

Ap	Apte 1880
Cap	Cappeller 1891
CDSL	Cologne Digital Sanskrit Lexicon
HK	Harvard-Kyoto transliteration convention
MW	Monier-Williams 1899
PW1	Böhtlingk and Roth 1855-1875
PW2	Böhtlingk 1879-1889
Sch	Schmidt 1928

1 The digitization of Monier-Williams' Sanskrit dictionary (MW)

The selection of MW as the first dictionary to be digitized for CSDL project was prompted by several reasons:

- MW is the last and therefore presumably the most up-to-date and complete of the large 19th century Sanskrit dictionary productions.

- All Sanskrit words in MW are written in transliteration making it possible to use OCR. It is the most compact of all the large Sanskrit dictionaries and has the further advantage that the target language is English.

- An earlier project to digitize Indian lexical resources at the University of Chicago in 1985 (which failed for lack of funding) also included MW.

After beginning the project it was found that the complicated structure of MW and its many abbreviations do not lend themselves easily to the process of a digital conversion. But in view of our earlier experience with the handling of Tamil lexical material we were quite convinced that a satisfactory result could be achieved within an acceptable period of time.

By the end of 1994 we produced a sample page MW288 in which the transliteration and the encoding of the structure could be successfully demonstrated. It was then decided to use a Kurzweil OCR system (K5000) available at the computer center of Cologne University to finalize a first complete conversion run on MW. This work was finished by J. Tümmers in the middle of 1995 with an accuracy rate of ca. 70%, as much as could be achieved considering the extremely small type of the 1964 reprint used. This resulted in a 15 MBytes computer file, which even though it was full of misreadings, errors and omissions contained the more or less complete structure of MW and was processed further with an editing programme to eliminate many of the `systematic' mistakes and to tentatively insert tags to represent the overall structure of the entries. The result was a file that obviously still needed a lot of proofreading but was a quite recognizable copy of the original MW. The vice-chancellor of Cologne University by the end of 1995 agreed to fund the correction and final creation of a machine-readable version of MW to be produced till the end of 1996. The MW computer files were then sent to India for proofreading and correction according to our specifications. These included the HK transliteration scheme and ASCII tags. From these corrected file a printed version of nearly 4000 pages was produced in Cologne which was used for normal proofreading on the printed pages. This turned out to be very laborious and was done partly twice. The resulting corrections on paper were again transferred to the computer files. A complete version containing ca. 17MB of data was received in Cologne in September 1996.

The final editorial process resulted in the identification of the 166,446 main entries of

MW. Apart from the continuing work of correcting typographical errors, several major tasks were completed:

- The tagging of the three levels--each alphabetically ordered--of Sanskrit main entries (see 1.3 for details).

- The expansion of Sanskrit abbreviated forms in the main entries of the 2nd and 3rd levels given in the printed MW with preceding bold hyphens (**-**) and little circles (°) to their full forms.

- The expansion of English abbreviated forms indicated by the same little circles to their full forms.

- The complete tagging of citations, grammatical information, all other Sanskrit words, and etymological references; preliminary (incomplete) tagging of different meanings; tagging of verb forms.

1.1 The codification of MW's structure

As Peter Schreiner (1996) points out in the introduction to his digital version of Mylius (1992)(Peschmyl) ``an electronic dictionary should actually be marked according to the guidelines developed by the Text Encoding Initiative (TEI)". Practical considerations, as for example the size of computer main memory (RAM), have led to our using tags consisting of a single (usually upper) sign of the ASCII code (IBM code page 437) which are otherwise not used in the text instead of the usually much longer SGML tags. {In the present notation it is estimated that after adding Cap PW2 and Sch to the existing MW the CDSL will contain ca. 50 MB of data.} In this paper they are shown by their positional number preceded by d' and enclosed by square brackets, e.g. ``[d'243]" indicating the ASCII position 243 in decimal notation. As problems are bound to occur in the transfer of such data to other computer systems these signs are finally to be replaced by characters of 7-bit (lower ASCII) code.

1.2 List of upper ASCII characters

The following upper ASCII characters have been used in the computer files to tag structural elements within main entries.

[d'020] = Page and column numbers of the printed MW
[d'175] = References to other entries in MW
[d'182] = Etymologies
[d'238] = Citations
[d'240] = Replaced meanings (given once with variants)
[d'241] = circled English abbreviated words expanded
[d'243] = circled Sanskrit abbreviated words expanded
[d'247] = Verbs
[d'248] = Abbreviated Sanskrit words not expanded
[d'250] = Grammar
[d'251] = root sign used in MW

1.3 The transliteration of Sanskrit

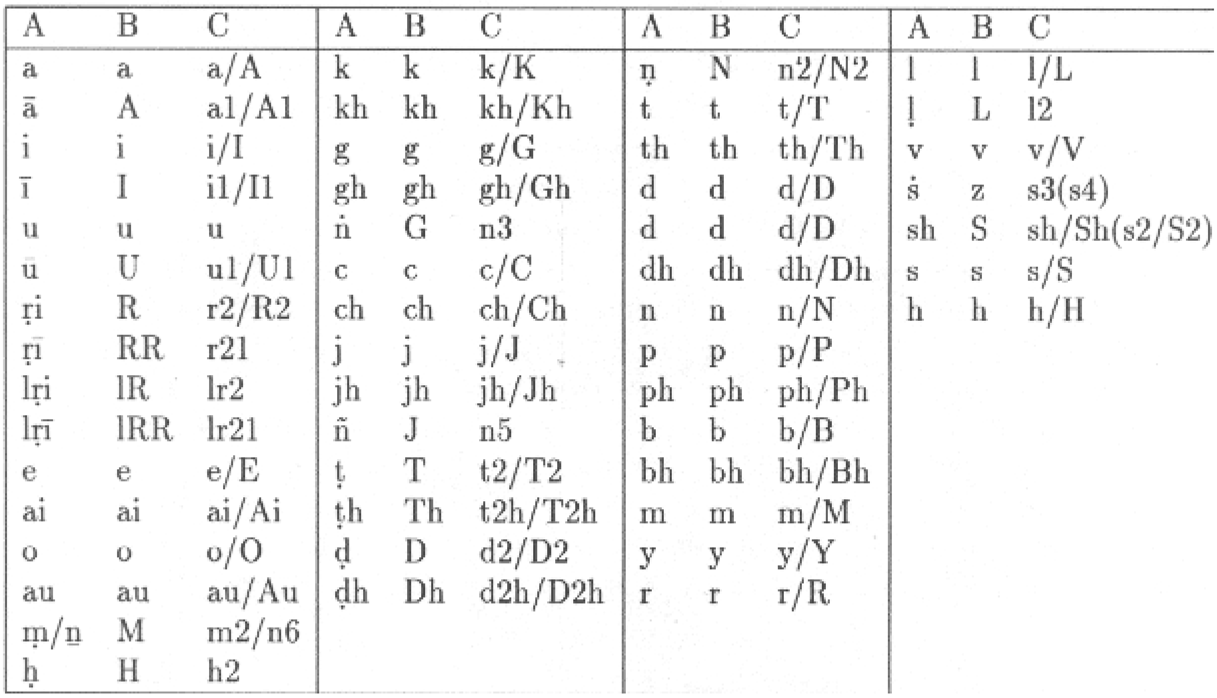
The transliteration of Sanskrit in the computer file is done exclusively in 7-bit ASCII code. It has three levels: the letters (vowels and consonants) themselves; the indication of accents and further diacritical marks; the indication of language (script).

The representation used for the DevanAgarI script and Roman transcription of Sanskrit with diacritical marks is based on the Harvard-Kyoto (HK) convention, where ordinary small and capital letters are mainly used. This system is not only economical but also quite readable. The following letters and signs are used:

```
        a A i I u U R RR lR lRR e ai o au M H
           k kh g gh G j jh J T Th D Dh L N
            t th d dh n p ph b bh m y r l v
              z S s h ' - -- 4 7 8 9 0 ° @
{The sign @ indicates a space between Sanskrit words.}
```

Transliteration systems for MW

A = MW print
B = HK adaptation
C = Anglicized Sanskrit



A	B	C	A	B	C	A	B	C	A	B	C
a	a	a/A	k	k	k/K	ṇ	N	n2/N2	l	l	l/L
ā	A	a1/A1	kh	kh	kh/Kh	t	t	t/T	ḷ	L	l2
i	i	i/I	g	g	g/G	th	th	th/Th	v	v	v/V
ī	I	i1/I1	gh	gh	gh/Gh	d	d	d/D	ś	z	s3(s4)
u	u	u	ṅ	G	n3	d	d	d/D	sh	S	sh/Sh(s2/S2)
ū	U	u1/U1	c	c	c/C	dh	dh	dh/Dh	s	s	s/S
ṛi	R	r2/R2	ch	ch	ch/Ch	n	n	n/N	h	h	h/H
ṛī	RR	r21	j	j	j/J	p	p	p/P			
lṛi	lR	lr2	jh	jh	jh/Jh	ph	ph	ph/Ph			
lṛī	lRR	lr21	ñ	J	n5	b	b	b/B			
e	e	e/E	ṭ	T	t2/T2	bh	bh	bh/Bh			
ai	ai	ai/Ai	ṭh	Th	t2h/T2h	m	m	m/M			
o	o	o/O	ḍ	D	d2/D2	y	y	y/Y			
au	au	au/Au	ḍh	Dh	d2h/D2h	r	r	r/R			
ṃ/ṉ	M	m2/n6									
ḥ	H	h2									

Apart from the transcription of DevanAgarI letters care has to be taken of accents, UdAtta and Svarita, both represented by the digit 4. Furthermore in MW the indication of vowel sandhi (``blending of short and long vowels") by circumflex is represented here by the number 7 if placed above a single vowel and by 9 if spanning two vowels. The (rare) combination of two separate vowels in MW is represented by adding the number 0 to the second vowel.

To indicate the beginning and end of Sanskrit strings opening and closing braces `{' and `}' are used. The percentage sign % is placed before the opening brace to indicate italicized secondary Sanskrit strings which occur embedded in English meanings in MW.

A particular problem is posed by proper names of Indian origin in the printed MW as they are not distinguished from the surrounding English by font change but which can have diacritical marks. These proper nouns may be called `Anglicized Sanskrit' and have been indicated in PW2 by spacing of letters. Diacritical marks in all these words are marked in CDSL by adding numerals to letters. Whereas it is quite easy to identify these words if they carry any diacritics, e.g. `DUrvA grass (Du1rva1 grass)' or `RAma (Ra1ma)', cases where no diacritics occur, have to be identified and marked `manually', names like e.g. `Apsaras' or `Yoga'. A further complication is introduced by the fact that these words may carry English grammatical suffixes, for example plural *s* or that many of them may be considered even as proper English loan words. No solution has been attempted for this problem and these words have so far remained unmarked.

1.4 The structure of the main entry

1.4.1 The three levels of main entries

In MW there are three levels {Cf. Monier-Williams 1899, Introduction, p. xiv.} (L1, L2, L3) of what may be called ``main entries''. {For an example overview of the structure of entries in MW see Appendix. [If not found in the online version please contact ami01@uni-koeln.de.]}

In print a level one entry (L1) Sanskrit headword is given in DevanAgarI followed by its italicized Roman transliteration. The DevanAgarI string is ignored in the digitized version as it is merely a repetition of the Romanized form (or the other way round). A level two (L2) entry is always written in bold font Roman transliteration. In the printed book these two are indicated by indentation of a paragraph.

Level three (L3) headwords are bold faced transliterated forming subentries to L1 or L2 entries. Many of them have a bold hyphen **-** in front of them which indicates that the word has to be compounded with the preceding L1 or L2 entry. This bold hyphen is represented in the digital version by double hyphen -- following the expansion of an compound word.

L1 and L2 level entries are separated from the preceding entry by an empty line. They are always ended by the lower hyphen `_' . Thereby a complete L1 or L2 entry group (including their L3 entries) is explicitly marked.{This marking is preferred to the implicit marking by empty line separation or reference to the beginning of L1 or L2 entries.}

Each Sanskrit headword is written twice adjacent to each other with both members enclosed with braces. The entry levels are indicated by the digits 1, 2 or 3 placed between the members of each headword. The whole form is preceded by a dot which marks them formally and unambiguously as main entries in search precedures. To give an example L3 entry (see also Appendix):

```
.{kuJjakuTIra}3{kuJja--kuTIra\}
```

This headword can be followed immediately after the last brace by another numeral to indicate a homonym.
The first member contains letters only, whereas the second member contains also accents, double hyphens or other marks.
After the Sanskrit word the meanings in English are given together with grammatical explanations and citations, which are usually tagged.

1.4.2 Coding of the structure of a main entry

Senses: The different English meanings of a Sanskrit word given in the MW print are not clearly separated from each other. The only indication is a semicolon placed between the different senses of a word, but this is not an unambiguous sign. As a preliminary measure these semicolons have been marked by us with a preceding dot but this needs further clarification, as grammatical information, especially verb forms are also separated by semicolons. Verbs have been marked, to the extent that they have as yet been identified (ca. 13,000) with [d'247].

No indication has been given by what method multiple meanings of word have been ordered but it can be presumed that either the most prevalent meaning or the earliest occurence of a particular meaning in the text sources has been given first. The matter has remained unclear so far and needs also more clarification.{See the remarks on PW1 by Ghatage 1975, Introduction, p. vii f. It may be surmised that MW has largely followed the order of meanings in PW1.}

Citations: Textual and other sources for the meanings of Sanskrit words are given frequently in MW but there is no indication for the procedure adopted as to when and how their inclusion has been decided upon. The citations have been marked by [d'238]. Note that the source ``ib.'' (ca. 10,000) which refers to the immediately preceding textual source has also been marked with [d'238].

Grammar: The indication of gender (m., f., n.) of nouns and in the case of verbs conjugation has been marked as far as possible by a preceding [d'250].

Crossreferences: Due to MW's method of partial non-alphabetic ordering of entries the user of his dictionary is frequently referred to other entries.{Almost 40,000 times.} This is done by either referring to an entry of a different level, or by pointing to a particular page and/or column. Four types of crossreferences are used.

- The most frequent reference term is ``see'' (ca. 10,000) or the equal sign ``=''{Note that ``='' can be also used in other contexts.} followed by a Sanskrit word or by `next', `preceding' (ca. 14,000).

- ``q.v.'' is placed after the Sanskrit word referred to. This either replaces the meaning of the word in that entry or gives additional information on the meaning or the etymology of a word.

- ``cf.'' as a term of reference points to additional information found elsewhere in MW.

These four terms have each been prefixed by [d'175].

Etymologies: Etymological references to cognate words from Indo-European and other languages are scattered throughout MW. These words ared tagged with a [d'182] and follow the abbreviation of the language in question.

1.5 Work that remains to be done

Though the main structure of MW has been marked and many corrections have already been carried out much work remains to be done on certain structural pecularities which often cannot be done globally. A decision has to be taken case by case.

- Foreign language words, etymologies.

These have been tagged to a large extent but Arabic Persian and Greek words have been marked only by dollar $ signs and their transliteration is still to be inserted.

- Hyphenation at the end of a line.

While correcting the computer files, hypenation of English words at the end of a line was removed. This often did not taken into account the hyphenation of compounds, e.g. `dice-board' or `barley-corns' which occur rather frequently. Thus hyphens have been removed erroneously.

- Quotation marks.

MW very frequently uses single quotation marks to indicate the literal translation of a word from Sanskrit. These have often been misinterpreted in the scanning process and been wrongly replaced by a comma. Subsequent proofreading has still left many of them uncorrected.

- Tagging grammatical information.

Even tough MW has given much grammatical information with verb entries, it is still diffcult to see how this can be coded homogeneously. One way to solve this problem could be to add grammatical information from outside the dictionary.

- Tagging of verbs.

As mentioned above many verbs have been marked by [d'247]. This has been done so far only on the basis of the English verb meanings being given in the infinitive with `to'.

- Expansion of abbreviated words. The sign <_ [d'243] has been used where abbreviated Sanskrit words have been expanded from a main entry.

In the case of English where the abbreviated form in printed MW is marked with ° ([d'248]) the sign +- [d'241] has been used to mark the expansion. In the case of English words thus replaced this was done manually and needs checking for correctness.

- Reference to other entries.

Refererence to a page number in MW, to a textual source which has not been repeated (when only its page or verse numbers are given), or reference to a preceding source by `ib.' have to be checked and proper reference labels to be inserted.

- Multiple Sanskrit head words with one sense.

Often main entries followed merely by comma, the word 'and', 'or', or '=' followed by variant Sanskrit words to which meaning is given only once have been marked partially by == ([d'240]) and have to be checked for correctness and consistency.

- Vedic accents on expanded forms.

If a main entry carries a Vedic accent this accent is automatically carried over to L3 compounds. These accents have to be removed in appropriate cases.

- Vowel sandhi of compounds (``blending of vowels'').

Vowel sandhi, which is indicated in the printed MW by four kinds of circumflex signs placed above vowels is a speciality of MW (see Introduction, pp. xix and xxxiv). All four signs have been represented by a single numeral (7). It should be possible to find a programmed solution to the distinction of these four circumflexes.

A number of further desirable taggings or changes have still to be undertaken, namely:

- Tagging of ``Anglicized Sanskrit''.

If such words carry diacritical marks it is quite easy to tag them. If not they have to be selected amd marked one by one, possibly by using a word list which excludes such word of Indian origin which can be considered as loan words. Additionally these words may be repeated without diacritics or in standard transliteration to facilitate search.

- Orthography of English words.

The spelling of English words in MW is not always consistent and often antiquated. To enable users nowadays to search for these words, such cases have to be identified and marked.

- Further corrections and tags.

It would be desirable to classify words according to certain categories, e.g. botanical, geographical etc. This can be achieved partly by using the structure of the printed book, where, for example, botanical names have been indicated by capital letters.
The ``Additions and Corrections" to MW given on pp. 1308-1333 and containing ca. 4000 entries have to be incorporated. These entries are to be marked as additions and then to be merged with the main corpus. Corrections given in reviews (e.g. Winternitz 1900) or private correction lists should also be considered for inclusion.
To retain the structure or sequence of Sanskrit entries in MW when merging or sorting them with other lexicographical material sequential numbers will be added to each of the entries combined with single sense units. This will facilitate later corrections, as these numbers can be used for reference to a particular entry and also provide a base for a general Sanskrit word catalogue where each sense unit is assigned a fixed number independent of the language in which that sense of a word is expressed.

1.6 Alphabetical order and sorting of Sanskrit

Although standard alphabetical ordering of Sanskrit is clear, if not always adhered to {Cf. the ordering of words in Mylius 1992, see esp. pp. 497ff.} it may be useful to give here the sorting sequences, especially as the sequence is not quite straightforward. For computer searching sorting is less imoprtant as processor power increases, but for merging different word lists as well as for checking/correcting of entries it is quite useful. The sorting sequence is:
Vowels:
a A i I u U R RR lR lRR e ai o au
AnusvAra + [yrlvzSs]:
My Mr Ml Mv Mz MS Ms
Visarga:
H
AnusvAra in final position:
M
Consonants:
k kh g gh G j jh J T Th D(L) Dh(Lh) N t th d dh n p ph b bh m y r l v z S s h

AnusvAra before the following consonants have to be converted to homorganic nasals before sorting:
Mk=Gk Mg=Gg Mc=Jc Mj=Jj MT=NT MD=ND Mt=nt Mn=nn Mp=mp Mb=mb Mm=mm

AnunAsika is equivalent to AnusvAra.
Avagraha is equivalent to *a* or is ignored(?).

Combinations of primary vowels a-i, a-u, A-i, A-u are sorted as single primary vowels (thus MW {prau0ga} (p. 652,3) before {prauga} (p. 714,2) {prakaGkata} and {mAu0tha} before {mAkanda}).

2 Other dictionaries to be incorporated in CSDL

The digitizing and merging of three other Sanskrit dictionaries (Cap, PW2, Sch) is planned as a second stage of the CSDL project. This stage is to be finalized by the end of 1998 and will constitute the main corpus of the CDSL.

Cappeller's Sanskrit-English Dictionary (Cap) which contains about 50,000 Sanskrit entries is explicitly based on PW but as the author in the preface, (p. v.), says, he has incorporated ``all primary words of well-settled meaning''. Therefore Cap can be used to reduce the sometimes huge number of meanings given with one entry in MW and PW2 and will thus allow an easy extraction of the basic senses of words.

Böhtlingk's shorter Sanskrit dictionary (PW2) is a destillation and enlargement of PW1 and at the same time the fundamental bilingual lexicographical work of the 19th century. The fact that the meanings are given in German may not be considered detrimental to its usefulness. The combination of PW2 with MW will make using PW2 easier for those who only know English. In fact the combination of MW and PW2 can be considered to some extent as a substitute for a translation of PW2 into English. The entries in PW2 are much better structured by the use of different type faces than those of MW (e.g. what has been called `Anglicized Sanskrit' in MW is printed spaced out in PW2) and can therefore be tagged more easily.

The third of the dictionaries to be incorporated is Schmidt 1928 (Sch) which contains not only Böhtlingk's own additions to PW2 (ca. 12,000) but also a further 12,000 words and meanings and concludes the comprehensive lexicographical compilations of the period.

3 Further outlook

By marking and numbering the entries in these four dictionaries independently so as to reflect their different structures, these can be retained even after merging the entries of all the four dictionaries. It will still be possible to extract the structural information given in each one as well as have an exact reference to the original printed source. Not only can the original content of these productions be determined but also a duplication of effort in the compilation of new Sanskrit dictionaries be avoided. Instead of including citations in the dictionary base of CDSL itself it will be sufficient to add pointers to external digital Sanskrit texts, where quotations can then be used in a much larger context.

As regards methods of long term corrections and improvements to CDSL these should be seen in the light of the possiblities of electronic communications.

MW, p. v: ``The words and meanings of the words of a Dictionary can scarcely be proved by its compilers to belong exclusively to themselves. It is not the mere aggregation of words and meanings, but the method of dealing with them and arranging them, that gives a Dictionary the best right to be called an original production.'' To allow free access to the data compiled it is necessary to impose as little as possible restrictions on lexical databases. In the light of the development in the protection of database in the EU and WIPO with the danger that commercial publishers may want to appropriate rights it seems desirable to to have high quality Sanskrit databases freely availabe to all researchers.

4 Bibliography

- Apte, V. S. 1890: The Practical Sanskrit-English Dictionary. Poona.
- Böhtlingk, Otto 1879-1889: Sanskrit-Wörterbuch in kürzerer Fassung. St. Petersburg.
- Böhtlingk, Otto and Rudolph Roth 1855-1875: Sanskrit-Wörterbuch. St. Petersburg.

- Cappeller, Carl 1891: A Sanskrit-English Dictionary. Strassburg.
- Ghatage, A.M. 1976-: An Encyclopaedic Dictionary of Sanskrit on Historic Principles. Poona.
- Ingalls, Daniel H.H. and Daniel H.H. Ingalls 1985: The MahAbhArata: Stylistic study, computer analysis and concordance. In: Journal of South Asian Literature 20:17-46.
- Jachertz, Thomas 1983: Beiträge zu einer bibliographischen Übersicht über die textliche Basis unserer europäischen Sanskritwörterbücher, vorzüglich des grossen Petersburger Wörterbuches (PW) und des kleinen Pete rsburger Wörterbuches (pw). Unpublished MA thesis, Cologne University.
- Monier-Williams, Monier 1899: Sanskrit-English Dictionary, Oxford.
- Mylius, Klaus 1992: Wörterbuch Sanskrit-Deutsch. Leipzig.
- Schmidt, Richard 1928: Nachträge zum Sanskrit-Wörterbuch in kürzerer Fassung von Otto Böhtlingk. Leipzig.
- Schreiner, Peter 1995: Arbeitsbericht zum Skannen des Sanskrit-Deutsch-Wörterbuchs von Klaus Mylius. Unpublished MS.
- Schreiner, Peter 1996: Prefatory matter to {Peschmyl} (the electronic version of the Sanskrit-German Dictionary by Klaus Mylius), Zürich.
- Sperberg-McQueen, C.M. and Lou Burnard (eds.) 1994: Guidelines for Electronic Text Encoding and Interchange. TEI P3, Text Encoding Initiative, Chicago, Oxford. [esp. chapter 12: Print dictionaries (p3di.doc)]
- Winternitz, M. 1900: Sir M. Monier-Williams: {A Sanskrit-English Dictionary.} New Edition Oxford 1899. In: WZKM 14:353-360.
- Wujastyk, D. 1988: Report on the Sanskrit Text Archive Conference. Austin, Texas, October 28-29.
- Wujastyk, D. 1996: Transliteration of DevanAgarI. http://www.ucl.ac.uk/ucgadkw/t/t.html
- Zgusta, Ladislav 1988: Copying in Lexicography: Monier-William's Sanskrit Dictionary and other Cases (Dvaikosyam). In: Lexicography 4:145-164.

HOME

Webmaster: T. Ghosh-Beverborg.