

A Proposal for the Digital Encoding of Palm-Leaf Tamil Manuscripts

Jean-Luc Chevillard

▶ **To cite this version:**

Jean-Luc Chevillard. A Proposal for the Digital Encoding of Palm-Leaf Tamil Manuscripts. International Forum for Information Technology in Tamil, 2003, Chennai, Tamilnadu, India. pp.109-121. halshs-00642780

HAL Id: halshs-00642780
https://shs.hal.science/halshs-00642780

Submitted on 16 Dec 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

17

A Proposal for the Digital Encoding of Palm-Leaf Tamil Manuscripts

Dr. Jean-Luc Chevillard

[jlc@ccr.jussieu.fr],
CNRS, University Paris 7
History of Linguistics Research Team
[UMR 7597, HTL])
<http://www.linguist.jussieu.fr/~chevilla/>

Introduction

While dealing with texts that have been handed down to us by tradition, a constant concern should be to understand the exact conditions of the successive material embodiments through which these texts have been existing, before being transferred to a new support, like what happens for instance when a collection of Classical Tamil poems, that was available at a time on a palm-leaf MS, is transferred onto a book by an editor, and eventually becomes a digital file, kept on a server, on the Internet.[1] Drawing inspiration from a recent book on Early Tamil Epigraphy [2003] by the great scholar Iravatham Mahadevan, I would like to examine here the implication for all who are interested in Tamil MSS, of the distinction which is to be done between Apparent Reading (henceforth AR) and Intended Reading (henceforth IR),[2] and emphasize the importance of dealing with these two dimensions separately, in two different phases of the decipherment.

To make clear to the readers what I designate here by AR and by IR, I will first of all reproduce (See **Fig. 1, Appendix 2**) a small photographic sample from a palm-leaf MS, that was kindly communicated to me by my colleague, Dr. Eva Wilden, formerly from Hamburg univer-sity in Germany, and now shortly to be attached to the EFEO center in Pondicherry, after a recent visit to the Tiruvāvaṭuturai Ātīnam. For convenience sake, the scanned image will be divided into 3 parts: R (right-hand side), L (left-hand side) and M (middle part), and those will be put on top of each other, although, as should be clear from the position of the holes, they follow each other horizontally from left to right (with overlapping), in the LMR order.

[1] Such is the case with the "Project Madurai" file server, at < http://www.tamil.net/projectmadurai/pmfinish.html>.

[2] This distinction is adapted from the one used by I. Mahadevan, who himself distinguishes between an "Apparent reading" and an "Actual reading" (see for instance pp. 227, 229, 236, 238, and other similar passages). I do not however use that second expression, because the expression "Intended reading" seems to me to convey more clearly the notion that the existence of distinctions is always relative to the ability of someone to perceive them.

I have placed the R segment in Fig. 1 on top because the poem which shall be used for the purpose of this demonstration has its beginning on its second line, after a long hyphen, looking approximately like this:

--- உயாவி{ளை}[3]யுபபினகொ

The continuation of this sequence is then found on the 3rd line of the L segment, where it looks approximately like this:

ள{ளை}சாறறியதாபடுபூழியசெடபுலமபடருந

Only the beginning of this 3rd line is reproduced here, and the reader can see for himself that the text con tinues, with some overlapping, on the 4th line of M segment, then (again) in the R segment (3rd line), and so on, until the end of the poem is found at the end of the 7th line of the R segment, as the following text:

... பனமாணபெதைககொழிநததெனஞசெ ---

This poem is known to belong to the Akanāṉūṟu anthology, where it has (in modern editions) the number 390, and the reader can compare the above text with what is found in the Murray Rajam edition (NCBH reprint), on p.213:

உவர் விளை உப்பின் கொள்ளை சாற்றி

அதர் படு பூழிய சேண் புலம் படரும்

...

பல் மாண் பேதைக்கு ஒழிந்தது, என் நெஞ்சே.

Another possible standard of comparison is the Kazhagam version of the same text, edited by Po. Vē. Cōmacuntaraṉār, where we see:

உவர்விளை உப்பின் கொள்ளை சாற்றி

அதர்படு பூழிய சேட்புலம் படரும்

...

பன்மாண் பேதைக் கொழிந்ததென் நெஞ்சே.

Basing my argument on these different elements, I should now proceed to illustrate with concrete examples the **AR** and **IR** concepts that were first mentioned in my opening paragraph (respectively "Attested Reading" and "Intended Reading"), and one look at the following chart (See **Chart 1**) – especially a comparison between the IR as explicited by Murray Rajam and the IR as explicited by Po. Vē. Cō. – should give the reader an idea of the complexity involved. But first of all, it also falls on me to mention two concepts which are very important for the editor of a text and which we continuously meet with, one being the concept of "Variant text" and the other one being the concept of "Error", although I must immediately say that the distinction between the two can very often appear as a blurred one.

AR	IR	Comment
கொள{ளை}சாறறி	கொள்ளை சாற்றி	

[3] The {ளை}combination stands for what is a single glyph in the writing system of the MS. Other instances of sequences of characters between "{" and "}" should also be understood as referring to a single glyph.

செடபுலம	சேண் புலம்	Murray Rajam
செடபுலம	சேட்புலம்	Po. Vē. Cō.
உயாவி{ளை}	உவர் விளை	Variant or error?

Chart 1.

However, not paying too much attention, at this early stage, to the variants, I shall now concentrate on those aspects that appear as the most significant differences between the AR and the IR, because they make the reading of a MS especially difficult. The three main differences are:

- the fact that the MS does not make use of puḷḷi, so that, for instance, the same glyph (க, ச, ட, த, ப, ற, ...) has sometimes to be read as having an "inherent a" (க, ச, ட, த, ப, ற, ...) and sometimes as being a pure consonant (க், ச், ட், த், ப், ற், ...).
- the fact that the glyph ெ must be read sometimes as modern ே and sometimes as plain ெ.
- the fact that the glyph ர has three possible intended readings: ர ("ra"), ர் ("r") and ர (sign for long ā), so that for instance the sequence கர can be read (at least) in 3 different ways as கர ("kara"), கர் ("kar") or கர ("kā").

For a modern reader of a living language that possesses a standard writing system, there is of course no point in distinguishing between AR and IR levels. He (or she) sees a written form and (in normal cases!) immediately understands an intended meaning. But for a text in a classical language, with an unfamiliar vocabulary, chances are that a casual (and untrained) reader will often stumble in his/her decipherment and could hesitate between several interpretations (supposing he/she can find at least one). Therefore, preserving the text which is recorded on a MS, should probably best be done in two phases:

- STEP 1. Recording the AR
- STEP 2. Proposing one IR (relatively to some agreed writing system)

One advantage is that both parts of the work are not necessarily being performed by the same person, as the level of proficiency required for STEP 1 is less than the one required for STEP 2 (becoming a *pulavar* is probably more difficult than becoming a good copyist).

One second advantage is that a corpus of text consisting solely of AR can probably be already used in some automatic processes, like "searching occurrences" of a word: what a trained human reader can do, a clever program can try to emulate. And such computerized automatic processes of the AR can probably help very much in the task of proposing an IR. One wonders indeed whether a "neural network" style learning algorithm could not be trained (on the basis of existing editions, with the IR they propose) and used in the search for the original IR which the Sangam poets had in mind. Still, leaving aside (for the time being) this speculation, the rest of this paper will, henceforth, be devoted to enumerating the necessary elements that should be present in a minimum encoding scheme, which could then be applied for faithfully recording the AR of the text on a MS, as it is.

AR, IR and the problem of variants

However, before we proceed to this descriptive task, it appears necessary to come back to the question of variants which we mentioned earlier. The reader of this paper might indeed wonder why it is so important to preserve the AR while all a modern reader is interested in is the IR. It must therefore be emphasized that what we believe we know about the IR is only an hypothesis, and that several IR-s can sometimes be associated to the same AR, because the writing system which is used in MSS is inherently ambiguous. I now give a few examples to illustrate this statement.

The 1st line of poem 8 of *Kuṟuntokai* appears in U.V.S.'s edition as:

கழனி மாத்து விளைந்துகு தீம்பழம் (*lectio* 1)

and 2 variants are mentioned by U.V.S.: they are the following:

"கழனி மரத்து" (*lectio* 2) and "கழனி மாஅத்து" (*lectio* 3).

However, we must admit that the distinction between *lectio* 1 and *lectio* 2 exists only in terms of IR, because one and the same AR has to be reconstructed for both, namely:

கழனிமாததுவி{ளை}நதுகுதீமபழம (*lectio* 1 and *lectio* 2)

whereas *lectio* 3 corresponds to a different IR

கழனிமாஅததுவி{ளை}நதுகுதீமபழம (*lectio* 3)

If we examine one single edition of a work, such situations where the editor indicates two distinct IR corresponding to the same reading are rather rare, but if we compare different editions of the same work, the frequency of the phenomenon increases. As noted by Mu. Caṇmukam Piḷḷai in his 1985 edition of *Kuṟuntokai* (=KT), we have a number of IR discrepancies between the KT texts as proposed by U.V.S. and S. Vaiyāpuri Piḷḷai (=S.V.P.), but we can observe that in a number of cases the AR must have been the same, as is illustrated by the following chart:

KT ref.	U.V.S.'s IR	S.V.P.'s IR	postulated AR
30:6	தமியேன்	தமியென்	தமியென
30:6	அளியேன்	அளியென்	அளியென
53:7	மகளிரொடு	மகளிரோடு	மகளிரொடு
94:3	மருள்வேன்	மருள்வென்	மருளவென
148:4	குருந்தோடு	குருந்தொடு	குருநதொடு
173:7	உளெனே	உளேனே	உளெனெ
181:6	பலகடம்	பல்கடம்	பலகடம
190:6	இயங்குதோறு	இயங்குதொறு	இயஙகுதொறு
191:1	இதுஎன்	இதுஎன	இதுவென

200:2	தாஅய	தாஅய்	தாஅய
211:2	நேர்ந்துநம்	நொந்தும்நம்	நொநது...[4]
236:2	நேர்ந்தனை	நொந்தனை	நொநத{னை}
270:5	உள்ளமோடு	உள்ளமொடு	உளளமொடு
301:5	இனமணி	இன்மணி	இனமணி
318:8	கள்வனும்	களவனும்	களவனும
320:1	கொள்மீன்	கோள்மீன்	கொளமீன

Chart 2

This is of course only a part of the discrepancies which have been listed by Mu. Caṇmukam Piḷḷai, who also compares other editions of KT, but if we take into consideration the fact that KT is one of the anthologies that have received a very high level of attention, it is to be expected that such variant IR-s based on the same AR should be quite frequent in general in the case of text that are ancient and not well understood.

Inventory of elements to be seen in the AR

We now proceed with a chart (see Chart 3, below) of the "alphabetical"[5] elements that can be met with inside a MS like the one that is partially reproduced on Fig. 1.

[4] The final portion of the AR which has to be postulated is not exactly the same in both cases, but still this appears as a good example of potential ambiguity.

[5] This means that we leave aside here the numerals and the (rare) punctuation elements. We have also not tried to include inside the chart such very rare elements as ஹி ஈ, ஞி, ஜு, etc.

அ	ஆ	இ	ஈ	உ	ஊ	எ	ஒ	ஐ	ஔ	ஃ
க	கா	கி	கீ	கு	கூ	கெ	கொ	கை		
ங										
ச	சா	சி	சீ	சு	சூ	செ	சொ	சை		
ஞ										
ட	டா	டி	டீ	டு	டூ	டெ	டொ	டை		
ண	{ணா}	ணி	ணீ	ணு	ணூ	ணெ	{ணா}	{ணை}		
த	தா	தி	தீ	து	தூ	தெ	தொ	தை		
ந	நா	நி	நீ	நு	நூ	நெ	நொ	நை		
ப	பா	பி	பீ	பு	பூ	பெ	பொ	பை		
ம	மா	மி	மீ	மு	மூ	மெ	மொ	மை		
ய	யா	யி	யீ	யு	யூ	யெ	யொ	யை		
ர	ரா	ரி	ரீ	ரு	ரூ	ரெ	ரொ	ரை		
ல	லா	லி	லீ	லு	லூ	லெ	லொ	{லை}		
வ	வா	வி	வீ	வு	வூ	வெ	வொ	வை		
ழ	ழா	ழி	ழீ	ழு	ழூ	ழெ	ழொ	ழை		
ள	ளா	ளி	ளீ	ளு	ளூ	ளெ	ளொ	{ளை}		
ற	{றா}	றி	றீ	று	றூ	றெ	{றா}	றை		
ன	{னா}	னி	னீ	னு	னூ	னெ	{னா}	{னை}		

Chart 3

This is of course an idealized chart, because it is based on a modern type face. The reader will get a closer idea of what the actual glyphs on many MSS look like if he compares Chart 3 with the following Chart 4, where a typeface is used which is based on an actual MS, namely the one from which a sample was given on Fig.1.[6] Its examination reveals that the glyphs it contains do not differ very much from the glyphs used in present day handwriting, one notable difference concerning the {ற, றா, றி, றீ, று, றூ, ...} family, for which the shapes {ர, ரு, ர, ரீ, ரூ, ...}[7] were rather significantlly different, not only from modern printed forms, but also from modern hand-written forms. Other possible differences concern the 2 series of conjunct forms {ஷ, ஸா, ஸை, ஸை} and {ஞ, ஜ, ரு}, which have been eliminated from printing by a script reform, but remain fairly frequent in modern handwriting. The reader will remark that the difference between ரு (=ரு) and ர (=றா) is very small in this MS (and sometimes seems unascertainable, if we are not helped by the context), which fact could easily be a cause for reading mistakes.

அ	ஆ	இ	ஈ	உ	ஊ	எ	ஒ	ஐ	ஔ	ஃ
	சா	சி	சீ	சு	சூ	செ	சொ	சை		
ஞ										

[6] This is the reason why not all characters look alike, since some were unavailable on the MS. Part of the missing ones have been substituted by characters from a modern typeface.

[7] No specimen of ரீ was found on the MS.

Chart 4

Another chart, giving some glyph samples from still another MS, and illustrating a slightly different handwriting, is also to be found in the Appendix **(Fig. 2)** to this article**.** The most notable trait, which seems to be frequent in MSS, is the characteristic form of the initial part of the conjunct forms {ஒ, கூ, சை, ஞை}, like for instance in . There are of course other particularities. Each MS has to be carefully examined for itself at the beginning, in order for the reader to become familiar with its peculiarities.

A possible 7-bit encoding

We have until now stayed away from technicalities and dealt only with the underlying concepts. One of the fonts actually used in this paper follow the TSCII standard. Another one has to extend this standard in order to be able to encode such conjunct forms as {ஒ, கூ, சை, ஞை}. But the encoding used has no consequences on the argument which is made here, that one should be able to reproduce Classical Tamil texts exactly as they were transmitted to us, in order to understand the vicissitudes of their transmission. It appears that it is convenient to have an extendable 7-bit method for representing the data as it stands in a MS. I briefly present here a possible extendable system, illustrating it with the fragment already used at the beginning of this paper, accompanied by a fragmentary chart (see **Chart 5**) of the possible entities. The fragment runs as follows:

{_u}{va}{#}{vi}{Lai}{yu}{pa}{pi}{n2a}{-e}{ka}{#}
{La}{Lai}{ca}{#}{Ra}{Ri}{ya}{ta}{#}{pa}{Tu}{pU}{zhi}{ya}

{-e}{ca}{Ta}{pu}{la}{ma}{pa}{Ta}{ru}{na}....

The entities that appear in this fragment can easily be identified in the following chart, for ease of reference:[8]

அ={_a}	ஆ={_A}	இ={_i}	ஈ={_I}	உ={_u}	ஊ={_U}	எ={_e}	ஒ={_o}	ஐ={_ai}
	ஃ={#}					ஒ={-e}		�ை={-ai}
க={ka}	கா={ka}{#}	கி={ki}	கீ={kI}	கு={ku}	கூ={kU}	கெ={-e}{ka}	கொ={-e}{ka}{#}	கை={-ai}{ka}
ஙa={Ga}								
ச={ca}	சா	சி	சீ	சு	சூ	செ	சொ	சை={-ai}{ca}
ஜ={Ja}								
ட={Ta}	டா	டி	ட	டு	டூ	டெ	டொ	டை
ண={Na}	ண={NA}	ணி	ணீ	ணு	ணூ	ணெ	ணொ	ணை={Nai}
த={ta}	தா	தி	தீ	து	தூ	தெ	தொ	தை
ந={na}	நா	நி	நீ	நு	நூ	நெ	நொ	நை
ப={pa}	பா	பி	ப	பு	பூ	பெ	பொ	பை
ம={ma}	மா	மி	மீ	மு	மூ	மெ	மொ	மை
ய={ya}	யா	யி	யீ	யு	யூ	யெ	யொ	யை
ர={#}	ரா	ரி	ரீ	ரு	ரூ	ரெ	ரொ	ரை
ல={la}	லா	லி	லீ	லு	லூ	லெ	லொ	லை
வ={va}	வா	வி	வீ	வு	வூ	வெ	வொ	வை
ழ={zha}	ழா	ழி	ழீ	ழு	ழூ	ழெ	ழொ	ழை
ள={La}	ளா	ளி	ளீ	ளு	ளூ	ளெ	ளொ	ளை={Lai}
ற={Ra}	ற={RA}	றி	றீ	று	றூ	றெ	றொ	றை
ன={n2a}	ன={n2A}	னி={n2i}	னீ={n2I}	னு={n2u}	னூ={n2U}	னெ={-e}{n2a}	னொ={-e}{n2A}	னை={n2ai}

Chart 5

Conclusion and openings

The present work is of course only a sketch, and can only take its full relevance in a computer implementation. The logical implications might be considered in two different directions, depending on whether one works on an already deciphered work or on a MS which has not yet been deciphered. In the latter case, the implications of the methodology advocated here would be to do separately the tasks of ascertaining the AR (Apparent reading) of the text and of proposing an IR (Intended Reading) for it. In the former case, the implications would be that one can take a text already available in a digital form (for instance one of the texts which is available on the file server of "Project Madurai") and remove the extra information which was added by the editors of the text, in order to make it conformant with today's writing systems. This could

[8] The 7-bit notation has not been completely explicited here, because the chart is probably easier to read like this, and the reader can easily fill in the blanks if needed.

have two useful benefits: (a) one could use it as a training tool for reading MSS; (b) one could in some cases be able to propose some new readings, which were not selected by the editor of the text, and which were masked by the anachronous presentation of the text with today's writing system. The appendix to this paper contains a sample Akam poem, presented according to this guideline.

Bibliography:

- U.V.S. *Kuṟuntokai* edition, குறுந்தொகை, உ.வே. சாமிநாதையர், அண்ணாமலை பல்கலைக்கழகம் [1983 reprint of 1937 edition]
- S.V.P.= S.Vaiyāpuri Piḷḷai, Caṅka Ilakkiyam, சங்க இலக்கியம், பாரி நிலையம், சென்னை [1967 reprint of 1940 edition]
- Akanāṉūṟu, 1981 [1957], Murray Rajam Edition, New Century Book House reprint.
- Akanāṉūṟu, 1983 [1970], Po. Vē. Cōmacuntaraṉār (Ed.), The South India Saiva Siddhanta Works.
- Naṉṉūl Viruttiyurai, Ca. Taṇṭapāṇi Tēcikar (Ed.), 1957, Tiruvāvaṭuturai Ātīṉam.
- Mahadevan, Iravatham, 2003, Early Tamil Epigraphy, Cre-A (Chennai, India) & Harvard University Press (Cambridge, MA, U.S.A.)
- Chevillard, Jean-Luc, 1996, Le Commentaire de Cēṉāvaraiyar sur le Collatikāram du Tolkāppiyam, French Institute of Pondicherry, Publication Number 84.1.

Appendix 1:

Idealized Facsimile, of a Tamil MS (AN390)

(1st version: TSCII compatible)

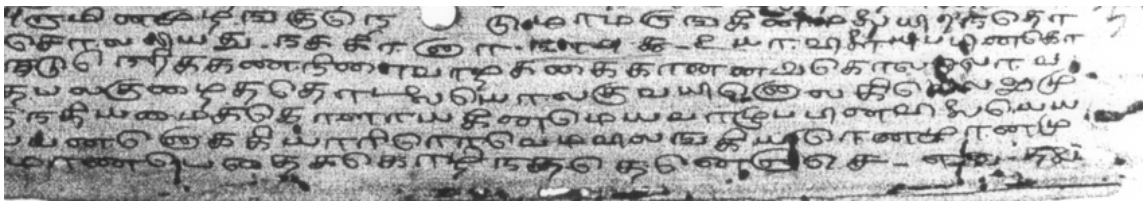
உவாவி{ளை}யுபபினகொள{ளை}சாறறி
யதாபடுபூழியசெடபுலமபடருந
ததாகொலுமணாபதிபொகுநெடுநெறிக
கணநினாவாழககைதானனறுகொலலொ
வணாசுரிமுசசிமுழுதுமனபுாள
வைதகலலகுலகவினபெறபபு{னை}நத
பலகுழைததொட{லை}யொலகுவயி{னா}லகி
நெலலுமுபபுமநெளியூரீர
கொளளீளொவெனசெரிதொறுமநுவலு
மவவாஙகுநதியமைததொளாயநின
மெயவாழுபபினவி{லை}யெயயாமென
சிறியவிலஙகினமாகபெரியதன
னரிவெயுணகணமாததனெ{ணா}ககி
யாரீளொவெமவிலஙகியீ இனென
மூானமுறுவலளபொவணணினற
சினனினாவாலவ{ளை}பபொலிநத
பனமாணபெதைககொழிநததெனனெஞுசெ

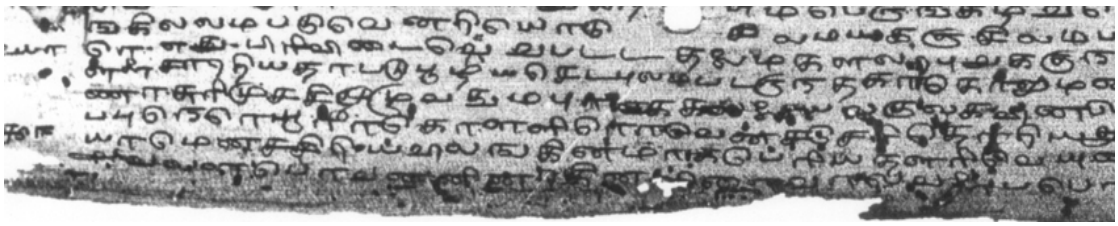
(2nd version: TSCII " half-compatible")

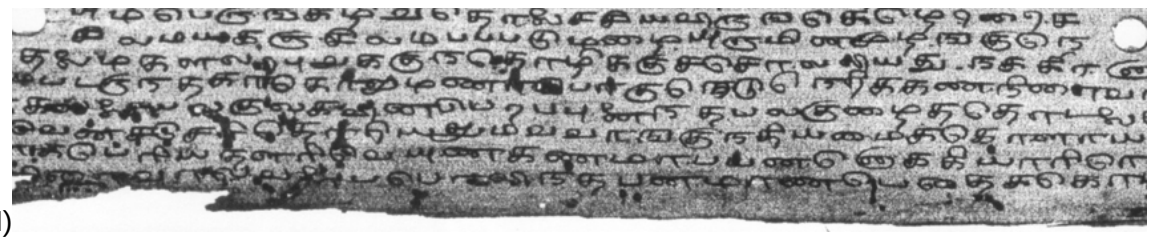
உவாவிகீயுபபினளொளகீசாறி
யதாபடுபூழியசெடபுலமபடருந
ததாளொதுமணாபதிபொகுநெடுநெறிக
கணநினாவாழகளைதானனஉகொலலொ
வணாசுரிமுசசிமுழுதுமனபுாள
வைதகலலகுலகவினபெறபபுனைநத
பலகுழைததொடயொலகுவயிளாலகி
நெலலுமுபபுமநெளியூரீர
கொளளீளொவெனசெரிதொறுமநுவலு
மவவாஙகுநதியமைததொளாயநின
மெயவாழுபபினவியெயயாமென
சிறியவிலஙகினமாகபெரியதன
னரிவெயுணகணமாததனெளகி
யாரீளொவெமவிலஙகியீ இனென
மூானமுறுவலளபொவணணினற
சினனினாவாலவகீபபொலிநத
பனமாணபெதைகளொழிநததெனனெஞுசெ

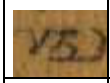
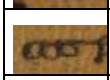
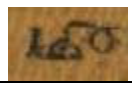
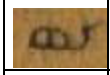
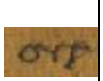
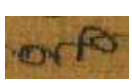
APPENDIX 2: Figures

Figure 1 (in 3 parts)

(R) 

(L) 

(M) 

										
	சா	கி	கீ	கு		செ	சொ	சை		
										
	சா	கி	கீ	கு		செ	சொ	சை		
										
	டா	டி		டு	டூ	டெ	டொ	டை		
	ணா	ணி	ணீ	ணு		ணெ	ணொ			
	தா	தி	தீ	து		தெ	தொ	தை		
	நா	நி	நீ	நு		நெ	நொ	நை		
	பா	பி		பு		பெ	பொ	பை		
	மா	மி		மு		மெ	மொ	மை		
	யா	யி	யீ	யு		யெ	யொ	யை		
	ரா	ரி	ரீ	ரு		ரெ	ரொ	ரை		
	லா	லி	லீ	லு	லூ	லெ	லொ			
	வா	வி	வீ	வு		வெ	வொ	வை		
	ழா	ழி	ழீ	ழு		ழெ	ழொ	ழை		
	ளா	ளி			ளூ	ளெ	ளொ			

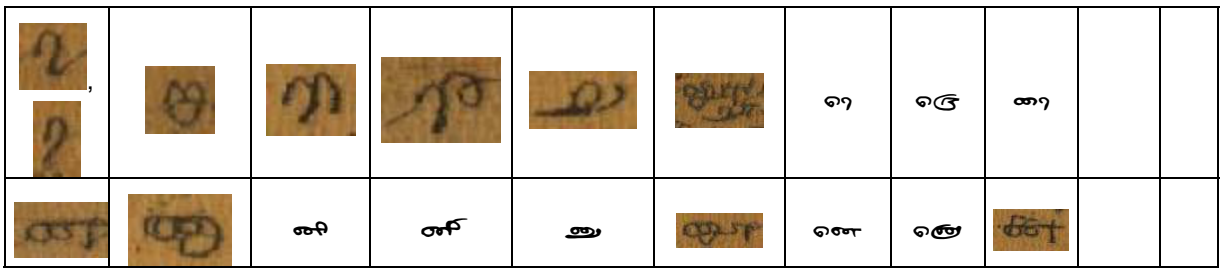


Figure 2