History of Language Technology from Grapheme to Distinctive Features

P.Sreekumar Dept: of Linguistics University of Kerala

Language Technology as itself is human species specific instinct to transcend time and space through speech, writing, printing and digital. Now we are passing through the third phase of language technology that is digital language technology and have already passed first language technology, writing and the second printing. Each phase have its own distinctiveness but evolving through the same paradigm of time and space with the same instinct of transcending. It is an initiation up to some extent a preparadigmatic initiation to read the philosophy of language technology through each phases of development.

1 Gene, the language of life

All living beings including man transcending as itself as a species through time by the very medium of gene. The encoded genotype are decoded as phenotype in the very womb of next generation, consequently no exception for any biotic beings no rapture for any species, it is the calibration of life and it is an unending procession of the perpetual continuity of life. We don't know from where it started and where it will end so let it be answered by genealogy and eschatology. In spite of the biotic instincts of transcending time with the substance of life, man, just one among the procession is alone in the sense s/he have to transcend not only his biotic substance but what s/he have acquired as culture, knowledge etc so he speak ,write, print and digitalize.

2 Speech, it is suspended by its medium

Speech as itself is depending sound as the medium of its existence, so as sound as itself depending time as its canvas of existence, it is perishable and momentary within the moment of articulation and limited by the boundary of transmition. Sound cannot exist beyond the moment of articulation and cannot transmit beyond the space in which it realizes. So from the moment of articulation speech is suspended by its medium, sound. By speech, man cannot transcend what s\he have acquired beyond the boundary of time and space. But he has to transcend what he has acquired as a social being.

3 Writing, the first language technology

S/he writes, by replacing sound with visual called graphic form. It can survive beyond the time of articulation due to the imperishability of visuals beyond the time at which it is encoded. By definition writing is a "device for expressing linguistic elements by means of visible marks" .The imperishable visible mark called writing have induced a qualitative jump in the history of human beings by transcending what s/he has acquired beyond time and space. Writing supplemented speech but not just as an extension but as a new existence of speech beyond time and space. But in the context of social history, the practice of writing is quantitatively limited to few, so only few can transcend what they have acquired but the rest remained as other in the archeology of knowledge, they have no history but only memories.

3.1 Technology of writing

The technology of writing replaces phone to graph, grapheme to phoneme and allophone to allograph by an active tool, which is the archetype of pen and a passive canvas which was the archetype of paper, both was in the same paradigm of sound and time.

linguistic units ⟷ graphemes ⟷ graphic forms

Different linguistic units, phoneme represents grapheme in alphabetical writing(English),syllable represent diagrapheme in syllabic writing(Malayalam) and pictoriographeme represents in pictorial writing (Japanese) and ideographeme correspondence in idiographic writing(Chinese).

4 Printing, a quantitative revolution

In the essence of technology printing is not so much differ from writing, but in its social consequence printing, the second language technology was a quantitative revolution. It begot reading, literacy and illiteracy also.

4.1 Technology of printing

Printing represents graphemes as glyphs and collection of typical glyphs as font. English graphemes a b c d e f etc are represented by different fonts as follow.

Name of font	**Type of glyphs**
Times New Roman	A b c d e f g h I j k
Bookman Old Style	A b c d e f g h I j k
Arial	A b c d e f g h I j k
Book Antiqua	A b c d e f g h I j k

By the emergence of printing as an industry, typography has evolved and different types of fonts were designed for each language for printing.

5 Digital Language Technology, beyond time and space

Digital language technology is in essence and in its consequence entirely different from writing and printing. It is a paradigm shift from the history of language technology; it transformed the idea of time and space and transcends the medium of language itself. By the frame of reference it is translingual and global and transcend the boundary of language it self.

Technology

Philosophy of language technology is graphocentric not phonocentric. Whether writing or speech is primary is not the question but digital language technology primarily deals with distinctive graphic forms which can be represented by digital binary.
Three components are primary requisites of digital language technology

1. **Encoding**: Text to be represented in the memory of a computer

2. **Typing:** Text to be typed at the keyboard of a computer.

3. **Rendering**: A way to present text on the screen of a computer and the printing of text on paper

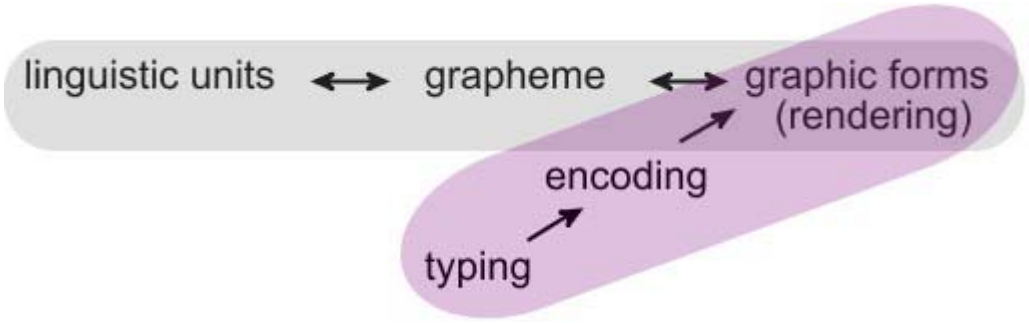
Figure shows the process by which the input from the keyboard is mapped into the appropriate encoding which is then mapped into the appropriate rendering.

typing → encoding → rendering

Malayalam in digital language technology

Typing	Encoding	Rendering
Key A	Unicode 0D05	A

The encoded substance in digital language technology is known as character. One character is represented as a binary number in the memory of a computer; it can be manifested as different glyphs on screen or print according to the context of distribution.



The linguistic unit can be articulated by speech by the performance of bundle of features of it, same time it can be typed and rendered on screen or a paper as graphs. In speech, distinctive features are manifested as phone but in digital language technology, distinctive features of a grapheme manifested as graphs on screen or paper.

See how Malayalam grapheme B is realized in different contexts

Character **glyphs**

(Grapheme)---------- (allograph no-1] --- [allograph no-2]

B ---------- B is in initial and m is in non initial positions

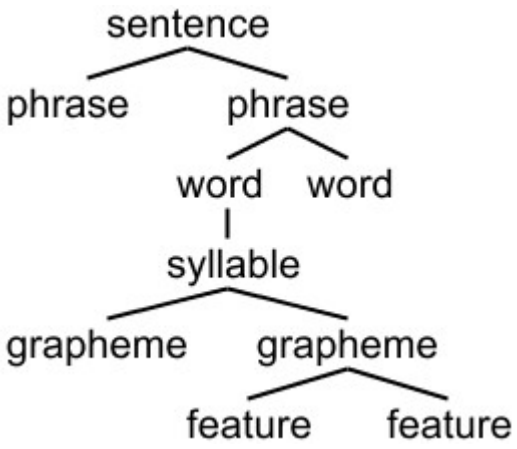
Bundle of features
[+independent grapheme]
[-final]
[+only initial]
[+vowel
[+long]
[+central]

[+unrounded] etc.
Graphemic features of a grapheme include its phonemic features also. So a new paradigm can be formulated by the assemblage of phonemic and graphemic features, it can be conceptualized as textual features of a grapheme.

Textual feature of a grapheme:

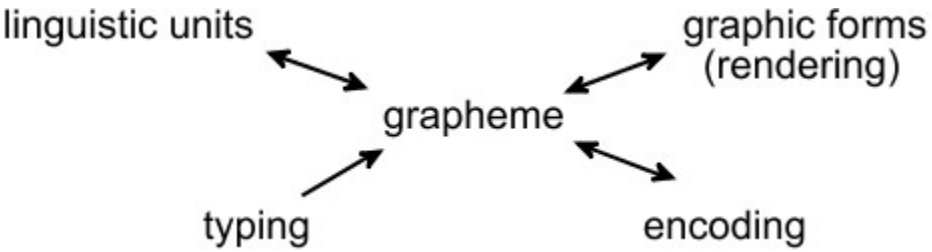
Bundle of [Graphemic feature + phonemic feature]+contextual variation

Each grapheme of a language can be decomposed as bundles of phonemic and graphemic features. Even now the most sophisticate technology of textual encoding Unicode deals distinctive visible marks in a language only as character. So we can encode each grapheme as abstract distinctiveness of phoneme and grapheme with its distributional variants and graphotactics. Such computational textual paradigm can generate all and only visible and phonable text of a language.



So a sentence can be rewritten as bundles of features in its terminal string. In short centre of textual process in digital language technology should be shift from grapheme to features.

Graphocentric model



The above graphocentric model is not considering distinctive features of graphemes by which it is characterizing as a text in language. If we map all distinctive graphic and phonemic features of a language with its distributional rule it shall be called text generation device in a language

Computational text generation device

A computational text generation device is an application model of text generation device. It will encode each graphemes of language as features by a two dimensional phono-graphic algorithm and generate possible and only grammatical words in a language. Followings are the input to generate text of a language by the device.

1. Phonological distinctive features
2. Graphological distinctive features
3. Distributional and combination possibility of each features (graphotactics)
4. Contextual variation of each bundle features when it realize as visual and verbal in all other contexts (allographic and allophonic variation)

Testing of modal

The modal can be tested to generate verbal and nominal root forms of Malayalam for the development of lexical database.

Testing procedures

1. Define all phonological and graphological distinctive features of Malayalam
2. Generate all graphemes and allograph of Malayalam by the combination of features
3. Test a set of corpora by distinctive feature algorithm
4. Generalize combination and distribution possibility of features from the corpora
5. Use the generalized text generation of Malayalam to generate root base of noun and verb.

(Acknowledged to Computational Linguistic Team C-DIT)