# Normalizing Cologne Digital Dictionaries

Dr. Dhaval Patel

drdhaval2785@gmail.com

Abstract

Cologne Digital Sanskrit Dictionaries site maintains 36 Sanskrit related dictionaries as on 31th October 2016. In Sanskrit NLP, these dictionaries are the main source for lexical data. Discounting 3 English – Sanskrit dictionaries, there are 33 dictionaries in total which have headwords in Sanskrit. As these dictionaries are compiled over a vast period of time, different conventions were followed by their authors. When we want to align these digital lexica, we need to understand these conventions and arrive at a standard convention, so that these different databases can communicate to one another.

Examples would make it more evident. Majority of dictionaries tend to use the uninflected form of the headword e.g. धर्म, but rarely inflected form is also used e.g. धर्मः, Similarly there is a tendency to use ‘अर्‌’, ‘ऋ’ or ‘आ’ in different dictionaries for words ending with ऋ e.g. पितर्‌, पितृ or पिता. The data referred to by headwords धर्मः / धर्म or पितर्‌ / पितृ / पिता are the same. When we want these dictionaries to communicate seamlessly with one another, such differences need to be ironed out. Greater consonance can be brought between these dictionaries by some form of standardization. We have tried to analyse different conventions followed by these authors / editors and come out with a standardized convention in the present paper, so that the headword list is normalized to a standard format.

# Normalizing Cologne Digital Dictionaries

Dr. Dhaval Patel

[drdhaval2785@gmail.com](mailto:drdhaval2785@gmail.com)

Cologne Digital Sanskrit Dictionaries website[[1]](#footnote-0) has a rich treasure of Sanskrit related dictionaries. There are 36 dictionaries hosted there as on 31th October 2016. There are three English – Sanskrit dictionaries (AE, BOR and MWE)[[2]](#footnote-1). Even after discounting these dictionaries whose headwords are in a foreign language, there are 33 digital dictionaries whose headwords are mainly in Sanskrit[[3]](#footnote-2) and should be of utmost interest to Sanskrit NLP researchers.

There are many generative and analytical computational tools in Sanskrit NLP, which require annotation or extraction of lexical data in various degrees. If these tools are to be supported with lexical data in long run, we need to standardize various conventions used in various dictionaries into a predefined format.

Examples would make the need for such standardization in dictionary headwords more evident. AP tends to use inflected form of the headword e.g. धर्मः, whereas MW tends to use the uninflected form of the headword e.g. धर्म. Similarly PW has tendency to use ‘अर्‌’ for words ending with ऋ e.g. पितर्‌, whereas MW has tendency to use ‘ऋ’ e.g. पितृ and SKD has tendency to use पिता. The data referred to by headwords धर्मः / धर्म or पितर्‌ / पितृ / पिता are the same. For any NLP related program to link properly to the data, we need to iron out these differences. Even if the input is पितृ, data from पितर्‌ of PW or पिता of SKD should be accessible to that user.

Because of difference of conventions in various dictionaries, sometimes the user is not able to land on the desired dictionary entry. This calls for a study of various conventions followed by various dictionaries and need to come out with a standardized convention for headword normalization. The suggested standard conventions proposed in this paper are open to discussion and correction, if need be. After due deliberation, they can be finalized.[[4]](#footnote-3)

# **Convention 1 - Treatment of अनुस्वार before consonants and terminal अनुस्वार**

There are at least 6 conventions of handling अनुस्वारs before consonants in झर्‌ प्रत्याहार[[5]](#footnote-4) and terminal अनुस्वारs[[6]](#footnote-5) in various dictionaries. They are described below. When अनुस्वार occurs before any of the letters य,र,ल,व,श,ष,स or ह, it is uniformly depicted as अनुस्वार in all dictionaries. The conventions mentioned below are not applicable to such अनुस्वारs.

## **Option 1.1**

Treat it as fifth letter of the वर्ग of following letter, when in between a word e.g. अङ्क, चञ्चल, खण्ड, प्रशान्त, नितम्बः.

Dictionaries: ACC,AP,BEN,BHS,BOP,BUR,CAE,CCS,GRA,GST,IEG,INM,MCI,MD,MW,MW72,PD,PE,PGN,PUI,PW,PWG,SCH,SHS,SKD,SNP,STC,VCP,VEI,WIL,YAT

## **Option 1.2**

Treat it as अनुस्वार, when occurring in between a word (other than the cases where the first member of compound ends with म्‌) e.g. अभिशंका, कुंचित, पुंड्रः, तंद्रा, विडंबित.

Dictionaries – AP90

### Note regarding 1.1 and 1.2 –

KRM is a special dictionary, whose headwords have only verbs. Therefore, the options 1.1 and 1.2 are not applicable to it.

## **Option 1.3**

Treat it as अनुस्वार when first member of compound ends with म्‌ and the second member of compound starts with झर्‌ letters e.g. संगीत, प्रतिसंधानम्‌.

Dictionaries: ACC,AP,AP90,BEN,BHS,CAE,CCS,MCI,MD,MW,PD,PW,PWG,SCH, STC,VEI,WIL

## **Option 1.4**

Treat it as fifth letter of the वर्ग of following letter, when occurring in the cases where the first member of compound ends with म्‌. e.g. सङ्गीत, निःसन्धि .

Dictionaries: BOP,BUR,GRA,GST,KRM,IEG,INM,MW72,PGN,PUI,SKD,VCP,YAT

### Notes regarding 1.3 and 1.4 –

(1) PE is inconsistent (1.3 and 1.4). See गङ्गासरस्वतीसंगम and वरदासङ्गम.

(2) SHS is also inconsistent. See सङ्ग्रह and कर्म्मसंग्रह.

(3) SNP doesn’t have any such case prima facie, hence excluded from this list.

## **Option 1.5**

Use अनुस्वार at the end of a word to denote neuter gender e.g. अकौटिल्यं, जघनं, तैलं.

Dictionaries: AP90,SKD

## **Option 1.6**

Use अनुस्वार at the end of a word to denote अव्ययs mostly (not to denote neuter gender), where usually म्‌ is supposed to be. e.g. अनुकामं, अहर्निशं.

Dictionaries: YAT

## **Standard convention 1**

1. Convert every nasal followed by consonant to अनुस्वार.

2. Convert every terminal अनुस्वार to म्‌.

It is pertinent to note that majority of dictionaries tend to prefer convention 1.1. So statistically we should lean towards 1.1, but there are a few problems with that approach computationally. Many of the dictionaries following 1.1 also follow 1.3. Therefore, deciding whether to keep it as fifth letter of a वर्ग or अनुस्वार depends on the knowledge of whether it is last letter of a compound or not. Therefore, uniformly converting every internal nasal to अनुस्वार is computationally easier choice.

# **Convention 2 – Treatment of अनुस्वार of verb**

अनुस्वारs in verbs are handled a little differently than convention 1 in dictionaries, so they are treated here separately.

## **Option 2.1**

Verbs are presented with removal of अनुबन्ध and with conversion to fifth letter. e.g. स्तम्भ्‌.

Dictionaries: AP,BEN,BOP,BUR,CAE,CCS,GRA,GST,MD,MW,MW72,PD,PW,PWG,SCH,SHS,STC,YAT

## **Option 2.2**

Verbs are presented as in धातुपाठः e.g. स्तन्भ.

Dictionaries: KRM,PD,SKD,VCP,WIL

## **Option 2.3**

Verbs are presented with removal of अनुबन्ध but without conversion to fifth letter i.e. with अनुस्वार e.g. स्तंभ्‌.

Dictionaries: AP90

### Notes regarding options 2.1 to 2.3 –

(1) ACC, BHS, IEG, INM, MCI, PE, PGN, PUI, SNP and VEI do not have enough headwords to decide this convention decisively.

(2) PD tends to give two separate headwords, one following options 2.1 and the other following option 2.2 e.g. अङ्क्, अकि. Therefore, it is included in both categories.

## **Standard convention 2**

Option 2.3[[7]](#footnote-6).

# **Convention 3 - Duplication of consonants after 'r'.**

## **Option 3.1**

Duplication is not done e.g. पूर्व.

Dictionaries: ACC,AP,AP90,BEN,BHS,BOP,BUR,CAE,CCS,GRA,GST,IEG,INM,KRM,MCI,MD,MW,MW72,PD,PE,PGN,PUI,PW,PWG,SCH,SNP,STC,VCP,VEI

## **Option 3.2**

Duplication is done e.g. पूर्व्व.

Dictionaries: SKD,WIL

### Note –

(1) SHS and YAT are inconsistent in this convention. See निर्विघ्न / निर्व्विकल्प in SHS and दुर्वच / दुर्व्वचस्‌ in YAT.

(2) VCP highly leans towards option 3.1, but there are a few inconsistent entries which follow convention 3.2 e.g. अग्निपर्व्वत c.f. पर्वत (3.1)

## **Standard convention 3**

No duplication.

# **Convention 4 - Uninflected / inflected forms**

## **Option 4.1**

Dictionaries present headwords in inflected forms (प्रथमा विभक्तिः एकवचनम्‌ forms) e.g. धर्मः.

Dictionaries: AP,AP90,SKD

## **Option 4.2**

Dictionaries present headwords in uninflected forms e.g. धर्म.

Dictionaries: BEN,BHS,BOP,BUR,CAE,CCS,GRA,GST,IEG,INM,MCI,MD,MW,MW72,PD,PE,PGN,PUI,PW,PWG,SCH,SHS,SNP,STC,VCP,VEI,WIL,YAT

### Note regarding options 4.1 and 4.2 –

(1) ACC is inconsistent e.g. अनन्ताचार्यः / अचलाचार्य.

(2) KRM has only verbs as headwords, so it is not included here.

## **Standard convention 4**

Uninflected form

# **Convention 5 – Treatment of ऋकारान्त words**

## **Option 5.1**

Uses ऋ at the end e.g. कर्तृ

Dictionaries: ACC,AP,AP90,BEN,BOP,BUR,CAE,GRA,GST,IEG,INM,MD,MW,MW72,PD,SHS,STC,VCP,VEI,WIL,YAT

## **Option 5.2**

Uses अर्‌ instead of ऋ at the end e.g. कर्तर्‌.

Dictionaries: BHS,CCS,PW,PWG,SCH

## **Option 5.3**

Uses inflected form with आ at end e.g. कर्ता

Dictionaries: PUI,SKD

### Note –

(1) KRM, MCI, PE, PGN, SNP do not have enough data to decide the convention conclusively.

(2) There is a possibility that a slightly different convention is followed in ऋकारान्त निपातित words e.g. जामातृ. They have not been examined in the current paper.

## **Standard convention 5**

Use ऋ at the end.

# **Convention 6 – Treatment of words ending with ‘at’**

## **Option 6.1**

Treat words derived from शतृ suffix as ending with अत्‌ e.g. गच्छत्‌.

Dictionaries: AP,AP90,BOP,BUR,GRA,GST,MD,MW,MW72,PD,SHS,VCP,WIL,YAT

## **Option 6.2**

Treat words derived from शतृ suffix as ending with अन्त्‌ e.g. अनागच्छन्त्‌.

Dictionaries: BEN,BHS,CAE,CCS,PW,PWG,SCH,STC,VEI

## **Option 6.3**

Treat words derived from शतृ suffix as ending with अन्‌ e.g. पश्यन्‌.

Dictionaries: SKD

### Note regarding options 6.1 to 6.3 –

(1) ACC, IEG, INM, KRM, MCI, PE, PGN, PUI and SNP don’t have enough words with this suffix to decide conclusively conventions 6.1 to 6.3.

## **Option 6.4**

Treat words derived from वतुप्‌ / मतुप्‌ suffix as ending with वत्‌ / मत्‌ e.g. भगवत्‌.

Dictionaries: ACC,AP,AP90,BOP,BUR,GRA,GST,IEG,INM,MCI,MD,MW,PD,SHS,VCP,WIL,YAT

## **Option 6.5**

Treat words derived from वतुप्‌ / मतुप्‌ suffix as ending with वन्त्‌ / मन्त्‌ e.g. भगवन्त्‌.

Dictionaries: BEN,BHS,CAE,CCS,PW,PWG,SCH,STC

### Note for options 6.4 and 6.5 –

(1) KRM has only verbs as headwords, so it is not included here.

## **Standard convention 6**

Treat all words mentioned above as ending with अत्‌.

# **Convention 7 – Treatment of वस्वन्त words**

## **Option 7.1**

Use वस्‌ at end e.g. विद्वस्‌.

Dictionaries: AP,AP90,BOP,BUR,CCS,GRA,GST,INM,MCI,MW,MW72,PD,PE,SHS,VCP,WIL,YAT

## **Option 7.2**

Use अंस्‌ at the end e.g. विद्वंस्‌.

Dictionaries: CAE,PW,PWG,SCH

## **Option 7.3**

Use आंस्‌ at end e.g. विद्वांस्‌.

Dictionaries: BHS,MD,STC

## **Option 7.4**

Use आन्‌ at the end e.g. विद्वान्‌.

Dictionaries: PUI,SKD

### Note –

(1) ACC, BEN, IEG, PGN, SNP and VEI do not have enough data to decide the convention conclusively.

(2) KRM does not have such headwords, so it is excluded from this list.

## **Standard Convention 7**

Use वस्‌ at the end.

# **Future research directions**

# Convention 8

Analysis of conventions regarding the तकारान्त words like महत्‌ / महन्त् / महन्त्‌ / महान्‌ etc. is not yet done. A dump of a sample word महत्‌ is noted below for reference. This requires closer look, because at least PUI, PW and SKD follow inconsistent convention.

a.महत्:AP,AP90,BHS,BOP,BUR,GRA,INM,MD,MW,MW72,PUI,PW,SHS,SKD,VCP,WIL,YAT

b. महन्त्‌:BEN,CAE,CCS,IEG,PW,PWG,SCH

c. महान्त्‌:STC

d. महान्‌:PE,PUI,SKD

# Convention 9

Analysis of सकारान्त words is pending.

# Convention 10

Analysis of रेफान्त words is pending.

# 

# Abbreviations

Dictionary Year Full form

|  |  |  |
| --- | --- | --- |
| ACC | 1962 | [Aufrecht's Catalogus Catalogorum](http://sanskrit-lexicon.uni-koeln.de/scans/ACCScan/2014/web/index.php) |
| AE | 1884 | [Apte Student's English-Sanskrit Dictionary](http://sanskrit-lexicon.uni-koeln.de/scans/AEScan/2014/web/index.php) |
| AP | 1957 | [Practical Sanskrit-English Dictionary, revised edition](http://sanskrit-lexicon.uni-koeln.de/scans/awork/homepage/ap-sample.php) |
| AP90 | 1890 | [Apte Practical Sanskrit-English Dictionary](http://sanskrit-lexicon.uni-koeln.de/scans/AP90Scan/2014/web/index.php) |
| BEN | 1866 | [Benfey Sanskrit-English Dictionary](http://sanskrit-lexicon.uni-koeln.de/scans/BENScan/2014/web/index.php) |
| BHS | 1953 | [Edgerton Buddhist Hybrid Sanskrit Dictionary](http://sanskrit-lexicon.uni-koeln.de/scans/BHSScan/2014/web/index.php) |
| BOP | 1847 | [Bopp Glossarium Sanscritum](http://sanskrit-lexicon.uni-koeln.de/scans/BOPScan/2014/web/index.php) |
| BOR | 1877 | [Borooah English-Sanskrit Dictionary](http://sanskrit-lexicon.uni-koeln.de/scans/BORScan/2014/web/index.php) |
| BUR | 1866 | [Burnouf Dictionnaire Sanscrit-Français](http://sanskrit-lexicon.uni-koeln.de/scans/BURScan/2013/web/index.php) |
| CAE | 1891 | [Cappeller Sanskrit-English Dictionary](http://sanskrit-lexicon.uni-koeln.de/scans/CAEScan/2014/web/index.php) |
| CCS | 1887 | [Cappeller Sanskrit Wörterbuch](http://sanskrit-lexicon.uni-koeln.de/scans/CCSScan/2014/web/index.php) |
| GRA | 1873 | [Grassman Wörterbuch zum Rig Veda](http://sanskrit-lexicon.uni-koeln.de/scans/GRAScan/2014/web/index.php) |
| GST | 1856 | [Goldstücker Sanskrit-English Dictionary](http://sanskrit-lexicon.uni-koeln.de/scans/GSTScan/2014/web/index.php) |
| IEG | 1966 | [Indian Epigraphical Glossary](http://sanskrit-lexicon.uni-koeln.de/scans/IEGScan/2014/web/index.php) |
| INM | 1904 | [Index to the Names in the Mahabharata](http://sanskrit-lexicon.uni-koeln.de/scans/INMScan/2013/web/index.php) |
| KRM | 1965 | [Kṛdantarūpamālā](http://sanskrit-lexicon.uni-koeln.de/scans/KRMScan/2014/web/index.php) |
| MCI | 1993 | [Mahabharata Cultural Index](http://sanskrit-lexicon.uni-koeln.de/scans/MCIScan/2014/web/index.php) |
| MD | 1893 | [Macdonell Sanskrit-English Dictionary](http://sanskrit-lexicon.uni-koeln.de/scans/MDScan/2014/web/index.php) |
| MW | 1899 | [Monier-Williams Sanskrit-English Dictionary](http://sanskrit-lexicon.uni-koeln.de/scans/MWScan/2014/web/index.php) |
| MW72 | 1872 | [Monier-Williams Sanskrit-English Dictionary](http://sanskrit-lexicon.uni-koeln.de/scans/MW72Scan/2014/web/index.php) |
| MWE | 1851 | [Monier-Williams English-Sanskrit Dictionary](http://sanskrit-lexicon.uni-koeln.de/scans/MWEScan/2013/web/index.php) |
| PD | 1976 | [An Encyclopedic Dictionary of Sanskrit on Historical Principles](http://sanskrit-lexicon.uni-koeln.de/scans/awork/homepage/pd-sample.php) |
| PE | 1975 | [Puranic Encyclopedia](http://sanskrit-lexicon.uni-koeln.de/scans/PEScan/2014/web/index.php) |
| PGN | 1978 | [Personal and Geographical Names in the Gupta Inscriptions](http://sanskrit-lexicon.uni-koeln.de/scans/PGNScan/2014/web/index.php) |
| PUI | 1951 | [The Purana Index](http://sanskrit-lexicon.uni-koeln.de/scans/PUIScan/2014/web/index.php) |
| PW | 1879 | [Böhtlingk Sanskrit-Wörterbuch in kürzerer Fassung](http://sanskrit-lexicon.uni-koeln.de/scans/PWScan/2014/web/index.php) |
| PWG | 1855 | [Böhtlingk and Roth Grosses Petersburger Wörterbuch](http://sanskrit-lexicon.uni-koeln.de/scans/PWGScan/2013/web/index.php) |
| SCH | 1928 | [Schmidt Nachträge zum Sanskrit-Wörterbuch](http://sanskrit-lexicon.uni-koeln.de/scans/SCHScan/2014/web/index.php) |
| SHS | 1900 | [Shabda-Sagara Sanskrit-English Dictionary](http://sanskrit-lexicon.uni-koeln.de/scans/SHSScan/2014/web/index.php) |
| SKD | 1822 | [Sabda-kalpadruma](http://sanskrit-lexicon.uni-koeln.de/scans/SKDScan/2013/web/index.php) |
| SNP | 1974 | [Meulenbeld's Sanskrit Names of Plants](http://sanskrit-lexicon.uni-koeln.de/scans/SNPScan/2014/web/index.php) |
| STC | 1932 | [Stchoupak Dictionnaire Sanscrit-Français](http://sanskrit-lexicon.uni-koeln.de/scans/STCScan/2013/web/index_fr.php) |
| VCP | 1873 | [Vacaspatyam](http://sanskrit-lexicon.uni-koeln.de/scans/VCPScan/2013/web/index.php) |
| VEI | 1912 | [The Vedic Index of Names and Subjects](http://sanskrit-lexicon.uni-koeln.de/scans/VEIScan/2014/web/index.php) |
| WIL | 1832 | [Wilson Sanskrit-English Dictionary](http://sanskrit-lexicon.uni-koeln.de/scans/WILScan/2014/web/index.php) |
| YAT | 1846 | [Yates Sanskrit-English Dictionary](http://sanskrit-lexicon.uni-koeln.de/scans/YATScan/2014/web/index.php) |

1. <http://www.sanskrit-lexicon.uni-koeln.de/> [↑](#footnote-ref-0)
2. For details of abbreviations, refer<http://sanskrit-lexicon.uni-koeln.de/> or abbreviation section of this paper. [↑](#footnote-ref-1)
3. There are some Tamil, Arabic headwords too, especially in IEG. [↑](#footnote-ref-2)
4. This is an ongoing project and the latest version of this classification can be seen at<https://github.com/sanskrit-lexicon/hwnorm1>. Scholars are requested to send their suggestions or comments at<https://github.com/sanskrit-lexicon/hwnorm1/issues>. [↑](#footnote-ref-3)
5. 100686 / 431178 headwords fall in this category. Data of sanhw1.txt accessed on 31st Oct 2016. [↑](#footnote-ref-4)
6. 39007 / 431178 headwords. [↑](#footnote-ref-5)
7. See explanation in convention 1 for reason of this choice. [↑](#footnote-ref-6)