# Reviewer 1

## Strenghts

The paper provides a historical report on the "Cologne Digital Sanskrit Lexicon Project," a crutial online digital repository of Sanskrit dictionaries. The paper is informative in many respects, but also anecdotal, which makes it appear more like a blogg entry than an academic publication. Moreover, the paper falls short of academic writing standards such as providing bibliographical data for referenced academic publications (for example to Scharf's "Ramopakhyana, p.1 Monier-Williams Sanskrit-English Dictionary, Cappeller's Sanskrit Wörterbuch etc.). It also lacks a list of abbreviations, and the "Key References" are listed inconsistently. Sentences like "It's the most advanced UI feature by far. IDs are all capitals now as well. Cap has become CCS, PW2 is PW, and only Sch remains as SCH. PW2 was never used outside Cologne, it's PW or PWK now" are hardly intelligibly for the uninitiated. In its presented form, the paper is not publishable. However, after careful revision, it has the potential to become an informative article.

## Weaknesses

An abbreviated version of this paper could be presented orally, but it is impossible to publish the paper as it is

## Reproducibility

NA

## Overall recommendation

2.5

## Questions to the authors

## Missing references

## Typos and presentation

# Reviewer 2

## Strenghts

This paper gives a detailed and spirited overview of a central resource for anybody working on Sanskrit texts: the Cologne Digital Sanskrit Lexicon. A wealth of interesting behind-the-scenes information is provided about the project, as it went through various transitions in the 31 years so far of its existence. An overview of the current states and future plans of the project will be of interest to most participants in the conference.

## Weaknesses

none

## Reproducibility

5

## Overall recommendation

4

### Questions to the authors

Please go into greater detail of the prospects for linking this lexical resource to a full-text source corpus, which is a key point of theoretical interest.

### Missing references

### Typos and presentation

==> The English is good, but needs polishing in places (e.g. on p. 4, "thanks for … Brown University" is not clearly connected to the sentence and should probably be "thanks to … ," and "know at that time errors" should be "errors known at that time." Terms such as "L-based" (p. 7) and "SL multidictionary" (p. 14) need to be explained. On p. 23, "reference works, Whitney Roots" should presumably be item 3.3.

## Reviewer 3

### Strenghts

The paper tries to give an account of the history and current state of one important, perhaps the most important, online collection and interface to Sanskrit Lexicon, the Cologne Digital Sanskrit Lexicon Project (CDSLP).

• Detailed account of the history of the CDSLP. • Interesting technical details for anyone interested in the Github repositories that contain the data: • Uses SLP1 data • Describes XML format (but see below) • Useful details about the workflow for corrections (section §4.2).

### Weaknesses

• Stylistic issues • The paper has not been proofread: • "It's the ever-updated structured data that remains of value, but we can't ignore ordinary users as well." • The style of the paper is not fitting for a conference. It often reads like a mix of a personal account and a manifesto, e.g.:

- "In the ever ongoing correction of mistakes and additional hyperlinks Cologne digital dictionaries will remain superior to other digital Sanskrit dictionary projects and it's our responsibility to keep up the standard and spread the word." (p2)

- "And now, after 20 years, we still have to fight for the freedom of books on Indology and dictionaries of rare languages. Times have changed. The time has come." (p9)

- There's also crude generalisations: "An Encyclopaedic Dictionary of Sanskrit on Historical Principles (reached the middle of the letter `a`), is made at Deccan College in an environment that has only an Intranet, and no Internet connection on intention. Dr. Prasad Joshi, the general editor since 2017, has a plan to prepare 200 pages for print per year so that it will not be published even around 3000 years from now. Infinity is the limit. Ending a dictionary is not a goal, it's a process. Something a German lexicographer will never understand." (p19)

- Sometimes clarity has been sacrificed for no discernible reason:

  - "The total number of normalized entries is 392600+ and it still

grows similar to the Himalayas at around 2 millimeters per year." (p1, Abstract)
  - Does that mean they enter 2 new entries a year?
- Section called "XML Structure" (§3.4.2) reads like random notes, none of which are about the XML Structure (of what?).

• Problems with content • I don't know the details about the CDSLP, but I can imagine that the Copyright infringement section (§3.2) is not objective. • §3.4: Encoding Standard • The section called "XML Structure" (§3.4.2) is actually an account of several different things, in a sort of • Section 3.4.3 (pp11f) is called "XML Structure." Yet what it contains is this: • Information on Thomas Malten's original input method (plain text, not XML, but we have to infer this). • That CSDLP moved to github, after being managed through emails. • Then there's this: "Where is the code? Can other people see the code? The issues quickly piled up between 2014 and 2024 not so quickly to be solved (csl-orig, CORRECTIONS20, cologne-stardict, csl-apidev, csl-corrections, csl-pywork21, csl-websanlexicon, MWS22, PWK, hwnorm1 repos being the most active ones)." The author forgets to mention where the code can be seen. • §3.4.3 "Data Structure" contains examples of XML code that are not well-formed (k1' andk2' tags are not closed). • §4.1 "Normalizing Headwords": this section really just tells us that they are not normalized. It's only at the end that a possible solution is mentioned (for cases where feminine forms are treated as separate headings): "Feminine forms as search forms could be considered as alternative headwords. Quite easy, if we identify them with the L-body technique." (p16) But there's no further explanation of this "L-body technique." • References need work, they are incomplete and the formatting is off: • Gasūns 2006 missing?

- Ingalls 1985 seems wrong.

- "Otto Böhtlingk an Rudolf Roth"

### Reproducibility

N/A

### Overall recommendation

1.5

### Questions to the authors

### Missing references

References need to be fixed:

- Gasūns 2006 missing?
- Ingalls 1985 seems wrong.
- "Otto Böhtlingk an Rudolf Roth"

### Typos and presentation

Please get this paper proofread after revising it.

## Reviewer 4

### Strenghts

A report on the state of the Cologne Digital Sanskrit Lexicon Project. Marcis's report regarding the CDSL is a welcome document that make the wider public aware of the history and recent advancement in the digitization of Sanskrit bilingual and monolingual dictionaries. Particularly instructive is the recent work

he describes on Sanskrit-Russian dictionaries which are less known in Western Europe, the Americas, and India. The report contains much information not available anywhere else.

## Weaknesses

The report is verbose and diffuse. It is not well researched, nor well written, and is based entirely on the author's personal knowledge.

## Reproducibility

N/A

## Overall recommendation

2.5

## Questions to the authors

Marcis assumes that CDSL developed independently of work done at The Sanskrit Library. Many of the deliberations he makes in this report were already deliberated and decided decades ago. For example, regarding section 2.1.2, it was decided that it is counterproductive to merge the dictionaries because it is instructive to know who added what meanings at what time. Secondly a combined interface was tried (at The Sanskrit Library) and ruled out. What we do now is look up once and allow automated lookup of the same headword immediately if another dictionary is clicked. This is less confusing and more efficient computationally than dumping the results of many dictionaries in the same page. It appears from this report that people who have volunteered to work on the CDSL made no effort to coordinate with those of us who have worked on these dictionaries for decades previously. Marcis contacted me with a draft of this report just a few weeks ago, and I just sent my reports written at the conclusion of preceding funded projects to him then. Concerning 2.2 Incomplete Tagging, 3.1.2 Evolution of Data Formats, 3.4.3 Data Structure, 5.2.2 Unique Tags, The ideal goal of further transformation of the structure of the dictionaries should not be haphazard, but should adopt the Text-Encoding Initiative dictionary guidelines for all dictionaries. We developed samples of such a transformation of MW twenty years ago. Jim Funderburk was hesitant to transition to this form two decades ago because of the complexity, the length of the TEI tags, and resultant illegibility of the source files. However, since then TEI has developed author modes that would make working with the marked-up XML files easier using standard XML editors such as Oxygen. Oxygen can automatically transform XML files to JSON and vice versa. Concerning section 2.2.2 Verb Form Tagging, a decade ago in 2013 we coordinated the morphological identification tags used by the University of Hyderabad (Amba Kulkarni), The Sanskrit Heritage website (Gérard Huet), and The Sanskrit Library (Peter Scharf). The complete list of such tags is displayed at https://sanskritlibrary.org/helpmorphids.html. Concerning section 2.3.3, mention should be made of new dictionaries not included in CDSL, in particular Gérard Huet's Sanskrit-French dictionary, and several specialized dictionaries at The Sanskrit Library. Concerning section 3.1.3 Integration of Dictionaries and Corpora, 4.1 Normalizing Headwords, the topic was discussed by Funderburk and Scharf decades ago. Scharf has taken the approach to develop a Paninian dictionary of Sanskrit and to use that as the base list against which to coordinate headwords. Such a coordination would be independent of any particular dictionary in a mapping file, not by the addition of fields to any of the existing dictionaries. The approach should be to preserve the original dictionaries in their original form, not to modify them randomly based on some arbitrary idea of uniformity. There are fundamentally different principles adopted by the original dictionary authors that render making the dictionaries uniform impossible without destroying their work. The two major differences are the following: (1) MW combines preverbs with verbal roots in its headwords while PW does not but enters the preverbs as subheadings. (2) SKD, Apte use the nominative singular as headword while most others use the stem. Concerning section 3.2.2 Current Copyright Status, the current copyright status is Creative Common share-alike, non-commercial, development, acknowledge, as declared in a TEI XML-header file accompanying every dictionary source file.

**Missing references**

**Typos and presentation**