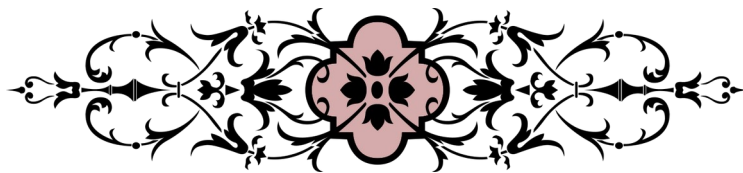


OPEN-SOURCE



PHILOLOGY

andrew ollett

DEFINITIONS

Philology:

Scholarship that attempts to **recover** the forms and meanings of texts composed before our time.

DEFINITIONS

Philology:

Scholarship that attempts to **recover** the forms and meanings of texts composed before our time.



I assume that this audience doesn't need to be convinced that this is important.

Two Visions of the Future



1. Technological change has meant that [philological] scholarship is getting better and better.

Two Visions of the Future



2. Technological change has meant that [philological] scholarship is getting worse, or at any rate not much better.

WHY?

Obviously there is no single explanation.

But I see two related developments:

- ✱ the **commercialization** of scholarship
- ✱ the **word-processorization** of scholarship

DEFINITIONS

Commercialization:

The tendency for scholarship to generate money for people other than the scholars themselves.

DEFINITIONS

Commercialization:

The tendency for scholarship to generate money for people other than the scholars themselves.



This is not necessarily a bad thing! (University presses need to recoup their costs, etc.)

DEFINITIONS

Commercialization:

The tendency for scholarship to generate money for people other than the scholars themselves.



But it does create a lot of **disincentives** to do philological work.

DEFINITIONS

Commercialization:

The tendency for scholarship to generate money for people other than the scholars themselves.



But it does create a lot of **disincentives** to do philological work.

→ editions “lower the value of the paper they’re printed on”

DEFINITIONS

Commercialization:

The tendency for scholarship to generate money for people other than the scholars themselves.



But it does create a lot of **disincentives** to do philological work.

- editions “lower the value of the paper they’re printed on”
- “impact” as a metric of research quality

DEFINITIONS

Commercialization:

The tendency for scholarship to generate money for people other than the scholars themselves.



But it does create a lot of **disincentives** to do philological work.

- editions “lower the value of the paper they’re printed on”
- “impact” as a metric of research quality → philology takes time

DEFINITIONS

Commercialization:

The tendency for scholarship to generate money for people other than the scholars themselves.



And it has contributed to a **restrictive** and **unfair** regime of property rights.

DEFINITIONS

Commercialization:

The tendency for scholarship to generate money for people other than the scholars themselves.



And it has contributed to a **restrictive** and **unfair** regime of property rights.

→ authors are usually expected to **completely alienate** all rights to their work

DEFINITIONS

Commercialization:

The tendency for scholarship to generate money for people other than the scholars themselves.



And it has contributed to a **restrictive** and **unfair** regime of property rights.

- authors are usually expected to **completely alienate** all rights to their work
- publishers **restrict** access to wealthy institutional customers

DEFINITIONS

Commercialization:

The tendency for scholarship to generate money for people other than the scholars themselves.



And it has contributed to a **restrictive** and **unfair** regime of property rights.

- it directly or indirectly **discourages** collaboration and sharing
 - authors are usually expected to **completely alienate** all rights to their work
 - publishers **restrict** access to wealthy institutional customers

DEFINITIONS

Word-processorization:

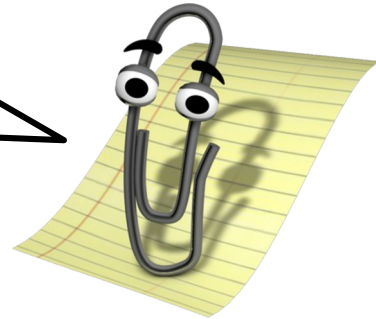
The tendency to **equate scholarly work in the humanities** with Microsoft Word™ documents.

DEFINITIONS

Word-processorization:

The tendency to **equate scholarly work in the humanities** with Microsoft Word™ documents.

I take it for granted that this *has in fact happened*, but we can talk about how, why, and when in the questions.



CONSEQUENCES *of* WORD-PROCESSORIZATION

Dependence on proprietary software and formats

You **have to** use Microsoft Word (or a clone).

The publisher will **probably** produce page proofs in Adobe InDesign™.

Both are **expensive, closed-source** software.

CONSEQUENCES *of* WORD-PROCESSORIZATION

Dependence on tedious outsourced labor

Just as in the pre-computer age, sending a document to a publisher usually involves having someone manually typeset it (and, if it is a book, index it)...

... if the publisher is reputable and responsible.

CONSEQUENCES *of* WORD-PROCESSORIZATION

Poor publication quality

If the publisher is less reputable and responsible, they're likely to simply run off the Microsoft Word document as-is, without bothering about typesetting (or anything else).

You all have seen books and journal articles like this.

CONSEQUENCES *of* WORD-PROCESSORIZATION

Critical editions are impossible

Editions with a critical apparatus are impossible in
Microsoft Word and an awful pain in Adobe InDesign.

CONSEQUENCES *of* WORD-PROCESSORIZATION

Fetishism of the printed page

- 1) Researcher writes something on a computer
- 2) Publisher converts it into a printed document
- 3) Researcher takes **photographs** of the printed document that are not computer-readable
- 4) Why?

CONSEQUENCES *of* WORD-PROCESSORIZATION

Fetishism of the printed page

The formats favored by publishers and scholars are **designed to make it difficult** to “go back” to the author’s original document.

→ Try extracting Devanagari text from a PDF sometime.

WHAT TO DO?

ЧТО ДѢЛАТЬ?

**Create the conditions — first of
all *technologically* — for *open-
source philology*.**

Create the conditions — first of
all *technologically* — for *open-
source philology*.



This is not meant to be scolding!

Like philology in general, open-source
philology will not be for everyone, and in our
small field, **all contributions are valuable.**

Create the conditions — first of
all *technologically* — for *open-*
source philology.



But it would be nice...

DEFINITIONS

Open-source:

The *source materials* of a project are made freely available for modification and redistribution. [[Wikipedia](#)]



Borrowed from *open-source software*.

DEFINITIONS

Open-source scholarship:

Basically putting stuff that you're working on online.

[More of a slogan than a real phenomenon.]



Kris Schaffer ([Hybrid Pedagogy](#))



Alex Fink ([Public Fragments](#))

DEFINITIONS

Open-source scholarship:

Academia.edu doesn't count.

It is a private company that makes money off of researchers without paying them anything.

DEFINITIONS

Open-source philology:

1. Ensuring that the **source materials** and the **products** of philological work are available **for free** to anyone who wants to use them, and available in **formats** that invites people to use them.

[i.e., making them available online with permissive licenses.]

DEFINITIONS

The idea behind this is simple: there is always **far too much** for any one person to do, and we want to encourage people to engage with the texts we study, even and especially in ways that we don't anticipate.

DEFINITIONS

In some subfields of South Asian Studies, this first aspect is already somewhat of a reality (e.g., people type up texts and share them on [GRETIL](#)).



aso2101 added ācārān'ga

a37e536 on Aug 15, 2020 10 commits



tei

added ācārān'ga

6 months ago



README.md

added samaraiccakaha

6 months ago

README.md

Prakrit Digital Text Project (PDTP)

This is an attempt to collect as much Prakrit literature as possible, using consistent and well-documented reference schemes and a uniform system of transliteration.

Project structure

All of the text data will be stored as TEI (P5) XML files. They are in the `tei` directory. (See below for current coverage and desiderata.) The point of using TEI is to be to (a) locate specific passages in the texts (using XPath), and (b) generate a number of secondary formats (e.g., plain text, HTML) if desired. Stylesheets for generating these secondary formats will be provided. Each file contains a `<teiHeader>` element that describes the source(s) on which the digital text is based.

Transliteration

Prakrit is written with different orthographies in Dēvanāgarī, and it is also transliterated in different ways. This corpus uses the [ISO 15919](#) system for transliteration, which means `ṁ` is used for *anusvāra* rather than `m̐` (for `e` and `o` see below). These texts **preserve** the orthography of their source editions in the following ways, which are diagnostic of

DEFINITIONS

Open-source philology:

2. Resisting Microsoft Word™ and switching to **free** and **open-source** software for philological work;
and acquainting ourselves with the **formats** and **tools** that will allow us to interact with text data more robustly and productively.

DEFINITIONS

Formats:

- **TEI** for representing texts (manuscript transcriptions, variant readings, editions, etc.)
 - Support for Indic texts much more robust in recent years thanks to the **DHARMA project**
- XSLT for going from TEI to other file formats (including LaTeX and HTML)

DEFINITIONS

Tools:

- **Python:** Every textual scholar *must* know it:
 - processing texts in and out of TEI
 - going between text-data (strings) and tabular data (spreadsheets)
 - running sophisticated searches and statistical analyses
 - transliteration and other language-specific tasks

DEFINITIONS

Tools:

- **LaTeX**: Typesetting software for producing *human-readable* critical editions.
 - reledmac/reledpar

DEFINITIONS

In the old days, you were *severely constrained* by what you could display in the footnotes or endnotes of a printed book.

But **now** there is **no limit** to the amount of data that you can put into your working files, even if the “human-readable” outputs have to represent a small selection of that data.

<lg xml:id="Lī.630">

<l>भणि<app><lem wit="#N">अं</lem><rdg source="#ed">यं</rdg></app>
<app><lem source="#ed">च</lem><rdg wit="#N">व</rdg></app> मए
पि<app><lem>अअम</lem><rdg source="#ed">ययम</rdg><rdg
wit="#N">असहि</rdg></app> ण हु एसो दुण्णओ सुहे<app><lem
source="#ed">ण</lem><rdg wit="#N">णण</rdg></app>म्ह ।</l>

<l>वो<app><lem source="#ed">ले</lem><rdg wit="#N">अ</rdg></app>इ
<app><lem wit="#ed">ज</lem><rdg wit="#N">झ</rdg></app>इ ण दे<app><lem
wit="#N">व्वो</lem><rdg source="#ed">वो</rdg></app> अणु<app><lem
source="#ed">कू</lem><rdg wit="#N">कु</rdg></app>लो होइ एत्ताहे ॥

<note type="totext">

<foreign xml:lang="pra-Deva">पिअअम</foreign>] ed (PJ);

<foreign xml:lang="pra-Deva">पिअसहि</foreign> N.

<foreign xml:lang="pra-Deva">देव्वो</foreign>] N;

<foreign xml:lang="pra-Deva">देवो</foreign> ed (PJ).

</note>

</l>

</lg>

That TEI code generated (via XSLT) the following plain-text output, which is what I gave to Harvard University Press for the edition of *Līlāvatī*:

६३० भणिअं च मए पिअअम ण हु एसो दुण्णओ सुहेणम्ह ।
वोलेइ जइ ण देव्वो अणुकूलो होइ एत्ताहे ॥ १५८

१५८ पिअअम] ed (PJ); पिअसहि N. देव्वो] N; देवो ed (PJ).

This **ignores** all of the in-line variation (which is mostly orthographic) and only presents the “substantive” variants (contained in a `<note>` element).

That TEI code generated (via XSLT) the following plain-text output, which is what I gave to Harvard University Press for the edition of *Līlāvatī*:

६३० भणिअं च मए पिअअम ण हु एसो दुण्णओ सुहेणम्ह ।
वोलेइ जइ ण देव्वो अणुकूलो होइ एत्ताहे ॥ १५८

१५८ पिअअम] ed (PJ); पिअसहि N. देव्वो] N; देवो ed (PJ).

But the data is **still there** and can be published separately.

That TEI code generated (via XSLT) the following plain-text output, which is what I gave to Harvard University Press for the edition of *Līlāvaī*:

६३० भणिअं च मए पिअअम ण हु एसो दुण्णओ सुहेणम्ह ।
वोलेइ जइ ण देव्वो अणुकूलो होइ एत्ताहे ॥ १५८

१५८ पिअअम] ed (PJ); पिअसहि N. देव्वो] N; देवो ed (PJ).

भणि<app><lem wit="#N">अं</lem>
 <rdg source="#ed">यं</rdg></app>
<app><lem source="#ed">च</lem>
 <rdg wit="#N">व</rdg></app>

PHILOLOGY *and* DATA

I will give some examples of preparing philological data from a project I'm currently working on: an edition and translation of a Prakrit work on poetics (mostly figures of speech like comparison, etc.) called the *Mirror of Ornaments* (*Alaṅkāradappanō*).

Nobody really knows about this text, but I believe it is the earliest surviving work exclusively concerned with poetics from South Asia.

PHILOLOGY *and* DATA

Philological scholarship is concerned with **texts**, and both works with and produces (either as final products or — more commonly — as “byproducts”) a lot of textual data:

 Manuscript images

PHILOLOGY *and* DATA

Philological scholarship is concerned with **texts**, and both works with and produces (either as final products or — more commonly — as “byproducts”) a lot of textual data:

- ✻ Manuscript images

- ✻ Transcriptions

PHILOLOGY *and* DATA

Philological scholarship is concerned with **texts**, and both works with and produces (either as final products or — more commonly — as “byproducts”) a lot of textual data:

- ✻ Manuscript images
- ✻ Transcriptions
- ✻ Word lists/glossaries

PHILOLOGY *and* DATA

I have a single TEI file for the text:



<lg n="28" met="gāthā" xml:id="ADa.28">

<l><app><lem type="conj" resp="#A0">सुसरिसअं जं खेवो</lem><rdg wit="#J" source="#N #Bh">सुसरिसमापखेवं</rdg><note xml:lang="eng">The restoration <foreign xml:lang="pra-Deva">पक्खेवो</foreign> is also possible.</note></app> विअलइ स <app><lem resp="#A0" type="em">च्चेअ</lem><rdg wit="#J" source="#N #Bh">च्चेव</rdg></app> होइ दरविअला ।</l>

<l><app><lem resp="#A0" type="em">एक्के<lb n="3"/>क्कमोवमाणेहिँ</lem><rdg wit="#J" source="#N">एक्क<lb n="3"/>क्कमोवमाणेहिँ</rdg><rdg source="#Bh">एक्कक्कमोवमाणेहिँ</rdg></app> होइ <app><lem resp="#A0" type="em">एक्केक्कमा</lem><rdg wit="#J" source="#N #Bh">एक्कक्कमा</rdg><note xml:lang="eng">For the spelling see Hēmacandra's <hi rend="italic">Dēśīnāmamālā</hi> (1.146).</note></app> णाम ॥ २८<supplied resp="#ed"> ॥</supplied>

<note type="chaya">

<lg>

<l>सुसदृशतां यत्क्षेपं विकलति सैव भवतीषद्विकला ।</l>

<l>अन्योन्योपमानैर्भवत्यन्योन्या नाम ॥</l>

</lg>

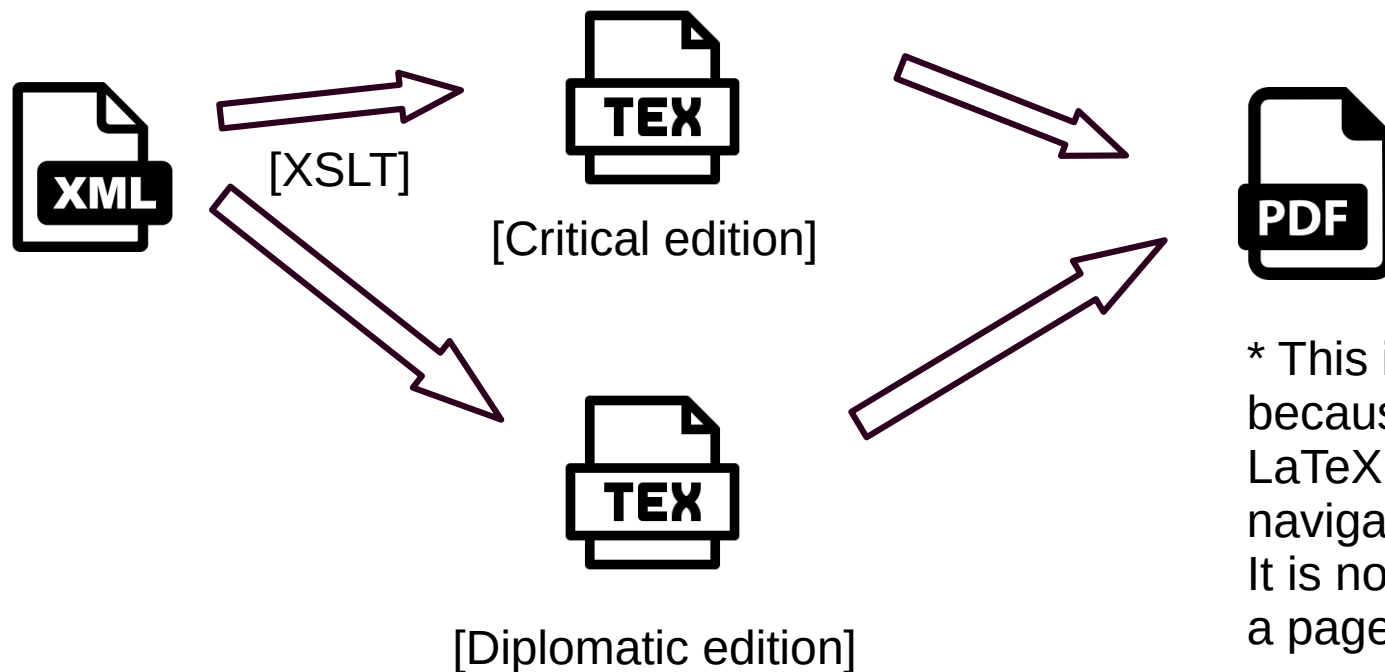
</note>

</l>

</lg>

PHILOLOGY *and* DATA

This is converted into a critical and diplomatic edition:



* This is a PDF. But because it's made with LaTeX, it has hyperlinks, navigable contents, etc. It is not just an image of a page.

- 65 सुसरिसअं जं खेवो विअलइ स च्चेअ होइ दरविअला । गाथा
66 एक्केक्कमोवमाणेहिं होइ एक्केक्कमा णाम ॥ २८ ॥
-

65 सुसरिसअं जं खेवो] AO; सुसरिसमापखेवं J
N Bh. The restoration पक्खेवो is also
possible.

65 च्चेअ] *em.* AO; च्चेव J N Bh.

66 एक्केक्कमोवमाणेहिं] *em.* AO; एक्कक्कमोवमाणेहिं
J N; एक्कक्कमोवमाणेहिं Bh.

66 एक्केक्कमा] *em.* AO; एक्कक्कमा J N Bh. For
the spelling see Hēmacandra's
Dēśīnāmamālā (I.146).

- 65 सुसरिसअं जं खेवो विअलइ स च्चेअ होइ दरविअला । गाथा
66 एक्केक्कमोवमाणेहिं होइ एक्केक्कमा जाम ॥ २८ ॥
-

65 सुसरिसअं जं खेवो] AO; सुसरिसमापखेवं J
N Bh. The restoration पक्खेवो is also
possible.

65 च्चेअ] *em.* AO; च्चेव J N Bh.

66 एक्केक्कमोवमाणेहिं] *em.* AO; एक्कक्कमोवमाणेहिं
J N; एक्कक्कमोवमाणेहिं Bh.

66 एक्केक्कमा] *em.* AO; एक्कक्कमा J N Bh. For
the spelling see Hēmacandra's
Dēśināmamālā (I.146).

```
<app><lem resp="#A0" type="em">चेअ</lem>  
  <rdg wit="#J" source="#N #Bh">चेव</rdg></app>
```

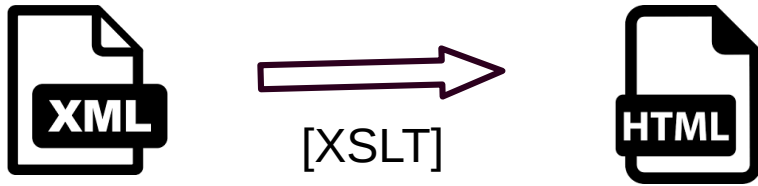


- 1 सिलेसाणसालेसा ॥२५॥ संखलोवमाजहा ॥ सग्गस्सवकणअगि
- 2 जहा ॥ सोसंझारोअसमोचलपेम्मोजो जणोसुहअ ॥ सोकिंभा
- 3 क्रमोवमाणेहिं होइएकक्रमाणाम ॥२८॥ दरविअलाजहा ॥ पी
- 4 एकक्रमाजहा ॥ पअई विमलाओदोणिविविबुहजणेणिवुईक

रीकंचणगिरिणुवमहिअलाहोओ ॥ महिवीढस्सविसरवरणपच्चलोतहतुमंचेअ ॥२६॥ लेसोवमा
 सइसंझाराएणवजोणरिंछोली ॥२७॥ सुसरिसमापखेवंविअलइसचेवहोइदरविअला ॥ एक
 णत्थणीसरूआपहपेसिअलोअणासटकंठा ॥ लिहियव्वदारलग्गाणचलइतुहदंसणासाए ॥२९॥
 राओअ ॥ एकेक्रमएकक्रमसरिसाओतुहकित्तीतिअससरिआअ ॥३०॥ णिंदाएसलहिज्जइउवमेओजत्थसा

PHILOLOGY *and* DATA

It could just as well be converted into HTML for a website:



Edition

A

B

C

Conspectus

Romanized

ಕನ್ನಡ

ಮನದೊಳಗಾಗಳುಂ ನಿಟಿಸದಾಮುಳಿಸಂ ಗತರಾಗಮೋಹನೊ
ಯ್ಯನೆ ಮುನಿನಾಥನೆಯ ವಮನಾಂತರಮಿಂತಿದು ದಾಕ್ಷಿಣಾತ್ಯಕಂ
ಮನದೊಳಗಾಗಳುಂ ನಿಟಿಸದೊಯ್ಯನೆ ಮೋಹಮನಂತೆ ರಾಗಮಂ
ಮುನಿಸುಮನೊಲ್ಲದಂ ಮುನಿಪನೆಯುದಿದಕ್ಕುಮುದೀಚ್ಯವಾಚಕಂ

Apparatus

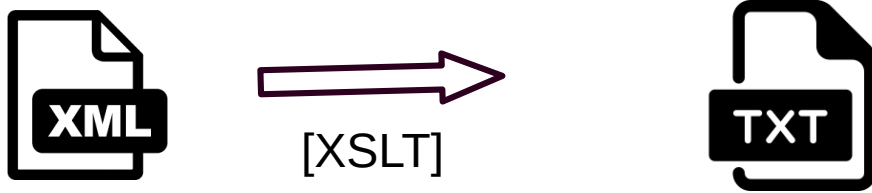
- oyyane BC] oykane A.
- vacanāntaram AB] vacanāntarav C.
- dākṣiṇātyakaṁ C] dākṣiṇātyavā AB.
- oyyane BC] oykane A.
- olladaṁ AB] alladaṁ C.

Translation

Here is another saying on the Southern Way:

PHILOLOGY *and* DATA

Or into plain text for (e.g.) GRETIL:*



* GRETIL now uses TEI.

PHILOLOGY *and* DATA

In addition, while I (?) work on a program that can semi-automatically **tokenize** and **lemmatize** Prakrit texts, my TEI file includes a **manually tokenized** version of each verse, which is then used to generate **word lists**.

(I don't know yet what I'll end up doing with these word lists, besides automatically creating a glossary and index for the text.)

```
<ab xml:id="ADa.28-a" type="analysis">
  <s>
    <w lemma="सुसरिसआ">सुसरिसअं</w>
    <w lemma="जं">जं</w>
    <w lemma="खेव">खेवो</w>
    <w lemma="विअल - v">विअलइ</w>
    <w lemma="तं">स</w>
    <w lemma="चिअ">चेअ</w>
    <w lemma="हो">होइ</w>
    <w>
      <w lemma="दर" msd="Hyph">दर -</w>
      <w lemma="विअल">विअला</w>
    </w>
  </s>
  <s>
    <w>
      <w lemma="एकेकम" msd="Hyph">एकेकम -</w>
      <w lemma="उवमाण">उवमाणेहिं</w>
    </w>
    <w lemma="हो">होइ</w>
    <w lemma="एकेकम">एकेकमा</w>
    <w lemma="णाम">णाम</w>
  </s>
</ab>
```

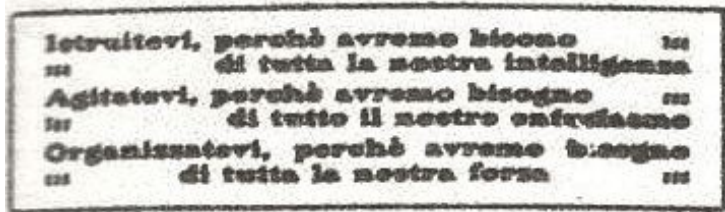
ADa.28ekkekka-1	ADa.28 ēkkēkkama-	ēkkēkkama				Hyph	,
ADa.28uvamARehi2	ADa.28 uvamāṇēhim ^u	uvamāṇa					,
ADa.28hoi3	ADa.28 hōi	hō					,
ADa.28ekkekka-4	ADa.28 ēkkēkkamā	ēkkēkkama					,
ADa.28RAma5	ADa.28 ṇāma	ṇāma					,
ADa.28-1dara-1	ADa.28-1dara-	dara				Hyph	,
ADa.28-1vialA2	ADa.28-1vialā	viala					,
ADa.28-1jahA3	ADa.28-1jahā	jahā					,
ADa.29pIRa-1	ADa.29 pīṇa-	pīṇa				Hyph	,
ADa.29tTanI2	ADa.29 tthanī	thaṇa					,
ADa.29sarUA3	ADa.29 sarūā	sarūva					,

Kernel: 5.4.0-050400-lowlatency
Lemma: POS Meaning
sumdara ADJ beautiful
User: andrew
paa NOUN food
Battery: 0h 10m 19s 26%
vinṇāsa NOUN placement
vimala ADJ spotless
alamkāra NOUN ornament
rēhia ADJ
sarīra NOUN body
SUI: 420MiB / 7.70GiB 5%
sui NOUN sound
dēvia NOUN goddess
DISKS
ca CONJ and
kavva NOUN poem / 227GiB
paṇama VERB bow
pavara ADJ best
vaṇṇaḍḍha
savva ADJ all
ja
hō VERB become

SSID: agn
Channel: Type Sanskrit PSM
AP MAC: ts sundara
Link quality: tb pada
tb vinyāsa
Local IP: 10.0.0.
Public IP: ts vimala 73.9.1
Up: ts alaṅkāra / 54
Down: tb śarīra
tb śruti
tb dēvī
ts ca
tb kāvya
CPU: tb praṇam
All Core: tb pravara 3.10Ghz
tb sarva
tb bhū

LICENSE IT!

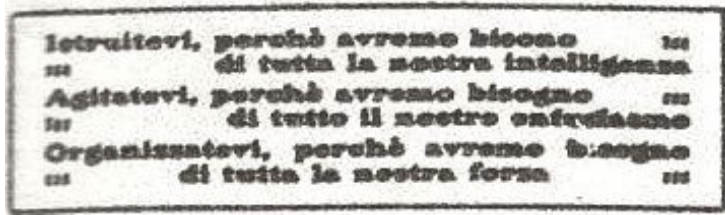




WHAT CAN I DO?

Educate yourselves.

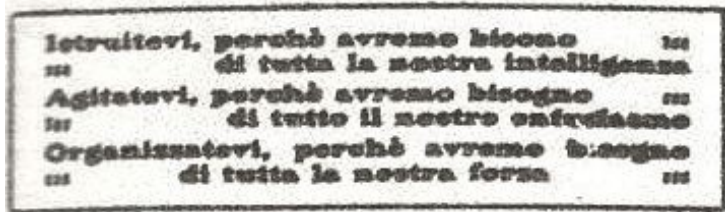
- [The Programming Historian](#)
- DHARMA [encoding guidelines](#)
- Read around and see what is possible with digital tools (e.g., [Oliver Hellwig](#)'s work)
- Take a class!



WHAT CAN I DO?

Fight.

- Insist on [free licenses](#) with colleagues and publishers.
- Seek out open-access venues for publication.



WHAT CAN I DO?

Organize.

- Start open-source projects with other colleagues.
- **Share** code, tools, and ideas.

THANK YOU!