Indus writing is multilingual: a part-syllabic system at work

S. Srinivasan, J. V. M. Joseph and P. Harikumar*

A solution to unravel the mystery behind the Indus writing and the underlying language is reported here. The binding between two phonemes present in a language can be quantified by analysing the bigrams of characters that represent them. The application of information theory to this vexed problem of deciphering Indus script leads one to discern the type of script and it turns out to be abugida type. All the vowels and consonants present in the Indus text are elucidated. The analysis shows that the language underlying the Indus text is both Dravidian and Aryan in origin and encompasses more than one language. The evolution of Tamil, Kannada, Telugu, Prakrit and Devanagari scripts is traced. It brings to light that the Indus Civilization had reached great heights in literacy.

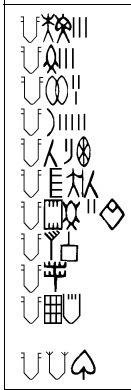
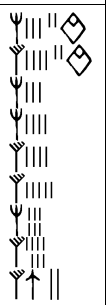
Keywords: Abugida, bigrams, Indus writing, information theory, vowels and consonants.

MANY attempts have been made since 1921 to decipher the Indus text based on rebus principle[1]. The first excavation carried out at Mohenjodaro[2] brought to light that an advanced urban civilization[3,4] (see Figure S1 in Supplementary material on-line) did exist in the Indian subcontinent during the bronze age 4300 years Before Present (ybp). An acceptable decipherment of Indus text is yet to be found. The reason attributed to the failure in reading the Indus text is twofold[5–11]. The first is the brevity of texts found in the inscriptions whose average length does not exceed five characters per seal[6]. Second, there is no bilingual passage available that relates the extinct script to that practised in ancient India. We believe that the Indus Valley Civilization did not die, but took a new avatar in the form of multiple Indian languages and scripts numbering about 15 in all[12]. The cause for the apparent collapse of the Indus Civilization could be natural disasters like disease, earthquake, flood, fire, tsunami, etc. However, the people do not appear to have vanished en masse, but exodus took place to much safer zones within the Indian subcontinent as if a beehive was disturbed and forced to relocate to different places. They appear to have possessed tacit knowledge of the languages spoken at that time.

Akin to the Roman alphabets that find widespread use in Europe, the people of the Indus Civilization might have used a common script to communicate and record the multiplicity of languages that were in vogue. This inference is drawn from the following fact. The Indus Civilization had spread over an area of 1.5 million sq. km. This is about twice the area of present-day South India, where principally four major Dravidian languages are spoken, namely Tamil, Kannada, Telugu and Malayalam. The two major metropolitan cities of the Indus Civilization, namely Mohenjodaro and Harappa (M&H), geographically separated by 640 km, might have employed a common script and there were more than 50 identical texts found from these two sites[6] (Table 1).

The brevity of the Indus text shall be explained on the following lines. We believe that the short text-bearing Indus seals[6–11] were primarily used for propagating literacy among people in the form of teaching aids like the nursery school flash-card material. Hence nearly all sign combinations appear in these seal inscriptions and one obtains a satisfactory statistical information from

Table 1. Indentical looking texts found from Mohenjodaro and Harappa (M&H)

1	2	3	4	5
				

S. Srinivasan lives at Apartment #43, First Avenue; and J. V. M. Joseph lives at Apartment #98, Fifth Avenue, Pudupattinam Colony, Kalpakkam 603 102, India; P. Harikumar lives at Plot No. 202, Fifth Street, Rajeswari Nagar, Kelambakkam 603 103, India.
*For correspondence. (e-mail: indussrini@gmail.com)

these short texts. Also, these artefacts were found to be strewn in front of the houses and street corners. The Indus people might have discarded these materials as useless, carrying with them only the valuables. So far, about 3000 seal inscriptions have been unearthed from the Mohenjodaro and Harappan sites. A systematic inventory of the Indus signs was brought out by Iravatham Mahadevan (IM) in the form of a compendium[6].

In Indian epigraphy and palaeography, there were two scripts prevalent in ancient India, namely Brahmi and Kharosthi[13–17]. Rock edicts of Emperor Asoka were engraved in Brahmi script[13]. Though the script of Asokan edicts was the earliest specimen of writing in India, scholars have suggested its development from a still older script that could possibly be from the Indus Valley script[18]. What we know with certainty about the Indus text is its direction of writing. It was written from right to left and this conclusion[19] was reached by C. J. Gadd, S. Smith, S. Langdon, J. Marshall and G. R. Hunter. The Indus text consists of about 419 unique signs (see Table S1 in Supplementary material on-line) or characters. The words found in the Indus text span from 1 to 14 signs on a given line. The phonemic values of these signs and the language used are not known. In this work, information theory has been used to shed light on the nature of the Indus script[20–22]. To discern the Indus script type to be logographic, syllabic, part-syllabic, alphabetic or purely random, we employed the information theoretic approach of computing the first- and second-order entropy values from the pairwise combination of signs listed in IM corpus on Indus text. We have also computed these values for other languages as well.

Indus script a part-syllabic system of writing

To establish a possible relationship between the languages in currency today and the Indus language, sample texts were collected from Tamil, Kannada, Sanskrit and English[23–35]. Literary works and dictionary materials available in these languages form the source material for information theoretic analysis. The language text in alphabetic form was converted to syllabic form and the text containing part-syllabic form was converted to fully syllabic and alphabetic forms. From these expanded and reduced character sets, pairwise distribution of letters was obtained for a given language. This forms the basic input to the entropy calculations[22]. Table 2 lists the results obtained from this analysis.

The first- and second-order entropy values give an insight into the effective number of signs or characters, namely N_1, N_2 used at the beginning and successive positions in the words[22,33]. If the part-syllabic present-day Tamil characters are fully converted to syllabic form, it would consist of 247 distinct signs. Similarly, Sanskrit and English characters would occupy 594 and 153 signs respectively. The computed values of N_1, N_2 obtained for fully syllabic Cangam Tamil text are 92.6 and 14.9 respectively. The corresponding values for the Indus text are 76.8 and 12.6 respectively. The Indus text comprises 419 signs. Of these, only 399 signs figure in the pairwise distribution. The remaining 20 appear in isolation and form the solo members of the Indus signs. The computed value for N_1 is 76.8 for the Indus script and this means that the beginning letter of a word can be guessed from any one of the 76 predominantly used signs. Likewise, the computed value for N_2 is 12.6 and this implies that after guessing the occurrence of the first sign, the guess for the second sign is limited to 12 signs. The second millennium Tamil–Brahmi script is fully syllabic and consists of 208 distinct signs[36,37]. *The values of N_1 and N_2 obtained for the Indus text lie between the values for alphabetic and syllabic systems of writing*. Hence it can be established that the Indus script is part-syllabic. This implies that the Indus text must contain medial-vowel signs that usually follow the consonants and conjunct consonants. In the case of Kannada script the medial-vowel signs never occur at the beginning of words[38]. They appear in words at the medial and terminal positions only.

The script used to write Tamil also belongs to the part-syllabic system[39]. There are five medial-vowel signs (ா, ி, ீ, ை,, ா) present in Tamil, which combine with the consonants and make it a part-syllabic system. Likewise, we could identify ten medial-vowel signs in the Indus text, viz. [', '', ', , , , , , ,] and they seldom occur at the beginning position in words (Table 3).

If the nature of the language spoken by the Indus people was predominantly Dravidian, it could easily be discerned by identifying the presence of short vowels for 'e' and 'o' in the Indus text. However, they are absent in Aryan languages. Evidence for the Dravidian nature of the Indus text is revealed by the occurrence of text lines that differ by one terminal sign, namely the comb symbol . Fifteen such pairs of lines were identified from the M&H sites (Table 4). The text lines that terminate with the jar symbol are understood to be the short medial-vowel sign 'e' and those that end with to be the long medial-vowel sign 'ē'.

A majority of Indian languages were endowed with a script of their own. For instance, Tamil, Kannada, Telugu, Malayalam, Sanskrit, Gujarati, Punjabi, Bengali and Oriya have their own scripts for writing letters and numerals[12]. Though there are many scripts practised today, they all share one common feature, namely the count of the vowel and consonant sign inventory does not exceed 16 and 36 respectively. All these scripts begin with 'a' and 'k' to be the first members of the vowel and consonant lists respectively. The Dravidian language inventory of characters includes two more vowels, namely the short 'e' and short 'o' and two additional consonants, namely 'l̲' (ழ, ഴ) and 'r̲' (ற, ര).

Table 2. Entropy values for Indus and other languages

Script/ language	T	V	C	J	C_v	J_v	M	N	H_0	$N_0 < N$	H_1	N_1	H_2	N_2
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Indus signs	?	Iravatham Mahadevan						417	8.33	321	6.26	**76.84**	3.65	**12.56**
	?	Bryan Wells						694	9.01	514	6.51	**90.88**	3.72	**13.14**
Chinese	L	Huang Xuanjing *et al.*						10,000	12.4	5403	9.62	785.8	6.09	68.03
	R	pi value: first 10,000 places						10	3.32	10	3.32	9.995	3.32	9.966
Tamil Cangam literature	A	12	18	0	0	0	1	31	4.95	31	4.42	21.35	3.09	8.51
	S	12	18	0	216	0	1	247	7.78	220	6.53	**92.62**	3.90	**14.94**
	RS	12	18	0	90	0	1	121	6.85	115	5.84	***57.10***	3.78	***13.72***
	ES	12	18	75	90	375	1	571	8.54	373	6.18	**72.48**	4.00	**16.01**
	ES	12	18	75	216	900	1	1,222	9.69	824	7.19	145.9	4.29	19.58
Kannada Sarvagna	A	14	34	0	0	0	0	48	5.55	47	4.59	24.02	3.03	8.18
	S	14	34	0	476	0	0	524	8.0	256	6.71	**105.0**	3.54	**11.60**
Sanskrit Story of Nala	A	16	34	0	0	0	0	50	5.09	34	4.60	24.23	3.11	8.63
	S	16	34	0	544	0	0	594	8.22	298	6.65	100.1	4.39	20.97
	RS	16	34	0	170	0	0	220	7.43	172	5.84	**57.38**	4.03	**16.37**
English – Hamlet by Shakespeare	A	6	21	0	0	0	0	27	4.75	27	4.29	19.59	2.96	7.76
	S	6	21	0	126	0	0	153	7.11	138	5.87	***58.35***	3.35	***10.16***
Tamil Lexicon	A	12	22	1	0	1	1	37	5.21	37	4.31	19.85	3.08	8.46
	S	12	22	1	264	13	1	313	8.06	266	6.47	88.97	4.67	25.43
Kannada dictionary	A	14	34	0	0	0	0	48	5.58	48	4.71	26.18	3.30	9.82
	S	14	34	0	476	0	0	524	8.45	349	6.99	127.4	4.58	23.90
MW-Sanskrit dictionary	A	16	34	0	0	0	0	50	5.64	50	4.68	25.67	3.34	10.10
	S	16	34	0	544	0	0	594	8.64	399	6.92	120.9	5.17	36.07
English wordlist CornCob	A	5	21	0	0	0	0	26	4.70	26	4.16	17.88	3.49	11.26
	S	5	21	0	105	0	0	131	6.99	127	6.03	65.21	4.49	22.39

V, Vowel; C, Consonant; J, Conjunct consonant; C_V, Consonant–vowel; J_V, Conjunct–vowel; M; Mute; $N = V + C + J + C_V + J_V + M$; N, Distinct number of signs used in the corpus; R, Random; T, Type of script; L, Logographic; A, Alphabetic; S, Syllabic; RS, Reduced syllabic; ES, Extended syllabic; H_0, Zeroth-order entropy per character; N_0, Distinct number of signs/characters appearing in bigram; H_1, First-order entropy per character; N_1, Expectation value of signs/characters in the first order; H_2, Second-order entropy per character; N_2, Expectation value of signs/characters in the second order; MW, Monier Williams; CornCob, A word list available in www.mieliestronk.com/wordlist.html

Table 3. Positional and the frequency distribution of medial-vowel signs in the Indus text

Sign#	Sign	I	M	F	TM	KN	DV	RM
097		1c	87	3				
099		2+	625	22	ಾ		ಿ	
123		0	186	7		ಎ		ā
176		0	38	316		ೇ	ೆ	ī
211		0	42	184	ು	ು	ು	u
342		1c	420	971	ಒ	ೆ		e
347		1c	116	1	ಱ	ಖ	ೂ	ū
358		0	32	0*	ಐ	ೈ	ೈ	ai
374		0	0	9s		ಂ	ಂ	ṁ
321		0	1	12		ಃ	ಃ	ḥ

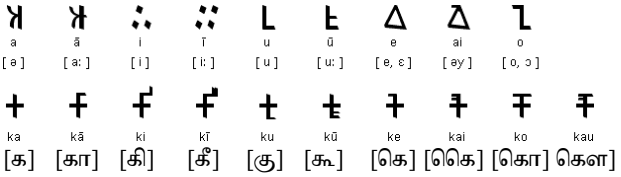
Sign#, Sign no. as listed in Mahadevan's corpus.
I, Initial; M, Medial; F, Final; TM, Tamil; KN, Kannada; DV, Devanagari; RM, Roman; c, Continuation from the previous line; +, Likely to be initial_vowel sign (sign 100); *Appears always with to the left; s, Listed to be solus, but appears in the final position.

Table 4. Identical looking Indus words that differ by short and long medial-vowel sign 'e'

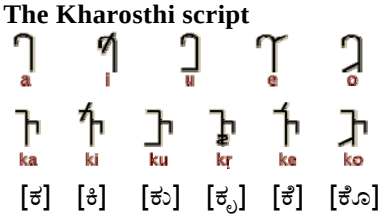
Site name	Text no.	Text line	Site name	Text no.	Text line
H	4497		M	2848	
M	2380		M	2444	
M	3238		M	3502	
M	3512		M	1620	
M	2499		H	5282	
M	2502		H	4487	
M	1226		H	4318	
M	1040		H	5276	
H	5286		H	4646	
H	4650		H	4161	
H	4094		H	4563	
H	5280		H	4338	
H	4074		H	4679	
H	5266		H	5306	
H	5470		H	5308	

The ancient rock inscriptions[13] found in India were written using Brahmi and Kharosthi scripts. They date back to 2500 ybp and belong to the syllabic system of writing. The texts found in Kharosthi script were written from right to left and the Brahmi script from left to right. The genesis of these scripts is not known. It is believed that all modern Indian language scripts have evolved from the Brahmi script. The consonant–vowels in Brahmi and Kharosthi scripts appear in conflated form.

The Brahmi script



The Kharosthi script



Early Tamil inscriptions that date back to 2200 ybp were written in southern Brahmi. Tamil emerges as a part-syllabic system of writing[37] about 1400 ybp. It looks easier to trace the genesis of Tamil, Kannada and Devanagari scripts directly from the Indus script rather than deriving it from the Brahmi route[40]. The Indus script seems to be the panacea for understanding the varied forms of medial-vowel signs and conjunct consonants present in modern Indian language scripts. We were able to identify from the Indus text the presence of scripts for five Indian languages, namely Tamil, Kannada, Telugu, Prakrit and Sanskrit.

Tamil

The ancient Tamil grammar treatise, *Tolkāppiyam*[41,42], describes the number of vowels and consonants present in Tamil. The two aphoristic propositions of interest have been culled out and are listed below.

1. *The twelve phonemes through /au/ are known to be vowels, so has it been laid down (verse-8 of 1602).*
2. *The eighteen phonemes through /n/ are called consonants, thus goes the usage codified (verse-9 of 1602).*

The initial-vowel sequence for Tamil is listed below:

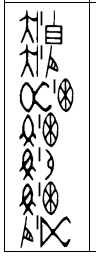
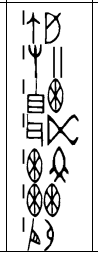
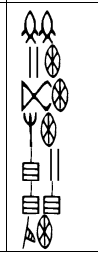
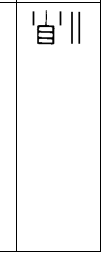
Roman	a	ā	i	ī	u	ū
Tamil	அ	ஆ	இ	ஈ	உ	ஊள
Indus	𑀁	𑀂	𑀃	𑀄	𑀅	𑀆
Roman	e	ē	ai	o	ō	au
Tamil	எ	ஏ	ஐ	ஒ	ஓ	ஔள
Indus	𑀇	𑀈	𑀉	𑀊	𑀋	𑀌

These vowels appear more frequently at the beginning of words and seldom occur at the terminal position. The 'rain' symbol 𑀆 stands for the last member (12th in the list) of the initial-vowels and denotes the phoneme 'au'. The initial-vowel that appears to the left of medial-vowel signs also acts as a word-divider in the Indus text. Tamil employs the special diacritic mark, 'dot' to depict basic consonants in final positions in the words. It helps to do away with ligaturing of two or more consonants and is also used to differentiate between the short and long vowels of 'e' and 'o'. The occurrence frequency of characters obtained from the Tamil corpus reveals that the medial-vowel signs {ெ, ே, ை} and ர are used most[39]. Likewise, the two most used signs in the Indus text are 𑀁 and 𑀂. Tamil-like feature deducible from the Indus text is given below.

Primary medial-vowel sequence					
c-a	c-ā	c-i	c-ī	c-u	c-ū
க	கா	கி	கீ	கு, ஜு	மூ,தூ,ஜூ
□	□𑀂	□𑀄	□𑀄	□𑀅,□𑀆	□𑀅,□𑀆,□𑀆

Secondary medial-vowel sequence					
c-e	c-ē	c-ai	c-o	c-ō	c-au
கெ	கே	கை	கொ	கோ	கௌ
□𑀁	□𑀁𑀂	□𑀁𑀃	□𑀁𑀅	□𑀁𑀅𑀂	□𑀁𑀅𑀆

The symbol □ denotes a consonant invested with the inherent medial-vowel sign /a/. Table 5 gives the complete list of primary and secondary medial-vowel sequences. These medial-vowel sequences do not employ the sign 𑀍. Instead the short single vertical stroke sign 𑀇 is used. The medial-vowel notation for Tamil in the Indus text resembles that of the Bhattiprolu inscriptions (2100 ybp)[36]. There is a separate marker for the medial vowels /a/ and /ā/ in this system. Examples for the presence of basic consonant and the medial-vowel /a/ consonant that appear in the Indus text are listed below.

c_ca	ca_c	ca_ca	c_c
			

Examples for words beginning with a given consonant and followed by other medial-vowel signs for a, ā, i that appear in the Indus text are listed below.

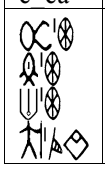
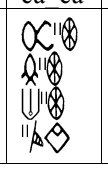
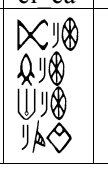
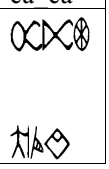
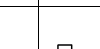
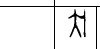
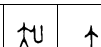
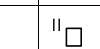
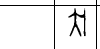
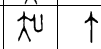
c_ca	cā_ca	ci_ca	ca_ca
			

Table 5. Tamil-like consonant–vowel sequences found in the Indus text

				□															
c-a	□	□(+)	□																
c-ā	□ா	□ா	□																
c-i	□ி	□ி	□																
c-ī	□ீ	□ீ	□																
c-u	□ு		□																
	embedded-u பு, து, மு		□																
c-ū	□ூ		□																
			□																
	embedded-ū பூ, தூ, மூ		□																
			□																
c-e	ெ□	ே□	□																
c-ē	ை□	ெ□	□																
c-ai	ை□	ைெ□	□																
			□																	
c-o	ெ□ா	ே□ா	□																
c-ō	ொ□ா	ெ□ா	□																
			□																
c-au	ெ□ள	ெ□ൗ	□																
			□																
c	◌̇	◌̇	□																

c, Consonant; a, i, u, e, o, Short medial-vowel signs; +, Prior to script reform by C. J. Beschi (year 1730); ; ; ; ; □, Consonant invested with the inherent medial-vowel /a/.

Bryan Wells[9], in his corpus does not differentiate between the short single vertical stroke signs that were top and middle aligned (,) in the text. The same holds good for the two short vertical stroke signs (,) as well. We feel that these signs need to be differentiated in order to depict words that have identical meaning and are written with and without the word beginning with consonant as they appear in Tamil. Examples for such occurrences are: camai=amai (to construct), capai=cavai=avai (an assembly, a court), cāl=āl (a banyan tree), cāṉṟōr=āṉṟōr (a noble person), yāṉai=āṉai (elephant), yāmai=āmai (tortoise), yāṟu=āṟu (river), yāy=āy (mother), yār=ār (who), yāṇṭu=āṇṭu (year). Here a and ā can be equated to the Indus signs and , cā to be and yā to be . The signs and stand for the medial-vowels and the signs and stand for initial-vowels. Also, there is a likelihood that the initial-vowels and shall follow the medial-vowels and in word conjunctions. The bigram appears 15 times in the IM corpus. The possibility of such an occurrence is mentioned in verse 226 of *Tolkāppiyam* that reads as

There occurs /a/ distinct
After the word-final /ā/ preceded by a short vowel,
And after the one-letter [/ā/] ending word
[Before the succeeding word].

For example, [Tamil] palā + kōṭu → palāakkōṭu;
kā +kuṟai → kāakkuṟai and [Indus] … ; … ; … ; … .

The bigrams  and appear to be one-letter word and the signs and stand for the two basic consonants.

The consonants conflated with medial vowel sign /u/

Like the presence of medial-vowel sign /ā/ (or ⊓), there is no separate medial-vowel sign /u/ found in Tamil. This kind of feature is also noticed in the Indus text. We have identified the following signs to represent the conflated form of consonants with the medial vowel sign /u/. They are:  . The medial-vowel sign does not follow the above conflated signs. However, medial-vowel signs such as , , , , do appear after the above conflated signs. The bigrams such as and stand for the consonant conflated with medial-vowel signs /o/ and /ū/ respectively.

Kannada

The initial-vowel sequence for Kannada is given below.

Roman	a	ā	i	ī	u	ū	ṛ	e	ē
Kannada	ಅ	ಆ	ಇ	ಈ	ಉ	ಊ	ಋ	ಎ	ಏ
Indus									

Roman	ai	o	ō	au	aṃ	ḥ
Kannada	ಐ	ಒ	ಓ	ಔ	ಅಂ	ಅಃ
Indus						

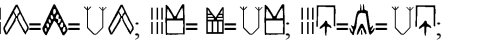
Kannada-like features deducible from the Indus text are given below.

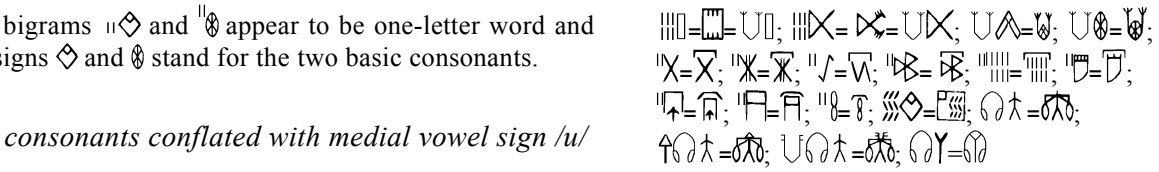
Primary medial-vowel sequence					
c-a	c-ā	c-i	c-ī	c-u	c-ū
□	□	□	□	□	□
□	□	□	□	□	□

Secondary medial-vowel sequence					
c-e	c-ē	c-ai	c-o	c-ō	c-au
□	□	□	□	□	□
□	□	□	□	□	□

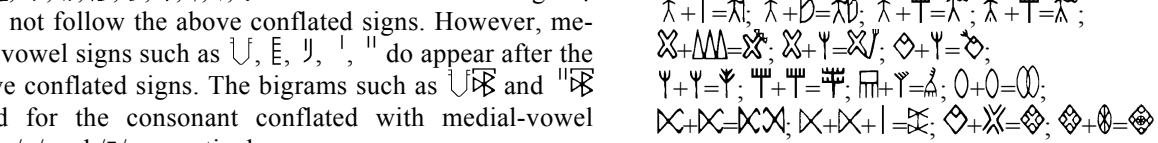
These medial-vowel sequences do employ the sign , but not the short single vertical stroke sign . A full list of primary and secondary medial-vowel sequences is given in Table 6.

Some instances of conflation of initial-vowel sign with consonants are listed below.





We identify the signs  , to be Kannada letters o(ಒ) and ṅ(ಙ). Likewise, the signs , are the Kannada letters i(ಿ) and ñ(ಞ) respectively. Some instances of conflation of consonants that form conjunct consonants are listed below.



The medial-vowel signs do follow these conjunct consonant signs indicating that they are made up of more than one consonant. The number of unique signs that precedes the medial-vowel sign in the Indus text can also be accounted based on grouping the medial-vowel sequence described. The number of conflated consonants that precedes the medial-vowel sign also adds to this tally (Table 7).

The grouping for medial-vowel signs , and for the Indus and Kannada scripts is shown below.

The basic words in Kannada indicating parts of the human body and numeral names do end with medial-vowel 'e' and 'u' in addition to the medial-vowel 'a'. A few of them are listed below.

Words ending in 'e'

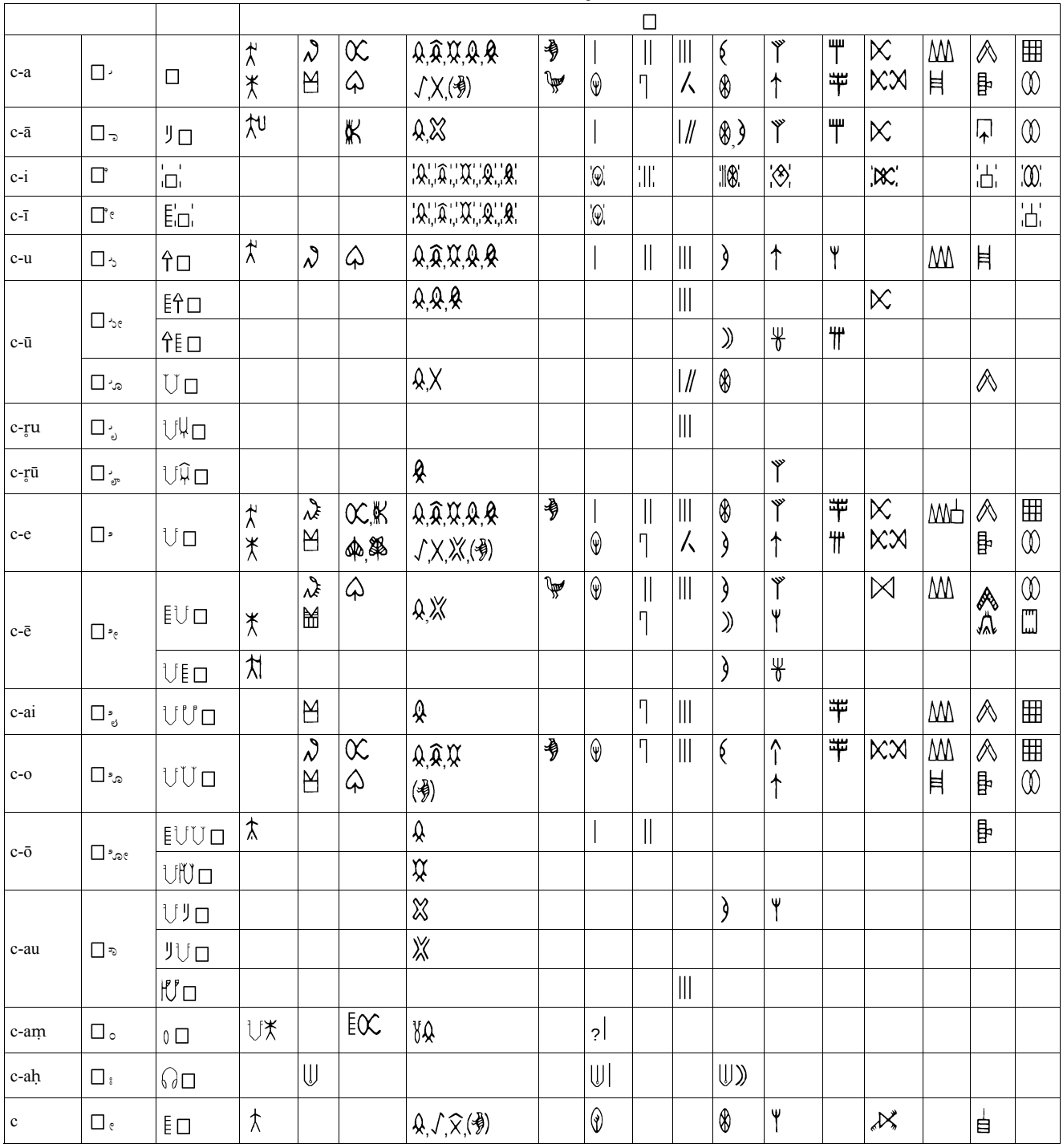
Kuttige (neck), ede (heart), nālige (tongue), hoṭṭe (stomach), haṇe (fore-head), ondūvare (one and half).

Words ending in 'u'

Mūgu (nose), kaṇṇu (eye), hallu (tooth), kālu (leg), bennu (back), beraḷu (finger), hubbu (eye-brow).

Ondu (one), eraḍu (two), mūru (three), nālku (four), aidu (five), āru (six), ēlu (seven), eṇṭu (eight), ombattu (nine), hattu (ten).

Table 6. Kannada-like consonant–vowel sequences found in the Indus text



c, Consonant; a, i, u, e, o, Short medial-vowel signs; □, Consonant invested with the inherent medial-vowel /a/.

The words listed in columns 1 and 2 of Table 1 that end with the sign 𖵰 represent words ending in 'e', and the words listed in column 4 of Table 1 that end with the sign 𖵬 represent words ending in 'u' from the Indus text.

Telugu

There is a lot of similarity between Kannada and Telugu scripts[43,44]. The medial-vowel signs resemble each other in many respects. However, these languages differ from

Table 7. *n*-Distinct occurrence of consonant–vowel combinations from the IM corpus

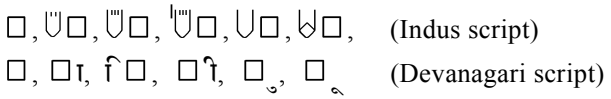
Indus	Roman	Actual value (*n*)		Anticipated maximum (*n*)
c_v	c_v	Indus	Kannada	Kannada
	□ā	**32**	**43**	2*36 = 72
	□ī	**63**	**80**	4*36 = 144
	□u	22	45	2*36 = 72
	□ū	6	19	1*36 = 36
	□e	**135**	**131**	6*36 = 216
	□ē	31	54	2*36 = 72
	□ū	37	62	3*36 = 108
	□o	31	43	2*36 = 72
	□ai	12	20	1*36 = 36

the semantics point of view. Many nouns in Telugu end with the letter 'ka'. Words of this nature are: ciluka (a parrot), jinka (an antelope), kukka (a dog), mēka (a goat), nakka (a jackal) and piccuka (a sparrow). This characteristic feature differentiates Telugu words from Tamil and Kannada words. The word 'eluka' in Telugu meaning 'rat' corresponds to 'eli' in Tamil and 'ili' in Kannada. The Indus text abounds with such words that end with the sign denoting the letter 'ka'. We could identify the likely consonants for Telugu from the Indus strings that terminate with trigrams of the type  . Table S2 (see Supplementary material online) lists all the primary and secondary medial-vowel sequences for Telugu.

Prakrit and Sanskrit

Prakrit-like features deducible from the Indus text are given below.

Primary medial-vowel sequence:



(Indus script)

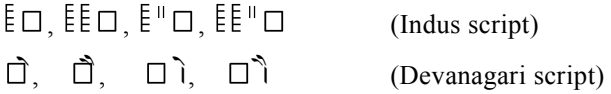
□, □ɪ, □ि, □ी, □ु, □ू (Devanagari script)

Secondary medial-vowel sequence:

(Indus script)

□ै, □ौ, □ो, □ी (Devanagari script)

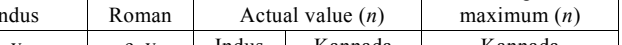
Doubling of medial-vowel sign े (Indus sign) is not a feature of the Kannada script. However, the presence of can be construed to represent the diphthongs 'ai' and 'au' present in the Devanagari script.



(Indus script)

□ै, □ौ, □ो, □ी (Devanagari script)

Much refined and alternate form for these four medial-vowels is given below.

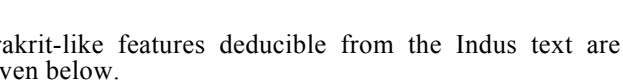


(Indus script)

A full list of primary and secondary medial-vowel sequences for Prakrit is given in Table S3 (see Supplementary material online).

An alternative form to indicate the presence of long medial-vowels 'ē'() and 'ō'() and diphthong 'ai'() exists in the Indus text. Their forms are:  ,  and  . In Mohenjodaro text and appear in the ratio 4 : 1, and  and  in the ratio 11 : 1. Likewise for the Harappan text,  and  appear in the ratio 1 : 4 and  and  in equal measure. In the Mohenjodaro text,  and  appear in equal measure and for Harappan text in the ratio 3 : 1. The above statistics leads one to believe that the Harappan people tried to simplify their script by gradually eliminating the usage of additional signs like  , and  . This feature observed in script reformation indicates the maturity level attained by the Harappan people in literacy.

Also the initial-vowels a, u, ū, ā, au, i, ī appear as follows:



A full list of primary and secondary medial-vowel sequences for Sanskrit is given in Table S4 (see Supplementary material online).

Brahmi

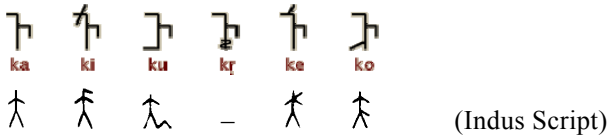
In Brahmi script[12,36], the primary medial-vowel sign conflates with the consonant at the right side and the secondary medial-vowel sign conflates to the left side of the consonant. This feature is reflected in Prakrit-like consonant–vowel sequence obtained from the Indus text, for example:

Roman	r-ē	ś-ē	t-ē	ṭh-ai	th-ē	m-ē	b-ē
Brahmi							
Indus							
Roman	r-ā	ś-ā	t-ā	ṭh-ā	th-ā	m-ā	b-ā
Brahmi							
Indus							

This feature of medial-vowel sign appearing on either side of the consonant sign suggests that the Brahmi script could have evolved from the archaic script practised for writing Prakrit language. The absence of bigrams , , , and presence of bigrams , , , , in the Indus text indicate that Prakrit and Kannada belong to different language families.

Kharosthi

The genesis of the fully syllabic nature of the Kharosthi script can be traced from the Indus script.



(Indus Script)

The Indus sign √ bears semblance with the nasal consonant 'ṇ' the Kharosthi script. It does not appear at the initial position in words and selectively combines with the sign 𐩨 to form the bigram √𐩨 in the Indus text. We assign the phonemic value 'r' to the Indus sign 𐩨.

Aryan and Dravidian presence in M&H

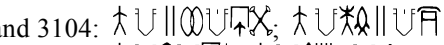
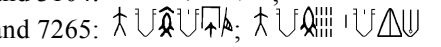
Having discerned the existence of four radically different scripts for archaic Tamil (aTamil), archaic Kannada (aKannada), archaic Prakrit (aPrakrit) and archaic Sanskrit (aSanskrit) languages from the Indus text, it is possible to obtain demographic details on the metropolitan cities of M&H. The secondary medial-vowel modifiers ˡ,𖭁 and 𖭁𖭁 play a crucial role in aTamil. Likewise, the secondary medial-vowel sign 𖭂 plays a vital role in aKannada. There is no secondary medial-vowel sign akin to ˡ or 𖭂 in aPrakrit. Simultaneous occurrence of signs such as {ˡ and 𖭂}, {ˡ and 𖭁}, {ˡ and 𖭃}, {𖭂 and 𖭁}, {𖭂 and 𖭃} within the same seal is seldom found. The absence of bigrams 𖭂𖭁, 𖭂𖭃 in the Indus text clearly indicates that the signs 𖭁, 𖭁, 𖭃, 𖭃 do represent the long vowels. They act both as initial and medial-vowels. The Prakrit language contains the short /e/ and short /o/ medial vowels. They appear as the bigrams ˡ𖭁 and ˡ𖭃 in the Indus text. The occurrence of the doublet 𖭂𖭂 denoting the phoneme /ai/ is a feature of aSanskrit. Akin to the presence of secondary medial-vowel ˡ in aTamil, there is a need for the presence of the medial-vowel ˡˡ in aSanskrit to denote the phoneme /au/. The trigram 𖭂𖭂ˡˡ stands for the medial-vowel /au/ in aSanskrit.

The lines of text gathered from M&H sites alone constitute nearly 90% of the volume of Indus writing. We have culled out three independent datasets from the Indus text that would reflect aTamil, aKannada and aPrakrit features. The Indus seals having the medial-vowel sign ˡ and initial-vowel sign ˡ form the aTamil dataset. It consists of 100 seals from M, 23 seals from H and 20 seals from other sites (O). Likewise, the seals having the medial-vowel sign 𖭂 form the aKannada dataset. It consists of 77 seals from M, 234 seals from H and 15 seals from O. Similarly, seals having the vowel signs 𖭁, 𖭁, 𖭃, 𖭃 form the aPrakrit dataset. It consists of 152 seals from M, 56 seals from H and 42 seals from O. An ensemble of these three datasets was also generated and this forms a micro-corpus of the Indus text. The size of this micro-corpus was found to be 25% that of the macro-corpus. It consists of 338 texts from M, 314 from H and 77 from O. These numbers are nearly proportional to the number of objects unearthed from those sites. They are 1540, 985 and 381 respectively. A closer examination of these three datasets reveals demographic information pertaining to M&H cities. The lines of text obtained for languages aTamil, aKannada and aPrakrit scale in the ratio 4 : 3 : 6 for Mohenjodaro and 1 : 10 : 2 for Harappa. The occurrence of seals pertaining to a given language can be construed to represent the number of speakers of that language as well. Grouping aTamil and aKannada as a Dravidian language family and aPrakrit as the Aryan language family, one gets the statistics pertaining to Dravidian (dark-skinned) and Aryan (light-skinned) speakers. It turns out that 52% of the inhabitants from Mohenjodaro were Dravidians and 48% Aryans. Likewise, for Harappa the Dravidian population was 82% and the Aryan population 18%. The presence of antique Dravidians in all the Indus settlements was 67% of the total population and the antique Aryans presence was 33% of the total population. This observation corroborates with the estimate obtained from a genetics-based study on the Indian population[45,46]. They show that the ancestral North Indian (Aryan) component ranges between 39% and 71% in most Indian groups.

The purity of Tamil language is vouchsafed by the presence of the medial-vowel sign ˡ, the presence of medial-vowel doublet sign 𖭁𖭁 and the absence of medial-vowel sign 𖭂 and that of Sanskrit by the presence of the medial-vowel sign 𖭂, the presence of medial-vowel doublet sign 𖭂𖭂 and the absence of the medial-vowel sign 𖭁. An intriguing feature found in these languages was brought out by Thirumala Ramachandra of Prakrit Academy, Hyderabad[47]. The 'Dative case' marker takes the form 'ku' in Tamil, 'ge' in Kannada, 'k-ē' in Prakrit and 'ki' in Telugu. In the Indus text, these correspond to the signs 𐩱𖭁, 𐩱𐩨, 𐩱𐩧𐩱 and 𐩨 respectively. These signs occur at the terminal position in Indus texts with a frequency 1, 65, 1 and 1 respectively. Also, the inferences drawn by Ramachandra are the following: Sanskrit is based on a pure Aryan tongue and Tamil on a pure Dravidian tongue. The admixture of these two tongues led to the formation of Prakrit. This presumption is corroborated by the occurrence of sign 𖭁, a Dravidian characteristic symbol that appears in 63% of lines in aPrakrit set. According to Chilukūri Narayana Rao, the Telugu language evolved from Prakrit[47]. There is some veracity in his statement by gauging from the findings obtained from genomic studies. Sengupta *et al.*[48] made the following observation: 'Only the L1 Subclade of haplogroup L occur among Indians and even Subclade L1 is only found at high frequency among certain castes that speak languages of Southern Dravidian group such as Tamil, Kannada and Malayalam. It is a big mystery why negligible amount of

L1 present among castes of Telugu, the biggest Dravidian Linguistic group in India.'

The following Indus text that reflects Telugu-type words which terminate with bigram of the type  and also includes the sign  is suggestive of the view of Chilukūri Narayana Rao that Telugu language was an offshoot from Prakrit. They are:

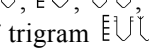
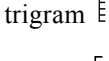
Seal nos 1536 and 3104: 
Seal nos 5086 and 7265: 

Indus writing Dravidian in origin

Radiocarbon dating[11] of the artefacts found from Indus Valley fixes the period of existence of this urban society to be 4500–3700 ybp. The findings of Pitchappan[49] for the Genographic project, about the migratory history of the human species in India, by analysing DNA from 10,000 males are astounding (see Figure S2 in Supplementary material on-line). The pattern of gene flow obtained from various linguistic groups reveals that the ancestors to the present-day Dravidian language speaking people might have inhabited the Indian subcontinent 10,000 ybp. It is apt here to quote from the writings of Sri Aurobindo on Tamil and Sanskrit[50]: '*I found myself continually guided by words or by families of words supposed to be pure Tamil in establishing new relations between Sanskrit and its distant sister, Latin, and occasionally, between Greek and Sanskrit. Sometimes the Tamil vocable not only suggested the connection, but proved the missing link in a family of connected words. And it was through this Dravidian language that I came first to perceive what seems to me now the true law, origins and, as it were, the embryology of the Aryan tongues*.'

The words 'rice' and 'paddy' found in English may have etymological leanings from Tamil. The word 'rice' could have originated from the Tamil word 'arici' probably also for 'oryza' in Latin and the word 'paddy' that sounds like 'paḍi' refers to the receptacle with which paddy is measured.

Some important observations

1. Like the presence of bigrams  , the bigrams  do not occur.
2. Like the presence of bigrams  the bigrams do not occur.
3. Like the presence of bigrams  , the bigrams  do not occur.
4. Like the presence of trigram  , the trigram  does not occur.
5. Like the presence of trigram , the bigram  could also occur.
6. Like the presence of trigram  , the bigram could also occur.
7. The presence of bigram  is indicative of Aryan and the presence of bigram  is indicative of Dravidian families.

Entropy computation

The basic ingredient to this information theoretic approach is the frequency of the pairwise combination $L(I, J)$ of all the Indus signs. The first- and second-order entropy values are computed based on the following formula[22].

$$N_1 = 2^{h1}; \quad h1 = \sum_{I=1}^{N_0} P(I)B(I); \quad B(I) = \log_2[1/P(I)];$$

$$P(I) = \sum_{J=1}^{N_0} L(I,J) \Big/ \sum_{I=1}^{N_0}\sum_{J=1}^{N_0} L(I,J),$$

$$N_2 = 2^{h2}; \quad h2 = \sum_{I=1}^{N_0} P(I) \sum_{J=1}^{N_0} P(I,J)B(I,J);$$

$$B(I,J) = \log_2[1/P(I,J)]; \quad P(I,J) = L(I,J) / \sum_{J=1}^{N_0} L(I,J),$$

where $L(I, J)$ is the bigram frequency, N_0, the distinct number of signs/characters appearing in the bigram; $h1$, the first-order entropy per character; N_1, the expectation value of signs/characters in the first order; $h2$, the second-order entropy per character and N_2, the expectation value of signs/characters in the second order.

Conclusion

The Indus seals were largely recovered from materials strewn in front of the houses and over the streets. These are mostly inscriptions and might have been produced on a large scale to serve as learning material for the beginners and for disseminating the script to the masses. Even though the count of the Indus seals is about 3000 only and the inscribed text in the seals was brief, the statistics obtained from them is not sparse, but good enough to represent a continuous text or a long passage. The decipherment of Indus text based on pairwise distribution of Indus signs and information theory reveals that the Indus script could be classified as a part-syllabic system of writing. In present-day India, if one travels from south to north or east to west about 1000 km or less, the spoken language and the script differ. Likewise, in the erstwhile Indus region that spanned over a million square kilometres in area, the people might have spoken more than one language. Hence it is prudent to arrive at the conclusion that the Indus text is multilingual based on the structural semblance of characters present in Tamil, Kannada, Devanagari and Indus scripts.

1. Fairservis, W. A., The script of the Indus Valley Civilization. *Sci. Am.*, 1983, vol. 248, No. 3, 58–66.
2. Rajagopalan, R., *The Secrets of Indus Valley*, Children's Book Trust, New Delhi, 1992.
3. http://www.archaeologyonline.net/artifacts/photo-gallery.html
4. Lawler, A., Unmasking the Indus. *Science*, 2008, **320**, 1276–1285.
5. Lawler, A., The Indus script – write or wrong? *Science*, 2004, **306**, 2026–2029.
6. Mahadevan, I., *The Indus Script – Texts, Concordance and Tables*, Memoirs of the Archaeological Survey of India, New Delhi, 1977, No. 77.
7. Parpola, A., *Deciphering the Indus Script*, Cambridge University Press, Cambridge, 1994.
8. Suryanarayana, K., *The Origin of Human Speech, Writing and Religion*, Vavilla Ramaswamy Sastrulu & Sons Press, Chennai, 1955, pp. 3–4.
9. Wells, B., *Epigraphic Approaches to Indus Writing*, Institute of Mathematical Sciences, Chennai, 2009.
10. http://caddy.bv.tu-berlin.de/indus/welcome.htm
11. Rao, S. R., The Indus script – methodology and language. In *Radiocarbon and Indian Archaeology* (eds Agrawal, D. P. and Ghosh, A.), Tata Institute of Fundamental Research, Mumbai, 1973, pp. 323–340.
12. Kannaiyan, V., Scripts – in and around India. First edition 1960, Latest Reprint 2000, Publications of the Government Museum, Chennai; http://www.chennaimuseum.org/draft/publn/publn.htm, 1960.
13. Daniels, P. T. and Bright, W., *The World's Writing Systems*, Oxford University Press, New York, 1996.
14. http://www.omniglot.com/writing/brahmi.htm
15. http://www.omniglot.com/writing/kharosthi.htm
16. http://www.ancientscripts.com/brahmi.html
17. http://www.ancientscripts.com/kharosthi.html
18. Subramanian, N., *The Tamils*, Institute of Asian Studies, Chennai, 1996, pp. 21–23.
19. Ross Alan, S. C., *The 'Numerical-Signs' of the Mohenjo-Daro Script*, Memoirs of the Archaeological Survey of India, New Delhi, 1938, No. 57, p. 10.
20. Young, J. F., *Information Theory*, Butterworth & Co., London, 1971, pp. 39–58.
21. Kondratov, A., Languages and codes, No. 2, *Breakthrough*, Journal on Science & Society, Calcutta, India, 1996, vol. 7, No. 2, pp. 23–30.
22. Bennet Jr, W. R., *Science and Engineering Problem-Solving with Computers*, Prentice Hall, New Jersey, 1976, pp. 132–146.
23. Lehmann, T. and Malten, T., *A Word Index for Cankam Literature*, Institute of Asian Studies, Chennai, 1993.
24. Vaiyapuri Pillai, S. (ed.), *Tamil Lexicon*, University of Madras, Macmillan India Press, Chennai, 1982.
25. http://kannadakasturi.com
26. http://www.vicharamantapa.net/vachana/sarvajna/sarvajna2.html
27. http://www.sanskritweb.net/sansdocs/nala-i.itx
28. http://www.sanskrit-lexicon.uni-koeln.de
29. http://sanskritdocuments.org/learning_tutorial_wikner/P058.html
30. http://www.shakespeare-online.com/plays/hamletscenes.html
31. http://www.mieliestronk.com/wordlist.html
32. http://starman.vertcomp.com/math/pi/PI.100.000.TXT
33. Tsou, B. K., Lai, T. B. Y. and Chow, Ka. Po., Comparing entropies within Chinese Language. *Lect. Notes Comput. Sci.*, 2005, **3248**, 466–475.
34. Payani, *Chinese Language – An Introduction* (ed. Vaidehi, S.), Sriperumbudur, 2002.
35. Srinivasan, S., Tamil and Chinese from the perspective of classical languages. *J. Tamil Stud.*, 2010, **78**, 29–40.
36. Mahadevan, I., *Early Tamil Epigraphy*, Harvard Oriental Series, Cre-A Publisher, Chennai, 2003, vol. 62, pp. 173–178.
37. Govindaraj, R., *Evolution of Script in Tamil Nadu*, Tamil Nadu Archaeological Society Special Issue No. 1, 1994, pp. 16–22.
38. Ramanandh, K. S. and Srinivasa Sarma, P., *Learn Kannada in 30 Days*, Balaji Publications, Chennai, 1970.
39. Srinivasan, S., The study of structure – property relationships of Tamil: an information theory approach. Ph D thesis, Tamil University, Thanjavur, 2000.
40. Venkatachalam, K., *Indus Civilization and Tamil Language* (eds Sridhar, T. S. and Marxia Gandhi, N.), Department of Archaeology, Government of Tamil Nadu, 2009, pp. 134–143.
41. Murugan, V., *Tolkappiyam in English*, Institute of Asian Studies, Chennai, 2001.
42. Meenakshi, K., *Tolkappiyam and Astadhyayi*, International Institute of Tamil Studies, Chennai, 1997.
43. Srinivasa Chari, K., *Learn Telugu in 30 Days*, Balaji Publications, Chennai, 1984.
44. Manickam, T. S., *Tamil and Telugu*, International Institute of Tamil Studies, Chennai, 1994.
45. Reich, D. *et al*., Reconstructing Indian population history. *Nature*, 2009, **461**, 489–494.
46. Balasubrahmanyam, S. N., On a genetics based study of the Indian population composition. *Sci. Cult.*, 2011, 32–37.
47. Jagannatharaja, M. G., *Tamil and Prakrit*, International Institute of Tamil Studies, Chennai, 1992, pp. 28–31.
48. Sengupta, S. *et al*., Polarity and temporality of high-resolution Y-chromosome distributions in India. *Am. J. Hum. Genet.*, 2006, 202–221.
49. Pitchappan, R. M., Origin of Dravidian and the genomic era. In Proceedings of the Silver Jubilee Celebrations of the International School of Dravidian Linguistics, Thiruvananthapuram, 2002.
50. Janaki, S. S., *Tamil–Samskrita Sambandhah*, Sanskrit Karyalaya, Sri Aurobindo Ashram, Puducherry, India, 2007.

ACKNOWLEDGEMENTS. We thank Thomas Malten, Institute of Indology and Tamil Studies, University of Cologne, Germany for generously offering e-texts for Cangam Tamil and Sanskrit corpus; Andreas Fuls, Berlin Institute of Technology, Germany for computing bigrams and trigrams from the ICIT database; A. G. Ramakrishnan, Indian Institute of Science, Bangalore for resolving doubts in connection with the Kannada script and orthography. S.S. had a personal discussion with Bryan Wells on Indus script and sign list while he served at MatScience, Chennai. Vanitha Vasu helped generate graphic images for all the Indus signs listed in Mahadevan's corpus. S.S. sought help from A. James and R. Griesh, Tamil Virtual University, Chennai for uploading Tamil characters and numerals on the computer system and took expert opinion from Shivaramu, R. S. Keshavamurthy, V. Sridhar and V. Gopalakrishnan for the views expressed on Kannada, Telugu and Sanskrit languages. We are grateful to S. Ramakrishnan, Cre-A Publisher, Chennai for providing a print copy of the book *Early Tamil Epigraphy*.

Received 24 June 2011; accepted 30 April 2012