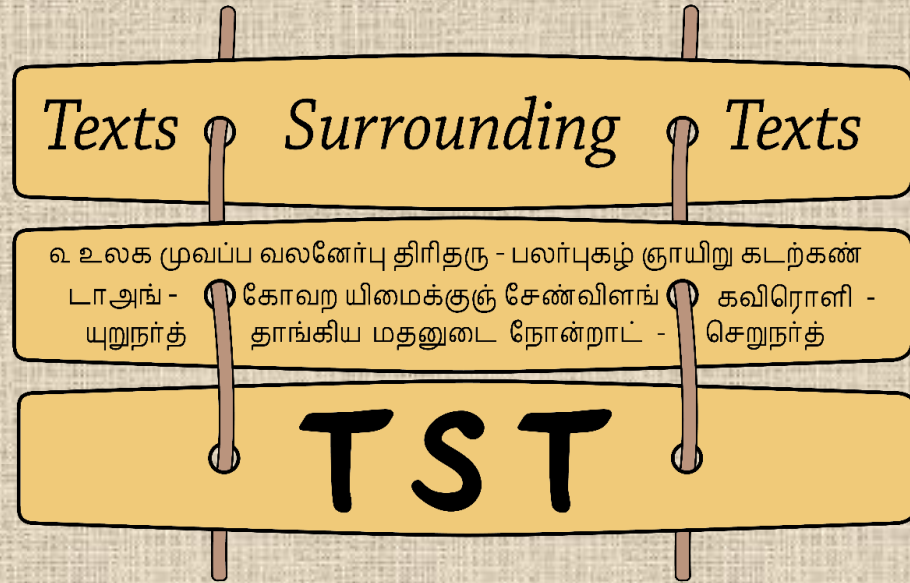


Transcribing and Transliterating Tamil Manuscripts



E. Francis (CNRS)

TST Webinar # 4

January 19th, 2021

All photos of BnF manuscripts by courtesy of BnF

2 Options

- Transcribing the original Tamil text in Tamil script
 - ! Use unicode font !
 - ! Use proper Tamil numerals (not letters similar to numerals) !
- Transliterating the original Tamil text in Roman script

In any case the conventions observed — whether in transcription or in transliteration — should be stated (e.g. *ō* and *ē* restored or not, overshoot *u* restored or not, etc.) and will be recorded in a dedicated field in the online form (with suggestions to choose from). This will allow automatic TEI-markup.

Display

- Two modes of displaying the Tamil text are implemented by Charles Li:

- (1) Tamil script display mode
- (2) Transliteration display mode

Whatever the option chosen (transcribing or transliteration), the Tamil portions will be displayed in Tamil transcription AND in Roman transliteration.

<https://tst-project.github.io/mss/>

https://tst-project.github.io/mss/Indien_0001.xml

A

click here to
shift to Tamil
script mode

Attuvaitānupavam

Record edited by Emmanuel FRANCIS (CNRS, CEIAS UMR
8564, EHESS/CNRS).

Published in 2020 by TST Project.

அ

click here to
shift to Tamil
transliterated
mode

அத்துவைதானுப-
வம்

Record edited by Emmanuel FRANCIS (CNRS, CEIAS UMR
8564, EHESS/CNRS).

Published in 2020 by TST Project.

The Advantages of Transliteration

- Searching a multi-script corpus, e.g. South and Southeast Asian inscriptions of the DHARMA project (inscriptions in various scripts, with various scripts used for one and the same language).

NB: Tamil manuscripts often contains Sanskrit portions/words (in Grantha, but also in other scripts used for Sanskrit).

- Representing graphic phenomena more easily and in a more detailed manner than in Tamil transcription.

Notably by using short-hand transliteration (see below), which will/can be automatically converted to proper TEI markup.

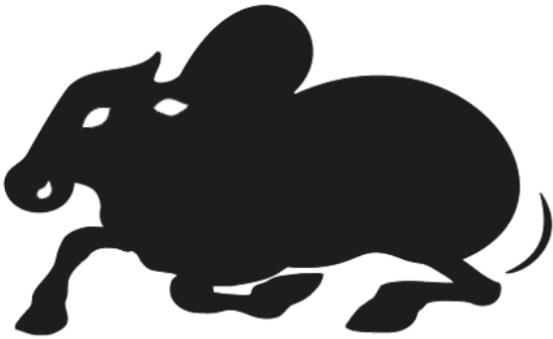
TST Transliteration Conventions

The TST transliteration conventions are conceived as a subset of the DHARMA transliteration and TEI-encoding conventions, for the sake of alignment and, hopefully, so as to spread a more detailed/granular and interoperable standard for the transliteration of South Asian and Southeast Asian texts.

See:

- DHARMA Transliteration Guide [⟨halshs-02272407v3⟩](#) = TG
- DHARMA Encoding Guide for Diplomatic Editions [⟨halshs-02888186⟩](#) = EGD
- DHARMA cheatsheet here: <https://erc-dharma.github.io/>

*Reading
suggestions
for long
winter
evenings ...*



dharmā
Transliteration Guide

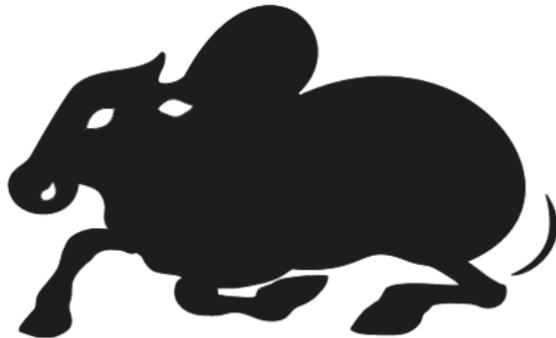
Dániel Balogh & Arlo Griffiths

Release Version 3, 2020-07-05

27 pp.



This project has received funding from the European Research Council (ERC)
under the European Union's Horizon 2020 research and innovation programme
(grant agreement No 809994).



dharmā
**Encoding Guide
for Diplomatic Editions**

Dániel Balogh & Arlo Griffiths

Release Version 1, 2020-07-05

150 pp.



This project has received funding from the European Research Council (ERC)
under the European Union's Horizon 2020 research and innovation programme
(grant agreement No 809994).

Today's Presentation and Discussion

- Showing/Explaining DHARMA-derived suggestions for TST transliteration and TEI-encoding conventions.

NB: further guidance and feedback will be provided by Axelle Janiak (DHARMA TEI-XML data-manager).

- Discussing these suggestions, so as to improve and update DHARMA TG and EGD.
- Aligning as far as possible the transcriptions and transliterations in the TST project, in the same approach that we have adopted for the TST controlled vocabulary for the paratexts.

In this connection, I am aware that Giovanni Ciotti and Marco Franceschini have elaborate transliteration conventions for their South-Indian colophons.

I am also aware that Jean-Luc Chevillard has his own conventions for diplomatic transliteration of Tamil manuscripts.

For the present purpose, we might not want/need to adopt a very strict diplomatic transliteration scheme.

Editorial Conventions (aligned with DHARMA conventions)

<https://tst.hypotheses.org/conventions> (under construction)

display		TEI markup
(abc)	unclear reading	<unclear>
(a/b)	alternative unclear reading	
[abc]	editorial restoration of uncertain or lost text	<supplied> (lost)
[a/b]	alternative restoration	
<abc>	editorial addition of omitted text	<supplied> (omitted)
[...]	gap of unknown number of characters	<gap>
[X]	gap of 1 character	
etc.	etc.	

DHARMA cheatsheet (in progress)

<https://erc-dharma.github.io/>

Description	DHARMA markup	DHARMA display
Line beginning	<code><lb n="1"/>svasti śrī</code> <code><lb n="2"/>kōpparakēcari</code>	(1) (2)
Word divided across lines	<code><lb n="1"/>...dhar</code> <code><lb n="2" break="no"/>ma...</code>	(1)...dhar- (2)ma...
Tentative reading (letters ambiguous outside of their context)	<code>dha<unclear>rma</unclear></code>	dha(rma)
	<code>dha<unclear cert="low">rma</unclear></code>	dha(rma?)
Unclear, could be read either a or o	<code><choice></code> <code><unclear>a</unclear></code> <code><unclear>o</unclear></code> <code></choice></code>	(a/o)
Lacuna restored (supplied)	<code>dha<supplied reason="lost">r</supplied>ma</code> <code>dha<supplied</code> <code>reason="illegible">r</supplied>ma</code>	dha[rma] dha[rma]

Unmarked \bar{o} and \bar{e}

Restored by the editor, as it indicates to the reader how the text has been understood by the editor.

In a further step, provided the convention of restoration is recorded

— in the “conventions” field [\bar{o} and \bar{e} restored or not, etc.] and in the “script” field [specifying whether double-curled *kompū* is used or not in the original, discriminating, when necessary, the different units/hands of, e.g., a composite MS —

all restored unmarked \bar{o} and \bar{e} can be automatically converted to TEI-markup and thus explicitly marked as a modern editorial intervention.

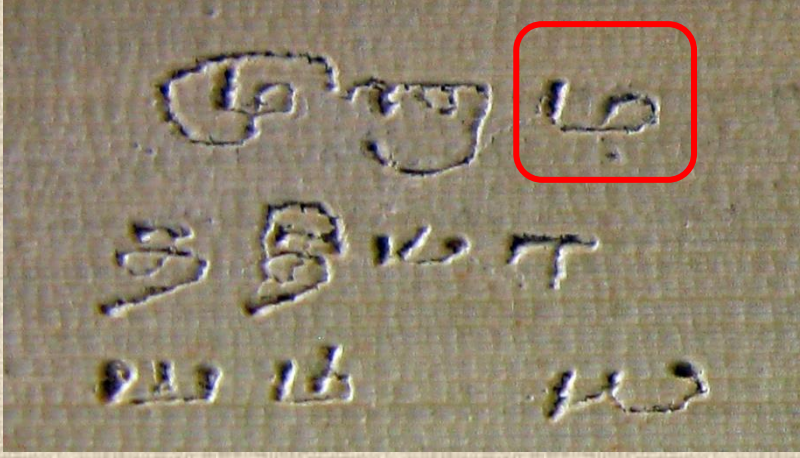
Possible to restore unmarked \bar{o} and \bar{e} in the transliteration display mode only and have a more diplomatic display (with original short *o* and *e* in the Tamil display mode).

Vowels

- Upper case for initial vowels
- Lower case for medial vowels

If primary encoding is in Tamil script, to be able to display space between words in the transliteration mode would require heavy markup ...

So it seems more practical to encode the original Tamil text in Roman transliteration, adding editorial space (not present in *scriptio continua*).

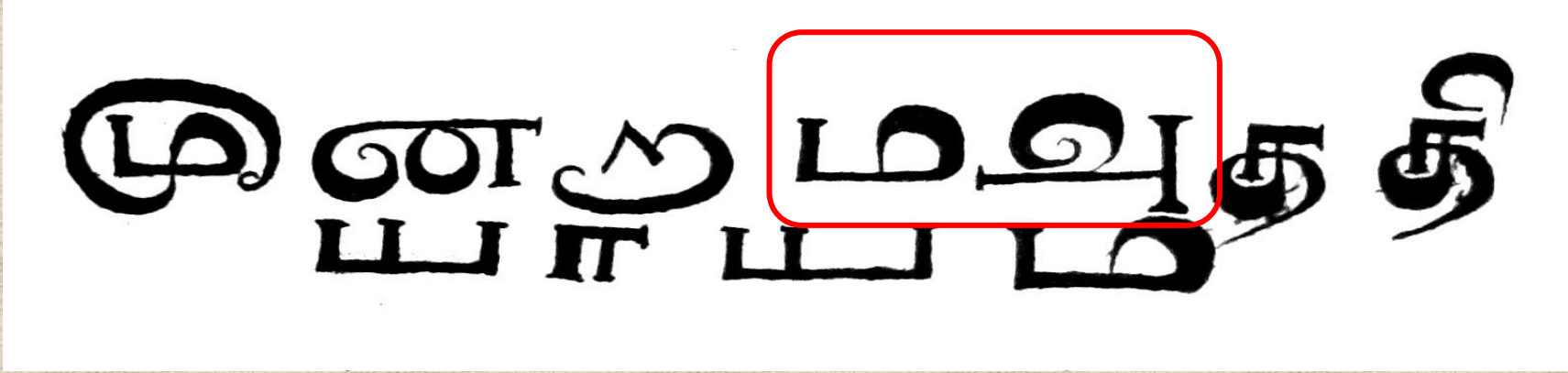


BnF Indien 262 script mode]

mūṇā^m attiyāyam [transliteration mode]

முனாமத்தியாயம் *or*

முனாமத்தியாயம் [Tamil



BnF Indien 339

mūṇrā^m Attiyāyam [transliteration mode]

முன்றாம் அத்தியாயம் [Tamil script mode]

திருவவதாரச்சருக்கம்

BnF Indien 297

tiru-v-avatāra-c-carukkam [transliteration mode]

திருவவதாரச்சருக்கம் [Tamil script mode]

திருஅவதாரச்சருக்கமுற்றும்

BnF Indien 297

tiru-Avatāra-c-carukkamurrum [transliteration mode]

திருஅவதாரச்சருக்கமுற்றும் [Tamil script mode]

Overshort *u*

Restored by the editor, as it indicates to the reader how the text has been understood by the editor.

Transliterated by

,

[apostrophe] followed by space



BnF Indien 1

ruḷ aṛiyā^{t'} uraittālum inmai [transliteration mode]

ரு ளறியா துரைத்தாலு மின்மை *or*

ருளறியாதுரைத்தாலுமின்மை [Tamil Script mode]

Puḷḷi (occasional in the original)

Transliterated by

- Middle Dot [U+00B7] followed by space

NB: the same middle dot can also be used (but with space before and after) for medial dot punctuation in the original.

In a further step, provided the convention of encoding the original *puḷḷis* in this manner is recorded — in the “conventions” field and in the “script” field — all unmarked restored *puḷḷis* can be automatically converted to TEI-markup and thus explicitly marked as a modern editorial intervention.

If the *puḷḷi* is consistently used in the MS or in a discrete part of it no need to encode it. Just record its use by ticking “*puḷḷi*” in the “Features” of “Script” in the field “Scribal hand” of the online form.

Scribal hand

^ v X

Codicological unit(s) Choose... Scope Choose... Scribe Choose...

Script Choose... Features Choose... Medium Choose...

Description of hand

Tamil

- ☐ modern ra
- ☐ long ō
- ☐ long ē

Devanagari

☐



puḷḷi will be added in this dropdown menu

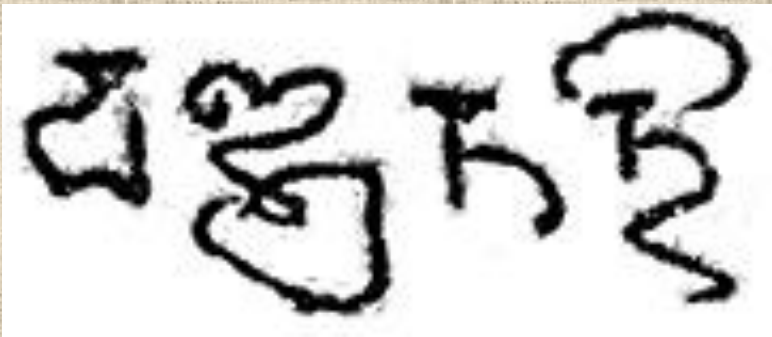
Grantha

TST shorthand = bold.

DHARMA markup = `<hi rend="grantha">`

NB: Letters common to both Tamil (Tamil Grantha) and Grantha scripts are not to be marked in bold or marked-up, e.g. *n*, *t*, *y*, *v*.

Even in a string evidently entirely in Sanskrit in a Tamil text, e.g.



is transliterated **vajranandi**

NOT **vajranandi**

(as the letter *va* is identical in the Tamil and in the Sanskrit Grantha portions of this inscription)

Virāma

Transliterated by

.

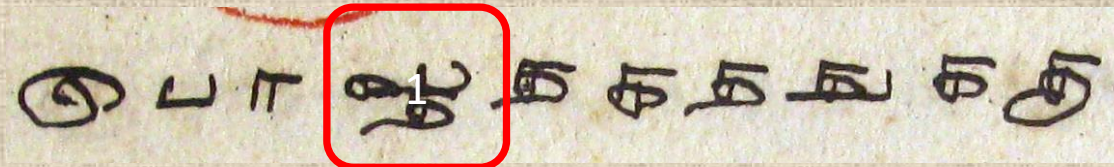
Middle Dot [U+00B7] followed by space (i.e. like for an original *pu!!i*, but in bold)

BnF Indien 549



Asmat[·]

BnF Indien 339



postta kattukku

Hyphen

- [hyphen]

Editorial hyphen. To split compounds, for word-split at the end of the line.

Displayed or not in Tamil script mode?

– [n-dash]

For original hyphen.

Numerical Values

See DHARMA conventions (EGD 7.1.1).

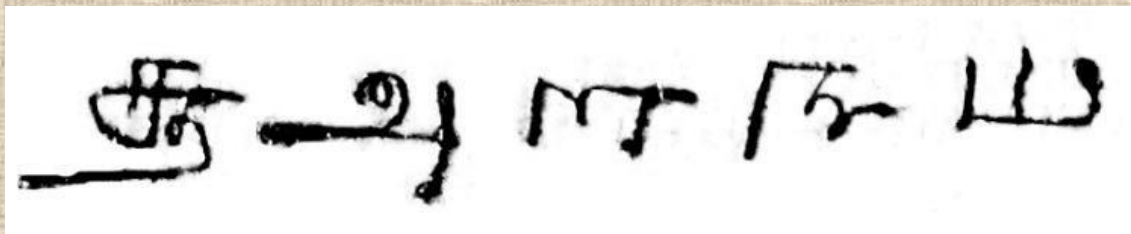
Decimal notation, transliterate without space.

௫௩ > 53

Additive notation, transliterate with space.

௫௩௧ > 5 10 3

- Shorthand TST: 5 10 3
- Shorthand DHARMA: 5 10+ 3



1000 8 100 3 10

Numeral 1830 is written as 1000 (plus) 8 (times) 100 (plus) 3 (times) 10

- Shorthand TST: 1000 8 100 3 10
- Shorthand DHARMA: 1000+ 8 100+ 3 10+

```
<num value="1830">  
  <g type="numeral">1000</g> 8 <g type="numeral">100</g> 3 <g  
  type="numeral">10</g>  
</num>
```

	Original	Encoding (shorthand transliteration)	Display (transliteration)
Decimal notation	௩௩	53	53
Additive notation	௩௩	5 10 (TST) 5 10+ (DHARMA)	5^0
Additive notation	௩௩௩	5 10 3 (TST) 5 10+ 3 (DHARMA)	5^03
Additive notation	௩௩௩௩௩	3 1000 4 100 3 (TST) 3 1000+ 4 100+ 3 (DHARMA)	$3^{000}4^{00}3$

See Indien 10: https://tst-project.github.io/mss/Indien_0012.xml **(display to be updated!)**

Symbols/Abbreviations

Shorthand

{...}

where “...” is the expanded form of the symbol, that is, encode in transliteration the expanded meaning between two curly brackets, e.g.

- {pc} *piḷḷaiyār cuḷi*, form undetermined.
- {pcs} *piḷḷaiyār cuḷi*, short form.
- {pcl} *piḷḷaiyār cuḷi*, long form.
- {mērpaṭi}, {varuṣam}, {mācam}, {tēti}, etc.

- {symbol} for unidentified symbol
- {symbol?} for what looks like a symbol, but is not identified

Other symbols:

- {cross}
- {...} create it freely but write to Charles and myself.

@ Gio+Marco: using {YK1}, {YK2}, etc. is thus OK, provided you send Charles values for @type and @subtype.

Magic Charles will automatically convert in the XMLs these shorthand encodings to TEI markup, using element <g> with various @type, e.g.

- {pcs} > <g ref="#pcs"/>
- {pcl} > <g ref="#pcl"/>
- {tēti} > <g ref="#tēti"/>

Displayed as transliterated in transliteration mode (except pc) and with corresponding font in Tamil script mode.

DHARMA shorthands and markup

A more elaborate markup could be adopted following the one currently developed in the DHARMA project.

Transliteration shorthand	DHARMA markup
	<g type="danda">.</g>
	<g type="ddanda">.</g>
//	<g type="ddandaOrnate">.</g>
~	<g type="dash">~</g>

Verse

Add ‡ at the end of an *aṭi*.

Add ‡‡ at the end of a stanza.

To be converted to TEI markup (if time permits).

```
<lg n="1" met="anuṣṭubh">
```

```
<l n="a"><lb n="31"/>bhūmi-dānāt paran dānaṃ</l>
```

```
<l n="b">na bhūtan na bhaviṣyati</l>
```

```
<l n="c">tasyaiva haraṇa-pāp<choice><sic>a</sic><corr>ān</corr></choice></l>
```

```
<l n="d">na bhūta<supplied reason="omitted">ṃ</supplied> na bhaviṣyati</l>
```

```
</lg>
```

Orthographic Peculiarities (1) Conjunct Letters

THIRUPPORUR AND VADAKKUPPATTU
EIGHTEENTH CENTURY LOCALITY ACCOUNTS

M. D. SRINIVAS
T. G. PARAMASIVAM
T. PUSHKALA



CENTRE FOR POLICY STUDIES CHENNAI
2001

கூ

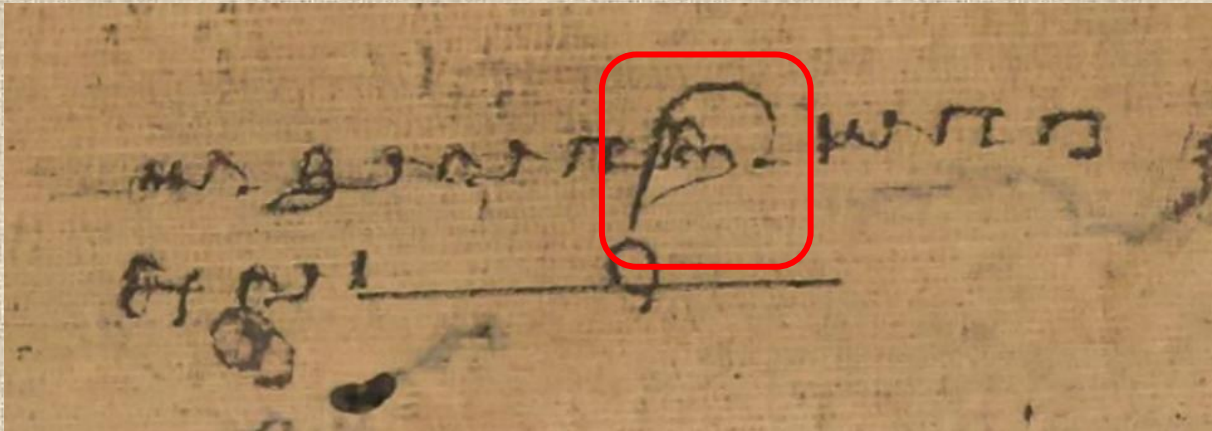
க் க

ந

க் கு

Srinivas et al. 2001, p. 47: kūṭṭeluttuka!

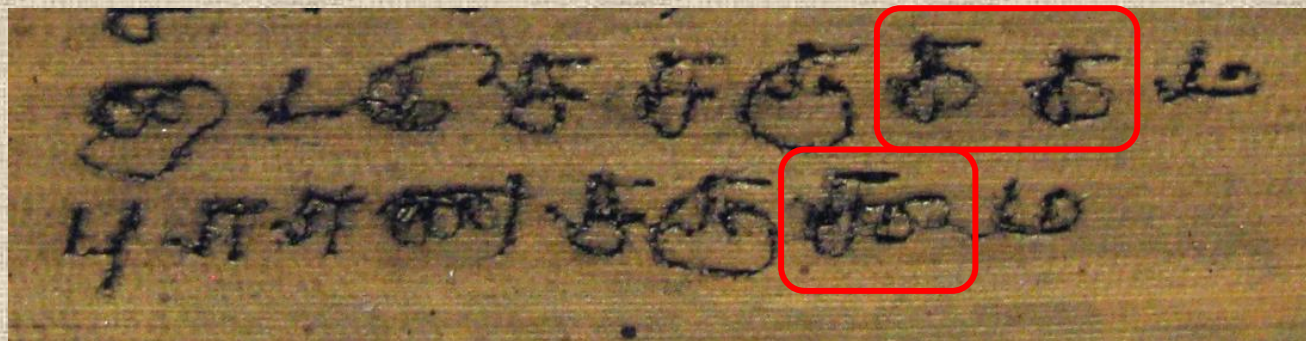
Conjunct double *t*



BnF Indien 3

y-itu vāt=tiyār
cuvaṭi

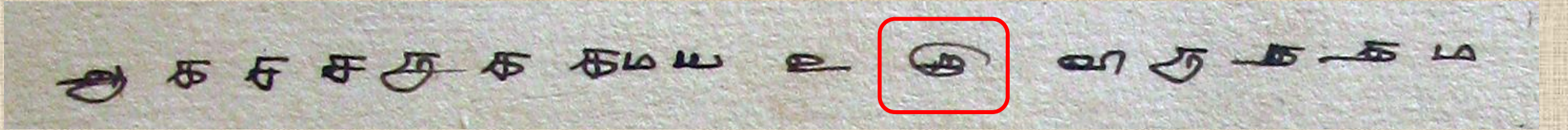
Conjunct double k



BnF Indien 310

nāṭṭu-c-caru**kkam**
purāṇa-caru**k=kam**

Conjunct double k



BnF Indien 297

Āka-c carukkam 10 2-**k=ku** viruttam

Orthographic Peculiarities (2) Final Grantha *m*

THIRUPPORUR AND VADAKKUPPATTU
EIGHTEENTH CENTURY LOCALITY ACCOUNTS

M. D. SRINIVAS
T. G. PARAMASIVAM
T. PUSHKALA



CENTRE FOR POLICY STUDIES CHENNAI
2001

ஆ

அம்

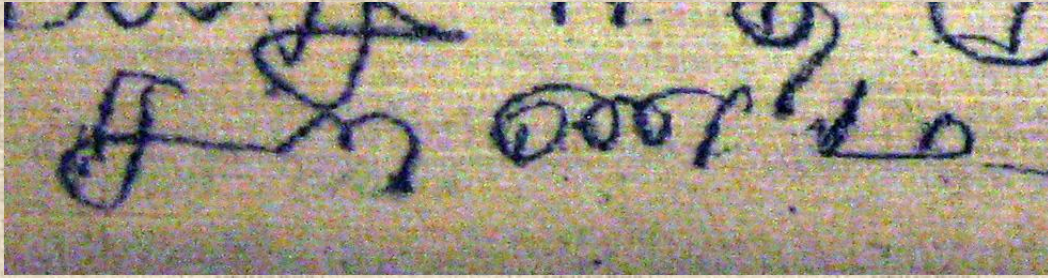
கூ

கம்

யூ

யம், யும்

Srinivas et al. 2001, p. 47: kūṭṭeluttuka!



BnF Indien
551



BnF Indien 974

கு

ளம்

ணு

ணம்

னு

னம்

Srinivas et al. 2001, p. 47: kūṭṭeḷuttukaḷ



BnF Indien 390

śrī-rāma-ceyam



BnF Indien 265

śrī-(r/r)ā=ma-ceyam

நு

நாம, நரம்

Srinivas et al. 2001, p. 47: kūṭṭeḷuttuka!



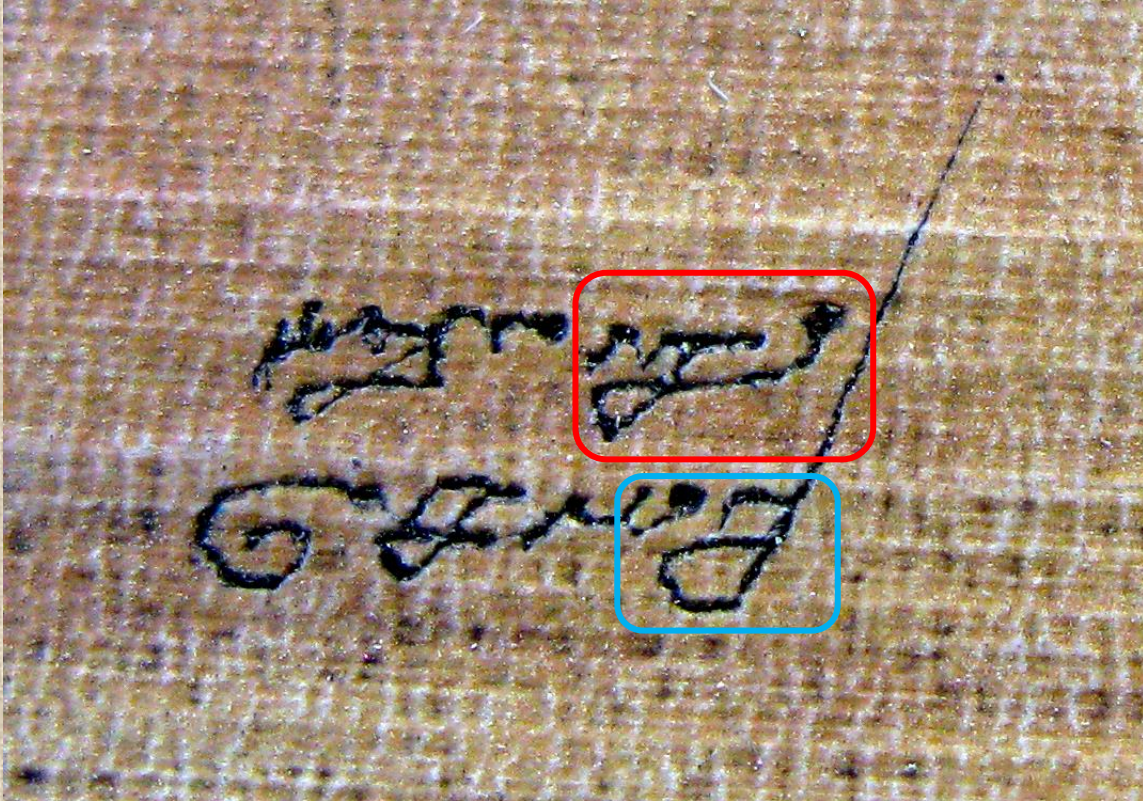
BnF Indien 1037

śrī-(r/ṛ)ā=ma-ceyam

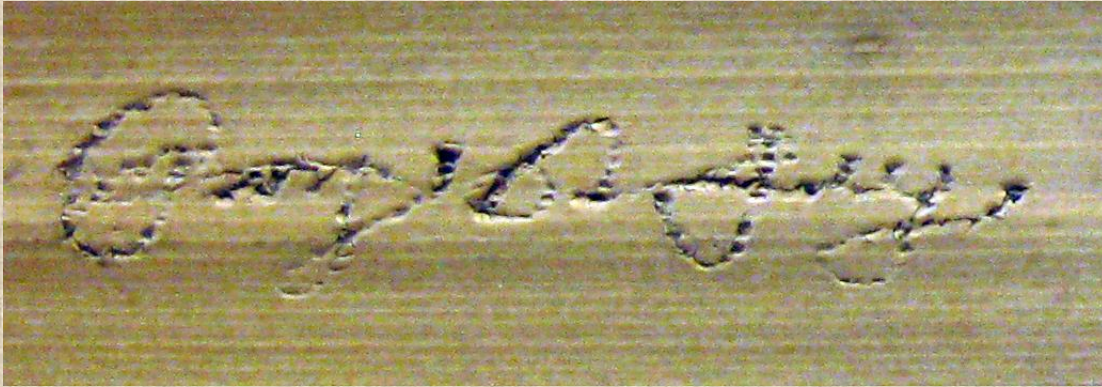
ॐ

கிரந்த மகரம்

Srinivas et al. 2001, p. 47: kūṭṭeḷuttukaḷ

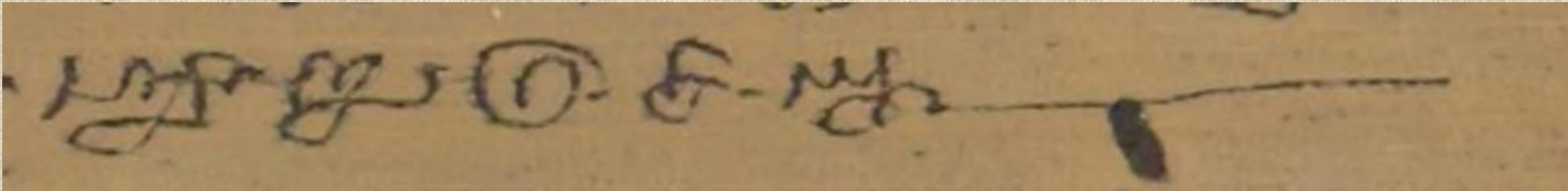


śrī-rā=ma-
ceyam.



BnF Indien 550

śrī-(r/r)ā=ma-ceya=m



Bnf Indien 3

śrī-(r/r)ā=ma-ceya=m={pcl}