OPEN ACCESS

Manuscript ID:
ASH-2023-11026519

Volume: 11

Issue: 2

Month: October

Year: 2023

P-ISSN: 2321-788X

E-ISSN: 2582-0397

Received: 30.07.2023

Accepted: 18.09.2023

Published: 01.10.2023

Citation:
Thanabalasingam, Uthayanan. "Death of Tamil? - A Necessary Call for Simplification." *Shanlax International Journal of Arts, Science and Humanities*, vol. 11, no. 2, 2023, pp. 9–22.

DOI:
https://doi.org/10.34293/sijash.v11i2.6519

Death of Tamil? - A Necessary Call for Simplification

Uthayanan Thanabalasingam
Independent Scholar, London, United Kingdom
https://orcid.org/0009-0002-8991-5698

Abstract
Adequacy and efficiency of Tamil language is examined within the context of cyberspace, primarily in terms of ease of use and the critical resource demand it places on digitisation. On identifying the nature and weight of the problem an ambitious proposal is made to change the transcription of Tamil letters.

Keywords: Tamil Script, Cyberspace, Digitisation, Unicode, Logograph, Tamil Brahmi, Phonetics

Introduction

"Tamil has a gravity of expression not found in any language" - Alexander Dubyanskiy, veteran Tamil scholar from Moscow State University.

A language with its measure of literacy rate has come to define the socio-economic development of people or country. For the languages that will survive tomorrow, it's essential to consider two key aspects namely the "ease of learning" and its "suitability in the technical world".

Why Ease of Learning?

A language that is easy to learn and write will be very advantageous. This is not to say that you can classify languages by complexity. Language complexity is a relative term. Our inherent ability to acquire a native language fluently- no matter what the complexity is – is one of cognitive wonder. However, research has shown that the ability to learn a second language deteriorates with age (Hartshorne). It is in the sphere of adult learning and second language learning that we could possibly get a glimpse of the complexity of a language.

If you consider adult learning, which is critically important for uplifting literacy rates desirably faster, one could easily deduce, it's not the spoken form but the writing system that essentially dictates the literacy and arguably the complexity.

In this context of literacy, *the difficulty of a language then is a measure of its writing system*.

A person's identity is his spoken language. Ideally, a person who speaks a language should be able to learn to write and read in a day. This is to say learning the symbol system that helps to transliterate the sounds of the language should be simple enough.

Unfortunately, one could argue older the language is, potentially complex is its writing system. Consider Tamil. Mere statement of fact that it has 247 unique symbols itself is daunting and may put off many who may show interest in learning it, no matter the enthusiasm that Alexander Dubyanskiy promotes.

There are ample examples how a simplified language could have a more profound impact on literacy rate thereby on socio economic development. There is no better and more visible example than the revolutionary work undertaken in China with pinyin, uplifting literacy rate from 20% to over 80% in phenomenal time (Yongbing; Xiao).

Why Suitability to the Technical World?

Another factor of importance of language is its suitability in the technological world. It could even determine the long-term survival of the language. A simple question "how easy and faster can you type a message on mobile using Tamil?" will help illustrate the importance of why one should consider how the language is used in the digital context. After all texting is currently the most popular form of communication (Noah) that has even overtaken verbal. The question is even more profound if you consider all activities related to language such as reading, writing (typing), and language manipulation cannot happen without involvement of technical interaction. Here language manipulation is taken to refer to things like translation and algorithmic manipulation for various applications including visual computer recognition to interact with future intelligent systems. Hence it should not come as a great surprise as to why it must be digitized and should be easy to use, for its survival.

Digitisation of a language has many factors. The most obvious one is the factor of encoding where characters of a language are encoded into digital form so that script of the language can be digitally represented in all digital platforms (to be part of Unicode). The effort to digitally encode has already led to simplification of many languages. Such changes to language have been observed along with significant revolutions in human history. As with many languages, reformation of script is not new to Tamil either. At various points in time, it went through several changes transforming itself from the oldest Tamil-Brahmi script to significant change with vattelluth (Wikipedia: Tamil Script; Subrahmanian). The recent and modern changes were popularly introduced by Veera-Ma-Munivar and Periyaar, driven primarily by the advent of print, making it more print friendly.

The other factor of importance is the proliferation and presence of wide variety information in digital form. Along with language models for translation of any content to and from other languages, availability of a large corpus of historical and current information is critical for digitisation of the language to drive its digital usage.

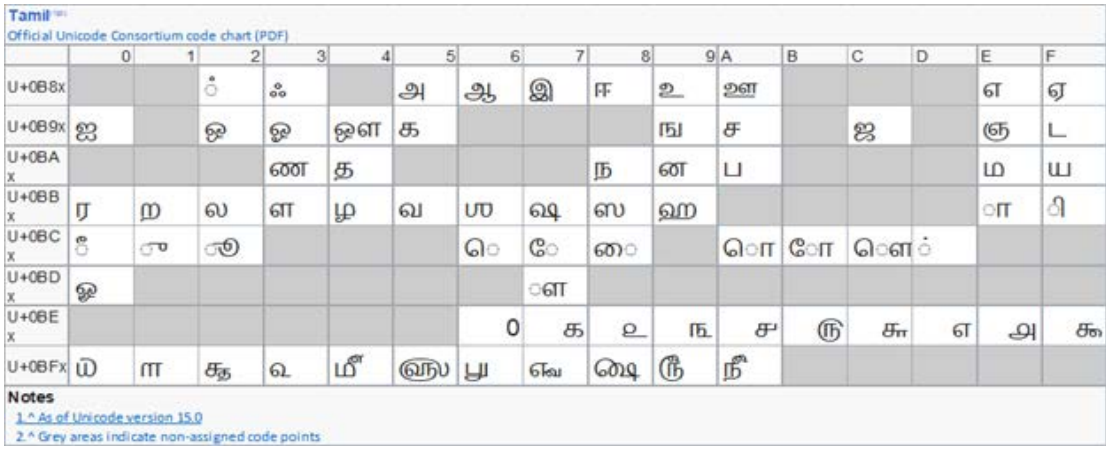
It's the third factor of digitisation that is more relevant to the discussion here, namely the aspect of everyday communication in that language. A language must offer a simpler, intuitive, and persuasive mechanism that doesn't demand relearning or realignment just for the purpose of digital communication.

The above is becoming increasingly important with the advent of AI. Aspects of entropy and complexity of language could become matters of life and death of that language. Entropy in this context is considered to mean both encoding and vocabulary-based entropy. Language may influence the speed of response when dealing with large data based neural systems. In a life critical system, be it medical or say self-driving car, choice of language may end up making key difference in microseconds of importance.

Consider the Tamil Unicode Table Below (Wikipedia: List of Unicode Characters)

It presents a problem. A Tamil writer not only have to contend with 247 letters to select from, but also has to do with quirks of writing that is however introduced with good intention such as ஶ or ஷ. It really requires relearning. Note that for simplification of the keyboard, along with vowels it's the compound letters (க: combined with the first vowel அ) that's given not the consonants (க்). It demands a mental model of printed form kept in mind, in order to reproduce it as per this simplification. While it solves the problem of not having to encode all 247 charecters, it blatantly breaks the intuition and in fact adds additional complexity of learning curve. It's likely to have a detrimental effect on the digitisation of Tamil and its use for communicative purposes like texting.

Hence, for the benefits of education, communication and better technology adoption, sections below look at the problem from the nature of Tamil script perspective.



Tamil[1]
Official Unicode Consortium code chart (PDF)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+0B8x			◌	ஃ		அ	ஆ	இ	ஈ	உ	ஊ				எ	ஏ
U+0B9x	ஐ		ஒ	ஓ	ஔ	க				ங	ச		ஜ		ஞ	ட
U+0BAx				ண	த				ந	ன	ப				ம	ய
U+0BBx	ர	ற	ல	ள	ழ	வ	ஶ	ஷ	ஸ	ஹ					◌ா	◌ி
U+0BCx	◌ீ	◌ு	◌ூ				ெ	ே	ை		ொ	ோ	ௌ	◌்		
U+0BDx	◌							◌ௗ								
U+0BEx							0	க	உ	ங	ச	ரு	சூ	எ	அ	கூ
U+0BFx	ய	ா	கூ	ெ	மீ	ஶ்ரீ	பூ	ஸ்ரீ	ஷ்ரீ	ஜ்ரீ	ஹ்ரீ					

Notes
1.^ As of Unicode version 15.0
2.^ Grey areas indicate non-assigned code points

Tamil Script & Adaptation problem

Tamil has 12 Vowels (refer to as life letters) and 18 consonants (refer to as body letters) and 216 explicit compound letters as shown below.

One of the key innovative milestones in language formation is the understanding that all the sound a language makes can be primarily reduced to vowels and consonants. *Hence the function of a script is to offer perfect transcription as possible*. This is to say script are used to write down the sounds so that it can be read or reproduced sounding the same. In this regard Tamil with its script offers one of the near perfect examples for phonetically least ambiguous language. A written word in Tamil always forms a clearly predictable one sound. Hence there is no need for separate phonetic alphabetic representation (like IPA) for learners to guide pronouncing or reading. Contrast this with English where, as an example, sound "See" could be transcribed as "Sea" or "See". While this feature is neither a strength nor weakness in the inherent potential of a language, it help to show ambiguity in transcription vividly. Consider the word "cacophony". Without phonetic alphabet it could be pronounced as ka-ko -fo-ny, while the correct pronunciation is kuh•ko•fuh•nee. While English deviated from phonetic orthography (graphemes (written symbols) correspond to the phonemes (spoken sounds)), Tamil remains one of the best examples of phonetically orthographic language.

Tholkapyam consonants		Vowels												
		அ a	ஆ ā	இ i	ஈ ī	உ u	ஊ ū	எ e	ஏ ē	ஐ ai	ஒ o	ஓ ō	ஔ au
க்	k	க	கா	கி	கீ	கு	கூ	கெ	கே	கை	கொ	கோ	கௌ
ங்	ṅ	ங	ஙா	ஙி	ஙீ	ஙு	ஙூ	ஙெ	ஙே	ஙை	ஙொ	ஙோ	ஙௌ
ச்	c	ச	சா	சி	சீ	சு	சூ	செ	சே	சை	சொ	சோ	சௌ
ஞ்	ñ	ஞ	ஞா	ஞி	ஞீ	ஞு	ஞூ	ஞெ	ஞே	ஞை	ஞொ	ஞோ	ஞௌ
ட்	ṭ	ட	டா	டி	டீ	டு	டூ	டெ	டே	டை	டொ	டோ	டௌ
ண்	ṇ	ண	ணா	ணி	ணீ	ணு	ணூ	ணெ	ணே	ணை	ணொ	ணோ	ணௌ
த்	t	த	தா	தி	தீ	து	தூ	தெ	தே	தை	தொ	தோ	தௌ
ந்	n	ந	நா	நி	நீ	நு	நூ	நெ	நே	நை	நொ	நோ	நௌ
ப்	p	ப	பா	பி	பீ	பு	பூ	பெ	பே	பை	பொ	போ	பௌ
ம்	m	ம	மா	மி	மீ	மு	மூ	மெ	மே	மை	மொ	மோ	மௌ
ய்	y	ய	யா	யி	யீ	யு	யூ	யெ	யே	யை	யொ	யோ	யௌ
ர்	r	ர	ரா	ரி	ரீ	ரு	ரூ	ரெ	ரே	ரை	ரொ	ரோ	ரௌ
ல்	l	ல	லா	லி	லீ	லு	லூ	லெ	லே	லை	லொ	லோ	லௌ
வ்	v	வ	வா	வி	வீ	வு	வூ	வெ	வே	வை	வொ	வோ	வௌ
ழ்	ḻ	ழ	ழா	ழி	ழீ	ழு	ழூ	ழெ	ழே	ழை	ழொ	ழோ	ழௌ
ள்	ḷ	ள	ளா	ளி	ளீ	ளு	ளூ	ளெ	ளே	ளை	ளொ	ளோ	ளௌ
ற்	ṟ	ற	றா	றி	றீ	று	றூ	றெ	றே	றை	றொ	றோ	றௌ
ன்	ṉ	ன	னா	னி	னீ	னு	னூ	னெ	னே	னை	னொ	னோ	னௌ

It can be argued, most such phonetically orthographic languages are aided by the fact that even syllabic construct for words is explicitly transcribed. The compound letters of Tamil by combining a vowel

and consonant with all its combinations (subject to strict rule of consonant followed by a vowel), are indeed such transcription. As example, க் (ik) + அ (a) = க (ka)) and hence form the unit of syllable. It then naturally leads to rules like, a word can only start with a vowel or a combined letter. Combined letters essentially help avoid double consonant in transcription such as அம்ம்ஆ (aa-mm-mm-aaa = amma for mother). Instead, the correct transcription அம்மா provides for combination of ம்ஆ as one compound letter மா. This in fact inherently capture the important pronunciation rule for the phonetic sound of the language clearly in its transcription. However, there are words with double consonants like தீர்ப்பு (theerpu). It can be seen in such cases where omission of consonant that captured by the compound letter (ப் captured in பு) will not have a significant impact on the sound of the word and can be considered silent.

In corollary, we can also see there is always a possibility to form combine letters with vowel followed by a consonant. Lets say represents அ and ம் combined. Then the word for mother (அம்மா) can be written as "மா". Such explicit transcription would have added another 216 symbols to Tamil (making the total count to 463!). Should one follow this it may have shorten the words while potentially eliminating the need for consonant letters appearing in words.

The above discussion also helps illustrate an important point. Larger the character set, simpler can be the words with lesser no of characters (hence lesser footprint) but with higher level of complexity in initial learning and transcript letters of the language. In essence transcriptions can evolve in any number of ways with varying pros and cons. In the case of Chinese its simplified to one character that represent the meaning. The question is which direction a language should take, for its digital survival? Should it expand its body of symbols for the benefit of compressing words with potential faster read with lower footprint? Or should it canonicalize the symbols instead for the benefit of faster acquisition and simpler encoding purposes?

It's often hard to propose and argue a case for change to a script, especially in the context of number of alphabets, against a system that has been time tested and in adopted usage. There must indeed be a timely persuasive need. It's context of the time and the need that essentially drives the changes to the transcriptions that can be justified. As opposed to being a vanity project with ambitions of either assumed simplification or beautification of a language based on logic or otherwise.

A language is a fluid concept especially in the context of phonetic as it changes with time, often deviating from the transcriptions or stricter form of phonetic orthography. However, its desirable, changes don't induce any structural changes and are well rooted in the tradition and structure of the language.

However, we must recognise Tamil must address the "technical adaptation problem" for its digital survival.

Adaptation problem is not much of a problem of digitisation in the sense of encoding. As described earlier, it's a problem where reintroduction of learning or adaptation need to happen for transition to digital usage. It touches on much wider aspects.

The most obvious being the disruption and inefficiency it causes for casual or otherwise textual communication, which has become paramount in the current digital world. Consider a simple case of typing அம்மா (amma) meaning mother.

Assuming a Tamil 99 Standard (Tamil elibrary), Following Steps will have to be Taken

Step 1: Select a vowel அ – first letter.

Step 2: Select the compound letter (ம) corresponding to the consonant you need (ம்). (Since consonants are represented by the first compound letters that's combined with a default vowel)

Step 3: From that compound letter select the consonant, based on selection provided or by other similar mechanism. (In some cases, its obtained by long or double pressing the key)

Step 4: For the next compound letter (மா), select the related compound letter that represents the related consonant letter (ம).

Step 5: Then select the compound letter (மா) from the prompted set. In some cased select the associated vowel that need to be combined with the consonant to form the compound needed compound letter!

The above requires special keyboard arrangement with additional computational cost to pop characters

in context. That is in addition to the thinking or organisational cost in a non-trivial manner. A typical writer in Tamil does not need to think of related consonant, vowel to write compound letters.

In contrast to the above, to type “mother” in English you don’t need any explanation of any steps. There is only one step explanation to type in the characters that are always present.

This is indeed a significant hurdle as users will often opt for easier language option should as per their proficiency.

Since Tamil uses templating based Unicode system for bare consonants and compound letters (except for the default compound letters that is formed with அ), it requires selection of multiple Unicode items to form a single character.

◌், ◌ா, ◌ி, ◌ீ, ◌ு, ◌ூ, ◌ெ, ◌ே, ◌ை, ◌ொ, ◌ோ, ◌ௌ, ◌், ◌ள

As a result of these modifiers, a character may appear as one but takes more than 16 bit to represent as it combines multiple Unicode. This is highly inefficient, especially given the high frequency of compound letters in the language. It also severely hinders programmability using Tamil, making it utterly inadequate for the digital world. Ex: திருவள்ளுவர் looks like it should have seven letters. However, according to Unicode, this name has twelve characters: த ◌ி ர ◌ு வ ள ◌் ள ◌ு வ ர (Wikipedia: Tamil All Character Encoding) for wider issues with current Tamil Unicode encoding)

Another observation that contributes to the problem is its relatively large logograph. While it has an impact on typeface and the space it takes on a key, it has also been shown language with larger logographs may take longer time to master (Miton and Morin).

The modifiers shown above that form the template also require superscript and subscript style that goes above and below the normal line of typeface.

Vowels	இ	உள	ஐ	ஓ	ஔள
Consonants	ஞ்	ழ்	ங்		

In addition, selected vowels and consonants shown above show multiple strokes (3 or more) to write and complete the letter.

இ, ஈ
ண்/ந்/ன்
ர்/ற்
ல்/ழ்/ள்

Another general aspect is the significant variation that exist in the transcript letters for the similar sounding vowels or consonants as shown below.

Larger number of transcribed letters also plays into information theory by shannon (Shannon). Using this principle, it can easily be shown that lower the canonical symbols in a language, higher will be the compression rate. This is to say with lesser characters any data (a written page or say a book) will have higher repetitive characters hence the potential for high compression rate. One immediate outcome of this will be in terms of energy needed for both processing and storage- in other words- there will be lesser number of trees that will have to be compensated for. If one extends this further to the context of machine learning and human machine interaction, this property of high entropy with low character system will start to make even more significant impact.

For this discussion, a preliminary study undertaken with sample texts that translated in three languages -Tamil, English, Korean- shows Tamil currently requires average 30% more file storage size.

The combined effect of the above stated problems makes Tamil inadequate in the technologically fast advancing world. Imagine a time when there will not be any papers or printed medium where all transactions must be undertaken digitally; with interactive systems that will require finer programmatic manipulation of the language; with every bit that will have an associated financial and environmental cost. That is not a farfetched situation. If Tamil is not prepared for that time, its highly likely other languages will be prefeed in its place, reducing Tamil to an obscure locally spoken language.

Inspiration from Tamil Brahmi

Surprisingly, within the inheritance of Tamil, there are indeed hidden and obvious gems that aid

with alleviating some of the issues, should one be brave enough to embrace. Consider the Tamil Brahmi script.

If we are to hold the premise true that the script is simply to provide transcription to the sound of language, then any script that sufficiently demonstrates the variation of sound, primarily of vowels and consonants, then can fulfil as the alphabets for that language. However, there are two key fundamental aspects we can expect such transcriptions to fulfil.

Orthography and Structural Support for the Phonetics

Intention here is to say there should be sufficiently separate and number of symbols to capture all variation of the spoken sound of the language (I.e: number of graphemes with clear different identities must exist to map phonemes of the language). Naturally, one can also expect these symbols to carry some level of similarity reflecting the similarity in sound. Ex: அ, ஆ are transcribed as similar looking symbols because they well aid the capture of the same vowel sound with short and long variation. The other key aspect is the support for syllabic structure that aids with the execution of the pronunciation of the word. In other words, there can also be rules at the symbolic level (as in compound letters of Tamil) to support and enhance the orthography. Essential this superficially shows the potential path script may have evolved with the sounds of the language.

Simplicity

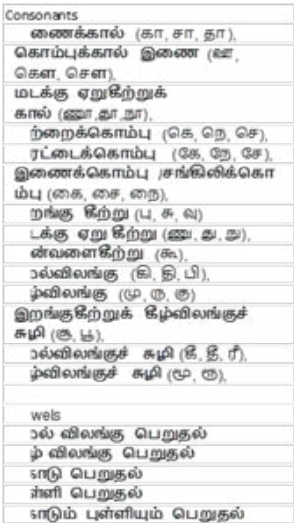
Simplicity or complexity is a relative concept and the evolution of a language can be stated, by definition, choses path of least resistance and simplicity. However, based on a chosen framework different language can be compared to propose an idea of simplicity. In this context one could propose a definition of simplicity based on the following.

Typographic Simplicity

The graphic appearance of the symbols must be simple. We can define that by number of strokes one as to do in the form of curve or line to complete a letter. Letter “A” can be stated to have three strokes. Letter “ஊ” can be said to have five strokes.

Simplicity in Symbol Variation

Additionally, we can also state no or minimum need for modifiers as other measure of typographic simplicity. Tamil has several modifiers as shown below (Wikipedia: தமிழ் எழுத்து முறை)



The above modifiers only illustrate how a symbol is manipulated to take different letter form to represent the sound of the language. When such modifiers are not intuitive then increases the typographic complexity. For example Tamil has syllabic letters that are compound letters (க) of consonants (க்) and vowels (அ) that are explicitly transcribed. In the case of same consonant (க்) combined with another vowel (ஓ) the compound letters is transcribed as (கோ). The modification of this nature, while it’s the result of the language evolution, is extensive and is not intuitive as to the meaning of such modification. The current Unicode system provided the following templates for the modification of the consonants, while representing the vowels as such.

◌, ◌ா, ◌ி, ◌ீ, ◌ு, ◌ெ, ◌ே, ◌ை, ◌ொ, ◌ோ, ◌ௌ, ◌், ◌ள

As shown earlier these modifiers as they form superscript and subscript also breaks the line of writing, making the types otherwise more loaded.

Learning Simplicity

More complex typography and a large number of transcribed symbols in a language will add to the burden of learning (Miton and Morin). This

is a relative measure. if we restrict the definition of learning simplicity to identification and memorisation of symbols and immediate application of it to pronounce or read with ease then we can see the merit of this measure. A language will have a learning advantage, certainly among adult learners, if its symbols can be learnt and read, sustainably, within few hours or in matter of days. As will be shown, this doesn't have to be a pipedream. The idea that a native language speaker is illiterate because there is difficulty in reading or writing of their language must be challenged vehemently indeed. (By that definition of literacy Buddha must also have been illiterate, as are many other scholars!).

Symbol Entropic Simplicity

As illustrated earlier, this is a measure of the digital footprint of the language purely from the symbols that it hosts. It simply depends on two key factors, 1. Number of symbols 2. Frequency of such symbols. It suffices to say the first will weigh in heavily to influence the measure negatively.

Considering orthography and simplicity as key fundamental needs allows for an interesting comparison with the parent script of Tamil namely the Tamil -Brahmi.

Brahmi letters that correspond to Tamil vowels and consonants are given below.

Tamil Vowel	அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஒ	ஓ	ஔ
Tamil Brahmi	𑀳	𑀴	𑀵	𑀶	𑀷	𑀸	𑀹	𑀺	𑀻	𑀼	𑀽	𑀾

Tamil Consonants	க்	ங்	ச்	ஞ்	ட்	ண்	த்	ந்	ப்	ம்	ய்	ர்	ல்	வ்	ழ்	ள்	ற்	ன்
Brahmi equivalent	𑀓	𑀗	𑀘	𑀜	𑀝	𑀡	𑀢	𑀦	𑀧	𑀫	𑀬	𑀭	𑀮	𑀯	𑀰	𑀴	𑀶	𑀷

Below shows the evolution Tamil script from Brahmi script (Reg 10).

The compound forms are as below.

Following observations can be noted.

1. Although, language and its symbol system has been evolving, there is a consistency in the basic consonants and vowels.
2. Symbol structures seem to have preferring evolutionary tendency towards curvy symbols from linear symbols.
3. While there are exceptions, Tamil Brahmi symbols still visibly capture the essence of the modern Tamil symbol and is still visibly recognisable as such. For example, the essence of அ can be caught in its linear strokes like 𑀳, which is very similar to the Brahmi script.
4. Brahmi scripts are typographically simpler by the number of strokes involved in the letters.
5. Brahmi letters seem to capture variation and modifications in simpler way. For example, in the case of vowels, short and long form is differentiated by addition on a line as shown below.

உ	ஊ
𑀷	𑀸

6. As another example, the compound letter formed from consonant க் (Brahmi) and vowel உ, ஊ is simply a visual combination of respective consonant and the vowel in Brahmi as 𑀓𑀷 𑀓𑀸 This is unlike the Tamil version கு, கூ where the modification is done to the consonant varying

greatly from the inherent vowel. In other words, in majority of the cases in Brahmi, compound letters can be constructed from the existing basic line strokes by keeping to the same line of writing.

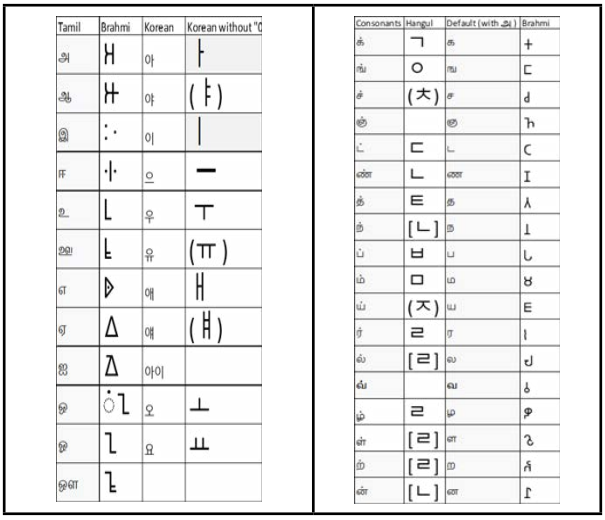
A Look at Korean – A Sister Language in Phonetic and Symbols

"Hangul must surely rank as one of the great intellectual achievements of Mankind" (Sampson p. 165).

Sampson continues, "[w]eknow this [to have a better alphabet] because there is an alphabet that is about as far along the road towards perfection as any alphabet is likely to get" (pp. 108–109). He goes on to indicate "[i]n its simplicity, efficiency and elegance, this alphabet is alphabet's epitome, a star among alphabets, a national treasure for Koreans…" (Sampson p. 109).

There is indeed no better language script worth looking into than Hangul, when considering any simplification or changes to a script of any language. Hangul is one of the most successful, simpler alphabet systems that's man made. This is not to say the hangul script can be adopted to successfully transcript other languages. Its elegance, simplicity and structure are the ones worth studying to guide the direction of change. In the context of Tamil, as will be seen, its relevance is even more important. While there is no research done demonstrating the potential commonality between Korean and Tamil, they do share significant phonetic correlation in addition to the basic language structures (Tulkens et al.). Moreover, as shown below hangul, following a line-based approach to letters, has striking similarities with Tamil-Brahmi script. It shows the likelihood that creators of hangul may have studied Brahmi scripts.

Hangul has 21 vowels and 19 consonants. Following also shows the structural similarities of both the languages..



Tamil	Brahmi	Korean	Korean without "ㅇ
அ	𑀅	아	ㅏ
ஆ	𑀆	야	(ㅑ)
இ	𑀇	이	ㅣ
ஈ	𑀈	오	ㅡ
உ	𑀉	우	ㅜ
ஊ	𑀊	유	(ㅠ)
எ	𑀎	애	ㅐ
ஏ	𑀏	에	(ㅒ)
ஐ	𑀐	아이	
ஒ	𑀒	오	ㅗ
ஓ	𑀓	요	ㅛ
ஔ	𑀔		

Consonants	Hangul	Default (with அ)	Brahmi
க்	ㄱ	க	+
ங்	ㅇ	ங	⊏
ச்	(ㅈ)	ச	d
ஞ்		ஞ	ϧ
ட்	ㄷ	ட	⊂
ண்	ㄴ	ண	I
த்	ㅌ	த	ʎ
ந்	[ㄴ]	ந	ɪ
ப்	ㅂ	ப	ᒐ
ம்	ㅁ	ம	४
ய்	(ㅈ)	ய	E
ர்	ㄹ	ர	ı
ல்	[ㄹ]	ல	ᔑ
வ்		வ	ᔑ
ழ்	ㄹ	ழ	ᔑ
ள்	[ㄹ]	ள	ϑ
ற்	[ㄹ]	ற	ᔑ
ன்	[ㄴ]	ன	ᔑ

Criteria	Korean	Tamil
linguistic syntactic typology?	SOV (Subject-Object-Verb)	SOV (Subject-Object-Verb)
Arrangement/pattern of Vowels and Consonants	1.Abugida language as it has a syllabic concept with an explicit case called bitchem	1. Abugida language as it has explicit compund letters that even specially transcribed
Phonemic orthography	Yes: Given the explicitly wriiten syllabic nature	Yes: Given the explicitly wriiten syllabic nature
Morphology: Agglutination or fusional	Tamil is an agglutinative language. Root word is combined with suffixes to form words	Korean is an agglutinative language. Root word is combined with suffixes to form words

Following can be Noted

1. It's designed from the ground up for people to learn and write easily. "A wise man can acquaint himself with them before the morning is over; even a stupid man can learn them in the space of ten days."

2. Hangul follows well intentioned line-based transcriptions that's said to mimics the shape of tongue when uttering the sound.
   Ex: ㄱ representing the [k] sound geometrically describes its tongue back raised
   representing the [m] sound geometrically describes a closed mouth.
3. Hangul follows predictable and consistent rule for transcribing similar sounding vowels or consonants (by adding an extra stroke, ex: ㅛ ㅑ ㅠ ㅕ) and by providing symmetrical opposites for other vowels sounds (ㅓ ㅏ, ㅏ for "aaa" ㅓ for "ooo")
4. Hangul also supports syllable blocks that's similar to compound letters in Tamil. Although these blocks are not transcribed explicitly by separate characters, they are pre-composed and added to Unicode character code.
   Ex: 가 where ㄱ "k" is the consonant, and ㅏ "ah" is the vowel. It is pronounced "Kah".
   고 where ㄱ "k" is the consonant, and ㅗ "oh" is the vowel. It is pronounced "Koh".

	ㅏ	ㅑ	ㅓ	ㅕ	ㅗ	ㅛ	ㅜ	ㅠ	ㅡ	ㅣ
ㄱ	가	갸	거	겨	고	교	구	규	그	기
ㄴ	나	냐	너	녀	노	뇨	누	뉴	느	니
ㄷ	다	댜	더	뎌	도	됴	두	듀	드	디
ㄹ	라	랴	러	려	로	료	루	류	르	리
ㅁ	마	먀	머	며	모	묘	무	뮤	므	미
ㅂ	바	뱌	버	벼	보	뵤	부	뷰	브	비
ㅅ	사	샤	서	셔	소	쇼	수	슈	스	시
ㅇ	아	야	어	여	오	요	우	유	으	이
ㅈ	자	쟈	저	져	조	죠	주	쥬	즈	지
ㅊ	차	챠	처	쳐	초	쵸	추	츄	츠	치
ㅋ	카	캬	커	켜	코	쿄	쿠	큐	크	키
ㅌ	타	탸	터	텨	토	툐	투	튜	트	티
ㅍ	파	퍄	퍼	펴	포	표	푸	퓨	프	피
ㅎ	하	햐	허	혀	호	효	후	휴	흐	히

Notice blocks are created by consonant vowel combination that is arranged in adjacent (가) if the vowel is vertical and or with vowel written under consonant (고) if the vowel is horizontal. In that way the compound syllables are formed by direct combination. Simplified keypad is formed with either 17 or 9 keys, simplifying the 40 letters. As a result, there is a contextual overhead during the typing of hangul using these keypads, although they are relatively simple as its based on identifiable symbols not based on templated approach.

Proposal to change the Tamil transcription.

The problem of current inadequacy of Tamil in the context of the fast emerging technological world, certainly warrants a rethink in terms of its transcriptions. Tamil is a rich and beautiful language with enviable traditions and literature. There is much more to discover in terms of its structure and potential applications. Hence the persuasion to make Tamil more relevant and simpler is even more important.

Based on the discussions, following simple but highly desirable goals can be set.

1. Make Tamil easy to read and write. Make it intuitive enough so that it can be learnt to read and write within a day. This is not a goal to teach vocabulary or mastery of the language. The aim is to simplify and optimize the letter system to the maximum possible extent so that it places lesser demand on the cognitive faculties.
2. Make Tamil one of the easiest languages to use in the technological world. This is to say, it should be simpler not just for faster texting, faster adaptation but also from a smaller footprint and computational (algorithmic friendly language) perspective.
3. Ensure the rigor and consistency of Tamil is preserved.

What is profoundly interesting is that one can set about achieving the above goals without having to resort to a radically new approach that may end up significantly deviating from the heritage of the language. Tamil-Brahmi scripts in its ancient simpler form gives sufficient clues to the possibilities. In other words, all one must consider is the natural reduction towards Brahmi script that in essence is indeed nothing but Tamil letters but are in a much simpler form. The approach considered here is to compare and if possible align such changes with the modern hangul scripts to borrow the proven innovations that underlies the hangul scripts.

Simplification of Vowel transcriptions

It's the vowels that modify the consonant to form the compound letters. In fact, all the modifiers (ெ, ா, ி, ீ, ு, ூ, ெ, ே, ை, ொ, ோ, ௌ, ◌், ◌ா) that are identified for Unicode purposes have direct correspondence to the respective vowels that modifies the consonants. It begs the question, why can't those modifiers

themselves be used as vowel transcriptions? In such a case compound letters will be identifiable as syllabic letters in its form as in the case of hangul, hence eliminating the additional explicit transcriptions of the compound letters. Unfortunately, these modifiers show a complex form as shown in the table below. If we are to avoid using such modifiers for vowels, any other proposal for transcription must also be simple enough to represent as modifiers for compound letters so that compounding can be visibly transparent.

Tamil	Brahmi	Korean	Korean without "0"	Proposed	Reasoning	Modifier based
அ		아			1. If we abandon the curvy aspects we end up with a vertical line and a horizontal line to represent அ 2. one horizontal line is used to symbolise the short form, so that double of such horizontal line can be used to show the long vowel with minimum change 3. Correspondence with hangul (avoiding conflict with ㅓ)	
ஆ		야	( )		1. The additional horizontal line (modifier) is used to show the long form 2. Hangul has "Y" initial iotation hence differs from the usage, although it does symbolise the long form of sound	◌ா
இ		이			1. This is an obvious simplification as it's the closest and most recognisable form. 2. Ending vertical aspect of the etter is used , removing all curvy parts 3. Unfortunately, the Brahmi letter is deviation from both tamil and the chosen letter. 4. Correspondence with hangul	◌ி
ஈ		으			1. For இ, ㅣ is used with any aspects that can be used to double up to indicate longer version, unlike ㅏ. It could have been given with single dot. It was avoided to avoid confusion to the letter இ. 2. In this case given Brahmi letters has shown very close and identifiable correspondence	◌ீ
உ		우			1. There is is a direct correspondence in hangul with ㅜ as in 우리, உரி (oo-ree) = our/mine 2. ㅜ only captures the horizontal aspect well while Brahmi captures the essence better	◌ு
ஊ		유	(ㅠ)		1. The above is modified with another horizontal line to indicate the long wevel, but while also being identical to Brahmi	◌ூ
எ		에			1. Eliminating curvy parts may give ㄱ. Since it conflicts with consonnats and hangul its avoided 2. The angle representation coinsides with Brahmi and romanised A 3. another possibility is to use , from lydian alphabet. Its not considered to avoid another modifier in the wave form	◌ெ
ஏ		애	( )			◌ே
ஐ		아이	(?)		1. ஐ=அ+ய , and the current shape is relatively a recent development(ref 16). Hence could have been used as it combines the ㅏ and ㅣ. However, it will confuse with ae sound of hangul and does not capture the ய் consonant, possibly giving the impression of double vowel as opposed to diphthong. 2. other possibility is  that combines the consonant for ய். 3. it will not alos be a good idea to indicate to include the short horizontal line as it indicates there may be a long form 4. Interestingly ஐ does look like the combination of அ and ய. hence its taken in that form. Also the shape "h" is avoided. 5. the given typy helps to corresponds to most of the words that starts with ஐ like ஐயனார் or chair . it also tries to visually preserves the vowel ㅏ and consonant ய present in it 6. Its only one of two diphthongs in tamil and hence its reasoned as acceptable to deviate from the rest of styles of the vowels	◌ை
ஒ		오			1. There are no line in ஒ. However, in keeping with a principle of hangul to ensure all vowels follow the same line style corresponding brahmi letter is chosen 2. Hangul letter ㅓ is not considered for its potential confusion with அ 3. Tamil Brahmi letter here is shown here by combining two brahmi unicode letters as there are no corresponding unicode letter	◌ொ
ஓ		요	(어		1. The above is modified with another horizontal line to indicate the long vowel, but while also being identical to Brahmi (also comoare with அி)	◌ோ
ஔ		여	(?)		1. ஔ=அ+வ. Here there are two different approach is possible a) symbol that shows the combination the vowel and consonant b) symbol that's closer to combines ஒ and ள. 2. Its mystical as to why ஒ and ள is used to represent the sound of "auv". It also has no correspondence to Brahmi letter and has no equivalent in hangul. 3. If we follow the same logic to expose the vowel and the inherent consonant sound then we arrive at the representation given. it also captures some sense of the ள along with ㅏ	◌ௌ

In summary above proposes the following changes

1. Similar sounds (short and long form) are separated with same symbol with simpler modifier differentiating them (ex: ㅏ, ㅑ for அ, ஆ). The modifier used in this case in the form of additional "-" is used mostly across all the short and long vowel forms (ex: உ,ஊ are given as ㄱ, ㄱ with extra "-" separating the form).
2. Typographic simplification by reducing letters by adopting Tamil Brahmi letters. (ex: இ, ஈ represented as ㅣ)
3. Proposed transcription is extensive since it is aimed in such a way the vowels themselves can also be modifiers. That is, they should be easier to combine with consonants to show compound letter forms. (ex: for க). The connecting rule here

simple where both letters are represented by just touching each other.

Simplification of Consonants

As shown above changes are minimized only to flatten the letters and to make similar sounding letters also look similar. These changes are indicative and principally identifies where the changes are needed. In this sense there is a graphic or significant topographic change exist between vowels and consonants.

Tamil		Hangul	brahmi-unicode	Proposed
க்	க	ㄱ	+	க(+)
ங்	ங	ㅇ	⊏	ங
ச்	ச	(大)	⊄	ச
ஞ்	ஞ		⊦	ஞ
ட்	ட	ㄷ	⊂	ட
ண்	ண	ㄴ	I	ண
த்	த	ㅌ	⅄	த
ந்	ந	[ㄴ]	⊥	ன
ப்	ப	ㅂ	⊔	ப
ம்	ம	ㅁ	⊻	ம
ய்	ய	(ㅈ)	ㅌ	ய
ர்	ர	ㄹ	≀	ர
ல்	ல	[ㄹ]	⊍	ல
வ்	வ		⊰	வ
ழ்	ழ	ㄹ	⊭	
ள்	ள	[ㄹ]	⊱	
ற்	ற	[ㄹ]	⊬	ற
ன்	ன	[ㄴ]	⊥	ன

Simplification of Compound Letters

As a possible illustration, the compound letters are formed by simply joining the vowels and consonants together. Since consonant letters and vowels can appear in a word on their own, the compound form is differentiated by ensuring both letters are touching each other.

Tholkapyam consonants			Vowels											
			அ ∤ a	ஆ ∤ ā	இ ∣ i	ஈ + ī	உ ∟ u	ஊ ∟ ū	எ ∧ e	ஏ ∧ ē	ஐ ⊨ ai	ஒ ∟ o	ஓ ∟ ō	ஔ ⊢ au
க்	க	k	க	கா	கி	கீ	கு	கூ	கெ	கே	கை	கொ	கோ	கௌ
			க ∤	க ∤	க ∣	க +	க∟	க∟	க∧	க∧	க⊨	க∟	க∟	க⊢

Following can be noted

1. Although a syllabic form like hangul is prescribed, it doesn't propose more complex syllabic block structure or batchim like arrangement. Following on from the compound of Tamil, it only makes compound nature more transparent by explicitly showing the vowels inherent in the consonant. This helps simplify the encoding as each vowel itself can act as modifiers, hence removing the need for separate modifier extensions. It's nothing but a ligature that occurs by graphemes or letters that are joined to form a single glyph.
2. It also simplifies different font creation processes. There is no need to ensure modifiers fit within the font representation, not to mention the simplicity gained in rendering the font (lower units per eM).
3. These compound letters can also be precomposed. In other words, they can also be treated as separate symbols on their own. In this way normalization needed for algorithmic manipulation can be simplified (Unicode collation algorithm for sorting and comparing etc).

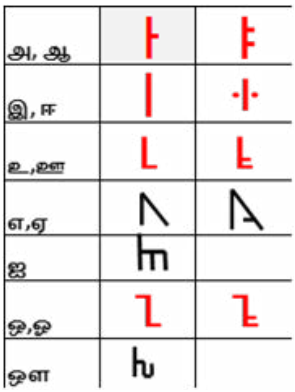
Summary Assessment

Learning to Read and Write is Simplified

Since compound letters are ligatures formed by explicitly combining the consonant with vowels, there is no need for the separate visual identification for the compound letters, albeit retaining the syllabic or combined nature. It immediately reduces the recognition or learning memory to 30 vowels and consonants.

Since vowels are further simplified by representing the similar sounds with similar symbols with consistent modification (additional dash) to show the variation, the symbol recognition can be reduced to 7. Similarly, consonants are simplified to 13 distinctive symbols as shown below. From this perspective distinctive symbols that need to be recognised are reduced to 20.

Furthermore, there are emerging research suggesting line based writing rather than curve based writing may improve the comprehension and response time in addition to favouring for dyslexic children (Guan).



Tamil	Proposed
க	க
ங	ங
ச	ச
ஞ	ஞ
ட	ட
ண, ந, ன	ண, ல , ன
த	த
ப	ப
ம	ம
ய	ய
ர, ற	ர, ற
ல, ழ, ள	ல,ள , ழ
வ	வ

Well Oriented for Digital Adoption

When a character and a modifier combine to form a new character, it effectively forms two letters in one.

For simplicity let's assume UTF-8 encoding. The English language will require 56 letters (including capital letters) to be encoded taking up 56 bytes. In Tamil, there are 30 vowels and consonant letters and 216 compound letters. Since compound letters are not precomposed but combined by using 14 modifiers, it takes 44 bytes to represent. However, Each compound letter is formed by two individual letters (combination of 18 consonants with 14

modifiers). Hence each compound letter will take up 2 bytes and 432 in all to render. That makes the total bytes to 462 bytes in all to represent Tamil.

This has a serious computational impact. Imagine if we must load the character set into an array for string manipulation, not to mention the collation algorithm that had to be run? It will have both memory and algorithmic complexity that many fold higher than for English.

Simplification sought character set doesn't necessarily negate all the problems. However, it offers alternative possibilities to treat the compound syllables as individual characters for all sort of algorithmic manipulation, potentially reducing the memory footprint required.

Moreover, the key simplification in vowels, while allowing for a variety of fonts and writing style, retains a much simpler form for recognition, not just from human learning but also from camera based artificial recognitions. And that is critically important in the world of things like driver less cars and other many upcoming intelligent solutions that increasingly rely on the flexibility of camera eye or visual recognition. For example orthographic neighborhood effects can be used to show the advantages of the suggested changes over existing scripts. It will also help avoid potential financial infrastructure changes, not to mention the criticality of performance. Simplified approach brings the language closer so leverage other technical innovation for things like parsing, machine learning , searching etc

From a typing perspective, very least it makes it more intuitive to deal with compound letters by negating the need for additional learning or so called "mental model" needed for compound letters. For example, a user doesn't need to know கா must be derived from க் and ஆ. A shift function that may be needed for bringing a compound letter can be drawn parallel with the capitalisation shift for English. Such solutions are not derived here.

Conclusion

Above is a direct bold imagination to rethink the Tamil alphabets, inspired from the need for change to make a digitally savvy language. To practically execute such a transition, it may require more studies and even multiple intermediary stages. The idea here is to demonstrate the inadequacy in comparison to other languages and to highlight the urgency to modernize to make it better suited to tomorrow's world from the perspective of Typing, editing, manipulating (employment of other language models and algorithms, reading and visual computer recognition).

Fortunately, Tamil Brahmi letters offer much needed simplicity while preserving the correspondence to the modern Tamil form. Hangul, while also similar, offers innovative possibilities and benchmarks for comparison. Within this context, possibilities of changes to the Tamil alphabet are proposed on the outset.

Vowels are considered for substantial changes for their need to also act as modifiers to represent the compound letters transparently. Consonants are taken as such except for simplifying them for phonetically similar sounds and to flatten the letters to preserve line of writing.

References

Barnhart, Anthony, and Stephen D. Goldinger. "Orthographic and Phonological Neighborhood Effects in Handwritten Word Perception." *Psychonomic Bulletin & Review*, vol. 22, no. 6, 2015.

Guan, Connie Qun, et al. "Curved vs. Straight-Line Handwriting Effects on Word Recognition in Typical and Dyslexic Readers Across Chinese and English." *Frontiers in Psychology*, 2021.

Hartshorne, Joshua K., et al. "A Critical Period for Second Language Acquisition: Evidence from 2/3 Million English Speakers." *Cognition*, vol. 177, 2018, pp. 263-77.

Miton, Helena, and Olivier Morin. "Graphic Complexity in Writing Systems." *Cognition*, vol. 214, 2021.

Noah, Sherna. "Texting Overtakes Talking as Most Popular Form of Communication in UK." *Independent*, 2012.

Sampson, Geoffrey. *Writing Systems*. Equinox Publishing Ltd, 2015.

Shannon, C. E. "A Mathematical Theory of Communication." *The Bell System Technical Journal*, vol. 27, 1948, pp. 623-56.

Subrahmanian, N. "Epigraphy: The Origin of the Tamil Script." *Tamil Studies*, vol. 2, no. 1, 1982, pp. 8-23.

Tamil eLibrary. "Tamil Font Encoding and Keyboard Layout standards of the Tamilnadu Government." *Tamil eLibrary*, https://tamilelibrary.org/teli/tnstd.html

Tulkens, Stephan, et al. "Orthographic Codes and the Neighborhood Effect: Lessons from Information Theory." *Proceedings of the 12th Conference on Language Resources and Evaluation*, 2020, pp. 172-81.

Wikipedia. "List of Unicode Characters." *Wikipedia*, https://en.wikipedia.org/wiki/List_of_Unicode_characters

Wikipedia. "Tamil All Character Encoding." *Wikipedia*, https://en.wikipedia.org/wiki/Tamil_All_Character_Encoding

Wikipedia. "Tamil Script." *Wikipedia*, https://en.wikipedia.org/wiki/Tamil_script

Wikipedia. "தமிழ் எழுத்து முறை." *Wikipedia*, https://ta.wikipedia.org/wiki/தமிழ்_எழுத்து_முறை

Xiao, Huimin, et al. "Pinyin Spelling Promotes Reading Abilities of Adolescents Learning Chinese as a Foreign Language: Evidence From Mediation Models." *Frontiers in Psychology*, 2020.

Yongbing, Liu. "A Pedagogy for Digraphia: An Analysis of the Impact of Pinyin on Literacy Teaching in China and Its Implications for Curricular and Pedagogical Innovations in a Wider Community." *Language and Education*, vol. 19, no. 5, 2005, pp. 400-14.

Author Details

Uthayanan Thanabalasingam, *Independent Scholar, London, United Kingdom,*
Email ID: *uthay@marginheight.net*