Normalizing headwords of Cologne digital dictionaries

Dr. Dhaval Patel

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. Some dictionaries tend to use first inflected form of the headword e.g. धर्मः, whereas some tend to use the uninflected form of the headword e.g. धर्म. Similarly some dictionaries have tendency to use 'अर्' for words ending with ऋ e.g. पितर्, whereas some have tendency to use 'ऋ' e.g. पितृ and some others have tendency to use आ e.g. 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 in this regards. Present paper tries to analyse the different conventions followed by these authors / editors and come out with a standardized convention, so that the headwords list is in standard format.

Normalizing headwords of Cologne digital dictionaries

Dr. Dhaval Patel

Cologne Digital Sanskrit Dictionaries website[1]has a rich treasure of Sanskrit related dictionaries. There are 36 dictionaries hosted there as on 31thOctober 2016. There are three English – Sanskrit dictionaries (AE, BOR and MWE)[2]. Even after discounting these dictionaries whose headwords are in a foreign language, there are 33 digital dictionaries whose headwords are in Sanskrit and are 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 has 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[3].

Convention 1 - Treatment of अनुस्वार before consonants

There are at least 6 conventions of handling अनुस्वारs 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 words.

Option 1.1

Treat it as अनुस्वार, when occurring in between a word (other than the cases where the first member of compound ends with म) e.g. अकुंठित.

Dictionaries – AP90

[1] http://www.sanskrit-lexicon.uni-koeln.de/

[2] For details of abbreviations, refer http://sanskrit-lexicon.uni-koeln.de/ or abbreviation section of this paper.

[3] 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.

Option 1.2

Treat it as fifth letter of the वर्ग of following letter, when in between a word (other than the cases where the first member of compound ends with म्) e.g. चञ्चल.

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

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

Option 1.3

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

Dictionaries: AP90, SKD

Option 1.4

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.

Dictionaries: YAT

Option 1.5

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

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

Option 1.6

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

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

Notes regarding 1.5 and 1.6– (1) PE is inconsistent regarding conventions 1.5 and 1.6. See गङ्गासरस्वतीसंगम and वरदासङ्गम. (2) SHS is also inconsistent. See सङ्ग्रह and कर्म्मसंग्रह. (3) MW is also inconsistent. See अनभिसन्धि and अभिसंधिकृत. (4) SNP doesn't have any such case prima facie, hence excluded from this list.

Standard convention

1.1. Convert every nasal followed by consonant to अनुस्वार.

1.2. Convert every headword ending with अनुस्वार to ending with म्.

It is pertinent to note that majority of dictionaries tend to prefer convention 1.2. Huet (2009) also favoured this convention[4]. So statistically we should lean towards 1.2, but there are a few problems with that approach computationally. Many of the dictionaries following 1.2 also follow 1.5. 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 - Duplication of consonants after 'r'.

Option 2.1

Duplication is done in all cases e.g. पूर्व्व.

Dictionaries: SKD, WIL

Option 2.2

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

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 2.2, but there are a few inconsistent entries as well e.g. पर्वत and अग्निपर्व्वत.

Standard convention

2. No duplication.

Convention 3 – Treatment of words ending with 'at'

Option 3.1

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 3.2

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

[4] See section 1.6 of Huet's paper.

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

Option 3.3

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

Dictionaries: SKD

Note regarding options 3.1 to 3.3– (1) ACC, IEG, INM, KRM, MCI, PE, PGN, PUI and SNP don't have enough words with this suffix to decide conclusively conventions 3.1 to 3.3.

Option 3.4

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 3.5

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 3.4 and 3.5– (1) KRM has only verbs as headwords, so it is not included here.

Standard convention

3. Treat all words mentioned above as ending with अत्.

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 अनुस्वार of verb

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

Option 5.1

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

Dictionaries: KRM, PD, SKD, VCP, WIL

Option 5.2

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 5.3

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 5.1 to 5.3– (1) ACC, BHS, IEG, INM, MCI, PE, PGN, PUI, SNP, VEI do not have enough headwords to decide this convention decisively. (2) PD tends to give two separate headwords, one following options 5.1 and the other following option 5.2 e.g. अकि, अङ्क्. Therefore, it is included in both categories.

Standard convention

5. Option 5.3[5].

Convention 6 – Treatment of ऋकारान्त words

Option 6.1

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

Dictionaries: BHS, CCS, PW, PWG, SCH

[5] See explanation in convention 1 for reason of this choice.

Option 6.2

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 6.3

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.

Standard convention

6. Use ऋ at the end.

Convention 7 – Treatment of words ending with वस्/यस् of क्वसु, वसु, ईयसुन् suffices

Option 7.1

Use वस्/यस् at end e.g. विद्वस्.

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

Option 7.2

Use वांस्/यांस् at end e.g. विद्वांस्.

Dictionaries: BHS, STC

Option 7.3

Use वान्/यान् at the end e.g. विद्वान्.

Dictionaries: PUI, SKD

Option 7.4

Use वंस्/यंस् at the end e.g. विद्वंस्.

Dictionaries: CAE, PW, PWG, SCH

Note– (1) ACC, BEN, BHS, 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/यस् at the end.

TODO

1. 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.

    - महत्:AP, AP90, BHS, BOP, BUR, GRA, INM, MD, MW, MW72, PUI, PW, SHS, SKD, VCP, WIL, YAT
    - महन्त्:BEN, CAE, CCS, IEG, PW, PWG, SCH
    - महान्त्:STC
    - महान्:PE, PUI, SKD

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.

3. Analysis of सकारान्त and रेफान्त words is pending.

4. Over and above these known differences, there may be numerous other conventions followed by different dictionaries. The present paper is not comprehensive by any means. A thorough discussion and review of literature needs to be made in order to take it near to a universally acceptable standard and comprehesiveness. A sequel of this paper may be necessary to complete this task.

5. This paper is based on empirical analysis. There is a need to design a framework by which such conventions can be derived computationally by some algorithm. It is absolutely necessary if the standard has to be comprehensive. This will ensure that even minor conventions are not missed.

References

Huet, Gérard. 2009. *Formal structure of Sanskrit text: Requirements analysis for a mechanical Sanskrit processor*. Topics in Sanskrit Computational Linguistics, Eds. G. Huet, A. Kulkarni & P. Scharf, Springer-Verlag Lecture Notes 5402.

Abbreviations

Dictionary	Year	Full form
ACC	1962	Aufrecht's Catalogus Catalogorum
AE	1884	Apte Student's English-Sanskrit Dictionary
AP	1957	Practical Sanskrit-English Dictionary, revised edition
AP90	1890	Apte Practical Sanskrit-English Dictionary
BEN	1866	Benfey Sanskrit-English Dictionary
BHS	1953	Edgerton Buddhist Hybrid Sanskrit Dictionary
BOP	1847	Bopp Glossarium Sanscritum
BOR	1877	Borooah English-Sanskrit Dictionary
BUR	1866	Burnouf Dictionnaire Sanscrit-Français
CAE	1891	Cappeller Sanskrit-English Dictionary
CCS	1887	Cappeller Sanskrit Wörterbuch
GRA	1873	Grassman Wörterbuch zum Rig Veda
GST	1856	Goldstücker Sanskrit-English Dictionary
IEG	1966	Indian Epigraphical Glossary
INM	1904	Index to the Names in the Mahabharata
KRM	1965	Kṛdantarūpamālā
MCI	1993	Mahabharata Cultural Index
MD	1893	Macdonell Sanskrit-English Dictionary
MW	1899	Monier-Williams Sanskrit-English Dictionary
MW72	1872	Monier-Williams Sanskrit-English Dictionary
MWE	1851	Monier-Williams English-Sanskrit Dictionary
PD	1976	An Encyclopedic Dictionary of Sanskrit on Historical Principles
PE	1975	Puranic Encyclopedia
PGN	1978	Personal and Geographical Names in the Gupta Inscriptions
PUI	1951	The Purana Index
PW	1879	Böhtlingk Sanskrit-Wörterbuch in kürzerer Fassung

PWG	1855	Böhtlingk and Roth Grosses Petersburger Wörterbuch
SCH	1928	Schmidt Nachträge zum Sanskrit-Wörterbuch
SHS	1900	Shabda-Sagara Sanskrit-English Dictionary
SKD	1822	Sabda-kalpadruma
SNP	1974	Meulenbeld’s Sanskrit Names of Plants
STC	1932	Stchoupak Dictionnaire Sanscrit-Français
VCP	1873	Vacaspatyam
VEI	1912	The Vedic Index of Names and Subjects
WIL	1832	Wilson Sanskrit-English Dictionary
YAT	1846	Yates Sanskrit-English Dictionary