ISSN 2313 -4887

தமிழ் இணைய மாநாடு 2017

மாநாட்டுக் கட்டுரைகள்

Tamil Internet 2017

August 25-27, Toronto, Canada

CONFERENCE PROCEEDINGS

Conference Chair: Prof. C.R. Selvakumar (Univ. of Waterloo, Canada)

International Forum for Information Technology in Tamil (INFITT)
16th Tamil Internet Conference 2017,
25-27 August 2017
University of Toronto, Scarborough Campus, Toronto, Canada

CONFERENCE PROCEEDINGS
Table of Contents

1. Technology - Phonology/Speech recognition/Voice synthesis & OCR

NLP & Machine Learning

Technology related to Content Development, Usage and Access

Digital Preservation, Digital Libraries/Archives, 62, 74, 96

Keynote Lecture

Enhancing Tamil document images for better OCR recognition

Ram Krishna Pandey, A G Ramakrishnan

MILE Laboratory, Department of Electrical Engineering,
Indian Institute of Science, Bangalore 560012, India
rkpandey@ee.iisc.ernet.in, ramkiag@ee.iisc.ernet.in

Abstract:

Spatial resolution is an important factor when passing a document image to a properly trained OCR for obtaining editable text. Experimental results in terms of character level accuracy on different resolutions of the same document image are different when passed to OCR for recognition i.e. the character and word level accuracy of document images is different when scanned at a different resolution. We have found that when the resolution is low the accuracy is also low and the accuracy increases steadily when the same image is scanned at higher resolutions. In this paper, we have studied the performance of a properly trained OCR on Tamil document images and found that the accuracy can be improved by increasing the resolution of the image. Through extensive experiments, we have obtained a convolutional neural network architecture which can take the low-resolution document images and can obtain high-resolution document images with a factor of two. We have shown a very good improvement in terms of PSNR, visual quality and character level accuracy of the reconstructed output images.

Keywords: document image, spatial resolution, convolutional neural network, super-resolution, document quality, OCR, PSNR, recognition accuracy.

I. Introduction:

Experimental results suggest that Optical character recognizers (OCR) character-level accuracy (CLA) reduces significantly with the decrease in the spatial resolution of document images. And in many situations, it is not possible to obtain HR image for e.g. supposes the document is scanned at low resolution and the original document destroyed. In this work, our objective is to construct an HR image, given a single LR binary image. The work reported in the literature mostly deal with super-resolution of natural images, whereas we try to overcome the spatial resolution problem in document images. Various methods proposed in literature are [3,4,6,9,14,16,17,20]. Note that all the works have shown their results on natural images. Since we are using only binary images for training and testing, we found that the above scheme does not improve the PSNR and OCR accuracy as much as our proposed simple and fast CNN architecture. In this work, we have used deep learning concepts to obtain a novel CNN architecture especially trained on binary Tamil document images with ReLU and PReLU activations, We have performed multiple experiments with various parameter settings and different architectures to obtain an effective approach for Tamil document image super-resolution.

II. Problem definition:

In digital world digitization of document images have important applications, consider a situation where the documents are scanned at a very low resolution to save memory

requirement and now, the ground truth is not available to perform the scanning again at high resolution. So we are left with only one choice of increasing the resolution of the already scanned document images.

Given a low-resolution image, we want to obtain a high-resolution image, for this, we have used a convolutional neural network and the architecture of the CNN is given in Fig 1. Since the images are document images (which has fewer features compared to natural images) we have optimally taken less number of filters for speed up and to avoid more redundancy because of the repeating nature of less varying filters. Once the model is trained means the weights are now stable, we can obtain a high-resolution image from the low-resolution input image. The size issue between the patches (LR and HR) have been taken care in the dataset. Loss function used in the CNN architecture is MSE and the model parameters are learned using standard back propagation and stochastic gradient descent with momentum.

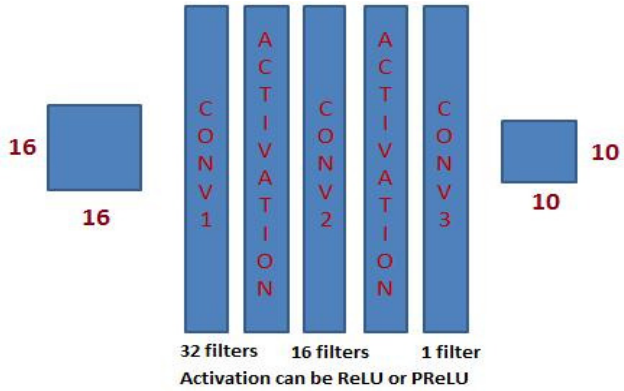


Fig. 1: CNN architecture used for obtaining high resolution image from a given LR image

III. Contributions:

Studied the performance of the OCR on different resolution of the document images and found the recognition can be improved by increasing the resolution of the image. This becomes the real problem of increasing the OCR accuracy on low-resolution Tamil document images. By performing extensive experiment on all recently available algorithms and techniques, we have tried to select (in terms of test time, higher PSNR and OCR accuracy) the CNN architecture suited for binary document image super-resolution. Our proposed CNN architecture for Tamil document image superesolution is fast and robust to resolutions. To show the robustness of the model the results of PSNR improvement and OCR CLA and WLA are given. We report the performance of our enhancement technique in terms of OCR accuracy for Tamil using MILE OCR [13]. It is important to note that the image is enhanced before passing it to the OCR and hence there is no need for any changes in the OCR design.

IV. Details of the CNN model:

Table I: 3-Layer convolutional neural network architecture designed for Tamil single document image SR with ReLU or PReLU

Layer	Type	Filter size and no.	stride	pad	output
conv1	$16 \times 16 \times 1$	$5 \times 5 \times 1 \times 32$	1	NIL	$12 \times 12 \times 32$
Act.fn.	$12 \times 12 \times 32$	$ReLU/PReLU$	1	NIL	$12 \times 12 \times 32$
conv2	$12 \times 12 \times 32$	$1 \times 1 \times 32 \times 16$	1	NIL	$12 \times 12 \times 16$
Act.fn.	$12 \times 12 \times 16$	$ReLU/PReLU$	1	NIL	$12 \times 12 \times 16$
conv3	$12 \times 12 \times 16$	$3 \times 3 \times 16 \times 1$	1	NIL	$10 \times 10 \times 1$

If the input to the convolution layer is of size $\mathbf{w_1 \times h_1 \times d_1}$ and at each layer we have four hyper parameters namely, the number of filters $\mathbf{n_f}$, the spatial extent filter $\mathbf{(s_e)}$, stride **(s)** and amount of zero padding **(zp);** then the output size is calculated according to the formula [24]:

$$\mathbf{w_2 = (w_1 - s_e + 2 \times z_p)/s + 1;\ \ h2 = (h_1 - s_e + 2 \times z_p)/s + 1 \text{ and } d_2 = n_f}$$

.

A. Initialization:

For the network to train well and not to become dead or irresponsive initialization is important. Suppose the network is initialized to a matrix of zeros the network will not update at all. And if it is initialized to a matrix of ones all neurons will try to extract similar features which are not desired. Out of the various initializations proposed we have used the initialization technique discussed in [21] : Var(Wi)=2/ni.

B. Activation function:

Various activation functions proposed in the literature in recent years, we are using ReLU [23] and PReLU [21] activations, which give better results for our specific task. Using ReLU as an activation avoids vanishing gradient problem in the network. Our network architecture is designed so that it avoids exploding gradient problem.

C. Dataset:

One of the important steps involved in training a deep neural network is the dataset availability, since Tamil document images super-resolution is not addressed in such a way, we have created our own dataset for this task. We wanted our super-resolution technique to work on multiple resolutions and hence we have taken a mixture of multiple resolutions of 135 document images for training and 18 for testing. From these images, we have randomly selected 5.1 million HR-LR patch pairs from the training images. LR image is obtained by down-sampling and up-sampling the HR image by a factor of 2. LR patches are 16x16 overlapping patches with a stride of 6 obtained from the LR image. And corresponding HR patches of size 10x10 taken from the HR image.

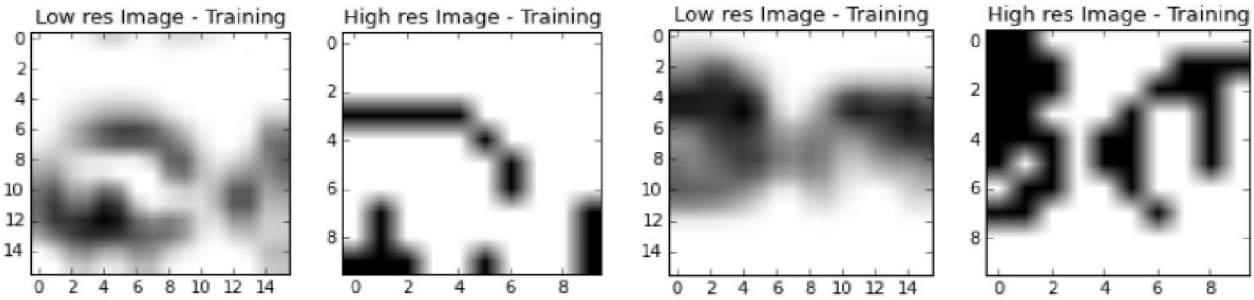


Fig. 2: Sample LR and HR patch pairs of dataset used to train the model.

We have taken proper care in such that HR-LR patches correspondence remains intact. The test data set and the training data set is different so that generalization is better.

Table 2 shows mean PSNR results for CNN_ReLU and CNN_PReLU methods on Tamil images scanned at different resolution. Figure 3 shows a segment of Tamil document image reconstructed using CNN_ReLU and CNN_PReLU. Table 3 shows the mean character level accuracy (CLA) and word level accuracy (WLA) of Tamil test images reconstructed using CNN_PReLU along with those of the LR input images. Figures 4 and 5 are the feature of a sample input Tamil character which is obtained by convolving the learnt filters of the CNN at the output of first and second layer of the CNN architecture. As can be visually perceived that the feature at first layer is less complex than the features at the second layer. The last layer has single filter which takes the weighted combination of the input and produces a single reconstructed HR output of the input features at this layer.

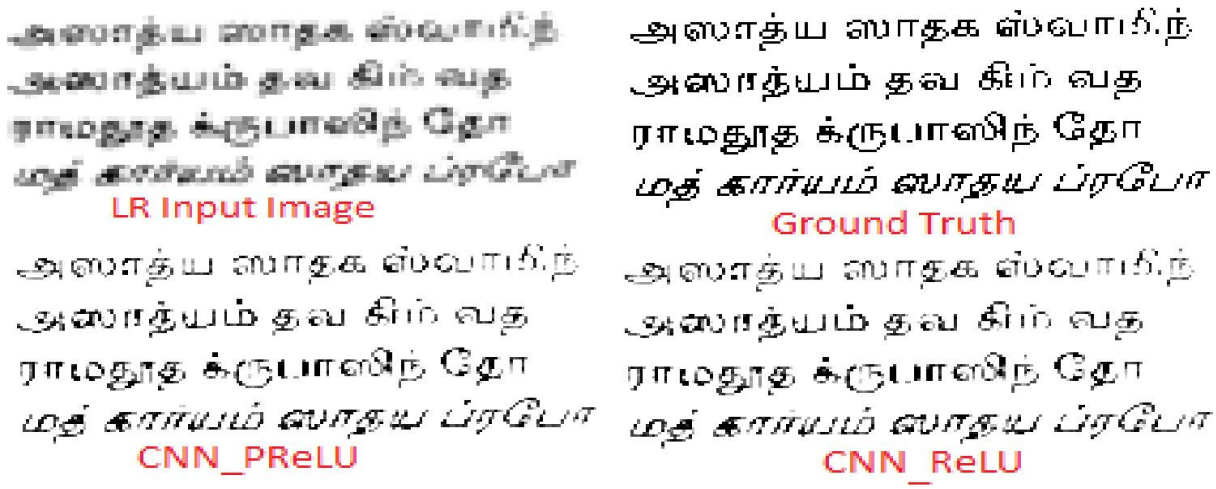


Fig. 3: Results with ReLU & PReLU activations on a segment of Tamil Binary image

Table 2: Mean PSNR of CNN_ReLU and CNN_PReLU

Resolution	LR_input	Bicubic	Yi.Ma [14]	CNN_ReLU	CNN_PReLU
Binary50dpi	14.17	14.74	14.95	**16.69**	**16.80**
Binary75dpi	15.08	15.62	15.88	**19.41**	**19.59**
Binary100dpi	16.40	17.28	20.10	**22.78**	**22.98**
Binary150dpi	17.55	18.36	21.65	**26.20**	**26.60**

Fig. 4: Total 32 features or the responses of filters at the output of first convolution layer.

Fig. 5: Total 16 features or the responses of filters at the output of second convolution layer.

Table 3: Mean WLA and CLA [14]

Metric	LR Input	CNN_PReLU
CLA	54.9	**90.2**
WLA	11.3	**54.0**

Table 4: Comparison of our results with Yi.Ma

Resolution	CNN_ReLU MET in sec.	Yi. Ma [14] MET in sec.	Speedup Factor
50 dpi	**1.6**	136	**85**
75 dpi	**3.1**	270	**87**
100 dpi	**3.9**	425	**109**
150 dpi	**14.8**	962	**65**

VI. Conclusion:

Our proposed approach to enhance the quality of Tamil document images resulted in mean PSNR improvements of 2.63, 4.51, 6.58 and 9.05 dB over LR images of 50, 75, 100 and 150 dpi, respectively. This improved the OCR character and word-level accuracy on these images as listed in Table 3. Our model is robust to different resolutions of input images and is lightweight the proof of robustness is the result listed in Table 4, and the proof of light weight is the execution time our model takes to obtain the HR image. Some of the input test images and the corresponding HR images are available at this URL [30]

REFERENCES

[1] Parker, J. Anthony, Robert V. Kenyon, and Donald E. Troxel, ”Comparison of interpolating methods for image resampling.” IEEE Transactions on medical imaging 2.1 (1983): 31-39.

[2] Yang, Chih-Yuan, Chao Ma, and Ming-Hsuan Yang, ”Single-image super-resolution: a benchmark.” European Conference on Computer Vision. Springer International Publishing, 2014.

[3] Chao Dong, Chen Change Loy, Kaiming He, Xiaoou Tang, ”Image super-resolution using deep convolutional networks.” IEEE transactions on pattern analysis and machine intelligence 38.2 (2016): 295-307.

[4] Jianchao Yang, John Wright, Thomas Huang, Yi Ma, ”Image super-resolution as sparse representation of raw image patches.” Computer Vision and Pattern Recognition, 2008.

[5] Nasrollahi, Kamal, and Thomas B. Moeslund, ”Super-resolution: a comprehensive survey.” Machine vision and applications 25.6 (2014): 1423-1468.

[6] Timofte, Radu, Vincent De Smet, and Luc Van Gool, ”Anchored neighborhood regression for fast example-based super-resolution.” Proceedings of the IEEE International Conference on Computer Vision. 2013.

[7] Timofte, Radu, Vincent De Smet, and Luc Van Gool. ”A+: Adjusted anchored neighborhood regression for fast super-resolution.” Asian Conference on Computer Vision. Springer International Publishing, 2014.

[8] Timofte, Radu, Rasmus Rothe, and Luc Van Gool. ”Seven ways to improve example-based single image super resolution.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[9] Huang, Jia-Bin, Abhishek Singh, and Narendra Ahuja, ”Single image super-resolution from transformed self-exemplars.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[10] Yang, Chih-Yuan, Chao Ma, and Ming-Hsuan Yang, ”Single-image super-resolution: a benchmark.” European Conference on Computer Vision. Springer International Publishing, 2014.

[11] Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, Thomas Huang, ”Deep networks for image super-resolution with sparse prior.” Proceedings of the IEEE International Conference on Computer Vision. 2015.

[12] Kim, Jiwon, Jung Kwon Lee, and Kyoung Mu Lee, ”Accurate image super-resolution using very deep convolutional networks.” Proc. IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[13] Shivakumar, H. R., and A. G. Ramakrishnan. ”A tool that converted 200 Tamil books for use by blind students.” Proc. 12-th International Tamil Internet Conf., Kuala Lumpur, Malaysia. 2013.

[14] Jianchao Yang, John Wright, Thomas S Huang, Yi Ma, ”Image super-resolution via sparse representation.” IEEE Transactions on image processing 19.11 (2010): 2861-2873.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron

Courville, Yoshua Bengio, ”Generative adversarial nets.” Advances in neural information processing systems. 2014. APA

[16] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, Thomas Huang, ”Coupled dictionary training for image super-resolution.” IEEE Transactions on Image Processing 21.8 (2012): 3467-3478.

[17] Christian Ledig, Lucas Theis, Ferenc Huszr, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, Wenzhe Shi, ”Photo-realistic single image super-resolution using a generative adversarial network.” arXiv:1609.04802 (2016).

[18] Johnson, Justin, Alexandre Alahi, and Li Fei-Fei, ”Perceptual losses for real-time style transfer and super-resolution.” European Conference on Computer Vision. Springer International Publishing, 2016.

[19] Hecht-Nielsen, Robert. ”Theory of the backpropagation neural network.” Neural Networks 1.Supplement-1 (1988): 445-448.

[20] Chao Dong, Chen Change Loy, Kaiming He, Xiaoou Tang, ”Learning a deep convolutional network for image super-resolution.” European Conference on Computer Vision. Springer International Publishing, 2014.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, ”Delving deep into rectifiers: Surpassing human-level performance on imagenet classification.” Proc. IEEE International Conference on Computer Vision. 2015.

[22] Glorot, Xavier, and Yoshua Bengio, ”Understanding the difficulty of training deep feedforward neural networks.” Aistats. Vol. 9. 2010.

[23] Nair, Vinod, and Geoffrey E. Hinton, ”Rectified linear units improve restricted boltzmann machines.” Proceedings of the 27th International Conference on Machine Learning (ICML-10). 2010.

[24] Karpathy, Andrej, F. Li, and J. Johnson, ”CS231n Convolutional Neural Network for Visual Recognition,” Online Course (2016).

[25] Team, The Theano Development, et al. ”Theano: A Python framework for fast computation of mathematical expressions.” arXiv:1605.02688 (2016).

[26] F. Chollet. Keras. https://github.com/fchollet/keras, 2015.

[27] Nielsen, Michael A. ”Neural networks and deep learning.” URL: http://neuralnetworksanddeeplearning. com/. (last visited: 25.07. 2017).

[28] Schmidhuber, Jrgen. ”Deep learning in neural networks: An overview.” Neural networks 61 (2015): 85-117.

[29] Ruder, Sebastian. ”An overview of gradient descent optimization algorithms.” arXiv:1609.04747 (2016).

[30] http://mile.ee.iisc.ernet.in/mile/SR_DocTamlImages.rar

Development of a Speech-Enabled Interactive Enquiry System in Tamil for Agriculture

S. Johanan Joysingh[1], M. Nanmalar[1], G. Anushiya Rachel[1], V. Sherlin Solomi1, V. Dhanalakshmi[2], P. Vijayalakshmi[1], T. Nagarajan[1]

[1]Speech Lab, SSN College of Engineering, Chennai, India,

[2]Tamil Virtual Academy, Chennai, India

Agriculture is one of the most prominent and quintessential occupations in the world. In India, more than half the population is involved in agriculture. Yet, there is a lack in productivity, measured in produce per hectare per year, compared to other countries such as Brazil, United States, and France [1]. This could be attributed to the fact that sufficient information about the state of the art methods and practices do not reach the farmers. Although information is available on various web portals, these cannot be fully utilized by farmers due to their lacking in technological aspects.

Speech technology aims at bridging the gap between man and technology by providing a natural interface between them. Providing such an interface to access agricultural data would mean that farmers can now obtain the latest information through a more natural way of communication. In the current work, we focus on building such an intuitive, time-independent and human-independent interface, to provide farmers in Tamil Nadu with agricultural information in their mother tongue, Tamil. Three crops are handled, namely rice, ragi, and sugarcane. Of these, rice has been given the utmost importance, since it is the most popularly grown crop.

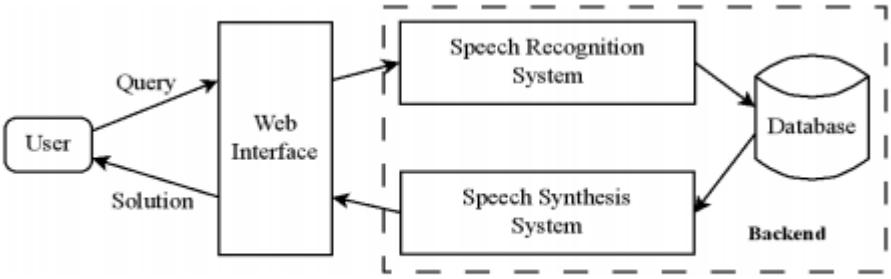


Fig. 1. Components of the enquiry system

Fig. 1 shows the components of the enquiry system. The main components are the speech recognition system that converts speech to text, and the speech synthesis system that converts text to speech. A conversational logic uses these two components along with a database to provide the user with the information they require. A web page acts as the interface between the user and the other components of the system. A typical flow would involve the following steps: (i) application is initialized by the user, (ii) a question in the form of speech is posed to the user, (iii) answer in the form of speech is obtained from the user, (iv) obtained answer is recognized, (v) if the speech is not recognized with confidence then the user is asked to confirm the recognized keyword, (vi) if recognized keyword is not the keyword intended by the user, process is repeated from step (ii) else system proceeds to the next step, (vii) recognized text is used to generate the next question, (viii) steps (ii) to (vii) are

repeated until the exact information required by the user is understood, (ix) finally the required information is synthesized. We refer to a sequence of one question and its corresponding answer to be a stage.

The questions are formulated such that they elicit one of 432 possible one-word responses. Therefore, in the current work, a triphone-based isolated word, limited vocabulary system is used. A triphone-based system provides better performance compared to a monophone system as it uses contextual information. To reduce confusion and increase recognition performance, multiple recognizers are used, one for each stage. To measure the confidence in recognition, as described in step (v) above, a likelihood-based threshold mechanism is used. In order to derive the threshold for each word, initially, all examples of each word (wi) in the training data are recognized by the appropriate recognition systems. The two best hypotheses (wi and wj) and the ratio of their corresponding log-likelihoods, p(Ow i,n |λwi) and p (Owi,n|λwj), is obtained. This ratio takes a value between 0 and 1, where a lower value indicates a higher level of confidence. The average of the ratios corresponding to all N examples of the same word is set as the threshold (Twi) for that word, as shown in Eq. 1.

$$T_{w_i} = \frac{1}{N}\sum_{n=1}^{N}\left[\frac{p(O_{w_{i,n}}|\lambda_{w_i})}{p(O_{w_{i,n}}|\lambda_{w_j})}\right]$$

Here, Owi,n|λwj) is the observation sequence corresponding the nth example of the word, wi, and λwi and λwj are the models corresponding to the words, w iand wj.

When a certain word is fed to the recognition system, the ratio between the first and second highest likelihoods is calculated. The first best recognition result is considered as the keyword. The ratio is compared with the threshold already calculated for that keyword. If the ratio is less than the threshold, then the decision is deemed to be correct and the system proceeds to the next stage. Otherwise, the system goes into the confirmation step as mentioned above. This approach is helpful in two ways. One, it avoids misdirection in the conversational flow as compared to a situation where there are no confirmations. Two, it reduces the time spent on each stage when the keywords are recognized with confidence, as compared to a situation where user replies are confirmed every time, irrespective of the confidence measure.

The user might answer in a phrase or sentence as well, to handle out of vocabulary words in such situations, garbage models [2] are used. A garbage model is a model that does not correspond to any particular word or its derivatives. Out of vocabulary words are usually random and are hence identified as garbage words and ignored.

A hidden Markov model (HMM)-based speech synthesis system (HTS) [3] developed with 5 hours of speech data is used to convert text to speech in this system. The advantages of using an HTS as opposed to recorded wav files are, (i) reusability - it can be used to synthesize any new sentence, hence modification of data is not an issue, and, (ii) reduced memory requirement - the size of the synthesizer is ~4MB, which corresponds to only approximately 4 wav files recorded at 48 kHz and 16-bit depth, and lasting ~10 seconds. HTS also offers better quality and lower memory requirement as compared to a unit selection synthesis (USS) system [4] which works by concatenating relevant contextual subword units to synthesize speech, instead of using a statistical parametric model.

In conclusion, the current work focuses on developing a speech-enabled interactive enquiry system to provide information related to agriculture. While existing systems described in [5] and [6] provide a speech interface for accessing agriculture commodity prices, in the current work, queries related to the production and protection of three crops,

namely, rice, ragi, and sugarcane are addressed. The current work also focuses on a conversational mode of accessing information. Further, the modularity of the application allows it to be expanded to other crops, with minimal effort.

References

1. "Agriculture in India", https://en.wikipedia.org/wiki/Agriculture_in_India (accessed on 15th April, 2017)
2. J. G. Wilpon, L. R. Rabiner, C.-H. Lee, and E. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden Markov models," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 38, no. 11, pp. 1870–1878, Nov 1990.
3. B. Ramani, S. Lilly Christina, G. Anushiya Rachel, V. Sherlin Solomi, M. K. Nandwana, A. Prakash, A. S. S, R. Krishnan, S. K. Prahalad, K. Samudravijaya, P. Vijayalakshmi, T. Nagarajan, and H. Murthy, "A common attribute based unified HTS framework for speech synthesis in Indian languages," in 8th ISCA Workshop on Speech Synthesis, Barcelona, Spain, pp. 311–316, Aug 2013.
4. B. Ramani, V. Sherlin Solomi, G. Anushiya Rachel, S. Lilly Christina, P. Vijayalakshmi, T. Nagarajan, and H. Murthy. "Development and evaluation of unit selection and hmm-based speech synthesis systems for tamil." In Communications (NCC), 2013 National Conference on, pp. 1-5. IEEE, 2013.
5. G. V. Mantena, S. Rajendran, B. Rambabu, S. V. Gangashetty, B. Yegnanarayana, and K. Prahallad, "A speech- based conversation system for accessing agriculture commodity prices in Indian languages," in Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA), pp. 153–154, May 2011.
6. S. Shahnawazuddin, K. Deepak, B. Sarma, A. Deka, S. Prasanna, and R. Sinha, "Low complexity on-line adaptation techniques in context of Assamese spoken query system," Journal of Signal Processing Systems, vol. 81, no. 1, pp. 83–97, 2015.

Deep neural network based medium vocabulary continuous speech recognition system for Tamil

Madhavaraj A, Ramakrishnan A G
Department of Electrical Engineering,
Indian Institute of Science, Bangalore
madhavaraj@mile.ee.iisc.ernet.in, ramkiag@ee.iisc.ernet.in

Abstract:

This paper presents our work on building a medium vocabulary continuous speech recognition (MVCSR) system for Tamil using neural networks. To build our automatic speech recognition (ASR) system, we have used 6.5 hours of Tamil speech recorded from 30 speakers covering a vocabulary size of 13,000 words. Of which, 4.5 hours of data was used for training, 1 hour of data for testing and 1 hour of data for cross-validation. We have built two independent recognition systems, one for phone recognition and the other for continuous speech recognition. Our system achieves phone-level accuracy of 75.1% and word-level accuracy of 96.5%.

I. Introduction:

In the last 40 years, we have seen steady progress in speech recognition research. This progress can be attributed to two factors: (i) the use of hidden Markov model (HMM) in modeling the temporal variations in speech and (ii) the increasing computational power of modern computers. In the past 10 years alone, we have seen many low-cost commercial interactive speech recognition applications developed by Apple, Microsoft, Amazon, etc. It is also well known that automatic speech recognition (ASR) research is mainly focused on English and other European languages. It can be said that no substantial progress has been made for Tamil speech recognition due to the unavailability of standard speech and text corpora. Our research focuses on overcoming these limitations to build a reasonably good Tamil LVCSR system.

Speech recognition researchers around the world have acknowledged the efficiency of deep neural networks (DNNs) in building ASR. DNNs trained on several thousand hours of speech, have reduced the word error rate significantly compared to traditional methods and achieve word-level accuracy of about 90% for vocabulary size of about 2,00,000 for English language. Such ASR systems are now widely used for commercial and entertainment purposes. Due to the lack of standard speech databases, ASR research in Tamil has not progressed at all. This motivated us to take up the task for building a domain and speaker-independent ASR system for a medium vocabulary task and later extend it for a larger vocabulary.

The rest of the paper is organized as follows; Section II describes the building blocks of an ASR system. Section III discusses about the tools and databases used in building our Tamil ASR system. In Section IV, we describe the steps in building a Tamil ASR. We provide the evaluation results of our phone recognition and continuous speech recognition

system in Section V. Finally, we conclude and briefly discuss our future research directions in Section VI.

II. Description of an ASR system:

The process of speech recognition involves many modules as described in Fig. I. The first step is to acquire speech signal through a microphone and convert it to digital format. The next step is pre-processing which involves noise removal and converting the signal to an overlapping frame sequence using windowing technique. The next step is to extract features from the frame sequence. Commonly used features are Mel-frequency cepstral coefficients and perceptual linear prediction coefficients. Then, these features are passed to the decoder which uses acoustic model (AM), language model (LM) and lexicon model (pronunciation dictionary) to decode the best possible word sequence.

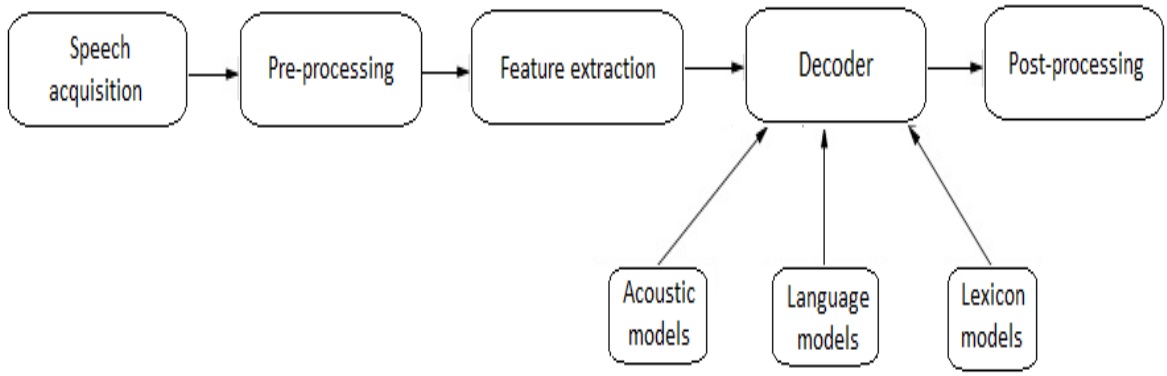


Fig. I: Block diagram of an ASR system

The acoustic model captures important attributes to characterize sound units (i.e., phones). Typically hidden Markov models (HMM) coupled with Gaussian mixture models (GMM) are used as acoustic models. Recently, deep neural networks (DNNs) have replaced GMMs providing a greater modeling power. Language model captures the likelihood of a word occurring after any other words. This is typically a bi-gram or tri-gram word model. Lexicon model bridges the gap between AM and LM. It is just a grapheme to phoneme conversion module, which maps each word in the vocabulary to its corresponding phone sequence.

Finally, the post-processing stage converts the recognized word sequence to human/machine readable format. For an ASR system to perform efficiently, we have to train AM on several hundred hours of transcribed speech corpus. Similarly, LM has to be trained on huge amount of text data.

III. Tools and databases used:

To develop our Tamil ASR system, we have used the state-of-the-art open source toolkit named Kaldi. It has numerous functionalities which can be used to build AM of our

desired choice. To build LM, we have used the toolkit named IRSTLM to compute the bi-gram word probabilities. We have built our own lexicon model (basically a grapheme to phoneme converter) to convert words to phone sequence. We have identified a total of 43 phones in Tamil language and our ASR system is built based on this phone set.

Learning an AM and LM requires transcribed speech corpus and text corpus respectively. We have obtained 6.5 hours of transcribed speech corpus from CIIL, Mysore. This corpus is a newspaper read-speech recorded from 30 speakers (18 male and 12 female). The entire corpus is divided into three chunks of size 4.5 hrs (training), 1 hour (development) and 1 hour (testing). The training set is used to learn the AM parameters and development set is used for validation purpose and the testing set is used for reporting the performance of our ASR system.

IV. Steps in building Tamil ASR:

To build an ASR system, we need speech utterances to be transcribed at phone-level but most often we have the utterances transcribed at sentence-level. As a flat-start, we convert the transcription to phone sequence and align them equally across time as shown in Fig. II. We then build a monophone system using these equal alignments and refine them iteratively. As the characteristics of a phone vary with respect to its preceding and succeeding phones, it is necessary to capture this context dependent variation for each phone and so triphone models are preferred to monophone models. The next step is to build a triphone model using the alignments obtained from monophone training. This step results in further refinement of the alignments. Then, we apply feature transformation on the input feature vector to further increase the accuracy of the system thereby further improving the alignments.

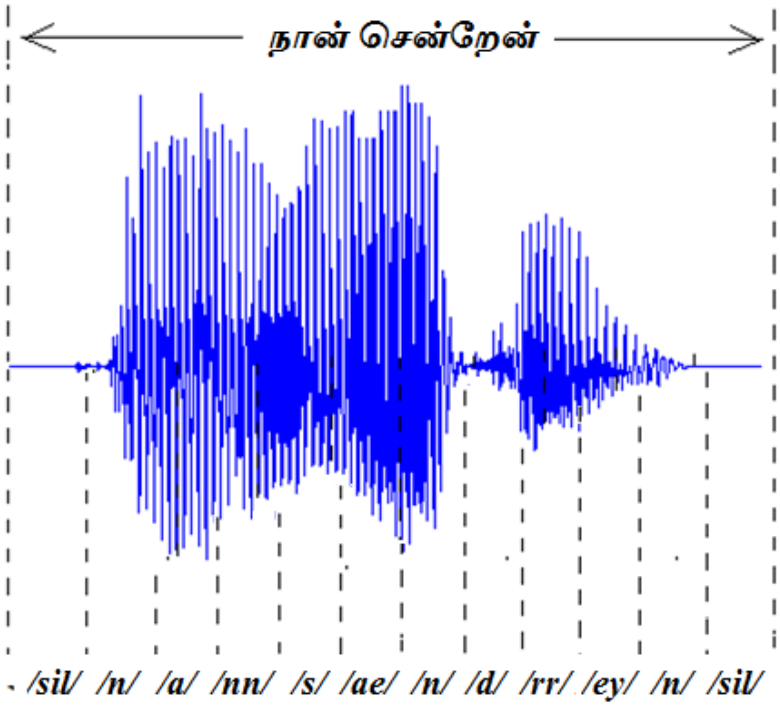


Fig. II: Converting text transcription to phone-level equal alignments

As the ASR task is intended to be speaker-independent, we need to suppress speaker information in AM and so a well known approach known as speaker adaptive training (SAT) is used to further refine the alignments. These alignments are then used to train the final deep neural network based AM. The DNN used here is a 7-layer fully connected feed-forward network, which is pretrained initially in an unsupervised manner

We have built two ASR systems: one for phoneme recognition and the other for continuous speech recognition and we have reported their performance in the next section.

V. Results:

We have achieved the best possible phone level accuracy of 75.1% using our phone recognition system and word level accuracy of 96.5% using our continuous speech recognition system.

VI. Conclusion and future work:

Thus, we have elaborated the steps involved in building a Tamil ASR system. With 6.5 hours of data, our system was able to achieve a word level accuracy of 96.5% for a vocabulary of 13,000 words. We plan to collect about 150 hours of speech and scale the vocabulary size to about 1,00,000 words to build a large vocabulary continuous speech recognition system. We are planning to use several hours of English training speech (which are readily available) for pre-training our Tamil ASR system to boost the accuracy further.

Acknowledgment:

The authors would like to thank Dr. L Ramamoorthy, CIIL, Mysore for providing us with Tamil speech and text corpus for our research purpose.

பார்வை மாற்றுத்திறனாளிகளின் தமிழ்மென்பொருள் பயன்பாட்டில் தேவைகள், சிக்கல்கள், தீர்வுகள்

விஜயராணி

ஆய்வுச்சுருக்கம் :

தகவல் புரட்சியுகத்தில் மிக வேகமாகப் பயணித்துக் கொண்டிருக்கிறோம். விரலசைவில் மொழிபேசும் பார்வை மாற்றுத் திறனாளிகள் கணினிசார் தொழில்நுட்பத்தைக் கையாளும் திறனையும் குறிப்பாக தமிழ்மென்பொருட்களைப் பயன்படுத்தும் நிலை பற்றிய நுட்பமான ஆய்வு இது. இதனைக் கணினி மொழிப் புரட்சி என்று கூறலாம். பா.மா.திறனாளிகள் தற்போதைய போட்டிச் சமுதாயத்தில் தங்களைத் தக்கவைத்துக் கொள்ள சமூகத்துடன் இணைந்து இசைந்து வாழ்ந்திட கணினியைப் பற்றி அறிவினைத் தங்களுக்குள் வளர்த்து வருகின்றனர். கணினிசார் மென்பொருட்களின் வருகை இவர்களின் வாழ்க்கைத்தரம் உயர உறுதுணையாக உள்ளது. அதில் குறிப்பாக தமிழ்மென்பொருட்கள், தாய்மொழிவழிக் கல்வி போல் பா.மா.திறனாளிகளிடம் தன்னம்பிக்கையை ஏற்படுத்தியுள்ளது. பா.ம.திறனாளிகளிடம் தொழிலடிப்படையில் பாகுபடுத்தி அவர்கள் பயன்படுத்துகின்ற தமிழ் மென்பொருட்களின் தேவைகள், சிக்கல்களும், சிக்கல்களுக்கான தீர்வுகளும் இக்கட்டுரையில் ஆராயப்பட்டுள்ளன.

குறிப்பு வார்த்தைகள் :

JAWS, NVDA, OCR, e.speak, talkback, shineplus, voiceover, webvisum

முன்னுரை :

காலை எழுந்தது முதல் இரவு துயிலும் வரை ஊடகங்களின் அசுரப்பிடிக்குள்தான் நாம் மூழ்கிக் கொண்டிருக்கிறோம். கணிப்பொறி குறித்த கல்வி தொடக்க காலத்தில் விந்தைமிகு பொருளாகத் தென்பட்டது. ஆனால் இன்று கணினிப்பொறி இல்லாத வீடுகள் இல்லை எனலாம். பார்வைத்திறன் படைத்தவர்களுக்கு நிகராக பா.மா.திறனாளிகள் பலர் கணிப்பொறி இயக்குவதில் வல்லன்மைத் திறனாளியாக இருக்கின்றனர்.

“பா.மா.திறனாளிகளை (Visually Challanged) 21 வகையாகப் பிரிக்கலாம்.”

(Vinodh Benjamin, Adjustment problem of partially visually sighted in their work part, M.Phil., School of Social Work, Chennai, 2001)

இவர்கள் இயல்பான பார்வைத் திறனுடையவர்களுக்கு (Sighted People) நிகராக, பல நிலைகளில் சிந்தனைச் சிதறல் இன்றிக் கவனக்குவிப்பின் காரணமாக மிகுந்த நினைவாற்றல் மிக்கவர்களாக உள்ளனர். பா.மா.திறனாளிகளில் பலர் கணிப்பொறியைக் கையாளும் திறனைப் பயிற்சியின் வாயிலாகப் பெற்றுள்ளனர். பா.மா.திறனாளி குழுக்கள் மற்றும் பல்கலைக்கழகங்கள் இவர்களுக்கு இப்பயிற்சியினை 3-5 நாட்கள் வழங்குகின்றன. இவர்களுக்கு என்று தனிப்பட்ட கணிப்பொறிகளோ திறன்பேசி (Smart Phone) களோ இல்லை. இவர்கள் சில மென்பொருட்களை உள்ளீடு செய்து கொண்டு பயிற்சியின் வாயிலாக ஆர்வமுடன், தேடலுடன் கற்றுத் தேர்ந்து வருகின்றனர். ஒவ்வொரு மாவட்டம் வாரியாக பா.மா.திறனாளிகளுக்கு என்று சட்டரீதியான குழுக்கள், அமைப்புகள் உள்ளன, ஒவ்வொரு கணிப்பொறி மற்றும் திறன்பேசியிலும் வன்பொருட்கள், மென்பொருட்கள் உள்ளன, பா.மா.திறனாளிகள் மென்பொருட்களை அதிலும் குறிப்பாக தமிழ் மென்பொருட்களின் பயன்படுத்துகின்றனர். இவற்றுள் மிக முக்கியமானவை திரைவாசிப்பான் (JAWS, NVDA), எழுத்துணரி (OCR), பேச்சொலிப்பான் (e.Speak) பயன்படுத்துகின்றனர்.

முதன்மைத் தகவலாளிகள் (Primary Source) :

இந்த ஆய்விற்கு தமிழகத்தில் பல்வேறு மாவட்டங்களைச் சேர்ந்த 64 பா.மா.திறனாளிகள் முதன்மைத் தகவலாளிகளாக இடம்பெற்றுள்ளனர். இவர்களில் சிலரை நேரில் சந்தித்து அவர்கள் கணிப்பொறி மற்றும் அலைபேசியை இயக்கும் திறத்தைப் பார்த்து தரவுகளைச் சேகரித்தும் பலரை அலைபேசி வாயிலாகப் பேட்டி கண்டும் தரவுகளைக் கட்டுரையாளர் சேகரித்தார். பா.மா.திறனாளிகளின் தமிழ்மென்பொருட்களைப் பயன்படுத்துவதில் உள்ள சிக்கல்களைக் கணிப்பொறி அறிஞர்கள் நிறைந்த அரங்கத்தில் எடுத்துரைத்து அதற்கான தீர்வுகளுக்கு வழிவகுக்க வேண்டும் என்ற எண்ணத்தின் அடிப்படையில் இக்கட்டுரை அமைகின்றது.

மேற்கோள் கட்டுரைகள் :

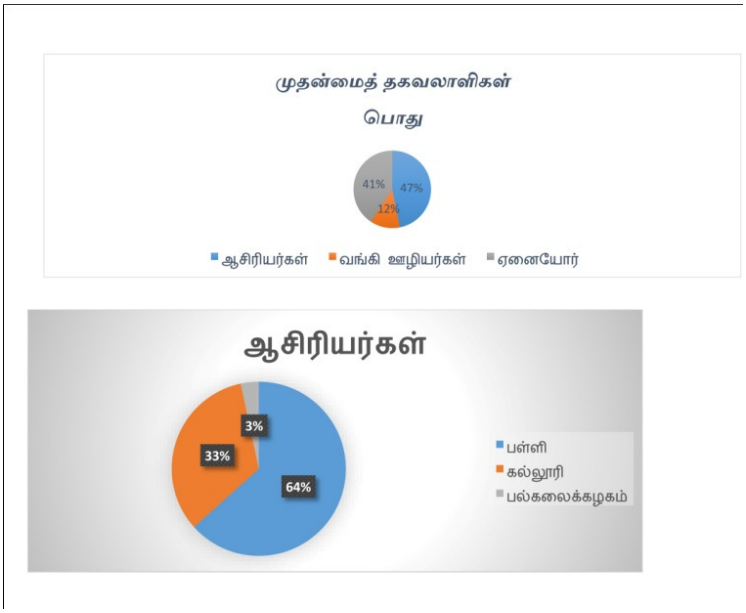
1. கு.ஞானகுரு, என்விடிஏ திரைவாசிப்பான் ஓர் அறிமுகம், 14.11.2015, தாகூர் கலைஅறிவியல் கல்லூரி, புதுச்சேரி.

2. பாண்டியராஜன், பார்வை மாற்றுத் திறனாளிகள் பயன்பாட்டில் தமிழ்க்கணினியும் செல்பேசியும், உத்தமம் 2013, புதுச்சேரி.
3. ஜி. ஞானவேல்முருகன், வெற்றிப்படிக்கட்டுகளில் பார்வையற்ற மாற்றுத் திறனாளிகள், விரல்மொழியர்-முகநூல்.
4. மகேஷ், பார்வையற்றவர்கள் கணினி பயன்படுத்துவது எப்படி?, விரல்மொழியர் 13.11.2015
5. ஷாஜகான், தொட்டால் பூமலரும், தொடாமல் தமிழ்மலரும், malaigal.com, 2014
6. பொன்விழி, ஒளிவழி எழுத்துணரி, 19.9.2005
7. அந்தகக்கவி பேரவை, பத்மசேஷாத்திரி மேல்நிலைப்பள்ளி, சென்னை, பா.மா.திறனாளிகள் கூட்டம், பிரதிஞாயிறு 1.30-3.00
8. www.thedroidlibrary.com/best-10android-apps-forthe-visually-impaired/
9. Visually impaired man makes software for blind http://igg-me/at/dft-org/x/12679366
10. http://www.nvaccess.org Non-visual Desktop Access Wikipedia
11. The Challenges of using technology when you're blind, Bruce Maguire

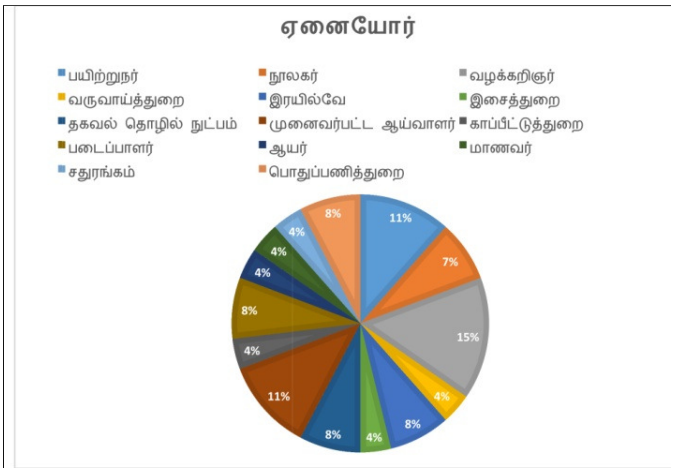
மேல்குறிப்பிட்டுள்ளவற்றுள் 1-7 வரையிலான கட்டுரைகள், பா.மா. திறனாளிகள் கணிப்பொறியை, சில மென்பொருட்கள் பயன்படுத்தும் விதம் பற்றிக் குறிப்பிடுகின்றன. ஏனையவை இந்த ஆய்விற்கு வலுசேர்க்கும் விதமாக தகவல்களைத் தந்து உதவுகின்றன. இவ்வாய்வின் பொருண்மை குறித்த கருத்திற்கு உத்தமம் அமைப்பின் வழி ஏதேனும் விடியல் கிடைத்திட வகை செய்திட வேண்டும் என்ற நோக்கத்தில் இங்கு இனிவரும் செய்தி பதிவு செய்யப்படுகின்றது.

முதன்மைத் தகவலாளிகள் பகுப்பு முறை :

கணினி பயன்படுத்துகின்ற பா.மா.திறனாளிகள் (ஆய்வுத்தகவலாளிகள்) அனைவரிடமும் முன்னதாகவே தயாரிக்கப்பட்ட வினாநிரல்களின் அடிப்படையில் தரவுகள் சேகரிக்கப்பட்டன. இவ்வாய்வில் பகுப்புமுறைத் திறனாய்வும், ஒப்பீட்டு முறைத் திறனாய்வும் தகவலாளிகள் கூறுகின்ற கருத்துக்களை விளக்குவதற்கு விளக்கமுறைத் திறனாய்வும் மேற்கொள்ளப்பட்டுள்ளன.



பா.மா.திறனாளிகள் 64 நபர்களைப் பேட்டி கண்டபோது அதில் கல்வித்துறை சார்ந்தவர்களே 47% உள்ளனர். இதற்கு அடுத்து வங்கி ஊழியர்கள் 12%, ஏனைய என்ற நிலையில் 14 துறைகளைச் சார்ந்த பா.மா.திறனாளிகள் 41% உள்ளனர். கல்வித்துறையில் பள்ளி ஆசிரியர்கள், தமிழ், ஆங்கிலம், வரலாறு, சமூக அறிவியல் ஆகிய பாடங்களைப் பயிற்றுவிக்கும் ஆசிரியர்களாகவே உள்ளனர். அறிவியல், கணிதம் கற்பிக்கின்ற ஆசிரியர்கள் யாரும் இல்லை. இதுபோலவே கல்லூரிகளிலும் தமிழ், ஆங்கிலம், வரலாறு, பொருளாதாரத்துறை சார்ந்தவர்களே பணிபுரிகின்றனர். பல்கலைக் கழகப் பேராசிரியர் தமிழ்த்துறை சார்ந்தவர். கற்பித்தல் பணிதான் இவர்களுக்கு ஏற்புடையதாக உள்ளது. (பள்ளி (17) கல்லூரி (10) பல்கலை (1)) இதற்கு அடுத்து வங்கிப்பணி செய்பவர்களில் பெரும்பான்மையும் தொலைபேசி இயக்கும் பணி (Telephone Operator-Mrs.Latha, Mr.Subharao) கடன் கொடுத்தல், கடன் வாங்குதல் குறித்த விவரங்களைத் தொலைபேசி வழியாகப் பரிமாறிக் கொள்ளும் ஊழியராக (Clerk) இணையம் மூலம் வாடிக்கையாளருக்குக் கடன் விசாரணைகளைச் செயல்முறைப்படுத்துபவர்களாகப் பணிபுரிகின்றனர். வங்கி கிளை மேலாளராக உள்ள இருவரும் கூட (முத்துச்செல்வி, விஜயராமன்) (Loan) கடன் விசாரணைப் பணியில்தான் அமர்த்தப்பட்டுள்ளனர். பணப்புழக்கம் யார்வசமும் அனுமதிக்கப்படவில்லை.



ஏனையோர் என்ற பாகுபாட்டில் பா.மா.திறனாளிகள் மேற்குறிப்பிட்டுள்ள பல துறைகளிலும் முக்கியப் பணிகளில் பணிபுரிகின்றனர். இந்திய ஆட்சிப்பணி (*IAS), இந்திய அயலகப்பணி (*IFS) – பா.மா.திறனாளிகள் பணிபுரிகின்றனர். *இவர்கள் இருவரையும் தொடர்புகொள்ள இயலவில்லை.

பா.மா.திறனாளிகளிடம் பெறப்பட்ட தகவல்கள் முதன்மை ஆதாரமாகவும், இணையத்தில் வெளிவந்துள்ள கட்டுரைகள், வலைப்பூக்கள், மின்குழுமம் அளித்த கட்டுரைகள், விக்கிபீடியா செய்திகள், துணைமை ஆதாரமாகவும் எடுத்தாளப் பெற்றுள்ளன.

தமிழ் மென்பொருள் பயன்பாட்டில் சிக்கல்கள் :

(களஆய்வின் பெறப்பட்ட தகவல்கள் **)**

1. **திரைவாசிப்பான் (Screen Reader)**

JAWS (Job Access With Speach) : சர்வதேச அளவில் அமெரிக்க, ஐக்கிய, ஆங்கில மொழிகளில் வாசிக்கப் பயன்படுகிறது.

குரலில் ஏற்றத்தாழ்வு, உச்சரிப்பு அழுத்த உணர்ச்சிக் குறியீடுகள், மனித குரலில் (Human Voice) உள்ளது. இடைவெளிவிட்டு வாசித்தல் இதில் உண்டு. கேட்டால் புரிந்துகொள்வது எளிது. அதிக விலைகொடுத்து வாங்க வேண்டும்.

‘வாணி’ தமிழ்மென்பொருள் மனித குரலில் ஏற்ற இறக்கமின்றிச் சலிப்புத்தட்டும் வகையில் சமீபத்தில் வெளிவந்துள்ளது.

NDVA (Non Visual Desktop Access) : இந்திய மொழிகளில் குறிப்பாகத் தமிழ்மொழியிலும் கிடைக்கிறது. ஆரம்பம் முதல் இறுதி வரை ஒரே ஒலிஅழுத்தத்தில்

உணர்ச்சிகளோ இடைவெளியோ இன்றிச் சலிப்புத்தன்மையுடன் ஒலிக்கின்றது. இயந்திரக்குரலில் (Robotic Voice) உள்ளது. ஆரம்பப் பயனாளிகள் புரிந்துகொள்வது சிரமம். இலவசமாகப் பதிவிறக்கம் செய்துகொள்ளலாம். JAWSஐ Crack செய்து பதிவிறக்கம் செய்யலாம்.

தீர்வு : NVDA பா.மா.திறனானிகளுக்குக் கிடைத்த வரப்பிரசாதம். மனிதக்குரலில் உச்சரிப்பு, உணர்ச்சி அடிப்படை வேறுபாடுகளுடன் கிடைத்திட வகைசெய்திட வேண்டும்.

Basha.indi.com : இதில் Indi input software தமிழில் உள்ளது. ஒலியியல் அடிப்படையில் வேலை செய்கிறது. மத்திய அரசு நிறுவனங்களில் பயன்படுத்தப்படுகிறது. அரசு ஆணை வேறுவேறு வடிவமைப்பில் (Format) வருவது வாசித்தறிய சிரமமாக உள்ளது. – ஜி.கண்ணன்.

2. **'Pdf'** ஆக உள்ள கோப்புகளை பா.மா.திறனாளிகளால் வாசிக்க இயலவில்லை. இதற்கு 'tts' பயன்படுத்த இயலவில்லை.

ஒருங்குறி எழுத்துக்களை மட்டும் இத்தமிழ் மென்பொருட்களின் வழி வாசித்து உணர முடிகின்றது.

⇒ pdf கோப்புகளை e.speak மூலம் வாசிப்பதற்கு வழிவகை செய்திட வேண்டும்.

3. ஒருங்குறி எழுத்துக்கள் மட்டும் அனைத்துத் துறைகளிலும் பயன்படுத்தப்பட வேண்டும்.

மதுரைத்திட்டம், தமிழ் இணைய கல்விக்கழகம், தம் தரவுகளை முற்றிலும் ஒருங்குறியில் பதிவிட்டிருப்பதால் பா.மா.திறனாளிகள் படிக்க, பயன்பெற மிகவும் உதவியாக இருப்பதாகக் கருத்துத் தெரிவித்துள்ளனர்.

⇒ அரசாங்கம் தற்போது நடைமுறைப் பயன்பாட்டில் வானவில், ஒளவையார் எழுத்துருக்களை மட்டும் பயன்படுத்தி வருகிறது, இதற்குப் பதில் 'லதா' 'பாமினி' போன்ற ஒருங்குறி எழுத்துருக்களைப் பயன்படுத்த அரசு சட்டரீதியாக நடவடிக்கை எடுக்க வேண்டும். 2010ஆம் ஆண்டு தமிழக அரசு அனைத்துத் துறைகளிலும் ஒருங்குறி எழுத்துக்களைப் பயன்படுத்த வேண்டும் என்று அறிவித்த போதிலும் கூட அரசாங்கமே இன்னும் அதைச் சரிவர கடைபிடிக்கவில்லை.

3. **Vocalizer :** இந்த மென்பொருள் 'JAWS' இல் உயர்ந்த மொழிநடையில் உள்ளது. தமிழ் மென்பொருளில் சில குரல் வேறுபாடுகளுடன் மட்டுமே உள்ளன.

⇒ அதிகபட்சமான குரல் வேறுபாடுடைய vocalizer தமிழில் நிறுவப்பட வேண்டும். இதற்கு அரசாங்கமும் கணினித்துறை சார்ந்த வல்லுநர்களும் முயற்சி செய்து புதிய பல மென்பொருட்களை உருவாக்கிட வேண்டும்.

4. எழுத்துணரிகள் எண்ணிக்கையை மிகக் குறைவாக உள்ளன. OCR-தமிழ் என்ற மென்பொருள் சீர்திருத்தம் செய்யப்பட்டு பல நிலைகளிலும் பயன்படுத்தப்பட வேண்டும். எழுத்துணரியில் கவிதை, உரைநடை, செய்யுள் ஆகியவற்றை வாசிக்கும்பொழுது புரிதலுக்குத் தகுந்தாற்போல வேறுபடுகிறது. மடிக்கணினியில் இந்த வேக வேறுபாட்டினை ஏற்படுத்திக் கொள்ளலாம். திறன்பேசிகளில் வேகமாறுபாடு செய்வது சிரமமாக உள்ளது. இந்தக் குறைபாடு சரிசெய்திடப்பட வேண்டும்.

Google OCR ஒவ்வொரு பக்கமாக மட்டுமே வாசித்தறிய முடிகிறது. தமிழ் OCR ல் தொடர்ந்து வாசித்திட வழிசெய்திட வேண்டும்.

⇒ அதிக எண்ணிக்கையிலான எழுத்துணரிகள் அதிகம் பயன்பாட்டிற்கு வர வேண்டும்.

5. **Shortcut Keys :** சுழலி/சொடுக்கும்பொறி கொண்டு தொடாமல் இருகரங்களையும் விசைப்பலகையைச் சுற்றி நகர்த்தி பா.மா.திறனாளிகள் shortcut keys பயன்படுத்துகின்றனர். கரூர் அரசு கலைக் கல்லூரி பேரா.முனைவர் K.சரவணன் விசைப்பலகையில் தானாக சில மாற்றங்களைத் தொழில்நுட்ப வல்லுநருடன் இணைந்து செயல்பட்டு இனஎழுத்துகளை எளிமையாக சுருக்கமுறையில் (Shortcut Key) உருவாக்கியுள்ளார்.

U - ங்க, O ^ - ஞ்ச, Y ^ - ண்ட, I - ந்த,

U ^ - ம்ப, 2 ^ - ன்ற, ; ^ -ய்ர், Ctrl ^ - .com

இவை போன்ற பல சுருக்கக் குறியீடுகள் அதிகம் உருவாக்கப்பட வேண்டும்.

6. தமிழ் எழுத்துக்களுக்கான இடஒதுக்கீடு, தமிழில் ஒருங்குறி எழுத்துக்கான இடம் மிகக் குறைவாக உள்ளது. (ஆங்கிலத்திற்கு அதிகம்) தமிழில் உள்ள எழுத்துக்களின் எண்ணிக்கை மற்றும் சொற்கள் மிக அதிக அளவில் இருப்பதாலும் பல எழுத்துக்கள் 3,4 ஒலியன்களைக் (தங்கம், மதன், காகம் – g, h, k) கொண்டிருப்பதாலும் இடம் அதிகம் ஒதுக்கினால் நலமாக இருக்கும்.

7. **குரல் தேடல் வசதி :** Google voice search என்ற வசதி ஆங்கில மொழியில் தற்போது கிடைக்கின்றது. இந்த வசதி தமிழிலும் கிடைத்தால் நன்றாக இருக்கும் என்று பல

பா.மா.திறனாளிகள் கருத்தினை முன்வைத்தனர். இதற்குத் தமிழில் லளழ, ரற, ணநந உச்சரிப்பு வேறுபாடு முக்கியமாகப் பயிற்சியளிக்கப்பட வேண்டும்.

8. **நூலகம் :** Kindle Amazon Library தமிழில் கொண்டு வரவேண்டும். bookshare.org அமெரிக்க நிறுவனம் இந்திய மாற்றுத் திறனாளிகளுக்கு இலவசமாக மின்புத்தகங்களை 'Digital Library' மூலமாக வழங்குகிறது. ஆனால் அமெரிக்கர்களுக்கு 50 டாலர் கட்டணம் வசூலிக்கிறது. நம் தமிழ்நாட்டிலும் பா.மா.திறனாளிகளுக்கு அரசாங்கம் இதுபோல் சலுகைகள் செய்திட வேண்டும்.

Webvisum : பா.மா.திறனாளிகள் பொருள் வாங்குதல், வரிகட்டுதல், பதிவுசெய்தல், படிவத்தினை நிரப்புதல் போன்ற சூழல்களில் இறுதியாக captcha என்ற பகுதி வரும். அப்போது கட்டத்திற்குள்ள எண் & ஆங்கில எழுத்தினை இவர்களால் வாசிக்க இயலாது. இதற்கு ஒரு சில நிறுவனங்கள் மாற்றுவழிகள் செய்துள்ளன. IRCTC-OTP, Mozilla Firebox-ல் மட்டும் webvisum சீரமைத்துப் புதுப்பிக்கப்பட்டுள்ளது. Firebox 35ல் இந்த வசதி உள்ளது. Firebox 55ல் இந்த வசதி இல்லை.

LIC : சில கணக்கீடுகள் மூலம் இப்பிரச்சினைக்குத் தீர்வு காணுகிறது. சான்றாக,
what is your product of 5 x 6 =□ பா.மா.திறனாளி 30 எனப் பதிவிட வேண்டும்.

பா.மா.திறனாளிகளுக்கு captcha சவாலாக உள்ளது. இதற்கான தீர்வு பொதுமைப்படுத்தப்பட வேண்டும்.

abbyefine reader இந்த திரைவாசிப்பான் ஆங்கிலத்தில் மட்டும் உள்ளது. இதன் வாயிலாக படங்கள் (ppt, scan, image) வாசித்து அறிந்திட முடியும். இதேபோல் தமிழிலும் ஒரு திரைவாசிப்பான் அவசியம் தேவை.

குறுஞ்செயலிகள் : ஆண்ட்ராய்ட் அலைபேசிகளில் பா.மா.திறனாளிகளின் பயன்பாட்டிற்கு என்று பல குறுஞ்செயலிகள் ஆங்கிலத்தில் கிடைக்கின்றன. ஆனால் ஒருசில மட்டுமே தமிழில் கிடைக்கின்றன எனத் தகவலாளிகள் குறிப்பிட்டனர். பா.மா.திறனாளிகள் மிக அதிக அளவில் பயன்படுத்தும் தமிழ் குறுஞ்செயலி talk back என்பதாகும்.

Eye-D : பா.மா.திறனாளிகளுக்கு அவர்களைச் சுற்றியுள்ள இடங்கள், பொருட்கள் ஆகியவற்றைத் தங்கள் அலைபேசி, புகைப்படக்கருவி மூலம் அச்சிடப்பட்ட எழுத்துக்களை வாசித்தறிய உதவுகிறது (Where am I, Around me, See object, Read

object). இது ஆங்கிலமொழியில் மட்டும் உள்ளது. தமிழ்மொழியிலும் வந்தால் நலமாக இருக்கும்.

முடிவு :

பா.மா.திறனாளிகளுக்கு PDF வடிவில் அமைந்த பக்கங்களை வாசித்தறிந்திட தகுதியான, தரமான Convertor தேவை. அரசாங்கம் எங்கும் எதிலும் தமிழ் என்று ஏட்டளவில் கூறாமல் செயல்முறையில் ஒருங்குறி எழுத்தினைக் கடைபிடித்திடக் கட்டாய அரசாணை பிறப்பித்திட வேண்டும். 'வாணி' மென்பொருள் மறுசீரமைவு செய்து பா.மா.திறனாளிகள் பயன்படுத்திட வழிவகை செய்திட வேண்டும். தமிழில் சிறந்த முறையில் இயங்கும் எழுத்துணரி (OCR) நடைமுறைப்படுத்த வேண்டும். நோத்தோ திட்டம் 800க்கும் மேற்பட்ட மொழிகளுக்கான எழுத்துருக்களை அறிவித்துள்ளது. http://sellinam.com/archives/1289. எந்த நூலை எடுத்தாலும் படிக்கும் அளவிற்குத் தொழில்நுட்பம் எளிமைப்படுத்தப்பட வேண்டும்.

நன்றியுரை :

இந்தக் கட்டுரை எழுதுவதற்குத் தேவையான தரவுகளை நேரிலும், அலைபேசி வாயிலாகவும் தந்து உதவிய பா.மா.திறனாளிகள் வினோத் பெஞ்சமின், சபாஷ், பார்த்திபன், ஆயர் மோகன்ராஜ் பீட்டர், பயிற்றுநர் சுப்பாராவ், தினகரன் (Indian Railway Service), A.நாகேந்திரன், பாலநாகேந்திரன் (Indian Revenue Service), முனைவர் கே.சரவணன், முனைவர் ஜி.கண்ணன், முனைவர் A.கண்ணன் ஆகியோருக்கு மனமார்ந்த நன்றியை உரித்தாக்குகிறேன். பல பா.மா.திறனாளிகளை எனக்கு நேரில் அறிமுகம் செய்து வைத்து ஆரம்பம் முதல் இறுதி வரை தரவுகள் திரட்டுவதில் உறுதுணையாக இருந்த சகோதரி கோமதி (TCS)க்கு ஆத்மார்த்தமான நன்றியைத் தெரிவித்துக் கொள்கிறேன்.

குறிப்பு : பா.மா.திறனாளிகள் – பார்வை மாற்றுத்திறனாளிகள்

References :

1. 64 தகவலாளர்களின் முழுவிவரப்பட்டியல் முழுக்கட்டுரையில் பின்னிணைப்பாக இடம்பெறுகிறது.

2. www.nvaccess.org
3. www.sethuparoa.blogspot.in
4. ksnauthusri.wordpress
5. முகநூல் பக்கம் : விரல்மொழியர்
6. en.wikipedia.org/wiki/non_visual_desktop_access
7. வள்ளுவன் பார்வை, இணைய தென்றல் – மின்குழுமங்கள்
8. www.thedroidlibrary.com/best-10android-apps-forthe-visually-impaired/
9. www.blindhelp.net

Enhancement of noisy Tamil speech for improved quality of perception for the hearing impaired

K V Vijay Girish, A G Ramakrishnan
MILE Laboratory, Department of Electrical Engineering,
Indian Institute of Science, Bangalore 560012, India
Email: kv@ee.iisc.ernet.in, ramkiag@ee.iisc.ernet.in

ABSTRACT

Noisy Tamil speech was enhanced using dictionary based source separation approach. The enhanced speech was perceptually evaluated by seven Tamil natives and judged to be significantly improved in intelligibility (MOS score of 3.5). Noisy speech is simulated by adding different noises from the NOISEX database to Tamil utterances at different signal to noise ratios. Noise dictionaries were built using random selection method. Tamil speech was derived from different speakers and speaker dictionaries were also built with 1000 atoms in each dictionary. First, the noise source was identified employing a selected subset of frame-wise features from the test data using signal to distortion ratio (SDR) measure with active set Newton algorithm (ASNA). Then, the speaker was identified, once again using a subset of high energy features of the test data from a concatenated dictionary comprising all the speaker dictionaries and the estimated noise dictionary.

Index Terms: Dictionary learning, cosine similarity, audio classification, source recovery, sparse representation.

INTRODUCTION

[17] Motivation for the present study

Most of the hearing aids [1] do amplification of the sounds that reach the ear, while some of them also compensate (equalize) for the specific auditory frequency response of the person with hearing disability (PWHD). Hence, when the speech contains a significant amount of background noise, the perception is further affected. While this issue is true also for the people with normal hearing, the effect is more pronounced for PWHD. Thus, a system is highly desirable, which possesses the capability of separating the speech from the noise. Here, we report our work on a system that identifies the noise type (among a set of few expected noise sources) as well as the speaker (from a known set of speakers that the person normally meets with during his daily life) and then extracts the speech from the noisy speech, known as speech enhancement. It has been observed that identifying noise and speaker type improves the speech enhancement performance.

B. Literature review

Dictionary learning is a method of represent-ing features from a large training data using a weighted linear combination of vectors called as atoms. Estimating weights corresponding to

these atoms is termed as sparse coding or source recovery. Audio signals are represented as a linear combination of non-negative dictionary atoms for audio source separation [2], [3], [4], recognition [5], [6], classification [7], [8] and coding [9], [10]. The simplest dictionary learning (DL) method is a random selection of features from the training data [11]. Matching pursuit [12], orthogonal matching pursuit (OMP) [13], focal underdetermined system solver (FOCUSS) [14] and basis pursuit [15] are some of the source recovery algorithms.

Noise classification can be seen as a first step in machine listening [16], which enables the system to know the background environment. Classification of noise types has been reported in the case of pure noise sources. Kates [17] addressed the problem of noise classification for hearing aid applications based on the variation of signal envelope as feature. Maleh et al. [18] used line spectral frequencies as features for classification of different kinds of noise as well as noise and speech classification. Casey [19] proposed a system to classify twenty different types of sounds using a hidden Markov model classifier and a reduced-dimension log-spectral features.

Sparsity based speaker identification using discriminative dictionary learning was done by Tzagkarakis et al. [20] while non-negative matrix factorization for feature extraction was explored by Joder et al. [21]. Representation of audio signals as a sparse, linear combination of non-negative vectors called as dictionary atoms has been used for audio source separation [2], [3], [4], recognition [5], [6], classification [7], [8] and coding [10], [9].

We have used active-set Newton algorithm (ASNA)[11] for source recovery. The training phase for the classification problem is DL from various speaker/noise sources. The dictionary atoms capture the variation in the spectral characteristics of the speech and noise sources.

PROPOSED APPROACH

Additive noise model is assumed and one of the noise sources is added at a time to each of the Tamil utterances with different signal to noise ratios, by scaling the energy of the noise segment appropriately.

$$s[n] = s_s[n] + s_n[n] \qquad (1)$$

where, $s_s[n]$, $s_n[n]$ and $s[n]$ are the clean Tamil utterance, noise signal and the simulated noisy speech, respectively. Figure 1 shows the speech utterance spoken by a Tamil speaker with a pitch frequency of 145 Hz, babble noise and noisy speech signal at an SNR of 0 dB.

A. Feature extraction and dictionary learning

From the training set of speech and noise sources, frames of 60 ms duration are extracted with a shift of 15 ms. The magnitude of the short-time Fourier transform of these frames are used

as the features. A dictionary is a matrix D ∈ $IR^{P\times K}$ containing K column vectors denoted as atoms, d_k, $1 \leq k \leq K$. A given feature vector y can be represented as a linear combination of a few dictionary atoms as $y \approx Dx$, where x is the vector of weights for the atoms. Features for each speaker and noise source are extracted separately and the corresponding dictionaries are built. Dictionary is learnt by random selection of N (= 1000) features as atoms. Noise and speaker dictionaries are built from the training data, using the random selection dictionary learning algorithm. $D_{is} = [d_{i1}\ d_{i2}..d_{iN}]$ and $D_{jn} = [d_{j1}\ d_{j2}..d_{jN}]$ are the i-th speaker and j-th noise dictionaries, respectively and each is a p X N matrix, with N atoms of dimension p.

B. Noise and speaker classification

Each test feature vector is explored for maximum cosine similarity with an atom of one of the noise dictionaries. The cosine similarity is computed with each atom of each of the noise and speech dictionaries. Then, a subset of features of the noisy speech are selected for noise classification based on the maximum similarity between the noisy feature y with any of the dictiona

$$\hat{k} = \arg\max_{k}\ y^T d_n^k \forall 1 \leq n \leq N \ and\ 1 \leq k \leq (M_s+M_n) \qquad (2)$$

where d is the n-th atom of the k-th source, Ms and Mn are the No. of speaker and noise dictionaries. If the atom with the highest correlation belongs to source $\hat{k}$, then the corresponding feature y is selected for noise classification, if and only if $\hat{k}$ corresponds to a noise source. Now, using this subset of test feature vectors, the noise source is estimated for each feature using the signal to distortion ratio (SDR) measure and ASNA recovery algorithm. The sum of SDRj over all the subset of features with the estimated noise index j is found as TSDRj. This is computed for each of the noise indices and the noise label for the utterance is estimated as $\hat{j} = \arg\max_{j}\ TSDR_j$

The speech component in noisy speech contains some silence segments, which become noise only segments having low energy. Hence, to determine the speaker, we use only the 60% of high energy test frames. Similarly, the top 80% atoms with the highest energy (before normalization) are chosen from the speech dictionary. The TDCS-0.8 algorithm (Cosine similarity based dictionary learning with $T_i = T_I = 0.8$ [22]) and ASNA with L1 normalization are used for speaker classification[23]. The sum of weights SW_k corresponding to each speaker dictionary index k is found and the total sum of weights for each speaker source is $TSW_k = PSW_k$ over all the features of the selected subset. The speaker label for the utterance is estimated as $\hat{k} = \arg\max_{k}\ TSW_k$.

C. Separation of speech and noise signals

The estimated noise index $\hat{j}$ and speaker index $\hat{k}$ are used to recover the noise and speaker components of the test features using a concatenated dictionary, $D=[D_n^{\hat{j}}\ D_s^{\hat{k}}]$ and recovery algorithm ASNA. The estimated features $\hat{y}$ and the noise and speech component $\hat{y}_n, \hat{y}_s$ are

$$\hat{y}=[D_n^{\hat{j}} D_s^{\hat{k}}][x_{\hat{j}}^T x_{\hat{k}}^T]^T$$

$$\hat{y}_n = D_n^{\hat{j}} x_{\hat{j}}^T,\ \ \hat{y}_s = D_s^{\hat{k}} x_{\hat{k}}^T$$

The speech component in the time domain s[n] is reconstructed using the estimated $\hat{y}_s$ and phase of the mixed audio signal using overlap and add method as $\hat{s}_s[n]$. The corresponding noise signal is estimated as $\hat{s}_n[n]=s[n]-\hat{s}_s[n]$.

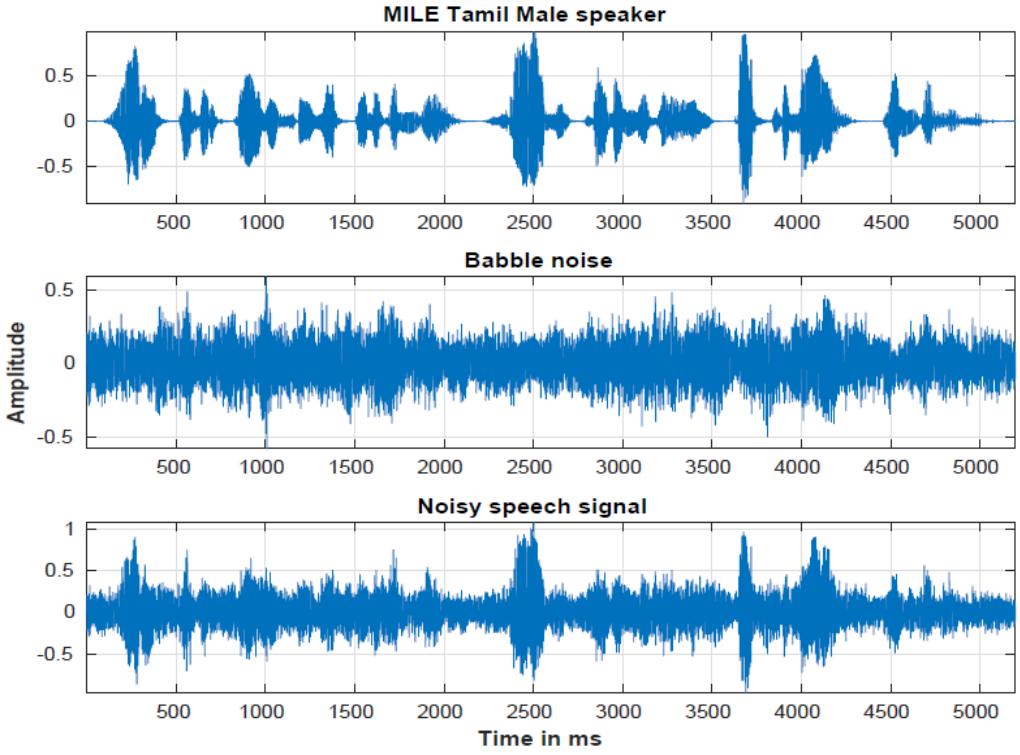


Fig. 1. Illustration of speech, noise and the noisy speech signal at an SNR of 0 dB

D. Measures for quantifying separation performance

The following measures are used to evaluate the performance of speech and noise separation:

- **Signal to distortion ratio (SDR)** [24]: between the original and the estimated speech signal is defined as:

$$SDR = 20\log_{10}\frac{\| s_s[n] \|_2}{\| s_s[n]-\hat{s}_s[n] \|_2}$$

SDR quantifies the deviation of the separated speech signal from the original signal; higher the SDR, better is the separation performance.

- **Mean Opinion Score (MOS)**: is the average of opinion score given by many human evaluators for the enhanced speech after separation. A subjective score (opinion score) ranging from 1 to 5, where 1 indicates unsatisfactory speech quality and annoying and objectionable distortion while 5 indicates excellent speech quality and imperceptible distortion [25] is used for evaluation on the enhanced speech.

RESULTS

The database for speech sources is taken from MILE Tamil speaker and noise sources (factory, babble) from NOISEX[26] and traffic noise from online. The noise classification accuracy is 100%. We have shown good speaker classification accuracy in [23] for English speakers. As we have only one Tamil speaker, we have not shown the speaker classification accuracy for Tamil speakers. The SDR obtained for speech and noise sources mixed at 0 dB SNR is shown is Table.I. It is seen that babble noise gives low SDR due to its speech like structure and traffic noise give the highest SDR.

TABLE I
SDR EVALUATION ON SPEECH MILE TAMIL SPEAKER MIXED WITH BABBLE, FACTORY AND TRAFFIC NOISE.

Speech source	Noise source	SDR
MILE Tamil speaker	Babble	5.4 dB
MILE Tamil speaker	Factory	7.3 dB
MILE Tamil speaker	Traffic	8.47 dB

A. Perceptive Evaluation

We validate the effectiveness of the system by 7 human evaluators. The subjective evaluation indicates a significant enhancement in the perceived quality of speech and the information communicated. All the human evaluators know Tamil and their average age is 24 years. The MOS score given the human evaluators is found to be 3.5 and the standard deviation is 0.84.

CONCLUSION

This methodology, when incorporated in hearing aids, will go a long way in improving the quality of life of the elderly and the people with hearing disability. It is also possible to embed this enhancement algorithm in mobile phones to improve the quality of the received speech, by estimating the noise source and enhancing it using a generic multispeaker dictionary, rather than a concatenated dictionary from a fixed set of speakers.

REFERENCES

[1] R. Turner, “Noises off: the machine that rubs out noise,http://www.eng.cam.ac.uk/news/noises-machine-rubs-outnoise-0, [Online] Accessed: 2017-04-12.
[2] T. Virtanen, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,” IEEE transactions on audio, speech, and language processing, vol. 15, no. 3, pp. 1066–1074, 2007.
[3] A. Ozerov and C. F´evotte, “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 3, pp. 550–563, 2010.
[4] G. J. Mysore, P. Smaragdis, and B. Raj, “Non-negative hidden Markov modeling of audio with application to source separation,” in International Conference on Latent Variable Analysis and Signal Separation. Springer, 2010, pp. 140–148.
[5] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, “Exemplarbased sparse representations for noise robust automatic speech recognition,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2067–2080, 2011.
[6] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, “Non-negative matrix factorization based compensation of music for automatic speech recognition,” in INTERSPEECH, 2010, pp. 717–720.
[7] Y.-C. Cho and S. Choi, “Nonnegative features of spectrotemporal sounds for classification,” Pattern Recognition Letters, vol. 26, no. 9, pp. 1327–1336, 2005.
[8] S. Zubair, F. Yan, and W. Wang, “Dictionary learning based sparse coefficients for audio classification with max & average pooling,” Digital Signal Processing, vol. 23(3), pp. 960–970, 2013.
[9] J. Nikunen and T. Virtanen, “Object-based audio coding using non-negative matrix factorization for the spectrogram representation,” in Audio Engineering Society Convention 128, 2010.
[10] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, “Sparse representations in audio and music: from coding to source separation,” Proceedings of the IEEE, vol. 98, no. 6, pp. 995–1005, 2010.
[11] T. Virtanen, J. F. Gemmeke, and B. Raj, “Active-set Newton algorithm for overcomplete non-negative representations of audio,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 11, pp. 2277–2289, 2013.
[12] S. G. Mallat and Z. Zhang, “Matching pursuits with timefrequency dictionaries,” IEEE Transactions on signal processing, vol. 41, no. 12, pp. 3397–3415, 1993.
[13] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, “Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition,” in Signals, Systems and Computers, Conference Record of The Twenty-Seventh Asilomar Conf. IEEE, 1993, pp. 40–44.
[14] I. F. Gorodnitsky and B. D. Rao, “Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm,” IEEE Trans. Sig. Proc., vol. 45, no. 3, pp. 600–616, 1997.
[15] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” SIAM review, vol. 43, no. 1, pp. 129–159, 2001.
[16] R. G. Malkin, “Machine listening for context-aware computing,” Ph.D. dissertation, Carnegie Mellon University Pittsburgh, PA, 2006.
[17] J. M. Kates, “Classification of background noises for hearing aid applications,” The journal of the Acoustical Society of America, vol. 97, no. 1, pp. 461–470, 1995.
[18] K. El-Maleh, A. Samouelian, and P. Kabal, “Frame level noise classification in mobile environments,” in Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on, vol. 1. IEEE, 1999, pp. 237–240.
[19] M. A. Casey, “Reduced-rank spectra & minimum-entropy priors as consistent & reliable cues for generalized sound recognition,” in Workshop for Consistent & Reliable Acoustic Cues, 2001, p. 167.
[20] C. Tzagkarakis and A. Mouchtaris, “Sparsity based robust speaker identification using a discriminative dictionary learning approach,” in Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European. IEEE, 2013, pp. 1–5.
[21] C. Joder and B. Schuller, “Exploring nonnegative matrix factorization for audio classification: Application to speaker recognition,” in Speech Communication; 10. ITG Symposium; Proceedings of. VDE, 2012, pp. 1–4.

[22] K. V. V. Girish, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Cosine similarity based dictionary learning and source recovery for classification of diverse audio sources," in India Conference (INDICON), IEEE Annual. 2016.
[23] K. V. V. Girish, A. G. Ramakrishnan, and T. V. Ananthapadmanabha, "Hierarchical classification of speaker and background noise and estimation of SNR using sparse representation," in INTERSPEECH, 2016.
[24] E. Vincent, R. Gribonval, and C. F´evotte, "Performance measurement in blind audio source separation," IEEE Trans. audio, speech, language processing, vol. 14, no. 4, pp. 1462–1469, 2006.
[25] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," IEEE Journal on selected areas in communications, vol. 10, no. 5, pp. 819–829, 1992.
[26] "NOISEX-92," http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html, [Online] Accessed: 2017-03-30.

Offline Tamil Handwritten Character Recognition: Challenges – An Analysis

M. Antony Robert Raj, S. Abirami
Department of Information Science and Technology
Anna University, Chennai – 600 025
antorobert@gmail.com, abirami_mr@yahoo.com

Abstract.

An impressive field of research, Offline Tamil Handwritten Character Recognition is one of the testing errands in the Optical Character Recognition field. Considering the intricacy, various works are accessible to give the answer for perceiving the Tamil written by hand characters, yet at the same time, it is open for the research. Maybe a couple of our works are giving the sensible solution in this field of research. The principle commitment of this works is, finding the indispensable difficulties behind the recognition of transcribed type of Tamil characters. This paper additionally dissects the effective algorithms of our attempts to direct the exploration in right strategy. Likewise, look at how the algorithmic methodology stands up to the handwritten character many-sided quality. The final analysis demonstrates the positive and negative of the works and gives the correct decision of the pattern to locate the privilege algorithmic strategy.

***Keyword**: Shape, Direction, Location,Tree Representation*

1 Introduction

In recent years, Offline Tamil handwritten character recognition assumes a critical part in the Optical Character Recognition (OCR). Computerizing the huge reports, for example, Palm scripts, old sonnets, chronicled records and so on, are fundamental for Tamil individuals. The expanding interest of this recognition framework persuades to include in this exploration by giving an important support and respectable commitment. The Tamil dialect contains rich character sets, which contain 247 characters incorporate (12 vowels, 18 consonants, 216 combinational characters and one unique character). The novelty of this paper lies in, distinguishing key the difficulties by analyzing the shape of the characters and furthermore the current commitments in this handwritten character frameworks to give the solution for this. Two center groupings of fundamental difficulties in Tamil handwritten character recognition are general multifaceted nature and author's many-sided quality.This part of the work portrays the most perceptible difficulties which made by the writers. The debate, for example, comparable shape, shape and angle variety, superfluous circles and curves, shape brokenness, pointless connectivity and character combination are basic testing variables recognized from those characters [1]. At first, the distinctive handwritten characters from various writers seem to be comparative in specific situations. This is one of the most noteworthy difficulties in this recognition framework. The accompanying difficulties are distinguished from the chosen characters (shown in **figure 1**).

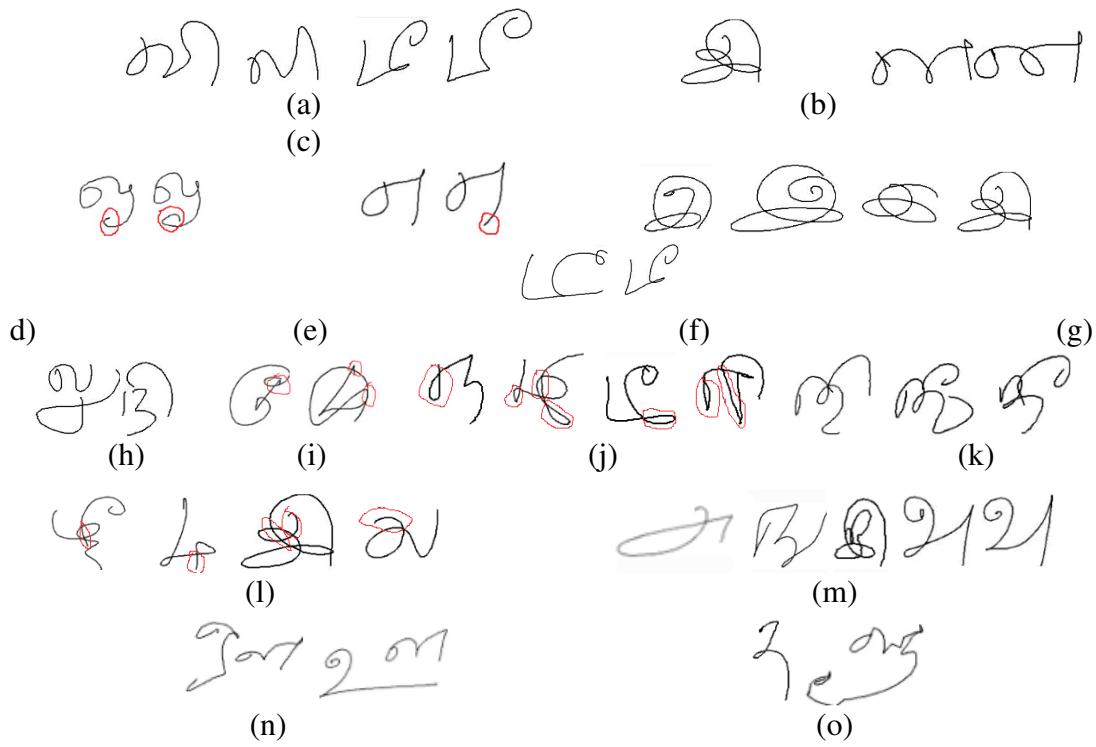


Fig. 1. (a) Similar shape introduced by writers (ஹி looks லி and டி looks பி). (b) Unnecessary curves (இ).

(c) Unnecessary loops middle of ள, looks like ன. (d) Similar in shape (ஒ and ஓ). (e) Miner variation among எ and ஏ.(f) Shape variation (இ). (g) Angle variation leads to shape variation (டீ looks பீ). (h) Character discontinuity (லு and நி). (i) Character structure connectivity (டு and மு). (j) Unwanted loops (நு, ழி, டீ and ரி). (k) Unknown shape character (றி, ஙீ and தி). (l) Unnecessary curves (ழி, ம, இ and ல). (m) Similar shape (அ looks எ, நு looks ங, கி looks இ, வு looks அ, and வு looks வி). (n) Double letters (ஔ and ஊ). (o) Writing issues (த and ஞ) The complexity introduced by writers such as variation in shape, similar in shape, needless loops and curves and the variation in location is giving more hardness in the recognition level. Different examinations have been taken in this exploration and various algorithmic systems were proposed on those characters. This work concentrates on the recognizable proof of the exceptional method for commitment to getting achievements in the examination on Tamil written by hand character recognition field.

The foremost phases of our research in this field are feature selection and pre-extraction, extraction and classification. The essential image processing techniques are utilized for the pre-preparing methods [3]. In the period of feature selection, the components are chosen by the zoning [3][6][7], quad [13], genuine shape with no division. In the following fragment, the algorithm to pre-extract the character portion was applied to the selected division of image in the previous procedure. For that, the chain code [9] travel has been utilized. Two methods for the feature extraction system are inferred on these chosen or pre-extricated feature, they are statistical [2][8] which depends on the pixel variety and structure which depends on the shape and direction [11][12][10].

While considering the statistical features, the pre-extraction procedures are not utilized at times in light of the fact that the pixel variety calculation can be applied specifically to the image. The pre-extracted pictures are fundamental for testing the shape and direction of character divide. A few feature extraction methods were proposed in this research field. This paper manages an examination of the imperative and best feature extraction systems for finding an ideal approach the exploration. With the analysis, five distinctive effective algorithmic strategies have been analyzed to proceed with the exploration in the correct bearing. They are the locational highlight which has been doing the statistical way by utilizing the quadtree [13] and zone [7][10], a directional element by utilizing the chain code [10] and the shape includes by utilizing the strip tree [12] and prism tree [11]. The multifaceted nature level of each character utilized as a part of the each work has additionally considered for this examination. Properties of each algorithmic idea and the execution of a similar when the unpredictability is expanded are dissected.

2 General Challenges: An Analysis

The Tamil handwritten characters are gathered impressively from different Tamil Writers in all age gathering and few of them from HP-India Tamil dataset [14]. Add up to 24700 examples which contain 100 samples for each character are gathered. In the investigation, the accompanying gatherings have been framed. They are, Good for Recognition (GR), Similar in Shape introduced by Writers (SSW), Shape Variation (SV), Location Variation (LV), Structure Discontinuity (SD), Unnecessary Loops (UL) and Unnecessary Curves (UC). With no issues, 30% of characters from all examples are gathered in the GR class. 5.4% characters are going under the gathering SSW. In the most minimal level, 2.0% characters come in the class SD. Three is no different issues are accessible in these three gatherings (GR, SSW and SD). At the largest amount, 22% characters contain SV issues and furthermore 64% of character are containing SV and different issues too. 9.7% characters are assembled with the issues LV yet 28% of characters are contained both LV and different issues.

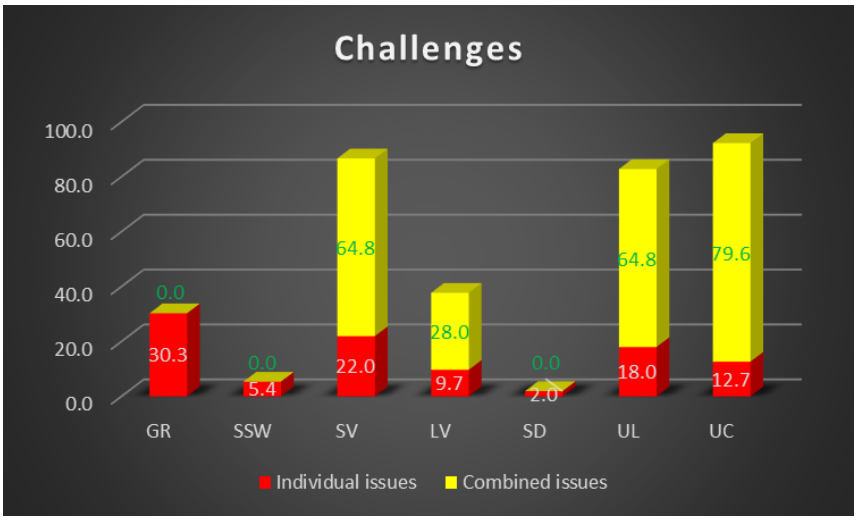


Fig. 2. Analysis on challenges

18% UL characters are accessible in the gathered examples and moreover 64.8 % of characters which contain UL and every other issue. 12.7% characters are chosen from the gathering UC. At the largest amount, 79.6% characters are contained UC and different issues. According to the examination, out of 24700 specimens, just 30% of characters are fitting to

perceive however 70% of characters are going under difficult issues, essentially shape variety, unnecessary loops, and unnecessary curves. 64 to 79% of issues contains the dangerous issues which require an uncommon algorithmic treatment to perceive. The portrayal of the same has appeared in figure 1.

3 Feature Analysis

The examination has been begun from few of imperative existing works.Conventional feature extraction[26] system was utilized to separate curve features, with the support of Neural network algorithm, 12 Tamil characters were recognized with the exactness 94.4%. The line and curve were separated by utilizing the normalized feature vectors[25]. With the assistance of Fuzzy approach, seven characters were perceived with the 88 to 100 % of precision. There are two phases were used in [23]. In thefirst stage, the picture was partitioned into 7*7 equivalent squares and component transition features vectors were extricated. Character Contour is acquired by utilizing Freeman's Chain code in thesecond stage. K-mean algorithm and multi-layer perception (MLP) support to get 92.77 and 89.66% precise in each stage. Row-wise, column wise and diagonal-wise feature variation [20]were extracted by utilizing the Encoding binary variation technique. With the assistance of comparison strategy, 10 characters (Tamil numerals) were classified. Kohonen'sSelf-Organizing Feature Map (SOFM) method [15] group the characters from aheight, width, thenumber of horizontal, vertical lines and curves, thenumber of circles and slant lines, image centroid and unique dots which were extracted from 8 letters as it were. Height and width [21] were scaled from the character images by utilizing bi-linear interpolation method. 67 characters were classified by the classified by the Self-organizing feature maps (SOP) calculation with the precision rate 98.50%. 10 diverse character is classified by MLP [16] with the assistance of the character highlights height and width. Endpoints, fork focuses, holes, length, shape, and flow of individual strokes were noted from theoctal graph in character image and projection profiles also included with these features. The same was taken as acontributionto feature matching methodology and accomplished over every one of the 82 % of accuracy rate [22]. Character stature, width, number of horizontal and vertical lines (long and short), horizontal and vertical curves, circles, slope lines, image centroid and dots[18] are separated and classified the characters by Support Vector Machine (SVM) with the precision result 97%. Statistical component[17], for example, cover vertical projection profile, word profile, background-to-ink transition separated from character image and the same was classified by the Hidden Markov Model (HMM) strategy. Pixel varieties were caught from thezone-based approach and classified the same by SVM[19]. Average of 82.04% (62.8 % for 3 characters, 98.9 % for 12 characters, every one of the 34 characters - 82.04) exactness rate was accomplished for 34 characters. Structure Boundaries (scale and shift invariant) [24] were extracted by eight-neighbor adjacent and MLP calculation utilized for perceiving the 30 characters with the accuracy rate 97%. The descriptions of the same have been listed in **Table 1.**

The issues has been attempted to solve in our past works. The solution which very adapting with Tamil Handwritten characters are examined here.

3.1 Locational Features

Two successful algorithms which were utilized as a part of the past works have been examined here, they are Quadtree [13] and the zoning [10] techniques. Slight changes have

been suggested on both the algorithmic technique to fulfill the need. The pre-extracted image was separated into four quads. Further, every quad has been subdivided into 4 sub-quads.

Table 1. Analysis on related works

Paper	Features	Pros	Cons
[26] T Paulpandianet al	Curve- Conventional Feature	It is useful to consider shapes of characters	Cannot address more variation.
[25] RM Sureshet al	Line and arc		Cannot address the printed form of characters also
[23] U Bhattacharyaet al	Component transition features vectors – Block (Zone)	Can address more variation	Failed to recognized if character count is increased with more variation
[20] R Bremananthet al	Row-wise, column-wise, and diagonal-wise feature variation	Good to address the characters with real shape	Not address the similar shape characters
[15] P Banumathi et al	Height, width, thenumber of horizontal, vertical lines and curves, thenumber of circles and slope lines, image centroid and special dots	Small changes are also recognizable	Failed,if the character count is increased
[21] R Indra Gandhiet al	Height and width	May be suitable for printed characters	Not enough to address any type of variation. Need a helps from other algorithms also
[16] Stuti Asthana et al	Height and width		
[22] R Jagadeesh Kannanet al	Endpoints, fork points, holes, length, shape, or curvature of individual strokes	Suitable for online and printed characters.	Failed to address similar shape characters.
[18] C Suresh Kumaret al	Character height, width, number of horizontal and vertical lines (long and short), horizontal and vertical curves, circles, slope lines,	It can addresslimited samples.	Cannot address shape variation if character count has been increased.

	image centroid and dots		
[17] AN Sigappi et al	vertical projection profile, word profile, background-to-ink transition	Matching with unique characters	
[19] N Shanthiet al	Pixel variations from zone	Highly suits for printed	Failed to address all shapes and variations
[24] J Suthaet al	Boundaries (scale and shift invariant)	Highly suits for printed	

Pixel density was figured from every quad (R1, R2, R3 and R4). The quad which had the most elevated density was taken as the principle feature vector (R1) as portrayed in figure 3. Whatever remains of the features were investigated from each sub-quad, which were taken as features if important. 28 Tamil characters were chosen from vowels and consonants for this test and accomplished 88.25% precision rate.

In zoning method, the pre-separated image was isolated into 4X4 zones [10] as appeared in figure 4 and the pixel density was ascertained from each zone. The zone which contains the most noteworthy pixel density esteem had considered as a feature. On the off chance that more than one zone contains a similar measure of pixel density, that zone was additionally considered for the feature portrayal. Thirty characters (vowels and consonants) were utilized for those examinations. The zoning strategy alone may not help for confronting all difficulties, in this way, these system has explored different avenues regarding directional components. The investigation of the same is recorded in the table2.

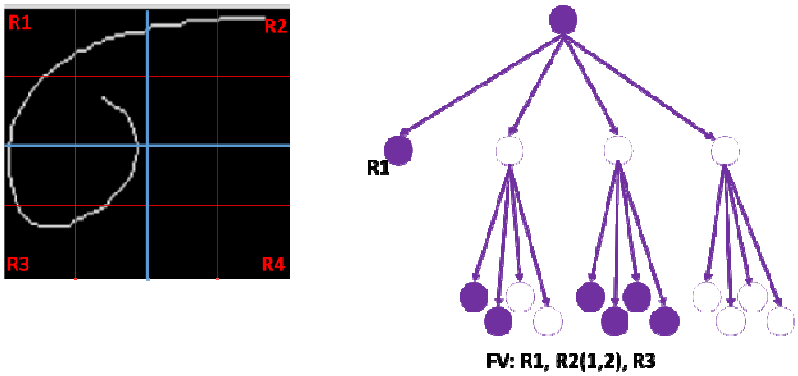


Fig. 3. Feature representation by Quadtree

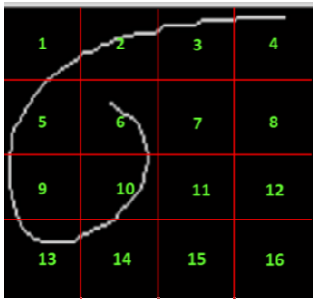


Fig. 4. Feature representation in zone (4X4)

Table 2. Analysis on locational points

Successive points	Issues
It can be highly helpful for recognizing the character with the real shape	Not withstands if character count has been increased
Slight changes inshape are also acceptable	Not suitable for similar shapes (ஒ and ஓ)
Highly fit for unique character even it has high variation	Cannot adaptable for minor changes in real shape (ஏ and ஈ)
Quadtree suits better than zone	
Solvable if unnecessary loops are introduced by the writer but a support is needed from direction and shape	Not suitable for location mismatch (ie, shape variation)

3.2 Directional Features

Chain code methodology was utilized here for finding the direction [10] of the character partitions (as shown in figure 5). It can be utilized in based on the writing order or the shape. Here, going to the pixel position of Tamil character is the hardest errand, so the pre-extraction systems were an essential one for this. On the off chance that the direction travel would not restore any direction but rather there was a pixel in the image then the point considered as a dot. If the directional points have returned all direction, at that point there may be a round shape feature.

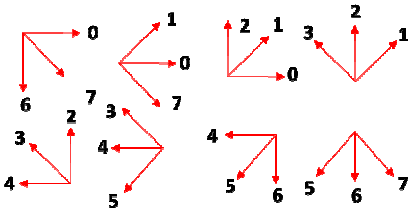


Fig. 5. Directional points

The directional point was a successful calculation while considering more samples. Moreover, to consider all kind of character issues, this calculation consolidated with zone system and accomplished 90.7 % precision rate. The examination about this feature is given in Table 3.

Table 3. Analysis of directional features

Successive points	Issues
It can solve the problem of location mismatch	Giving solution for similar shape, but not accurate
It can recognize the shape if the shape is changing	Raising issues when character count is increased

Giving solution for similar shape

3.3 Shape-based Features

It can catch the genuine shape of the character portion, here the two different tree strategies are utilized for getting the shape of the characters, and they are strip tree and prism tree. Here, the features are represented in the tree. Feature pre-extraction techniques are essential for this procedure.

3.3.1 Strip tree

Strip division [4] was occurred utilizing the formation of rectangular as appeared in figure 6. At first, the whole shape was taken in a rectangle which indicated as root. Further, a mid-line has been drawn by utilizing the touching points on the outskirt of the rectangle or end points, which isolates the rectangle into two. The same has been set apart as a sub-node under the root. The same was proceeded in all sub-rectangles until the point that no further rectangle can be subdivided. The level of the tree and the node esteems were considered as feature points. 88% precise outcomes have been accomplished for 10 characters were browsed vowels [12].

3.3.2 Prism tree

Prism tree [5] [11] is extracting the features by utilizing the triangle development as appeared in figure 7. At first, a line was drawn between two end points of the character parts. From the midpoint of the line, a perpendicular point was set apart on the character portion, this point was given the triangular association. Further, the triangle execution occurred from the adjacent points of the main triangle. It had been actualized for a specific level, which couldn't give a triangulated further. The principal rectangle was noted as root in the tree, whatever is left of them were signified as leaf nodes as appeared in figure 7. This tree execution was effective when compared with strip tree. 90.08% accuracy rate has been achieved for 28 chosen characters. The examination of the same has been portrayed in Table 4.

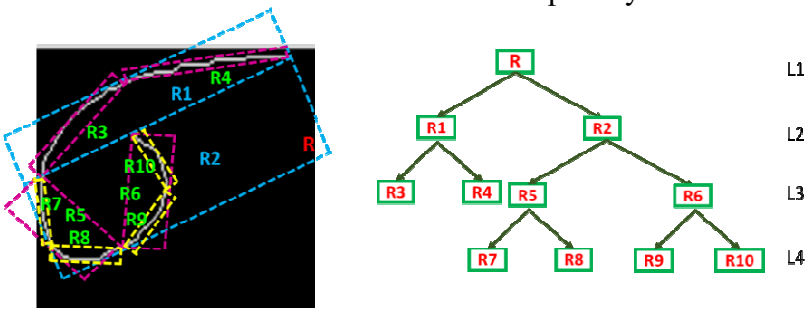


Fig. 6. Shape extracted by strip tree

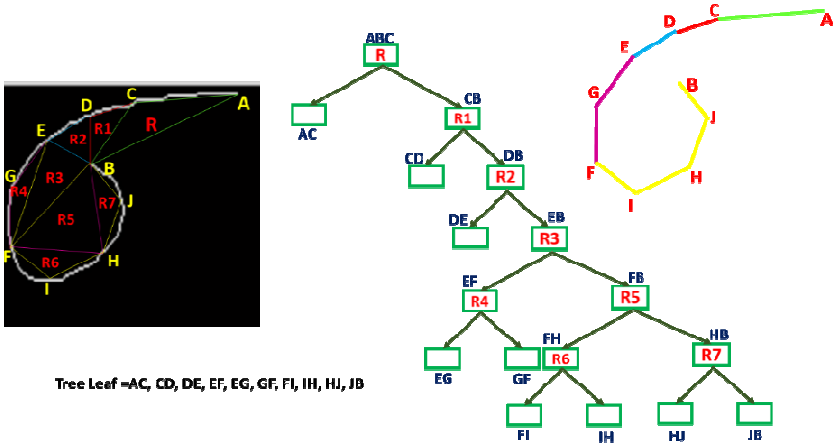


Fig. 7. Shape extracted by prism tree

Table 4. Analysis on shape-based features

Successive points	Issues
Location mismatch issues are solvable	Shape variation with location mismatch is not solvable
It can easily identify the shape of the character portion	Not solvable if the unnecessary portion has been introduced by writers

Giving solution for similar shape

4 Comparative Analysis

By and large, all algorithms are achievements in any of the viewpoints, particularly, the location-based methodology can address shape variety, theshape-based system can represent the location varies and the direction based technique can address both shape and location variety. The issue is, any of the arrangements can address the accompanying issues

- Similar in shape to the handwritten character set,
- Identify the uniqueness among the written by hand characters
- Should not have any attention on the undesirable circle and curves.
- High-end variety
- Style variations

Here, there is a solution for style variations and high shape variety, however, the calculation may not supplant the undesirable circle and curves. Rather, the solution must obstruct recognizing the uniqueness among the separate character sets and a minor variety which giving the comparative shape among irrelevant characters. The accompanying (Table 5) depicts the solution of every issue

Table 5. Feature Solution

S. No	Algorithm	Solution
1	zone	Pixel density in the zone may not help for representing all characters. If samples are increased it will be failed. Good for printed characters

2	Quadtree	Used for locating the character portion. It is also failing to represent all 247 characters. Comparing with zone, it is successful for unique character sets
3	Directional	Useful for perceiving more examples. May not suitable for similar in shape.
4	Strip tree	Better to represent the shape, but when representing in a tree, it is failed. The main reason is many characters return same tree representation sometimes.
5	Prism Tree	Good for representing shape when comparing Strip tree in this application. It is an exceedingly valuable calculation to speak to minor shape varieties likewise yet the top of the line variety may prompt disappointment when tests are expanded

According to this examination, all algorithmic strategy is essential for perceiving the written by hand character, where Prism tree can address shape consummately yet require a change in different sorts feature portrayal. Here, the shape extraction alone won't classify the transcribed characters. There must be a requirement for changing algorithmic technique in directional element extraction and location-based feature extraction. In the event that the methodology has a consolidated variant of all their strategy of feature extraction algorithm, at that point, this could be better to address all issues are found in Tamil handwritten character recognition.

5 Conclusion

This paper has been investigated different issues are found among the handwritten characters gathered from different perspectives. According to the examination taken from the gathered examples, 64 to 79% of the character contains all assortment of issues. Three different base algorithmic methods which gone under statistical and structural phases of feature extraction has been dissected here to distinguish the best route of this recognition framework. Shape-based prism tree methodology has been picked the best than strip tree technique. Changes must requirement for location and direction based systems. Noticeably, these three systems are essential for extracting their one of a kind method for feature extraction and draw out the uniqueness of Tamil transcribed character sets.

References

(1) M Antony Robert Raj, S Abirami, “A Survey on Tamil Handwritten Character Recognition using OCR techniques”, The Second International Conference on Computer Science, Engineering and Applications (CCSEA), 2012, 05, pp.115-127

(2) M Antony Robert Raj, S Abirami, “Analysis of Statistical Feature Extraction Approaches used in Tamil Handwritten OCR”, 13th Tamil Internet Conference-INFITT, 2013, pp.144-150.

(3) M Antony Robert Raj, S Abirami, “Offline Tamil Handwritten Character Recognition using Chain Code and ZoneBased Features”, 13th Tamil Internet Conference- INFITT; 2014, pp.28-34.

(4) Dana H. Ballard, "Strip Trees: A Hierarchical Representation for Curves", ACM, Graphics and Image Processing, 1981,ISSN : 0001 – 0782, Vol. 24, Pages: 310-321.
(5) Hanan, Samet, "Foundation of Multidimensional and Metric Data Structures". Book Published by Morgan Kaufmann, 2006, PP. 386-388.
(6) SV Rajashekararadhya, P VanajaRanjan, VN ManhunathAradhya, "Isolated Handwritten Kannada and Tamil Numeral Recognition: A Novel Approach". First IEEE International Conference on Emerging Trends in Engineering and Technology, 2008, Page(s): 1192-1195.
(7) SV Rajashekararadhya, P VanajaRanjan, "Zone-Based Hybrid Feature Extraction Algorithm for Handwritten Numeral Recognition of two popular Indian Script". World Congress on Nature & Biologically Inspired Computing, 2009, page(s): 526 – 530.
(8) M Antony Robert Raj, S Abirami, "Offline Tamil Handwritten Character Recognition Using Statistical Features", AENSI Journals, Advances in Natural and Applied Sciences, 9(6) Special, 2015, Pages: 367-374.
(9) SM Shyni, M Antony Robert Raj, S Abirami, "Offline Tamil Handwritten Character Recognition Using Sub Line Direction and Bounding Box Techniques". Indian Journal of Science and Technology, 2015, Vol 8(S7), 110–116.
(10) M Antony Robert Raj, S Abirami, "Hybrid Features based Offline Tamil Handwritten Character Recognition" 14th International Tamil Internet Conference, 2015, ISSN : 2313 - 4887 Pages: 360-370.
(11) M Antony Robert Raj, S Abirami, "Prism Tree Shape Representation Based Recognition of Offline Tamil Handwritten Characters", Springer AISC Series ISSN: 1867-5662, first International Conference on Computational Intelligence and Informatics, JNTU, Hyderabad, INDIA, 28-30 May 2016, ISSN: 2190-3018.
(12) M Antony Robert Raj, S Abirami, "Strip Tree based offline Tamil Handwritten Character Recognition", Springer SIST, International Conference on ICT for Intelligent Systems, Ahmedabad, INDIA, 28-29, November 2015, ISSN: 2190-3018, Pages: 367-374.
(13) M Antony Robert Raj, S Abirami, "Offline Tamil Handwritten Character Recognition using Statistical based Quad Tree" Australian Journal of Basic and Applied Sciences, 2016 ISSN 1991-8178,10(2) Special, Pages 103-109.
(14) http://lipitk.sourceforge.net/hpl-datasets.htm
(15) P Banumathi, Dr. GMNasira, "Handwritten Tamil Character Recognition using artificial neural networks", International Conference on Process Automation, Control and Computing (PACC), 2011, page(s): 1 – 5.
(16) Stuti Asthana, FarhaHaneef and Rakesh K Bhujade, "Handwritten Multiscript Numeral Recognition using Artificial Neural Networks", Int. J. of Soft Computing and Engineering, March 2011, ISSN: 2231-2307, Volume-1, Issue-1.
(17) AN Sigappi, S Palanivel and V Ramalingam, "Handwritten Document Retrieval System for Tamil Language", Int. J of Computer Application, 2011, Vol-31.
(18) C Suresh Kumar, Dr. T Ravichandran, "Handwritten Tamil Character Recognition using RCS algorithms", Int. J. of Computer Applications, October 2010, (0975 – 8887) Volume 8– No.8.
(19) N Shanthi, K. Duraiswami, "A Novel SVM -based Handwritten Tamil character recognition system", springer, Pattern Analysis & Applications,2010,Vol-13, No. 2, 173-180.
(20) R. Bremananth and A. Prakash, "Tamil Numerals Identification", International Conference on Advances in Recent Technologies in Communication and Computing, 2009, page(s): 620 – 622.

(21) R.Indra Gandhi, Dr. K. Iyakutti, "An attempt to Recognize Handwritten Tamil Character using Kohonen SOM", Int.Journal of Advance d Networking and Applications, 2009, Volume: 01 Issue: 03 Pages: 188-192.

(22) R. Jagadeesh Kannan and R.Prabhakar, "An improved Handwritten Tamil Character Recognition System using Octal Graph", Int. J. of Computer Science, 2008, ISSN 1549-3636, Vol 4 (7): 509-516.

(23) U. Bhattacharya, S.K. Ghosh and SK Parui, "A Two Stage Recognition Scheme for Handwritten Tamil Characters", Ninth International Conference on Document Analysis and Recognition,2007, Vol : 1 , page(s): 511 – 515.

(24) J. Sutha, N. RamaRaj, "Neural network based offline Tamil handwritten character recognition System" , International Conference on Conference on Computational Intelligence and Multimedia, 2007, Vol : 2, page(s): 446 – 450.

(25) R.M. suresh, S. Arumugam, LGanesan, "Fuzzy Approach to Recognize Handwritten Tamil Characters", Third International Conference, Proc. on Computational Intelligence and Multimedia Applications, 1999, page(s): 459 – 463.

(26) T. Paulpandian and V. Ganapathy , "Translation and scale Invariant Recognition of Handwritten Tamil characters using Hierarchical Neural Networks", Circuits and Systems, IEEE InternationalSymposium , 1993, vol.4, 2439 – 2441.

Challenges of Machine Learning with Tamil Texts from Ancient to Modern Tamil

Vasu Renganathan
University of Pennsylvania, Philadelphia, USA
(vasur@sas.upenn.edu)

Abstract

Digitization and implementation of machine learning algorithms with data from Tamil literature of the three genres namely Sangam, medieval and modern period has been a challenging task mainly due to their complex word structures, formation of multiple shades of meanings historically, and a host of others. The aim of this paper is not only to present methods of storing the existing digitized literature data using recent technologies, but also to devise suitable algorithm to manipulate them byplausible means, especially using relational database and JSON technologies. Without undermining the merits of any of the past technologies of Tamil information ages, including palm leaves, stone inscriptions, copper inscriptions and so on, this paper attempts to try to harness the power of digital technology in a number of noval ways. Our attempt will be to closely look into the power of the Java Script Object Notation (JSON) technology as well as Relational Database Structures along with other ways of manipulationof data using the Vue.Js technology. There have been many technologies to utilize the power of JSON format and relational databases, including Angular.js and others. But, we would like to show in this paper how the Vue.js technology along with the JSON and relational database structures has potentials to search and research vast amount of texts from Tamil's ancient traditions. The sites we would focus on illustrating and developing further in this paper include a) http://sangam.tamilnlp.com/mp/, which offers Tamil literature data in JSON format with a novel filter technique; b) http://sangam.tamilnlp.com/glossing.php, which provides a dynamic link between Tamil lexicon stored in a relational database with that of the texts through glossing technology; c) http://sangam.tamilnlp.com/read_poem.php, which offers a way to contextualize words among Sangam, medieval and modern texts and attempt to lay out a comprehensive historical information;d) http://sangam.tamilnlp.com/, which allows a wide-ranging search techniques across the three genres such as old, medieval and modern Tamil and finally plungingdeep into the morphological tagging possibilities as demonstrated in the site http://www.thetamillanguage.com/tamilnlp/tagit.html.

Tamil data in Java Script Object Notation (JSON) format:

As of now, we have many resources for e-texts of Tamil texts from Sangam, medieval, and modern Tamil in HTML format (cf. http://www.projectmadurai.org/, http://www.tamilvu.org/library/libindex.htm, http://ilakkiyam.com/and a host of others). Due to many constraints involved in HTML and text formats, using these sites for the purposes of machine learning and other novel ways analyzing the text including the use of advanced search techniques becomes harder, although not impossible. However, there are other resources including that of http://www.tamilpulavar.org/api.php, http://www.thetamillanguage.com/sangam etc., which use relational databases such as MySQL, Oracle etc., to store and retrieve data for the purposes of Natural Language Processing as well as search engines. Although the relational databases are popular and very powerful in many senses, they do still require a robust server technology as well as a front-end technology. In comparison to these two different data formats, the JSON format is considered to be

more popular than the other formats for the important reason that it can be used from a client side processing without having to rely on any server side technologies. The site http://sangam.tamilnlp.com/mp/ that is developed as part of this work consists of a number of Tamil Sangam literature works in JSON format.

A simple JSON record would be as below:

[{ "Number": 1,
"Line1": "அகரமுதலஎழுத்தெல்லாம்ஆதி",
"Line2": "பகவன்முதற்றேஉலகு.",
"Translation": "'The sound 'a' superceds all linguistic sounds; the primeval superceds the world",
"mv": "எழுத்துக்கள் எல்லாம் அகரத்தை அடிப்படையாக கொண்டிருக்கின்றன, அதுபோல உலகம் கடவுளை அடிப்படையாக கொண்டிருக்கிறது.",
"sp": "எழுத்துக்கள் எல்லாம் அகரத்தில் தொடங்குகின்றன; (அதுபோல) உலகம் கடவுளில் தொடங்குகிறது.",
"mk": "அகரம் எழுத்துக்களுக்கு முதன்மை; ஆதிபகவன், உலகில் வாழும் உயிர்களுக்கு முதன்மை",
"transliteration1": "akara mutala eḻuttellām āti",
"transliteration2": "pakavaṉ mutarrē ulaku"
}]

As can be seen from this example, each line from Thirukkural along with its other related information such as translation, transliteration, commentaries from different authors are noted with a key so, they can all be accounted for programmatically using the 'key' vs. 'value' pair of object notation. In principle, each of these records in this type of notations can further be illustrated with sub-notations in any imaginable recursive manner. More recursive any JSON string can be more indepth information they can be stored in. So, any program that is designed to read these lines would be capable ofprocessing them in a number of different complex ways possible.

The site under discussion employes the Vue.js technology (https://vuejs.org/) to manipulate and process data from Tamil literature stored in JSON format. JSON format, which adheres to object notation techniques, can relate data in a number of different relational possibilities and thus making the machine to identify the relations between words, phrases etc., in a coherent manner possible. Further, the search engine that is built in this site allows us to search both from lemma as well as inflected forms. To cite one example, searching "நுத" can fetch instances where this word is used in the forms such as நுதலும், நுதற், நுதி etc. Further, this format can hold comparatively large amount of data in a single client side page and searching through the text is also relatively faster.

This site can further be enhanced in such a way that it can be allowed to fetch records that are historically relevant by linking more JSON text of literature belonging to different periods of time. The three genres of Tamil namely Sangam, medieval and modern Tamil underwent a massive number of changes in word senses as well as in the context of formation of new words. In order for one to make an extensive research in such historical contexts, one needs to store text from the three genres as provided in the site http://sangam.tamilnlp.com/. Although this site uses Oracle back-end and PHP front-end, it performsthe fetching of records from many combinations within old, medieval and modern Tamil data. The word முகில், for example, when searched from all of the texts belonging to the three period, one can immediately notice that it occurs more in religious literature than in secular

literature such as the Sangam text. Similarly, when the word புணை is searched in all of these three genres, one can observe that this word is used more in Sangam literature and less in religious literature.Further this word may be found to be occurring with the meaning of 'boat' throughout in Sangam literature but only used in the meaning of 'tie together' in the context of medieval literature. What one may presume from these two simple examples is that the religious literature can be considered to make use of nature more extensively than the Sangam literature. Similarly, the use of the technology surrounding the word புணை is observed more with the livelihood of people during the Sangam period than those during the medieval period. Obviously, this type of analyses is possible only with e-texts stored either in JSON or RDBS formats, but neither the HTML or text formats would envisage such opportunities. For further details about historical research conducted for Tamil using this type of relational databases see Renganathan (2009) and Renganathan (2010).

Glossing technology:

Glossing is a process that is used widely in computer assisted learning and teaching. Additional interpretations of any word or phrase that occur within any e-text can be offered by linking the word with an electronic dictionary. Usually, the 'div' structure that is used in HTML technology is employed for this purpose extensively. When the pointer of a mouse is placed on any word within a text, theJavascript code written in AJAX technology is capable of consulting the electronic dictionaries stored in server side relational databases and fetch the dictionary entry in an asynchronous fashion. These technologies are used to read any Tamil text of all the three genres by consulting the Tamil lexicon stored in Oracle database. This is evident from the site http://sangam.tamilnlp.com/glossing.php. The advantage of this site is that it serves as an API to read any e-text from any site simply by adding the url in GET format as in 'http://sangam.tamilnlp.com/url_gloss.php?url='.

1)http://sangam.tamilnlp.com/url_gloss.php?url=export/etext_copy/aacaarakkoovai.txt
2)http://sangam.tamilnlp.com/url_gloss.php?url=http://www.projectmadurai.org/pm_etexts/utf8/pmuni0008_01.html

In this context, almost all of the e-text can be read consulting the Madras University Lexicon in a dynamic manner. Further, the glosses that are fetched in this page is not restricted to just the head entries of the lexicon, but all of the records where particular word occurs within the lexicon isalso fetched and displayed as part of the gloss. One advantage of this method is that when a particular word is inflected and the corresponding form is not available in the dictionary as a head word, fetching the related instances of using the inflected word in the other part of the dictionary can throw further light on the word. This way, even the inflected forms can still be glossed with the help of this technology. A comprehensive machine learning algorithm can still be devised to split inflected words into corresponding lemma so the corresponding entries can be fetched dynamically. Development of such a tagger can be possible for modern Tamil texts than forSangam Tamil for the reason that in Sangam Tamil texts the words occur in manyconvoluted combinations (ex.அவளிவளுவளெவள்)and in unusually split word forms based on meter(ex. இலனதுவுடையனிதெனநினை). (see Renganathan 2016 for a discussion on the limitations of tagging Sangam corpus).

Contextualized References of words and machine learning techniques:

As already mentioned, fruitful use of electronic Tamil data lies in the way how we store them and how we retrieve and synthesize them. Subsequently, there lies the element of machine learning by which the machine is made to learn from the algorithm we write, rather than following the algorithm

itself in a sequential manner. In other words, the fundamental idea behind machine learning is that the machine must be able to make new algorithms through the already available algorithms and constantly build its repertoire of knowledge. Although what we discussed so far in the context of JSON format and Glossing technology can not truly be defined as machine learning techniques as such, they can still be considered as the basis of machine learning for the fact that they offer a fundamental infrastructure to build such machine learning systems. Successful machine learning systems can be built more easily when the data is in a structured format, as in JSON or RDBM rather than in text or HTML format. In this context, this section attempts to discuss a developed system that is written in the PROLOG language using a list manipulation technique. The system in question can be tested in the url:
http://www.thetamillanguage.com/tamilnlp/tagit.html

The main idea behind this system is to use the concept of set theory to make the machine learn from a list structure. When a sentence in Tamil is given as input, this system parses the words and make a list structure with all the suffixes tagged using a predefined set of tags. For example, when the sentence **ஒருவகை சிவப்பு எறும்புகளுக்கு இறக்கைகள் கொண்டு பறக்கக்கூடிய வசதி வாய்ப்பு இருக்கும்.** is made as input, this system identifies the phrases, suffixes and other information such as noun, verb etc., using its dictionary and algorithm and consequently builds its database in a list form as in:
[["adj","oru"],["nom","vakai","noun"],["nom","civappu","tr"],["dat","eRumpu","tr","pl"],["nom","iRakkai","tr","pl"],["nom","koNTu","tr"],["pa_ajp","paRakkakkuuTu"],["nom","vacati","noun"],["nom","vaayppu","noun"],["pr","iru","neut.sg","conj"],["nom",".","period"],["nom",".","period"]]
With is structure in memory, when posed with questions such as **யாருக்கு வசதி வாய்ப்பு இருக்கிறது? எறும்புக்கு என்ன இருக்கிறது?** and so on, this system parses the input questions into corresponding list forms and attempts to match them with the existing database of list structures by employing PROLOG's logical operators such as 'subset', 'sublist' and so on. Subsequently, it gets the correct answers based on the matches it finds (see Gazdar and Chris Mellish 1989 for a discussion on list manipulation techniques using PROLOG). In a sense, this concept of responding to structures based on the list manipulation technique is identical to how human interprets natural language sentences (cf. Johnson-Laird 1983). Human's understanding of sentences, obviously, go beyond list structures, and uses many complex semantic analyses including presupposition, hyponymy, implications etc. The fundamental concept that is intended to illustrate here is that any successful machine learning algorithm for natural languages can or should begin in this type of list structures and build upon them successively by incorporating more semantic as well as syntactic knowledge in the form of list structures.

This system when posed with a simple question **என்ன?** would fetch all the information from the database and offer respective answers in Tamil sentences. What is unique about this system is that it is built with suitable algorithms to both decoding of Tamil words as well as generating sentences in a natural language format by employing necessary morphological operations encompasing Tamil morphology. In this respect, this system is adequeately built with morphological and sysntactic knowledge base of Tamil. However, what is lacking, perhaps can be developed further, is a sound semantic knowledge encompasing all possible word senses in the form of such semantic information like 'presuppostion', 'hyperonymy', 'hyponymy', 'synonymy', 'polysemy' etc., which mainly play a crucial role in the interpretation of natural language sentences. (see Renganathan 2016 for the interrelationship between syntax and semantics in the context of interpretation of human languages by machine).

Conclusion:

This paper, on the one hand, is an attempt to harness the power of the technologies of JSON, Vue.js, Relational database, PHP and others to the fullest extent possible, and on the other hand it lays out a comprehensive algorithm as to how these technologies can be exploited within the context of the data from the three genres of Tamil to store, retrieve and synthesize. In essence, this research is a continuation of my ongoing thrust to empower Tamil data with emerging digital technologies, and in no sense it can be considered complete. Particularly, the list manipulation technique that is described in detail in this paper and in my earlier works need fullest and continued consideration in order to foresee a comprehensive machine learning application that can interpret Tamil sentences like any human. Tamil is a complex language in many respects and its mirage of complexities can be accounted for only when the power of technology and the extensive and indepth linguistic knoweldge of Tamil can cross their paths in a productive manner.

References:

Gazdar, Gerald and Chris Mellish. (1989). *Natural Language Processing in Prolog*. Addison-Wesley Publishing Company: Wokingham.

Johnson-Laird, P. N. (1983). *Mental Models.* Cambridge: Cambridge University Press.

Renganathan, Vasu (2016) *Computational Approaches to Tamil Linguistics*. Cre-A., Chennai.

_____(2014). "Computational Phonology and the development of Text to speech application for Tamil". Paper presented and published in the proceedings of the Tamil Internet Conference, 2014. Pondicherry University: Pondicherry. (http://text2speech.tamilnlp.com/).

_____(2013). "தமிழை அறிய கணினிக்கு எத்தனை விதிகள் வேண்டும்". Paper presented and published in the Proceedings of the Tamil Internet Conference, 2013. University of Malaya: Kualalumpur, Malaysia.

_____ (2010) "Evolution of Tamil grammatical suffixes and writing Historical Grammar for Tamil" (In Tamil), Paper presented at the World Classical Tamil Conference, Coimbatore, India.

_____ (2009) "The Process of Grammaticalization and Evolution of Modern Tamil Forms". Paper presented at the Prof. Agesthalingom Commemoration conference, Annamalai University, India (August 19th to 21st, 2009).

_____ (2002) Interactive Approach to Development of English-Tamil Machine Translation System on the Web", in Proceedings of the Conference in Tamil and Internet, Foster City, San Francisco.

_____ (2001) "Development of Morphological Tagging for Tamil", In Proceedings of the International Conference on Tamil Internet 2001, Kuala Lumpur, Malaysia.

_____ (1997) "On Significance of Creation of Modern Tamil Corpora on the Web" paper presented at the First International Conference on Computerization of Tamil. National University of Singapore. May, 1997.

கணினி உதவியுடன் திருக்குறளில் வேற்றுமைகள் – ஓர் ஆய்வு

வ.மு. முத்துராமலிங்க ஆண்டவர்
PG & Research Dept. of Tamil,
Pachaiyappas College, Chennai – 30, Tamilnadu

Abstract:
Case markers in a language is the associative linguistic unit for conveying the meaning. Conveying an intended meaning is very necessary which otherwise may lead to unnecessary issues. This paper focuses on the analysis and interpretation of case markers with the combined efforts of linguistic and computational techniques for processing Thirukkural. A rule based system with Tholkapiyam to be the basis for most of the rules is being developed using python programming language for Tamil.

அறிமுகம்

உலகில் கிரேக்க நாட்டில் தான் தொன்மையான இலக்கணம் கி.மு. இரண்டாம் நூற்றாண்டில் தோன்றியது. ஆனால் இலக்கணம் தோன்றுவதற்கு முன்பே இலக்கண ஆய்வும் பல்வேறு இலக்கணக் கூறுகளாகக் கொண்ட ஆய்வுகள் தோன்றின [1]. பித்தா கோரஸ்(கி,மு.5ஆம் நூற்) ஹெரோகிளிட்டஸ், பிளோட்டோ (422 – 348) பலர் இலக்கணக் கூறுகளைக் கூறியுள்ளனர்.

தமிழின் முதல் இலக்கணம் தொல்காப்பியம் [2]. தமிழ் மொழியின் தனித்த அமைப்பை விளக்கவே தொல்காப்பியம் இயற்றப்பட்டது [3]. தொல்காப்பியத்தின் எழுத்து, சொல், பொருள் என 1611 நூற்பாக்கள் உள்ளன [4]. இந்நூல் தமிழ் மொழியின் இலக்கண இயல்புகளையும், தமிழ் இலக்கியக் கொள்கைகளையும், மொழி இலக்கிய வளர்ச்சி நிலைகளை ஆராய்ந்து உருவாக்கப்பட்ட தொன்மைத் தமிழின் மிகச் சிறந்த இலக்கண முதல் நூல் [5].

வேற்றுமைகள் - Case Markers

தொல்காப்பியத்தில் சொல்லதிகாரத்தில் கூறப்பட்டுள்ள ஒன்பது இயல்களில், மூன்று இயல்களில் தொல்காப்பியம் வேற்றுமை குறித்து இயல்களை அமைக்கிறார். மேலும் எழுத்ததிகாரத்திலும் வேற்றுமை பற்றிக் குறிப்பிடுகிறார் [5].

“ஐஓடு குஇன் அதுகண் என்னும்
அவ்வா றென்ப வேற்றுமை யுருபே”
“வல்லெழுத்து முதலிய வேற்றுமை யுருபிற்கு
ஒல்வழி ஒற்றிடை மிகுதல் வேண்டும்”

(தொல்காப்பியம் எழுத்து)

உரையாசிரியர்கள் வேற்றுமைகள் குறித்து சிறப்பாக விளக்குகின்றனர். சேனாவரையர் “ செயப்படு பொருள் முதலியனவாகப் பெயர்ப்பொருளை

வேறுபடுத்தலின், வேற்றுமையாயின என்பார். தெய்வச்சிலையார் என்ற உரையாசிரியர் இன்னும் சிறப்பாக விளக்குகிறார். " வேற்றுமை என்னும் பெயர் பொருளை வேறுபடுத்தினமையாற் பெற்ற பெயர். என்னை வேறுபடுத்தியவாறுஎனின் ஒரு பொருளை ஒரு கால் வினைமுதலாக்கியம் ஒரு கால் செயப்படுபொருளாக்கியும், ஒருகால் கருவியாக்கியும், ஒருகால் ஏற்பதாக்கியும், ஒருகால் நீங்க நிற்பதாக்கியும், ஒருகால் உடையதாக்கியதும், ஒருகால் இடமாக்கியும் இவ்வாறு வேறுபடுத்துவது என்க." என்ற உரையாசிரியர்கள் கூற்றுபடி மொழிக்கு வேற்றுமை என்ற இலக்கணக் கூறு இன்றியமையாதது என்பது புலப்படும்.

இவ்வாய்வுக் கட்டுரையில் தொல்காப்பிய வேற்றுமை இலக்கண விதிகளை அடிப்படையாகக் கொண்டு, திருக்குறளின் [6] வேற்றுமைகளை கணினி உதவியுடன் ஆராய்ந்து அரிய செய்திகளை வழங்குவதே இக்கட்டுரையின் நோக்கம். கணினி மொழியியல் துறையில் இலக்கண ஆய்வுக்கு தமிழ் ஏற்ற மொழியை எவ்வாறு பயன்படுத்துவது, கணினி நிரலாளர்கள் உதவியுடன் பைதான் (PYTHON) என்ற கணினி மொழி வழியாக திருக்குறளில் உள்ள வேற்றுமைகளை கண்டறிந்து ஆராய்ந்து இவ்வாய்வை நிகழ்த்த முயன்றுள்ளோம்.

ஒரு மொழியின் அடிப்படை அலகுகள் இரண்டு என மொழியியலாளர் குறிப்பர். 1.பெயர், 2.வினை, இந்த பெயரும் வினையும் இல்லாமல் எந்த மொழியும் இயங்குவது இல்லை. பெயரும் வினையும் தன்னியல்பிலிருந்து திரிந்து பொருளை வேறுபடுத்துகின்றன. பெயரோடு இணைந்து பொருளினை வேறுபடுத்தும் கூறு வேற்றுமை எனப்படும். பெயரும் வினையும் மட்டும் ஒரு தொடராக முடியாது. வேற்றுமை போன்ற இலக்கணக் கூறுகள் தான் இலக்கணத் தொடர்ப் பொருளை உணர்த்துகின்றன. வேற்றுமைகள் தான் தொடர்பொருள் உருவாக்கத்திற்கும் காரணமாகிறது, வேற்றுமைத்தொடர், வேற்றுமை அல்லாத தொடர், அல்வழித்தொடர், என்ற பகுப்பின் வழி இதனை அறியலாம் [7].

வேற்றுமையின் வகைகள் - Types of Case Markers

தொல்காப்பியம் வேற்றுமைக்கு நேரடியாக, இலக்கணம் கூறவில்லை. வேற்றுமை ஏழு என்று சொல்லிவிட்டு, விளி வேற்றுமையோடு எட்டு என்கின்றார். வேற்றுமை ஏழு என்று கருத்துடையவர்கள் உள்ளார்கள். ஆனால் தொல்காப்பியர் கருத்து விளியோடு எட்டு, என்றுகூறி அதற்காக விளிமரபு என தனி இயலை அமைப்பது குறிப்பிடத்தக்க ஒன்றதாகும்.

அவைதாம்,

பெயர் ஐ, ஒடு, கு, இன், அது, கண் விளி யென்ற ஈற்ற

என பெயர், ஐ,ஓடு, கு, இன், அது, கண், விளி எட்டு வேற்றுமைகளை தனித் தானியாக பிரித்து கூறுவது கணினி நிரலாளர்கள் பைதான் மொழிவழியை எளிமையாக நிரல்படுத்த, தொல்காப்பியத்தின் மொழி நிரல்படுத்தம் பெரிதும் உதவும்.

சாத்தன், சாத்தானை, சாத்தனொடு, சாத்திற்கு, சாத்தனின், சாத்தனது, சாத்தனதுகண், சாத்தா, எனப்படும். பெயரோடு எட்டு வேற்றுமைகள் பொருள் வேறுபடுத்தப்படுகின்றது.

திருக்குறள் மேற்கண்ட வேற்றுமைகளை கண்டறிந்து திருக்குறளில் உயர்ந்த இலக்கிய உச்ச நிலைக்கு வேற்றுமைகள் எவ்வாறு உதவுகின்றன என கணிணி வழி இவ்வாய்வை நிகழ்த்துவதற்கு சில முன்னோடி முயற்சிகளை முன்வைக்கலாம் ஏனெனில் கணினி மொழியில் துறையில் இலக்கண தொடர்களை பயன்படுத்தி கணினி உருவாக்கம் அறிந்த பலர் ஆய்வு நிகழ்த்தியுள்ளனர்.

வேற்றுமை தொடர்பான சூத்திரங்களை விளக்கி கொள்வது முதல்பணி சூத்திரங்களை எவ்வாறு கணினி மொழியாக மாற்றுவது, இரண்டாம் பணி அதன் வழி திருக்குறள் போன்ற தொன்மை இலக்கியங்களில் எவ்வாறு கையாளப்பட்டுள்ளன என்பதை ஆராய்வது மூன்றாவது பணி. இந்த மூன்று பணிகளில் இவ்வாய்வு செல்கிறது.

வினை முற்றுக்கு இடது புறமாக வரும் வேற்றுமைகளை வாய்பாட்டு முறையிலும், வகைப்பாட்டு முறையிலும் வகைப்படுத்தியுள்ளனர், என்கிறார் (டாக்ட்ர் பொற்கோ. 2001) உருபு அடிப்படையிலும் எண்ணுமுறை அடிப்படையிலும் பொருள் அடிப்படையிலும் தொல்காப்பியர் வகைப்படுத்துகிறார்.

தமிழ் இலக்கண மரபு, தொல்காப்பியமரபு, காக்கைபாடினியமரபு என்ற மரபுகளில் தொல்காப்பிய மரபு தமிழுக்கே உரிய தனித்துவமாக தெளிவான நோக்கோடு முறையாக அமைந்துள்ளது. வேற்றுமை பற்றி தொல்காப்பியரின் கொள்கை கோட்பாட்டு வழி தெளிவாக தெரிகிறது.

பகுதி, விகுதி, சந்தி, சாரியை, இடைநிலை, இடமிருந்து வலமாக தமிழ் இலக்கண சொற்கள் அமைகின்றன. வேற்றுமையும் பெயரோடு பொருந்தி ஈற்றில் வரும்.

வேற்றுமை வகை மாதிரிகள் Case Marker Classification with examples

வினைச்சொல்லின் வரைவிலக்கணம் பற்றி கூறிவந்த தொல்காப்பியர் வேற்றுமை பற்றியும் குறிக்கிறார்.

“ வினை எனப்படுவது வேற்றுமை கொள்ளாது
நினையுங் காலை காலமொடு தோன்றும்

வினை வேற்றுமை ஏற்காது என்றும் ஆனால் பெயர் வேற்றுமை ஏற்கும்.

இந்த வரையறை கணினி விதியாக மாற்றி நிரலாக்கம் செய்ய வேண்டும்.

பெயர் வினை

திருக்குறள் பெயருக்கு பின்னால் வருகின்ற வேற்றுமைகள்.

இரண்டாம் வேற்றுமை

இடிப்பாரை இல்லாத ஏமரா மன்னன் - ஐ
கெடுப்பார் இலானும் கெடும்.

இதனை இதனால் இவன்முடிக்கும் என்றாய்ந்து - ஐ – ஆல்
அதனை அவன்கண் விடல்

செய்வானை நாடி வினைநாடிக் காலத்தோடு -ஐ
எய்த உணர்ந்து செயல்

மடியை மடியா ஒழுகல் குடியைக் -ஐ
குடியாக வேண்டு பவர்

மூன்றாம் வேற்றுமை

வேலோடு நின்றான் இடுஎன் றதுபோலும் - ஒடு
கோலொடு நின்றான் இரவு

மதிநுட்பம் நூலோடு உடையார்க்கு அதிநுட்பம் - ஒடு
யாவுள முன்நிற் பவை

நயனொடு நன்றி புரிந்த பயன்உடையார் - ஒடு
பண்புபா ராட்டும் உலகு

பாலொடு தேன்கலந்து அற்றே பணிமொழி - ஒடு
வால்எயிறு ஊறிய நீர்

உடம்பொடு உயிரிடை என்னமற்று அன்ன - ஒடு
மடந்தையொடு எம்மிடை நட்பு

நாணொடு நல்லாண்மை பண்டுஉடையேன் இன்றுஉடையேன் - ஒடு
காமுற்றார் ஏறும் மடல்

காமக் கடும்புனல் உய்க்குமே நாணொடு - ஒடு
நல்லாண்மை என்னும் புனை

நான்காம் வேற்றுமை -கு

ஈதல் இசைபட வாழ்தல் அதுவல்லது
ஊதியல் இல்லை உயிர்க்கு -கு

பொச்சாப்பார்க்கு இல்லை புகழை அறிவினை -கு
நிச்ச நிரப்புக்கென்று ஆங்கு

அச்சம் உடையார்க்கு அரண்இல்லை ஆங்கு இல்லை -கு
பொச்சாப்பு உடையார்க்கு துணிவு

அனிச்சமும் அன்னத்தின் தூவியும் மாதர் -கு
அடிக்கு நெருஞ்சிப் பழம்

கருமணியின் பாவாய்நீ போதய்யாம் வீழும் -கு
திருநுதற்கு இல்லை இடம்

ஐந்தாம் வேற்றுமை -இன்

இகழ்ச்சியின் கெட்டாரை உள்ளுக தாம்தம் -இன்
மகிழ்ச்சியின் மைந்துஉறும் போழ்து
கடல் அன்ன காமம் உழந்தும் மடல்ஏறாப் -இன்
பெண்ணின் பெருந்தக்கது இல்
யாம்கண்ணின் காண நகுப அறிவில்லார்
யாம்பட்ட தாம்படா வாறு

ஆறாம் வேற்றுமை

உள்ளியது எய்தல் எளிதுமன் மற்றும்தான - அது
உள்ளியது உள்ளாப் பெறின்
பரியது கூர்கோட்டது ஆயினும் யானை - அது
வெரூஉம் புலிதாக் குறின்

ஏழாம் வேற்றுமை

மனத்துக்கண் மாசுஇலன் ஆதல் அனைத்து அறன் -கண்
ஆகுல நீர பிற
மனைமாட்சி இல்லவள்கண் மாண்புஆனால் உள்ளதுஎன் -கண்
இல்லவள் மாணாக் கடை
உடமையுள் இன்மை விருந்துஓம்பல் ஓம்பா -கண்
மடமை மடவார்கண் உண்டு
எழுமை எழுபிறப்பும் உள்ளுவர் தம்கண் -கண்
விழுமம் துடைத்தவர் நட்பு
கெடுவாக வையாது உலகதம் நடுவாக -கண்
நன்றிக்கண் தங்கியான் தாழ்வு
அழுக்காறு உடையான்கண் ஆக்கம்போன்று இல்லை -கண்
ஒழுக்கம் இலான்கண் உயர்வு
பிறன்கொருளாள் பெட்டுஒழுகும் பேதமை ஞாலத்து -கண்
அறம்பொருள் கண்டார்கண் இல்
பகைபாவம் அச்சம் பழிஎன நான்கும் -கண்
இகவாவாம் இல்இறப்பான் கண்
அருள்வெஃகி ஆற்றின்கண் நின்றான் பொருள்வெஃகிப் -கண்
பொல்லாத சூழக் கெடும்

திருக்குறளில் மேல்கண்டவாறு பல வேற்றுமைகள் உள்ளன. பொருள் வேற்றுமைகள் வழிதான் பொருள் வேறுபாடு உண்டாகின்றது. அதனைக் கண்டறிந்து அதற்கு பொதுவிதி அமைக்க வேண்டும். இவ்வாறு செய்யும் பொழுது ஏற்படும் சிக்கல்களை கண்டறிந்து அதற்கான தீர்வுகளை கண்டறிய வேண்டும்.

எடுத்துக்காட்டாக, ஏழாம்வேற்றுமை கண் என தொல்காப்பியம் காட்டுகின்றது. திருவள்ளுவர் நூற்றுக்கும் மேற்பட்ட குறள்களில் கண் என்ற சொல்லை பயன்படுத்துகிறார்.

மனத்துக்கண் மாஇலன் ஆதல் அனைத்தறன்
ஆகுல் நீர பிற

மனம்+ அத்து + கண் – இங்குசாரியையுடன் கண் என்பது ஏழாம் வேற்றுமை.

கண்ணொடு கண்ணினை நோக்கோக்கில் வாய்ச்சொற்கள்
என்ன பயனும் இல

கண்ணொடு – ஒடு என்ற மூன்றாம் வேற்றுமை "கண்" என்ற சொல் வேற்றுமையால் வருகிறது. கண் என்ற பெயரோடு ஒடு என்ற வேற்றுமை சேர்ந்து கண்ணொடு என்று வருகின்றது. கண் என்ற பெயரோடு இன் என்ற வேற்றுமை சேர்ந்து கண்ணினை என்று வருகின்றது. கண் என்ற பெயரோடு ஒடு இன் என்ற வேற்றுமை வழி பொருள் புலப்பாடு உன்டாகிறது. இதனை உள்வாங்கி இதற்கென விதிகள் அமைக்கப்படவேண்டும்.

ஆய்வு முறைமை

இந்த பொருண்மை வேறுபாட்டை கணிப்பொறிக்கு ஊட்டுவது மிகப்பெரும் சிக்கல் [8,9]. இதற்கென பைதான் (python) கணினி நிரலாக்க மொழிவழியாக முயற்சி செய்யலாம் [10]. இது தொடக்க நிலை முயற்ச்சிதாள் அமையும். மொழிகளில் இலக்கண அறிஞரும் இனைந்து கூட்டாக செய்யப்படவேண்டியது இது.

பைதான் (python) என்ற கணினி நிரலாக்க மொழியினை திருக்குறளில் இடம்பெற்றுள்ள வேற்றுமைகளைக் கண்டறிய பயன்படுத்தியதற்கு இன்றியமையா காரணம் உண்டு. மனிதர்கள் பேசும் இயற்கை மொழிகளின் உள்ளீடு(input) மற்றும் வெளியீடு பற்றி (output) ஆராய்வதற்கு மற்ற கணினி மொழிகளைக் காட்டிலும் எளிமையாக கட்டமைக்கப்பட்ட மொழி இம்மொழி [11]. இயற்கை மொழிஆய்விற்கு மற்ற கணினி மொழிகளை காட்டிலும் பைதான் மொழி நிரலாளர்களால் வடிவமைக்கப்பட்டுள்ளது.

பைதான் மொழியில் உள்ள உள் கட்டமைப்பு செயல்முறைகள் (inbuilt functions) மூலம் மொழிசார் இலக்கண கூறுகளை கண்டறிந்து ஆராய இம்மொழி அமைப்பு எளிதாக உதவுகிறது [12,13]. மேலும் மொழிசார் இலக்கண பெருந்தரவுகளை மற்ற கணினி மொழிகளில் ஆராய முற்பட்டோம் ஆனால் கால விரயம் ஏற்படும். ஆனால் பைதான் மொழியில் எத்தனை பெரிய மொழிசார் இலக்கணத் தரவுகளை ஆராய தகவமைப்புப் பெற்ற மொழியாகத் திகழ்கிறது. எவ்வளவு கடினமான இலக்கணப் பணிகளையும் பைதான் மொழிவழியாக எளிமையாகச் செய்யலாம். அத்தகைய திறனைக் கொண்டது இம் மொழி.

மேற்கண்ட காரணங்களால் பைதான் மொழியை தேர்வுசெய்து அதன் வழி திருக்குறள் வேற்றுகைளைக் கண்டறிந்து வகைப்பாடு செய்து வேற்றுமைக் கோட்பாட்டை உருவாக்க முயற்சிசெய்துள்ளோம். தொல்காப்பிய வேற்றுமை தொடர்பான விதிகள் வழி திருக்குறள் வேற்றுமைகளைக் கண்டறிய பைதான் மொழிவழியாக என்னென்ன வழிமுறைகள் உள்ளன என்பதை கண்டறிய முயன்றுள்ளோம்.

ஆய்வு முடிவு

கணினி நிரலாளர்களும் மொழியிலாளர்களும் இணைந்து கணினிமொழியியல் என்ற துறையில் இணைந்து பணியாற்றவேண்டிய பணிகளில் தொல்காப்பிய வழி திருக்குறள் வேற்றுமைகளைக் கண்டறிவதும் வகைப்பாடு செய்வதும், மிக இன்றியமையாப் பணி. இப்பணிக்கு ஆற்ற வேண்டிய தொடக்கநிலைப் பணிகளை இக் கட்டுரையில் குறித்துள்ளோம். .இந்த உருவாக்கத்தினை கல்வியாளர் மற்றும் மக்கள் பயன்பாட்டிற்கு ஏற்ற வகையில் வெளியிட விரிவான திட்டத்திற்கு முதற்பணியாக இப்பணி அமையும்.

ஆய்வுக்கு துணைநின்ற நூல்கள்

1. OLD GRAMMER OF TAMIL, AGASTHIYALINGAM
2. தொல்காப்பியம் தெய்வச்சிலையார் உரை
3. தொல்காப்பிய உருவாக்கம், டாக்டர் அகஸ்தியலிங்கம்.
4. தொல்காப்பிய அறிமுகம், டாக்டர் பொற்கோ.
5. தொல்காப்பியம் எழுத்து, சொல், பொருள், தமிழண்ணல் உரை.
6. திருக்குறள் பரிமேலழகர் உரை
7. மொழிநூல், மு. வரதராசனார்
8. Suzuki, Kristina Toutanova Hisami, and K. Toutanova. "Generating case markers in machine translation." Proceedings of NAACL HLT. 2007.
9. Schiffman, Harold/ F. "The Tamil case system." South Indian horizons: felicitation volume for Francois Gros on the occasion of his 70th birthday (2004): 293-322.
10. Steven, Bird, Ewan Klein, and Edward Loper. "Natural language processing with python." OReilly Media Inc (2009).
11. Lutz, Mark. *Programming Python: Powerful Object-Oriented Programming*. " O'Reilly Media, Inc.", 2010.
12. Smedt, Tom De, and Walter Daelemans. "Pattern for python." *Journal of Machine Learning Research* 13.Jun (2012): 2063-2067.
13. Van Rossum, Guido. "Python Programming Language." *USENIX Annual Technical Conference*. Vol. 41. 2007.

Sentence-medial pause identification for Tamil synthesis system

K. Mrinalini, G. Anushiya Rachel, T. Nagarajan, P. Vijayalakshmi
Speech lab, SSN College of Engineering, India
Email: (nagarajant,vijayalakshmip)@ssn.edu.in

Abstract:

Sentence-medial pause in synthetic speech is essential to make it sound more natural and intelligible. In the current work, part-of-speech (POS) based rules are used to identify the place-of-pause/phrase-breaks in any given Tamil text, to result in a pause-induced Tamil synthetic speech, which is expected to be natural. The place-of-pause is identified using unigram and bigram statistics from phrase-break annotated text corpora containing text from different domains such as tourism, sports, historic novels etc. The variation of phrase-breaks with respect to the domain of the text is also analysed and an overlap of 54.54% is observed across these domains. The unigram rule set for phrase-break introduces spurious breaks in a sentence which is reduced by the bigram rule set. Theplace-of-pause induced speech is synthesized using a hidden Markov model (HMM) based text-to-speech synthesis system (HTS) for Tamil. The quality and intelligibility for naturalness of the synthetic speech is evaluated using mean opinion score (MOS). The synthetic speech with pauses shows an improvement of0.93in MOS score over the synthetic speech without pauses.

Keywords: place-of-pause, phrase-break,bigram and unigram statistics, HTS

1. Introduction:

The right word may be effective, but no word was as effective as a rightly timed pause [Mark Twain].A sentence-medial pause(i.e., pause within a sentence) in speech or phrase-boundary in text have the power to add emotion and improve clarity of the intended message.For example,

Sentence without phrase-break/pause: Kiran likes to eat chocolate pizzas and cabbage.
Sentence with phrase-break/pause: Kiran likes to eat chocolate, pizzas, and cabbage.

The former sentence delivers the meaning that Kiran likes to eat pizzas and cabbage covered in chocolate while making use of comma in the latter gives the meaning that Kiran likes to eat three things namely, chocolate, pizza, and cabbage. It is evident from the above example that the sentence with pause is easily readable and understandable than the sentence without pause. Phrase boundaries in a text depict expected place-of-pause in speech. The pause or silence factor becomes more important in case of speech in order to incorporate a sense of naturalness and to improve intelligibility in it. The importance of pause in speech is attributed to the following factors:

- Pause helps in punctuating the words in speech to emphasize on important contents.
- Pause helps define emotion in speech.
- Proper pausing allows the listener to understand the speech content.
- Pausing helps the speaker catch some breath and gives time to build the next set of words to be spoken.

Thus, pauses in sentences (sentence-medial pauses) have the ability to improve the quality of the text and speech in terms of clarity of content.As given in [1], the sentence-medial pauses in any language depend on several factors such as, syntactic type of phrase that precedes a syntactic boundary (i.e., parts-of-speech information), phrase length, and distance between the current phrase and its dependent phrase. These factors are represented in written forms with the help of appropriate case-markers. Indian languages inherently lack case-markers or phrase-boundaries in its written form [2], [3]. Though modern writing include some case-markers such as, comma (,) and semi-colon (;), they are mostly artificial. Tamil does not make use of phrase boundaries except period (.) in its written form and identifying these additional phrase boundaries remain a research issue.

Speech synthesis systems aim at synthesizing highly natural and intelligible speech for a given text in any language. The state-of-the-art hidden Markov model (HMM) based speech synthesis system (HTS) for Tamil [4] results in synthetic speech of good quality in terms of intelligibility. However, incorporating naturalness in this synthetic speech remains a challenge. The above discussed phrase-break identification is necessary to induce phrase-breaks in Tamil text as a pre-processing stage before synthesizing the corresponding text. The phrase-breaks are converted to silence models in the synthesizer thus inducing sentence-medial pauses in the final synthetic speech. This is expected to improve the naturalness of the synthetic speech.

In [2], phrase-boundary prediction for Tamil and Hindi is carried out based on a decision tree developed on a manually annotated corpora. Morpheme tags occurring at the end of words having the need of phrase-breaks are identified manually. These morpheme tags are included as features and phraseboundaries are introduced after the occurrence of the morpheme tags. Though successful, this approach requires the laborious task of identifying the morpheme tags for the language. In [5], syllable level features i.e., word-terminal syllables are used to model phrase breaks. The terminal syllables serve to discriminate words based on syntactic meaning, and can therefore be used to model phrase breaks. The phrase-break model is developed based on the probability of various terminal syllables occurring before a phrase-break in the training text. This approach does not make use of any additional linguistic resources such as POS tagging. Improvement in terms of naturalness is achieved using phrase-prediction based on terminal syllables. However, relying only on the syllable feature can lead to deletion of required phrase-breaks or insertion of unnecessary phrase-breaks as it does not consider the purpose of the word in a particular context.

In the current work, phrase boundaries in Tamil text are identified based on parts-of-speech (POS) tags and these phrase-breaks are used to synthesize a pause induced speech in Tamil.Fig.1 shows the proposed system which consists of the proposed phrase-boundary

detector for Tamil to identify the place-of-pause in any given Tamil text, and the Tamil HTS system which converts the pause-induced Tamil text to pause-induced Tamil speech with improved naturalness.

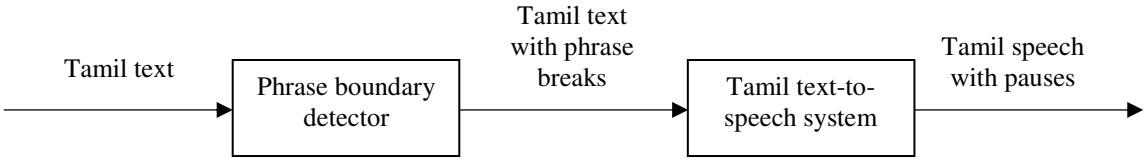


Fig.1 Pause-induced Tamil speech synthesis system

2. Phrase-boundary detector:

As described above, phrase-boundaries in text are assumed to be the expected place-of-pause in speech.POS tagging helps in defining the nature of the words in a given context and thus is a reliable source to identify the importance of the words as a phrase boundary. For example, Tamil is a verb-ending language where phrases and sentences predominantly end with verbs. The Tamil POS tagset [6] contains of 9 POS tags under the verb category, based on the context of the word, tense, and number. However, not all these 9 tags are essential phrase boundaries in Tamil. Thus, the phrase-boundary identification is carried out using unigram and bigram statistics of the POS tags from POS tagged Tamil texts having manually inserted phrase-breaks and covering various domains such as, tourism, agriculture, health, and general domain. The unigram statistics are taken by identifying the POS of the word following which there is a comma in the text, whereas bigram statistics are obtained by extracting the POS of two words before the occurrence of comma. In the current work, the unigram and bigram statistics refer to the frequency count distribution of the POS tags. Analysis of phrase boundaries in the texts of different domains is carried outto check the possibility of variation in phrase-breaks due to variations in domains.

2.1 Domain analysis:

Tamil texts with manually annotated phrase-breaks from 4 different domains is POS tagged [6] to analyse the variation in unigram phrase-break statistics across domains. The general domain text [7] contains 3732 sentences whereas the tourism, health and agriculture texts [8] contain 15200, 14999 and 3400 sentences respectively.**Table I** shows the top 6 frequently occurring POS tags (covers more than 75% of phrase-breaks) in various domains based on unigram statistics.

An overlap of 54.54% of top POS tags is observed across domains whereas the rest vary depending on the domain. While the occurrence of nouns and verbs as phrase-breaks across the domain remains consistent, other POS tags such as CC_CCS (conjunctions), RB (adverbs) and DM_DMQ (numerals) as phrase-breaks vary with respect to the domain. This variation can be credited to the fact that the structure of sentence and vocabulary of each domain is unique. In the current work, general domain is analysed further to derive phrase-boundary rules.

Table I: Distribution of POS based on unigram statistics across various domains

General Domain	Health Domain	Tourism Domain	Agriculture Domain
N_NN	**N_NN**	**N_NN**	**N_NN**
V_VM_VF	**V_VM_VF**	**V_VM_VF**	**V_VM_VF**
V_VM_VNF_VBN	**PSP**	**V_VM_VNF_VBN**	**V_VM_VNF_VBN**
N_NNP	CC_CCS	**PSP**	**PSP**
PSP	**V_VM_VNF_VBN**	N_NNP	RB
DM_DMQ	RB	RB	N_NNP

2.2 POS-based phrase boundary:

The general domain text is analysed and the frequency counts of POS of words preceding a comma in the text is obtained. It is observed that some POS tags (having frequency counts below 10) rarely occur before a phrase-break and hence can be neglected from the rule set.The POS tags are arranged in descending order of their frequency countand are used to frame the unigram rule set. The unigram rule set consists of 11 rules (i.e., 11 frequently occurring unigram POS tags).

For example,

Unigram rule: If (POS of the word == N_NN) then insert a comma after the word.

Sentence using unigram rule:வளர் இளம்பருவத்தில், தினமும், ஒரு முட்டையின், வெள்ளைக்கரு, சாப்பிடுவது, நல்லது.

It is observed that, phrase-breaking using the unigram ruleset induces spurious and unnecessary phrase-breaks in sentences. Thus to reduce this, bigram POS statistics is obtained where the frequency counts of POS of two words preceding a comma is used to frame better rules.21 bigram rules are framed using the bigram frequency counts. Apart from the POS derived rules, phrase length between two consecutive phrase-breaks is also considered to avoid frequent phrase-break insertions resulting in unnecessary pauses in the final synthetic speech.The distance between phrase-breaks was fixed to be a minimum of 3 words.

For example,

Bigram rule: If (POS bigram == JJ N_NN) and (dist. from previous break >= 3) then insert comma after the word.

Sentence using bigram rule: வளர் இளம் பருவத்தில், தினமும் ஒரு முட்டையின் வெள்ளைக்கரு சாப்பிடுவது, நல்லது.

Common noun (N_NN) is a high frequency POS tag in any text and using the unigram rule above results in spurious commas. This issue is reduced by the bigram rule which states that a noun preceded by an adjective (JJ) is a probable place of phrase-break. Similarly, bigram rules for other POS tags are also framed some of which are given below:

Rule 1: If (POS bigram == N_NN PSP) and (dist. from previous break >= 3) then insert comma after the word.(where PSP – postposition)

Rule 2: If (POS bigram == N_NN V_VM_VF) and (dist. from previous break >= 3) then insert comma after the word. (where V_VM_VF – finite verb)

The pause-induced Tamil text is synthesized using a hidden Markov model (HMM) based speech synthesis system (HTS) for Tamil.

3. Tamil text-to-speech synthesis system:

A text-to-speech (TTS) synthesis system converts any given text to the corresponding speech. Hidden Markov model (HMM) based text-to-speech synthesis system (HTS) [9] is a widely used approach for speech synthesis which consists of two phases namely, training phase and synthesis phase. In the training phase, HMMs are trained with features derived from one hour of speech data collected. Each feature vector includes Mel-generalized coefficients with its derivative (35*3) and excitation features namely, the log fundamental frequency with its derivatives (1*3), thus totalling to a 108-dimensional feature vector. Context-independent models are trained for all the phonemes in the data following which context-dependent pentaphone models are trained using tree-based clustering. For the current work, common phoneset in [7] and letter-to-sound rules defined in [4] are used. In the synthesis phase, text input is converted to a pentaphone sequence which is used to concatenate HMMs and form a sentence-level HMM. From the sentence HMM, the fundamental frequency and cepstral coefficients are generated using a speech parameter generation algorithm. The new parameters are used to synthesize the speech output using a Mel log spectrum approximation (MLSA) filter.

As shown in Fig.1, the phrase-break induced text is given to a Tamil text-to-speech synthesis system as input. In order to develop the Hidden Markov model (HMM) based text-to-speech synthesis system (HTS), five hours of Tamil speech data is collected from a femalespeaker in studio environment. The speech is recorded at 48 KHz using a carbon microphone. The phrase-breaks occurring in the input text are converted into silence models in the HTS system which results in a pause at that position in the output synthetic speech.

3.1 Pauses insynthetic speech:

Pauses in synthetic speech is expected to increase its intelligibility and naturalness. The phrase break process using bigram rule-setdiscussed earlier, is used to insert phrase-breaks in appropriate places in the input target language text. This text is synthesized using the above developed HTS for Tamil.

The naturalness of the synthetic speech with and without proposed pauses, is evaluated for 20 sentences using the mean opinion score (MOS) [10] given by 10 human evaluators. The 20 sentences are induced with phrase-breaks using the proposed method along with the already existing phrase-breaks in modern Tamil text.The MOS varies between 1 and 5 where 5 denotes that the synthetic speech is highly natural and 1 denotes poor naturalness in the synthetic speech. The average MOS for the systems are given in Table II.It is observed that, the speech with pause is more natural due to the pauses identified in the input Tamil text.

Table II. Performance evaluation of speech synthesis system in terms of MOS

Speech Synthesis System	**MOS**
Without proposed pauses	2.89
With proposed pauses	**3.72**

4. Summary and future work:

Sentence-medial pauses are essential in synthetic speech to make it sound more natural and intelligible. The sentence-medial pauses in written form is depicted using case-markers. The absence of case-markers in Tamil makes it difficult to identify the phrase boundaries in Tamil text. The current work, proposes the integration of phrase-boundary detector and speech synthesis system where the phrase-break induced by the former is converted to sentence-medial pauses in the latter.The phrase-boundaries are detected using rule framed from the statistics of unigram and bigram POS tags preceding a phrase-break. The phrase-breaks across domains is observed to have an overlap of upto 54.54%. The phrase boundary inserted text is synthesized using an HMM-based speech synthesis system (HTS) for Tamil. The output of the HTS is evaluated using MOS and compared with the MOS of speech synthesized using text without phrase boundaries. It is observed that pause-induced text shows an improvement of 0.93 in MOS score.

The phrase-boundary or place-of-pause depends on other factors such as utterance length, nativity of the speaker, domain knowledge of the speaker, end syllable characteristics etc. These features can be combined with the current POS features to improve the quality of phrase-boundaries identified in the current work which in-turn will improve the naturalness of the synthetic speech.

References:

[1] Fujisaki, Hiroya, Sumio Ohno, and Seiji Yamada. "Factors affecting the occurrence and duration of sentence-medial pauses in Japanese text reading." Proc. ICPhS. Vol. 99. 1999, pp.659-662.

[2] A. Bellur, K. B. Narayan, R. K. K, and H. A. Murthy, “Prosody modelling for syllable-based concatenative speech synthesis of Hindi and Tamil,” in National Conference on Communications (NCC), Bangalore, January 2011, pp. 1– 5.

[3] H. Verlag, “Tamil Language for Europeans”, Ziegenbalg’s Grammatica Damulica. Hubert & Co., 2010.

[4] G. Anushiya Rachel, V. Sherlin Solomi, K. Naveenkumar, P. Vijayalakshmi, and T. Nagarajan, “A small-footprint context-independent HMM-based synthesizer for Tamil,” International Journal of Speech Technology, vol. 18, no. 3, pp. 405–418, 2015.

[5] K. S. P. Anandaswarup Vadapalli, Peri Bhaskararao, “Significance of word-terminal syllables for prediction of phrase breaks in text-to-speech systems for Indian languages,” in Proceedings of 8th ISCA Speech synthesis Workshop, Barcelona, Spain, September 2013, pp. 89–194.

[6] L. Sobha, G. Sindhuja, L. Gracy, N. Padmapriya, A. Gnanapriya, and N. H. Parimala, “AUKBC Tamil parts-ofspeech corpus (aukbctamilposcorpus2016v1),” 2016.

[7] B. Ramani, S. L. Christina, R. G. Anushiya, V. S. Solomi, M. K.Nandwana, A. Prakash, S. A. Shanmugam, R. Krishnan, S. K. Prahalad,K. Samudravijaya et al., “A common attribute based unified HTSframework for speech synthesis in Indian languages.” in SSW8, 2013,pp. 291–296.

[8] “English-Tamil Parallel Text Corpus (tourism, health and agriculture) - EILMT.” LINGUISTIC RESOURCE, TDIL, MeitY, India, 2016.

[9] K. Tokuda, H. Zen, and A. W. Black, “An HMM-based speech synthesissystem applied to English,” in IEEE Speech Synthesis Workshop, 2002,pp. 227–230.

[10] Viswanathan, Mahesh, and Madhubalan Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale." Computer Speech & Language 19.1 (2005): 55-83.

Prefix Trees (Tries) for Tamil Language Processing

Elango Cheran

Enabling Tamil language computing on each new technology platform requires ensuring that each layer of the technology stack supports the language. While it is useful to assess what those layers are, and the progress that has been made for Tamil over the years, it is also important to look forward to solving future problems that need work at higher-level layers. Towards that goal, the prefix tree (trie) is an important data structure that can be used to enable basic Tamil language operations that enable more advanced work for Tamil to be done. The ways in which we can apply prefix trees for Tamil are general enough that they would very likely apply to other Indic languages, too.

A prefix tree is a special type of tree data structure that is used to efficiently store several strings that may share various prefix substrings in common with each other. Each letter of a string is stored as a node, with each subsequent letter stored in a child node of the previous letter's node. For example, if we constructed a prefix tree to hold the strings [bot, bow, be, bed, go, got], it would look like:

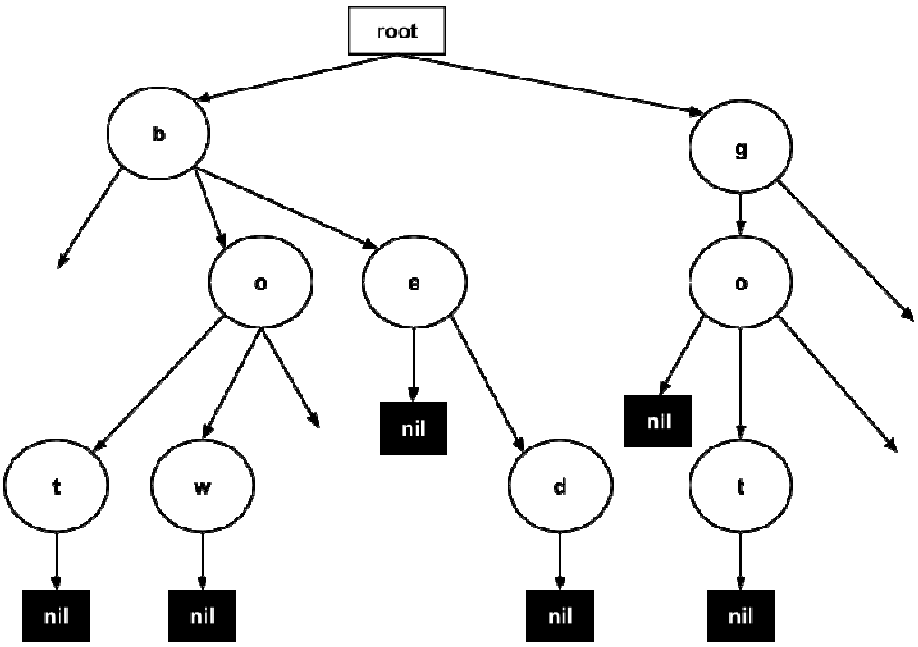


Figure 1: a prefix tree constructed from a list of strings

More generically, prefix trees can contain sequences made up of (a set of) elements. Most commonly, prefix trees are used to represent strings, and strings are merely sequences of characters. Following each path from the root of a prefix tree to a leaf node will contain the ordered elements of an input sequence. In **Figure 1**, a leaf node is represented as "nil", which can be considered like an end-of-word marker that is not a part of the tree's input sequences.

There are some easily solvable challenges in the basic processing operations of Tamil text for which prefix trees offer a natural solution. To explain those challenges, let's first examine the English example in Figure 1. In the Unicode character set, every English letter is represented by a single Unicode codepoint in the specification. In a programming language like Java which supports Unicode, these codepoints each map to a single Character value, where the Character refers to the data type as provided by the programming language. So an implementation of a prefix tree that only supports English could have each internal tree node hold a single Character value.

The Unicode specification for Tamil (and other Indic languages) represents the logical letters of the alphabet sometimes with more than codepoint. Letters like அ..ஔ (vowels, or *உயிரெழுத்து*) and க..ன (consonant+"a", or *அகரமெய்யெழ்த்து*) are all represented by one codepoint. Letters like க் (consonant, மெய்யெழுத்து) and கா..கௌ (C+V except C+அ, அகரமெய்யெழுத்து தவிர மெய்யெழுத்து) are represented by two successive codepoints in Unicode. **See Figure 2**. In the end, there is a one-to-one correspondence between a Tamil Unicode character sequence and the logical Tamil text that it creates, and that is the ultimate goal of any universal language encoding specification. Any challenges to deal with the text thereafter at a higher level are technical ones for software developers.

The above description of how to map Tamil language letters into the corresponding Unicode codepoint(s) already reflects the challenge somewhat. Figure 2 shows not only the difference between number of logical letters and codepoints, but it also shows the counter-intuitive property that the character sequence of வருக is a subset of வருகை. For these reasons, one of the first tasks that naturally arises when dealing with Tamil text in Unicode is

to convert the character sequence into a sequence of strings representing the logical letters. This parsing task, like any other stateful task involving transitions between states, can be represented by a finite state machine (FSM). The FSM represents all of the logic used to enumerate the states and determine the conditions required in order to transition from one particular state to another.

Figure 2: a list of words, letters, and Unicode points for English and Tamil

Text	**Logical letters**	**Number of logical letters**	**Text Unicode codepoints**	**Number of Text Unicode codepoints**
go	g, o	2	g, o	2
got	g, o, t	3	g, o, t	3
bot	b, o, t	3	b, o, t	3
bow	b, o, w	3	b, o, w	3
தணி	த, ணி	2	த, ண, ி (0BBF)	3
தணிகை	த, ணி, கை	3	த, ண, ி (0BBF), க, ை (0BC8)	5
வருகை	வ, ரு, கை	3	வ, ர, ு (0BC1), க, ை (0BC8)	5
வருக	வ, ரு, க	3	வ, ர, ு (0BC1), க	4

Parsers naturally can be described using FSMs, and in the case of parsing Tamil text, the FSM containing all valid transition paths is the prefix tree itself. Thus, the prefix tree containing the string representations of all logical Tamil letters is the simplest mechanism for writing code to parse Tamil text (**Figure 3**).

For example, a Tamil text-to-letter parsers written in more imperative code will require complex logic for consonants since there must be a peek ahead at the next character (if there is one) to distinguish அகரமெய்யெழுத்து (C+அ) from மெய்யெழுத்து (C) or any other

உயிர்மெய்யெழுத்து (C+V). In the case of வருகை, after parsing the வ, the next character in the stream is ர. Until we look ahead at the next character, we won't know if ர is the full letter (in the case of the word வரவு) or if the next character should be combined for the full letter (in the case of வருகை, where the next character after ர is ◌ு (0BC1)). The imperative style implementation also requires verbose nested if-else statement logic.

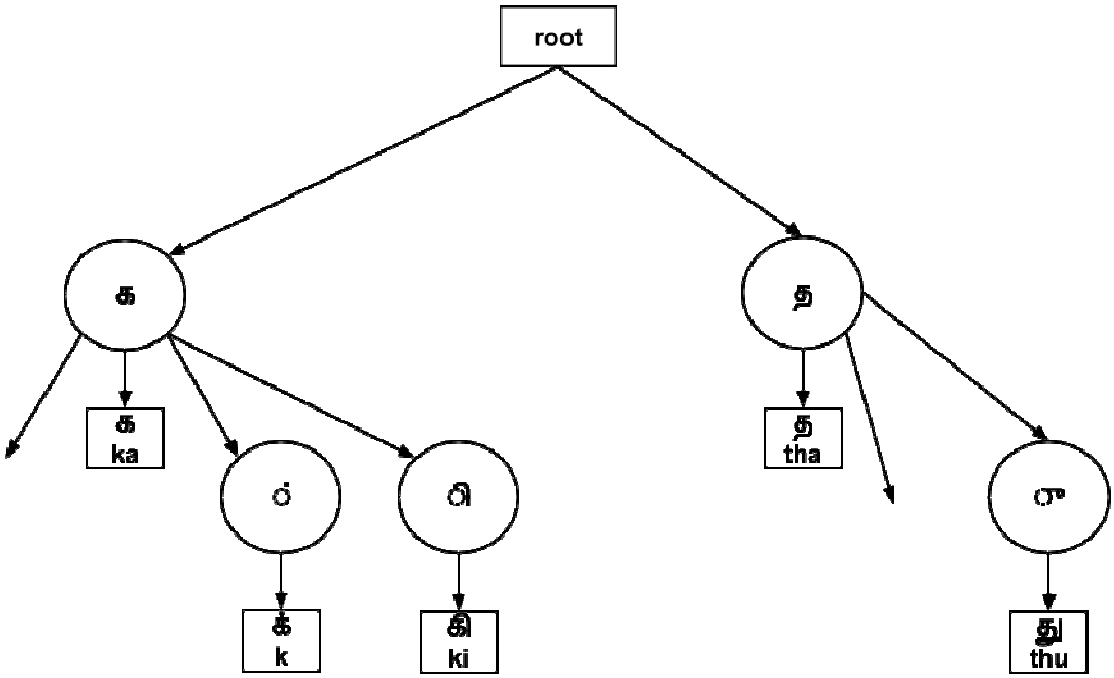


Figure 3: a prefix tree containing each Tamil letter's character sequence representation

However, a prefix tree offers the operation of longest shared prefix, meaning that given an input string, it will return the longest string in the prefix tree shared with the input. In the English example, with an input of "gothic", the prefix tree will return "got", even though "go" exists in the tree and is also a prefix. This operation is exactly all that is needed for Tamil text-to-letter parsing. **See Figure 4.**

The discrepancy in approaches is more apparent for letters with longer Unicode codepoint sequences. Some Grantha letters in Tamil text require up to 4 codepoints. An imperative style parser code may need a 4-level nested if-else block, but the prefix tree-based parser code remains unchanged. Dealing with letters requiring more than 2 characters is an infrequent case for Tamil text, but it is perhaps relevant and significant for other Indic languages.

Figure 4: the input and output of a prefix tree's longest shared prefix function

Input string	Output from Tamil letter prefix tree's longest prefix function
வருகை	வ
ருகை	ரு
கை	கை

It is important to take note that many useful Tamil language operations do not happen at the unit of a letter (எழுத்து) but instead at unit of a phoneme (ஒலியன்). Tamil, as an agglutinative language, encodes much grammatical information through word suffixes. Suffixes (விகுதி) are so common that the basic rules governing word changes when adding suffixes are given a name (சந்தி, "sandhi" - to meet). A simple case to show the importance of phoneme units over letter units for grammar would be to indicate "and" for a compound subject, indicated by adding "-உம்" to each subject. For the words மாமா and மாமி, the sandhi rules imply the changes மாமா + (வ்) + உம் = மாமாவும் and மாமி + (ய்) + உம் = மாமியும், resulting in "மாமாவும் மாமியும்". The inserted வ் and ய் is based on the vowel sound of the final letter in both words, not the final letter's consonant sound. The words அக்கா and அண்ணா also add வ், whereas தங்கை and தம்பி add (ய்).

Figure 5: letters vs. phonemes for Tamil words

Word	Letters	Final letter	Phonemes	Final phoneme	Sandhi change (-உம்)
மாமா	மா, மா	மா	ம், ஆ, ம், ஆ	ஆ	வ்
மாமி	மா, மி	மி	ம், ஆ, ம், இ	இ	ய்
அக்கா	அ, க், கா	கா	அ, க், க், ஆ	ஆ	வ்
தங்கை	த, ங், கை	கை	த், அ, ங், க், ஐ	ஐ	ய்
அண்ணா	அ, ண், ணா	ணா	அ, ண், ண், ஆ	ஆ	வ்
தம்பி	த, ம், பி	பி	த், அ, ம், ப், இ	இ	ய்

There are extra sandhi rules applied in the case of noun case suffixes (வேற்றுமை). For two words with the same last letter, மடு and காடு, adding the same case suffix "-இல்" operates differently. For மடு, the change is simpler: மடு + (வ்) + இல் = மடுவில். For காடு, an arithmetic on the word happens first: காடு + -இல் = (க், ஆ, ட், உ) - உ + ட் + (இ, ல்) = காட்டில். Because of the final -ட், the final -உ is dropped, the -ட் is doubled, and then the இல் is added.

We can use a prefix tree to split a Tamil word into phonemes by modifying the tree to allow a value to be associated with each leaf node (input string), much like a map/dictionary. Each Tamil letter string in the tree is associated with its phoneme sequence: கி -> [க், இ]; கூ -> [க், ஊ]; க் -> [க்] **(Figure 6)**. Linguistic operations in Tamil often operate on a sequence of phonemes and return a sequence of phonemes (**Figure 7**). Given phonemes, a prefix tree can be created that converts the phoneme sequence back into regular Tamil text (Figure 8).

Figure 6: the input and output of a Tamil letter prefix tree modified to be associative

Input string	**Output from Tamil letter prefix tree's associated phoneme sequence**
காடு	[க், ஆ]
டு	[ட், உ]

Figure 7: a function to prepare a noun for adding a case suffix (வேற்றுமை)

```
(வரையறு-செயல்கூறு வேற்றுமை-முன்-மாற்றம்
 [சொல்]

 (வைத்துக்கொள் [
        எழுத்துகள் (தொடை->எழுத்துகள் சொல்)
        ஒலியன்கள் (தொடை->ஒலியன்கள் சொல்)
        கள (கடைசி எழுத்துகள்)
        கஒ (கடைசி ஒலியன்கள்)]
```

```
(பொறுத்து
...
(= "டு" கள)
(செயல்படுத்து தொடை (தொடு (கடைசியின்றி எழுத்துகள்) ["ட்ட"]))

(= "று" கள)
(செயல்படுத்து தொடை (தொடு (கடைசியின்றி எழுத்துகள்) ["ற்ற"]))

:அன்றி
சொல்)))
```

Figure 8: the input and output of a prefix tree constructed as the inverse of Figure 6's tree

Input string	Output from inverse Tamil phoneme prefix tree's associated letter string
க்ஆட்ட்இல்	கா
ட்ட்இல்	ட்
ட்இல்	டி
ல்	ல்

Using prefix trees that have been modified to be associative, we can easily define conversions from Tamil text to the old pre-Unicode encodings (and vice versa). More importantly, once we start to think of Tamil text less in terms of the underlying character sequences and more as sequences of logical letters or logical phonemes, implementing other operations becomes clearer. For example, true lexicographical sorting can be achieved for Tamil by splitting a word into letters and combining with a lookup map that indicates the relative ordering of each letter. If we define the lexicographical ordering of Tamil letters as [அ, ஆ, ..., ஃ, க், க, கா, ..., கௌ, ங், ங, ..., ன், ன, ..., னௌ], our lookup map would be {அ 0, ஆ 1, ..., ஃ 12, க் 13, க 14, கா 15, ..., கௌ 25, ங் 26, ங 27, ..., ன் 235, ன 236, ..., னௌ 247}. Then sorting Tamil words becomes equivalent to sorting sequences

of numbers, which is straightforward. In a sense, sorting sequences of numbers is equivalent to sorting English text because of the one-to-one mapping of English letters and Unicode codepoints.

Once we use the appropriate data structures to model our domain more accurately, the functions we need to solve our basic problems become clear, and we can begin to solve more advanced problems. For example, an intelligent spell checker might use sequence alignment to measure closeness, for which the 2 most basic methods are global (Needleman-Wunsch) and local (Smith-Waterman). Using the phoneme representation of strings would not only fit the algorithms' designs, but it would also provide better results. There is much room to explore the implications of a phoneme-based modeling of Tamil text (as is done in Korean and other languages), but prefix trees offer a necessary first step in that direction, and they are general enough to be applicable to other Indic languages, if not more.

An open-source library that implements these ideas, along with example projects in Java and JavaScript using the library, including the above code, at: https://github.com/echeran/clj-thamil.

Quantifying shifts in language use among internet-using Tamil speakers

Vasanthan Thirunavukkarasu[1,2,*], Jonathan P. Evans[3], Sankar Raman[4], Sachit Mahajan[5], Mrinal Kanti Baowaly[5], Priyadharsini Karuppuswamy[1,2], and Sailesh Rajasekaran[6]

[1]Department of Engineering and System Science, National Tsing Hua University, Hsinchu, Taiwan
[2]Nano Science and Technology Program, Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan
[3] Institute of Linguistics, Academia Sinica, Taipei, Taiwan.
[4] Institute of Physics, Academia Sinica, Taipei, Taiwan.
[5] Social Networks and Human-Centered Computing Program, Academia Sinica, Taipei, Taiwan.
[6] Department of Material Science and Engineering, National Chiao Tung University, Hsinchu, Taiwan.
[*] E-mail: nanothamizhan@gmail.com

Abstract— Information is knowledge. Internet is the new-age tool of knowledge sharing. Amount of information available in internet as well as the amount of information that is dissipated through internet is immense. Thus, internet shapes human language and lives. Over the past, the way people interact and learn has changed enormously. In this age of information, it is necessary to learn how much language and communication is influenced in the living society. Thus, many new fields of studies such as natural language processing (NLP), speech recognition, language recognition, speaker recognition, computer-assisted teaching and learning are extensively researched. Tamil being one of the oldest surviving-classical language has adapted into various forms since its origin. But influence of other languages on Tamil was less significant in the past compared to now. English being one of the largest spoken languages in world, plays a key role in influencing native Tamil speakers today. In this work, through our research studies, we quantify the shifts in language use among the internet-using Tamil Speakers.

Index Terms— Internet, Social Networks, Human-Centered, Tamil-Computing.

INTRODUCTION

TAMIL is one of the longest-surviving classical languages in the world. Extensive efforts are implemented by researchers and Tamil diasporas around the globe to protect the fineness of this classical language. Recently Harvard Tamil Chair was established to elevate the Tamil language research. Scholars and experts in information technology are developing many ways by which classical Tamil can be preserved. In India more than one hundred thousand stone inscriptions were found. out of which around 60,000 stone inscriptions were in Tamil, which dates back to stone age period. This proves that Tamil is one of the oldest language spoken in India. Apart from this many epigraphs and petro graphs were found in various archeological excavation sites located in Tamil Nadu, India. Recent excavations in Keezhadi in Sivagangai district of Tamil Nadu provided evidences of ancient civilization that lived in banks of river

Vaigai. These findings dates back to Sangam era period. As humans evolved, languages were developed as a means of communication. These communications initially were passed on to generations only through sounds. Later these sounds were inscribed in stones as inscriptions in pictographic forms to represent a particular object. As thousands of years passed by, grammar was adapted. Early human civilizations recorded events, described things, in palm leaf manuscripts. As grammar usage got matured, excellent literatures were written. Many ancient Tamil literatures in palm leaf manuscripts aging thousands of years are being preserved by archeology department. Colonization and western printing technology converted many of these literatures into books. Millions of books were published in last couple of hundred years. The way Tamil language is spoken and the way Tamil script is written in stone-age inscriptions, palm-leaf manuscripts and books evolved with time. Today in this digital era huge amount of information is stored in digital-media. Especially social-media has become a platform were more information is being shared instantly.

In this work, we quantified the shift in language use among internet-using Tamil speakers. For a comprehensive analysis, we studied a total of 100 Tamil speakers under three age groups. Predominantly, we studied more people in age group 18 to 36 years of age to understand the use of Tamil in Internet among youth. We raised some key questions related to the language use in social networking sites and concluded our findings with some key points.

APPROACH AND METHODOLOGY

The study was conducted by individually surveying one hundred internet using Tamil speakers. We used Google forms to collect the data. We studied the response of young Tamil youth who use the internet most. We collected data from three different age groups. Fig. 1 illustrates the three different age groups. Individuals with 18-27 years of age, individuals with 27-36 years of age and individuals between 36-65 years of age participated in the study. Out of the 100 participants 30% of survey takers belonged to 18-27 years of age; 40% of survey takers belonged to 27-36 years age group; and remaining 30% belonged to 36-65 years of age group. Clearly, students and working Tamil youth constituted 70% of the population studied for this research. We also recorded the profession of the Tamils who took the survey. 27 of those who took the survey were Information Technology IT professionals. It is very important to analyze the impact of language shift among the software professionals who use computer at ease. PhD students and researcher scholars constituted the second largest professional group among the survey-takers. The other professions mentioned include but not limited to : doctors, managers - employers in business firms, photographer, house-wife, retired. This denote that we have covered a diverse set of group and not limited to a particular profession. The Male : Female gender ratio among the survey takers was 80 : 20 which shows the large number participation from Males as shown in Fig. 2. We are happy that almost 90% of the survey takers had one degree and are graduates. We made a Fig. 3 to show the educational qualifications of some of the survey-takers.

RESULTS AND DISCUSSIONS

Fig. 4, shows the medium of instruction in school/college. The 66% respondents were educated in English. and 34% had Tamil as their medium of instruction in school and college. We could also understand the impact of having English as the medium of instruction in education sector. We also infer that among current generation of youth if we consider surveying 100 native Tamil speakers, 66% of them are educated in English not in Tamil. Especially it is very important to note that those who are educated in Tamil are from 36-65 age group. Fig. 5 shows the volume of English usage in work place. 13% people use very

little English and more Tamil in their work. 34% people use average amount of English and Tamil a the work place. Almost 21% responded that they use more English and less Tamil. 29% responded that they use only English as their language of communication in work place. It is good know that almost 50% of the respondents still use Tamil as their language of communication in work place. The myth that English is the only language for bringing better communication in work environment might change among Tamil speakers. Majority of the respondents, as shown in Fig 6, responded that Tamil is the language, they use to communicate effectively with their friends and parents. They also voted that Tamil is easy to speak than English for better communication as shown in Fig 7. But when it comes to social media language use, more than 70% said English is easy to use. Only 29% felt Tamil is easy to use in social media such as Face book / Whatsapp. On raising questions about the script form used by Tamil speakers in social media some interesting results came. Only 26% of respondents said they use English to type messages in social media. 32% uses Tamil letters to type and most importantly, majority of Tamil speakers (almost 42%) use Tamil in Roman letters (Eg : Vanakkam, Nandri).

44% responded that they cannot express sentiments like love, anger, sadness well in English. 33% responded that may be. and only 22% responded that they can express sentiments like love anger and sadness well in English. We infer that even though 70% internet users feel that English is easy to use in social media, only 26% actually use pure roman English script to type in social media and even among those who use English they cannot properly communicate their sentiments in social media if they use English! Thus, many of the internet users prefer to use Tamil.

69% of the respondents replied that they have used Google Tamil (Indic) keyboard. 31% of the internet users have not yet tried Google Tamil (Indic) keyboard.

Around 75% responded that using roman English letters is more convenient and easy to type. Only 25% use Tamil letters to type Tamil in social media.

We questioned if the Tamil speakers knew the meaning of following Tamil words. **நல்குரவு, இயைந்த, பசப்புறு, புலவி, ஒற்றாடல்**; only 15% responded that they know the meaning of all 5 words. Almost 18% responded that they do not know the meaning of all 5 words. 13% replied they know meaning of one word. 14% replied they know the meaning of two words. 25% replied that they know the meaning of three words. 15% replied they know the meaning of 4 words.

70% of Tamil speakers responded that spelling Tamil is easy. So the popular theory, "spelling Tamil is difficult, that is why not many use Tamil in social media" is not valid anymore. 17% responded that they can type in Tamil fast using Tamil letters. 43% responded that they can type Tamil using Tamil letters in normal (average) speed. 26% said they can type in Tamil but slow. 15% responded that they do not know how to type in Tamil (never used a Tamil keyboard) !

By using Roman letters (English alphabets) to type Tamil 40% said they can type very fast. and 43% said they can type at normal speed. 17% said even if they use roman English alphabets to type Tamil their typing speed is slow.

64% responded that they can type in English very fast and 34% can type at normal speed. less than 2% said their English typing speed is slow.

100% of survey takers replied that they know the meaning of following Thirukkural (which are taught almost in all schools).

அகர முதல எழுத்தெல்லாம் ஆதி
பகவன் முதற்றே உலகு.

75% of survey takers replied that they know the meaning of following Thirukkural (which is not so commonly taught in all schools).

பரியினும் ஆகவாம் பாலல்ல உய்த்துச்
சொரியினும் போகா தம

Remaining 25% replied that they have not heard of this Thirukkural before.

Every day new Internet-words are being invented and popularly used, such as, Troll, Selfie, Meme, Smiley, Emoticon. 67% of Tamil speakers who use internet responded that they do not know the equivalent Tamil words for above words. For some commonly used English words like Car, Geometry, Printer, 65% responded that they know the equivalent Tamil words and 35% responded that they do not know the equivalent Tamil words for all above three English words.

92% of internet Tamil users use English for searching a query in Google. Only 8% use Tamil to search in Google. When questioned if Tamil is used does the search engine provide accurate answers that they sought, 40% of respondent said they did not get accurate result.

Almost 88% of the survey takers responded that their friends in social media use Tamil to express their thoughts.

Finally, we know the psychological mind set of Tamil speaking internet users we raised following question. "If Tamils don't use their language in internet/day-to-day life, Tamil letters (script) may deform in 200 years. Your thoughts?" 58% responded that they already started using Tamil and they will encourage their friends to use Tamil. 37% responded that "I won't let Tamil Script to deform. I will start to use Tamil Scripts from Now" and last but not least, 5% responded "It's OK to let Tamil script deform".

CONCLUSION

We have investigated, 100 Tamil speaking internet users. We tried to quantify the shift in language use among internet users. Majority of the Tamil speakers use Tamil to communicate in social media, either in Tamil letters or in roman alphabets form. It is very important to build more applications that will make Tamil language use more convenient and easy for the Tamil speaking internet users.

ACKNOWLEDGMENT

The authors would like to acknowledge Taiwan Tamil Sangam headed by President Dr. Yu Hsi for his support and support of Mr. Orissa Balu (Ocean Researcher), Mrs. Thamarai Selvi (Center for Classical Tamil, Tharamani, Chennai) and Prof. Udayasuriyan (Tamil University, Thanjavur) for encouraging the authors to conduct this research study.

REFERENCES

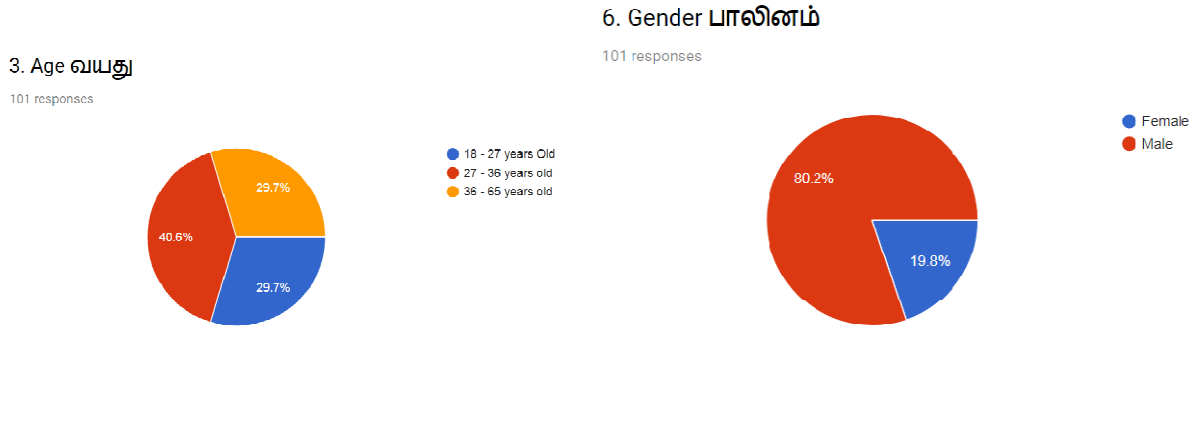
[1] Google KPMG analysis data, 2017.

[2] Thurairaj S, Hoon E P, Roy S S and Fong P K “Reflections of Students’ language Usage in Social Networking Sites: Making or Marring Academic English” The Electronic Journal of e-Learning Volume 13 Issue 4 2015. Appel, R., &Muysken, P. (2006). Language contact and bilingualism. Amsterdam: Amsterdam University Press.

[3] Auer, P. (2002). Code-switching in conversation: Language, interaction and identity. New York, NY: Routledge.
[4] Aydin, S. (2012). A review of research on Facebook as an educational environment. Education Tech Research Development.DOI 10.1007/s11423-012-9260-7.
[5] Cárdenas-Claros, M.S., & Isharyanti, N. (2009). Code-switching and code-mixing in Internet chatting: Between 'yes,' 'ya,' and 'si' - a case study. The JALT CALL Journal, 5. Retrieved from http://jaltcall.org/journal/articles/5_3_Cardenas_Abstract.html
[6] Craig, D. (2003). Instant messaging: The language of youth literacy. The Boothe Prize Essays 2003, 118-119.
[7] Cummings, A.B. (2011). An experience with language. ProQuest Dissertations & Theses: Literature & Language, 10.
[8] Dansieh, S.A. (2008). SMS testing and its potential impacts on students' written. International Journal of English Lingustics, 223(1), 222-229.
[9] Drouin, M.A. (2011). College students' text messaging, use of textese and literacy skills. Journal of Computer Assisted Learning, 27, 67-75.
[10] Eller, L.L. (2005). Instant message communication and its Impact upon written language. ProQuest Dissertations & Theses: Literature & Language, 13-16.
[11] Farina, F. & Liddy, F. (2011). The language of text messaging: "Linguistic ruin" or resources?. The Irish Psychologist. 37(6), 144-149.
[12] Felix, D. (2003). Important of study of English language in globalization era. Retrieved August 1, 2012, from http://english.ezinemark.com/important-of-study-of-english-language-in-globalization-era.-31f1ec50b22.html
[13] Grosseck, G. & Holotescu, C. (2008). Can we use Twitter for educational activities? Paper presented at the 4th International Scientific Conference, April 17-18, Bucharest
[14] Kabilan, M.K., Ahmad, N. & Zainol Abidin, M.J. (2010). Facebook: An online environment for learning of English in institutions of higher education? Internet and Higher Education, 13, 179-187.
[15] Md Yunus, M., Salehi, H. & Chen, C. (2012). Integrating social networking tools into ESL writing classroom - Strengths and weaknesses. English Language Teaching. Vol. 5, No. 8.
[16] Mphahlele M.L. & Mashamaite, K. (2005). The impact of short message service (sms) language on language proficiency of learners and the sms dictionaries: A challenge for educators and lexicographers. Proceedings of the IADIS International Conference Mobile Learning 2005, 161.
[17] Muñoz, C.L. & Towner, T.L. (2009). Opening facebook: How to use facebook in the college classroom. Paper presented at the Society for Information Technology and Teacher Education Conference 2009, Charleston, South Carolina.
[18] Muthusamy, P. (2009). Communicative functions and reasons for code switching: A Malaysian perspective. Retrieved on 5 August, 2011 from www.crisaps.org/newsletter/summer2009/Muthusamy.doc.
[19] Plester, B., Wood, C. & Bell, V. (2008). Txt msg n school literacy: Does texting and knowledge of text abbreviations adversely affect children's literacy attainment? Literacy. 42(3), 137-144.
[20] Roblyer, M.D., McDaniel, M., Webb, M., Herman, J. and Witty, J.V. (2010) "Findings on Facebook in higher education: A comparison of college faculty and student uses and perceptions of social networking sites". Internet and Higher Education.Vol. 13, pp. 134-140.

[21] Tagg, C. (2009). A corpus linguistics study of SMS text messaging. University of Birmingham Research Archive. Retrieved July 30, 2012, from http://etheses.bham.ac.uk/253/1/Tagg09PhD.pdf
[22] Thurairaj, S. , Mangalam, C. and Marimuthu, M., 2010. Reflection on Multiple Intelligences. International Journal of African Studies, 3, 16-25.
[23] Thurairaj, S. and Roy,S.S., 2012. Teachers' Emotions in ELT Material Design. International Journal of Social Science and Humanity, 2(3), 232-236. www.ejel.org 314 ISSN 1479-4403
[24] Saraswathy Thurairaj et al Thurairaj, S., Roy, S.S. & Subramaniam, K. (2012). Facebook and Twitter: A platform to engage in a positive learning. Proceedings of the International Conference on Application of Information and Communication Technology and Statistics in Economy and Education 2012, Bulgaria pp.157-167.
[25] Thurlow, C. (2002). "Generation TXT? The socio-linguistics of young people's text messaging" pp. 1-3. Retrieved July 18,2007, from http//extra.Shu.ac.uk/articles/vi/a3/thurlow 2002003-paper.htm

APPENDIX TABLES

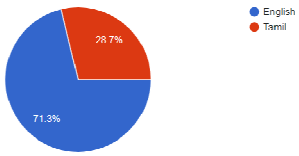


Doping Concentration (cm^{-3})	JL Mode	AC Mode	IM Mode
Source/Drain Doping	N : $1x10^{20}$ P : $1x10^{20}$	N :$1x10^{20}$ P :$1x10^{20}$	N :$1x10^{20}$ P :$1x10^{20}$
Channel Doping		N : $1x10^{18}$,N-Type P : $1x10^{18}$,P-Type	N : $1x10^{18}$,P-Type P : $1x10^{18}$,N-Type
Substrate Doping	N : $5x10^{18}$,P-Type P : $5x10^{18}$,N-Type		

Fig. 1 Device structure and important parameters of simulated 3-nm Gate Length (L_G) IM, AC & JL Germanium Bulk FinFET.

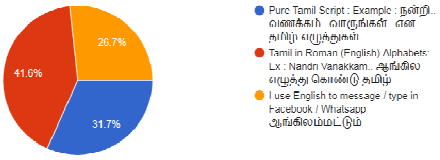
12. Which language you think is easy to use in Facebook / Whatsapp முகநூல் பகிரியில் எந்த மொழி பயன்படுத்துவது எளிமையானது ?

101 responses



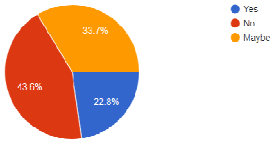
13. What language you use to type in Social Media (Twitter / Whatsapp / Facebook) ? நீங்கள் முகநூலில் பகிரியில் என்ன மொழி பயன்படுத்துகின்றீர் ?

101 responses



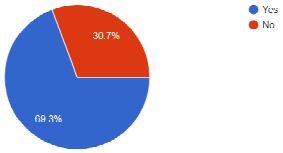
14. If you use English more in Facebook/Whatsapp do you think you can express sentiments like love, anger, sadness well in English? முகநூலில் பகிரும் பொழுது ஆங்கிலத்தை பயன் படுத்தி உங்கள் உணர்வுகளை முழுமையாக பிறர்க்கு உணர்த்த முடிகிறதா ?

101 responses



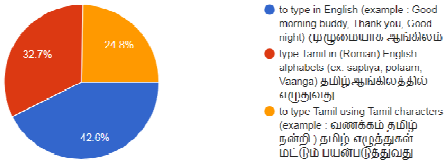
15. Have you ever used Google Tamil (Indic) keyboard in mobile or computer to type In Tamil? கூகுள் தமிழ் விசைப்பலகை பயன்படுத்தியதுண்டா?

101 responses



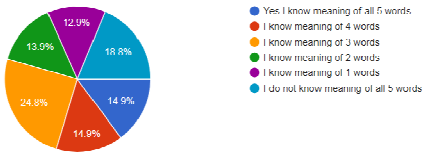
19. which is more convenient and easy for you எது உங்களுக்கு எளிமையாக இருக்கிறது

101 responses



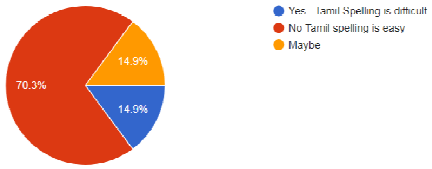
21. Do u know the meaning correct meaning of all the following 5 Tamil words? நல்குரவு, இயைந்த, பசப்புறு, புலவி, ஒற்றாடல் என்ற ஐந்து தமிழ் சொற்களுக்கு பொருள் தெரியுமா ?

101 responses



22. Do you think spelling (writing words in) Tamil is difficult தமிழில் பிழையில்லாமல்எழுதுவது கடினம் என்று உணர்கிறீர்களா ?

101 responses

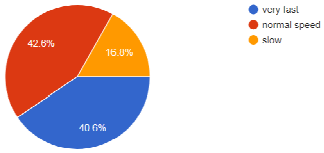


23. How fast you can type Tamil (using Tamil characters example :) தமிழில் எவ்வளவு வேகமாக தட்டச்சு செய்ய இயலும் ?

101 responses

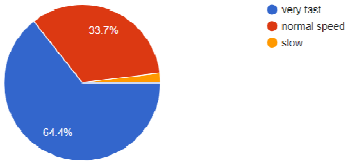
24. How fast you can type Tamil (using Roman (English) alphabets example : Nandri) தமிழை ஆங்கில எழுத்து கொண்டு எவ்வளவு வேகமாக தட்டச்சு செய்ய இயலும் ?

101 responses



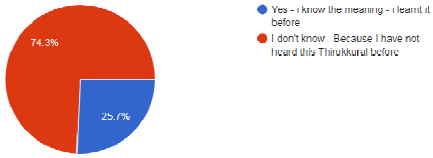
25. How fast you can type in English (Example : Thank you) ஆங்கிலத்தில் எவ்வளவு வேகமாக தட்டச்சு செய்ய இயலும் ?

101 responses



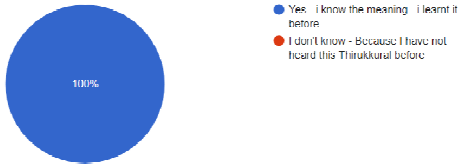
27. Do you know the meaning of this Thirukkural ? பரியினும் ஆகவாம் பாலல்ல உய்த்துச்சொரியினும் போகா தம. என்ற திருக்குறள் பொருள் அறிவீரா ?

101 responses



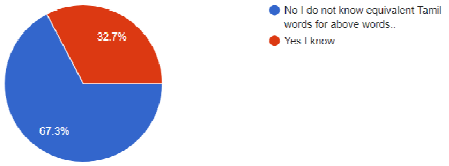
28. Do you know the meaning of this Thirukkural ? அகர முதல எழுத்தெல்லாம் ஆதிபகவன் முதற்றே உலகு. என்ற திருக்குறள் பொருள் அறிவீரா ?

101 responses



29. Do you know the Tamil words for Troll, Selfie, Meme, smiley, Emoticon? பின்வரும் வார்த்தைகளுக்கு நிகரான தமிழ் வார்த்தை தெரியுமா ? Troll, Selfie, Meme, smiley, Emoticon

101 responses

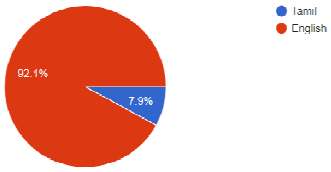


30. Do you know the Tamil words for Car, Geometry, Printer? பின்வரும் வார்த்தைகளுக்கு நிகரான தமிழ் வார்த்தை தெரியுமா ? Plastic, Car, Geometry, Printer?

101 responses

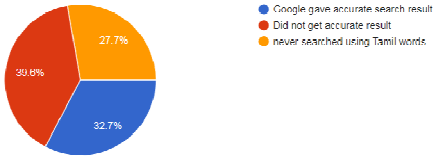
32. Which language you use in Google for searching? கூகுள் தேடலில் என்ன மொழி பயன்படுத்துவீர்கள் ?

101 responses



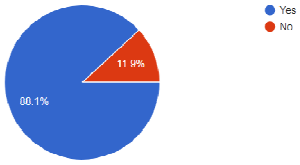
33. Have you used Tamil in Google search? If yes, how accurate was the search engine in returning the answers you sought? கூகுள் தேடலில் தமிழ் கொண்டு தேடியது உண்டா ? தேடலின் முடிவுகள் சரியாக பொருந்துகிறதா?

101 responses



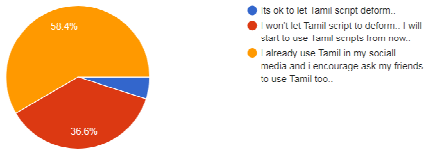
34. Do your Tamil friends use Tamil in social media to express their thoughts ? உங்கள் நண்பர்கள் சமூக ஊடகங்களில் தமிழை பயன்படுத்துகின்றனரா ?

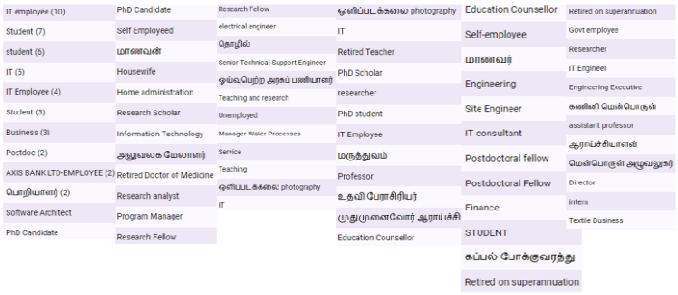
101 responses



37. If Tamils don't use their language in Internet/Day to day life, Tamil script may deform in coming 200 years.. தமிழை கணினியில் அலைபேசியில் அன்றாட வாழ்வில்பயன்படுத்தவில்லை என்றால் தமிழ் எழுத்து வடிவம்இன்னும் 200 ஆண்டுகளுள் சிதைவடைய வாய்ப்புண்டு. இந்த கருத்தினை எப்படி பார்க்கிறீர்கள்

101 responses





PhD (9)
BE (5)
PhD (Chemistry) (3)
Diploma (3)
MCA (3)
MF (?)
MCA-MASTER OF COMPUTER APPLICATION (2)
MSc (2)
M.Sc (2)
B.Tech IT (2)
Ph. D. (Chemistry), M. S. (Computer Science)
PhD (Energy materials)

PhD (Energy materials)
BE (Computers)
BE CSE, MA IR
Bachelors (Tamil literature)
BE with arrears
M.S.Ed
Masters in Comp. Science
MS (IT)
BE
Gastroenterologist (Post Graduate Medicin
Masters

MS Nanotechnology
இளநிலை பொறியியல்
PhD (chemistry)
BE MBA
M.Sc., MBA
B.E
B.E (Instrumentation),MBA(Project Management)
அறிவியலில் பட்டம் (அறி.தி)
BE civil
PhD Environmental Engineering
B.Sc computer science
Msc,M.Ed (mathematics)

BE Computer Science
PhD (Sociology)
M.Sc., M.Phil., Ph. D
முனைவர் PhD (Chemistry)
M.Com., M.Ed., M.Phil
MTech(Biotechnology)
MBA
BE computer science
Diploma, Bachelor Degree
BE(electrical and electronics)
BE (CS)
MSW

B.Sc computer science
Msc,M.Ed (mathematics)
I.ஆம் வகுப்பு
MS
BE CSE
M.A.B.Ed,
M. Sc (Physics)
phd
B.com, MBA
MS Software Engineering
MD
PhD (Industrial Engineering)

Mbbs
Bsc(chemistry)
B.Eng
Pursuing PhD
முனைவன்
B.Tech (Textiles)
BE(electrical)
BE(EEE)
B.Tech
MBBS
M.Tech.
Masters (ME)

M.E Thermal Engg
BE(Civil)
BE(Electronics & Instrumentation)
PhD(Chemistry)
PhD Biotechnology
PhD (Computer Science)
M.TECH(NANOTECHNOLOGY)
BE, MBA
B.Sc. (Physics)
ME(cse)
Mbbs
Bsc(chemistry)

இலங்கையில் அரசகரும்மொழிகள் நடைமுறையாக்கத்தின் ஒரு பகுதியான தமிழ்மொழி நடைமுறையாக்கத்தில் தகவற் தொழிநுட்பத்தின் வகிபாகம்

மு. மயூரன்
இலங்கை

இலங்கையின் அரசியல் யாப்பின்படி இந்நாட்டின் அரசுகரும மொழிகளில் ஒன்றாக தமிழ் ஏற்றுக்கொள்ளப் பட்டிருக்கிறது. நீண்டகால அரசியற் போராட்டங்களூடாகப் பெறப்பட்ட இவ்வுரிமை நடைமுறைப்படுத்துவதில் ஏராளமான சிக்கல்களும் போதாமைகளும் காணப்படுகின்றன.

அரசுகரும்மொழிகள் தொடர்பான அமைச்சர் முதல் பல்வேறு அரச துறைகளுடனும் தொடர்புற்ற அதிகாரிகள், ஊழியர்கள், பொதுமக்களிடமிருந்து நேர்காணல்கள் மூலமும் கேள்விக்கொத்துகள் மூலமும் பெறப்பட்ட தரவுகளின் அடிப்படையிலும் ஊடகங்களில் இது தொடர்பாக வெளிவந்த தகவல்களின் அடிப்படையிலும் பெறப்பட்ட அவதானங்களும் முடிவுகளும் இக்கட்டுரையில் தொகுப்படுகின்றன.

அரசுகரும்மொழிகள் தொடர்பான அமைச்சர் முதல் பல்வேறு அரச துறைகளுடனும் தொடர்புற்ற அதிகாரிகள், ஊழியர்கள், பொதுமக்களிடமிருந்து நேர்காணல்கள் மூலமும் கேள்விக்கொத்துகள் மூலமும் பெறப்பட்ட தரவுகளின் அடிப்படையிலும் ஊடகங்களில் இது தொடர்பாக வெளிவந்த தகவல்களின் அடிப்படையிலும் பெறப்பட்ட அவதானங்களும் முடிவுகளும் இக்கட்டுரையில் தொகுப்படுகின்றன.

அரசகரும மொழிகளில் ஒன்று என்ற அடிப்படையில் தமிழ் மொழியின் நடைமுறை-யாக்கத்தில் ஏற்படும் சிக்கல்களும் முட்டுக் கட்டைகளும் முதன்மையான காரணங்கள் தொழில் நுட்ப ரீதியானவை அல்ல, அரசியல் ரீதியானவையே. அந்த அடிப்படையில் தமிழ் மொழி நடைமுறையாக்கத்தில் உள்ள சிக்கல்களைக் களைவதில் அழுத்தமான வகிபாகத்தினை எடுப்பது அரசியல் ரீதியான மாற்றங்களாகவே இருக்க முடியும். இருப்பினும் தொழி நுட்ப ரீதியான தடங்கல்களும் தடைகளும் உள்ளன என்பதை இவ்வாய்வின் மூலம் கண்டரிய முடிந்த்து. தமிழ் மொழி நடைமுறையாக்கத்தில் தகவற் தொழி நுட்பத்துக்கென இரு வகிபாகம் ( ) இருப்பதை அடையாளம் காண முடிகிறது.

தமிழ் மொழி நடைமுறையாக்கத்தில் தகவற் தொழிற் நுட்பத்தின் வகிபாகம் இரண்டு கோணங்களில் ஆய்வு செய்யப்படுகிறது.

1. தற்போது தமிழ் மொழி நடைமுறையாக்கத்தில் தகவல் தொழில் நுட்பம் பங்கு பற்றும் இடங்களில் ஏற்படும் சிக்கல்களை இனம் காணுதலும் அச்சிக்கல்களைத் தீர்த்துகொள்ளும் வழிவகைகளும்.
2. எதிர்காலத்தில் தமிழ் மொழி நடைமுறையாக்கத்தினை குறை களைந்து மேலும் சீர்படுத்திக் கொள்ளத் தகவற் தொழி நுட்பத்தினைப் பயன்படுத்தக் கூடிய வழிவகைகளைக் கண்டறிதல்.

தற்போது பயன்படுத்தப்படும் தொழி நுட்பங்கள், ஏனைய துறைகளிலும் பயன்படுத்தப்படும் அடிப்படையான தகவற் தொழி நுட்ப வசதிகள் என்பதைத் தாண்டி சிறப்பான மொழிசார்ந்த தொழி நுட்பத் தீர்வுகள் எனுமளவில் பெரியளவில் விருத்தியுற்றவையாக இல்லை. ஊடகங்களில் வெளியாகும் தமிழ்மொழி நடைமுறையாக்கம் தொடர்பான முறைப்பாடுகளைப் பகுப்பாய்வு செய்யும்போது, அவற்றில் பல்வேறு முறைபாடுகள், அடிப்படையான தொழி நுட்பச் சிக்கல்களால் ஏற்படுவதைக் கண்டறிய முடிகிறது. ஒருங்குறி ஆதரவு, எழுத்துருச் சிக்கல்கள், மொழிபெயர்ப்பு மென்பொருள்களின் அடிப்படைகளாக அமைகின்றன. இவை தவிர, மொழிசார்ந்த கணிமையில் முறையான பயிற்சி வழங்கப்படாமை காரணமாக பொறுப்பான ஊழியர்கள் தமிழ் மொழி நடைமுறையாக்கத்தில் பல தவறுகளை இழைக்கின்றனர் என்பதையும் கண்டறிய முடிகிறது.

இப்போதாமைகளைத் தீர்க்குமுகமாக அரசு பல்வேறு நடவடிக்கைகளை இதற்கு முன்னர் எடுத்துள்ளது. குறிப்பாக, இலங்கைக்கான தமிழ்க் குறிமுறையாக யுனிகோடினையும்,, விசைப்பலகைத் தளக்கோலமாக ரெங்கனாதன் விசைப்பலகையும் ஸ் எல் ஸ் தர நிர்ணயமாக்கி, அரசு நிறுவன்ங்களில் அ ந் நியமங்களையே கட்டாயமாகப் பயன்படுத்த வேண்டும் என்றும் அறிவித்துள்ளது. விண்டோஸ், லினக்ஸ் இயங்குதளங்களுக்கான விசைப்பலகை இயக்கிகளில் ரெங்க நாதன் விசைப்பலகைத் தளக்கோலத்தை உள்ளடக்கியுள்ளது. தமிழ் மொழிக்கெனச் சில யுனிகோட் எழுத்துருக்களை உருவாக்கி இலவசமாக வெளியிட்டுள்ளது. இச்செயற் திட்டங்கள் இலங்கை தகவற் தொழி நுட்ப முகவர் நுறுவனத்தால் முன்னெடுக்கப்-பட்டன. அப்படியிருந்தும் இவற்றை நடைமுறைக்குக் கொண்டுவருவதிலும் உரிய பயிற்சிகளை வழங்குவதிலும் போதாமைகள் காணப்படுகின்றன.

எதிர்காலத்தை மனங்கொண்டு தமிழ் மொழி நடைமுறையாக்கத்தில் பயன்படுத்தப்-படக்கூடிய தற்போதய தகவற் தொழி நுட்ப முன்னேற்றங்களையும் கருவிகளையும் இவ்வாய்வு இனங்காண்கிறது. பயன்படுத்தப்படக்கூடிய மென்பொருட் கருவிகள்

சிலவற்றையும் அடிப்படை வசதிகளோடு உருவாக்கிக் காட்சிப்படுத்துகிறது. சில் பரிந்துரைகளையும் முன்வைக்கிறது.

அரசு நிறுவன்ங்களின் அதிகாரிகளுடைய நேர்காணல்களிலிருந்து மொழிபெயர்ப்பை வினைத்திறனுடனும் ஒருங்கிணைப்புடனும் செய்ய முடியாதிருப்பது தமிழ் மொழி நடைமுறையாக்கம் தற்போது எதிர்கொள்ளும் சவாலாக இனங்காணப்படுகிறது. கலைச்சொற் பயன்பாட்டில் ஏற்படும் முரண்பாடுகளைக் களைவது தொடக்கம் ஒரே பணியை மீளவும் பல முறை செய்யவேண்டி இருத்தல், மொழிபெயர்ப்பாளர்களிடையேயான தொடர்பாலும் ஒருங்கிணைப்பும் இல்லாதிருத்தல் என்பன வரை பல்வேறு சிக்கல்கள் மொழிபெயர்ப்பு விடயத்தில் காணப்படுகின்றன. இச்சிக்கல்களைக் களைந்து மொழிபெயர்ப்புப் பணிகளை ஒருங்கிணைப்பதற்கும் மொழிபெயர்ப்பாளர்களிடை-யேயான வலைமைப் பொன்றினை உருவாக்குவதற்குமான மென்பொருட் கருவி ஒன்று இக்கட்டுரைக்கென உருவாக்கப் பட்டுள்ளது. இணைய வழியாக இயங்கும் இம்மொன்பொருள் விக்சனரி முதலான கலைச்சொல் அகரமுதலிகளை உள்ளடக்கியிருப்பதுடன் ஏற்கனவே மொழிபெயர்க்கப்பட்டவற்றை மீள மொழிபெயர்க்கும் பணிச்சுமையினை இல்லாத்தாக்குவதுடன் மொழி-பெயப்ப்பாளர்களிடையேயான ஒருங்கிணைப்பையும் தொடர்பாடலையும் ஏற்படுத்துகிறது. இக்கருவி மூலம் வளர்த்தெடுக்கப்படும் மும்மொழி மொழிபெய்ர்ப்புகள் அடங்கிய தரவுத்தளத்தைக் கொண்டு செயற்கை நுண்ணறிவு கொண்ட மென்பொருள்களையும் திறன்மிக்க மொழிபெயர்ப்புக் கருவிகளையும் எதிர்காலத்தில் உருவாக்க முடியும்.

புகைவண்டி நிலையங்கள் முதல் அரசு நிறுவன்ங்கள் வரை குரல்வழி அறிவுப்புக்களைத் தமிழ் மொழியில் வழங்குவதில் குறைபாடுகள் காணப் படுகின்றன. வரையறுக்கப்பட்ட சொற்களஞ்சியத்தையும் வாக்கிய அமைப்புக்களையும் கொண்ட இவ் அறிவுப்புகள் எளிதாகவே மென்பொருள்களின் துணைகொண்டு தமிழ் ஏனைய மொழிகளிலும் வழங்கப்பட வேண்டும். இதற்குத் தமிழில் தற்போது விருத்தியடைந்துள்ள குரற் தொகுப்பி இக்கட்டுரைக்கென உருவாக்கப் பட்டுள்ளது.

அரசுகரும மொழிகள் திணைக்களம் மும்மொழி நடைமுறையாக்கலுக்கான அமைச்சு தமிழ் மொழி நடைமுறையாக்கத்தில் உள்ள முறைப்பாடுகளைத் தெரிவிக்கவெனத் தனியான தொலைபேசி எண் ஒன்றினை வழங்கியுள்ளன. இது மேலும் விருத்தி செய்யப்பட்டு, திறன் பேசிகளிலிருந்தும் இணையம் மூலமாகவும் படங்கள்-

ஒலிக்கோப்புகள்-வீடியோ மூலம் முறைப்பாடுகளைப் பதிவு செய்து அவற்றை பின் தொடர்வதற்கான வசதிகள் வழங்கப்பட வேண்டும்.

அரசு நிறுவனங்களின் அறிவுப்புப் பலகைகள், பெயர்ப்பலகைகள் போன்ற-வற்றைச் சரிபார்த்து அனுமதிப்பதற்கான கட்டமைப்பினை உருவாகுவது தொடர்பாக மும்மொழிகள் நடைமுறையாக்கத்துக்கான அமைச்சினர் நேர்-காணல்களின்போது தெரிவித்தனர். அவ்வாறான ஒரு கட்டமைப்பு உருவாகுமிடத்து, தமிழ் சொற் திருத்தி, இலக்கணத் திருத்தி போன்றவற்றைப் பயன்படுத்தி இணைய வழியாகவே முதற்கட்டச் சரிபார்ப்பினைச் செய்து கொள்ள முடியும். இவை தவிர, இலங்கையின அரசுகரும் மொழிகளைக் கற்பிப்பதற்கு தகவற் தொழி நுட்பத்தின் துணையைப் பெற்றுக் கொள்ளமுடியும்.

நீண்டகால அடிப்படையில் வளர்ந்து வரும் தொழி நுட்பச் செல் நெறிகளை ஒட்டி, தமிழ் –சிங்கள மொழிபெயர்ப்புக் கருவிகள், குரல் உணரிகள், எழுத்துணரிகள் மூலமான மொழிபெயர்ப்பு போன்றவற்றில் அரசுகரும மொழிகளை அமைச்சு முதலிடுவதையும் இவ் வாய்வு பரிந்துரைக்கிறது. ஏற்க்கனவே இத்துறையில் இடம் பெற்றுவரும் திறந்த மூல முயற்சிகளுடனும் அமைப்புகளுடனும் குழுக்களுடனும் வலைமைப் பொன்றினை உருவாக்குவது அவசியமானது.

பன்மொழி நடைமுறையாக்கத்தினைச் செய்துவரும் சிங்கப்பூர், கனடா போன்ற ஏனைய நாடுகளில் இத்துறையில் தகவற் தொழி நுட்பத்தின் வகிபாகம் பற்றிய அனுபவங்களையும் பெற்றுக் கொள்ளமுடியும்.

2016ம் ஆண்டு தமிழ்நாடு சட்டப்பேரவை தேர்தலும், தமிழக இளைஞர்களின் அரசியல் சார்ந்த சமூக இணையதளப் பயன்பாடும்

Sairam Jayaraman[1], Muruganandam Sundararajan[2]

[1] Madras Christian College, East Tambaram, Chennai – 600 059, Tamil Nadu, India E-Mail: gkj.sairam@gmail.com

[2] Chief Reporter, Hello Asia News Inc., 222-216 Crown Road NW, Edmonton-AB, T6J2E3, Canada E-Mail: muru@helloasianews.com

---

ஆய்வுச்சுருக்கம்

நாட்டின் பொருளாதாரத்திலும், கல்வித்துறையிலும் முன்னணி மாநிலமாக இருந்துவரும் தமிழ்நாட்டில், 2016ம்ஆண்டு நடந்த சட்டப்பேரவை தேர்தலின் போதுவாக்காளர்களை கவர்வதற்காக பல்வேறு புதியதொ ழில்நுட்பங்கள் அரசியல் கட்சிகளால் பயன்படுத்தப்பட்டன. குறிப்பாக சமூக இணைய தளங்களில் அரசியல் கட்சிகளால் மேற்கொள்ளப்பட்ட பிரச்சாரமானது மிகப்பெரிய தாக்கத்தை ஏற்படுத்தியது என்று பரவலாக நம்பப்படுகிறது. எனவே, இந்த ஆய்வானது 2016ம் தமிழ்நாடு சட்டப் பேரவை தேர்தலின்போது சமூக இணைய தளங்களை இளைஞர்கள் எவ்வாறு, எதற்காக பயன்படுத்தினார்கள், அதன் தாக்கம் உண்மையிலேயே வாக்குகளின் மூலம் பிரதிபலித்துள்ளதா என்ற கேள்விகளுக்கு பதிலளிக்கும் வகையில் நடத்தப்பட்டுள்ளது.

குறிப்புவார்த்தைகள்: Social Media, TN Election, Twitter, Facebook

முன்னுரை

சமீபகாலமாக நடந்துவரும் தொழில்நுட்ப புரட்சியானது பாரம்பரிய ஊடகங்களான செய்தித்தாள்கள், இதழ்கள் மற்றும் தொலைக்காட்சி என பலவற்றிற்கு சவால் அளித்து வருகிறது. அதிலும் குறிப்பாக சமூக இணையதளங்களின் வரவு பல்வேறு தரப்பினருக்கிடையே நிலவிவந்த பயன்பாட்டு மற்றும் தகவல் தொடர்பு சார்ந்த பிரச்சனைகளை வெகுவாக குறைத்துள்ளது எனலாம். சமூக இணைய தளங்களானது தகவல்கள் மற்றும் கருத்துக்களை உருவாக்கும், தெரிவிக்கும், பகிரும் மற்றும் ஒருவரது கொள்கை சார்ந்த பார்வையை உலகம் முழுவதும் ஏற்கெனவே தெரிந்த மற்றும் தெரியாத நபர்களுடன் விவாதிக்கும் புதிய ஊடகமாக உருமாறி வருகிறது.

தமிழகத்தின் முக்கிய அரசியல் கட்சிகள் புதிய தொழில்நுட்பங்கள், குறிப்பாக சமூக இணைய தளங்களின் அவசியத்தை உணர்ந்து, இந்த தேர்தலின்போது பேஸ்புக் மற்றும் ட்விட்டர் உள்ளிட்ட பல்வேறு சமூக இணைய தளங்களில் கணக்கைத்

தொடங்கி பதிவுகளை இடத் தொடங்கின. ஏற்கனவே, சில கட்சிகள் சமூக இணையதளங்களில் இருந்தாலும், தேர்தல் அறிவிப்பு வந்தவுடன்தான் அவை முழுவீச்சில் செயல்பட ஆரம்பித்தன என்பது குறிப்பிடத் தக்கது. தமிழக அரசியல் வரலாற்றில் முதல்முறையாக செய்தி அறிக்கைகள் பாரம்பரிய ஊடகங்களுக்கு அனுப்பப்படுவதற்கு முன்னதாக கட்சிகளின் அதிகாரப்பூர்வ சமூக இணைய தள பக்கங்களில் வெளியிடப்பட்டதோடு, கட்சிகளின் தேர்தல் அறிக்கைகள் மற்றும் வாக்குறுதிகள் ஆகியவை பேஸ்புக், ட்விட்டர் மற்றும் யூடியூப் உள்ளிட்ட சமூக இணையதளங்களிலும் மற்றும் கட்சிகளின் அதிகாரப்பூர்வ இணைய தளங்களிலும், இலட்சக்கணக்கான ரூபாய் செலவழித்து விளம்பரமும் செய்யப்பட்டது. "internetlivestats.com" என்னும் இணையதளம் அளிக்கும் தரவுகளின்படி இந்தியாவின் மொத்த மக்கள்தொகையில் 34.8% பேர், அதாவது கிட்டத்தட்ட 46 கோடிபேர் இணைய தளத்தை பயன்படுத்துகின்றனர். அதிலும் குறிப்பாக நகர்ப்புறங்களில் சமூக இணைய தளங்களை பயன்படுத்து-பவர்களின் எண்ணிக்கை 118 மில்லியனாகவும், ஊரகப் பகுதிகளில் 25 மில்லியனாகவும் உள்ளது. இணையதள பயன்பாட்டில் உலகிலேயே இந்தியா இரண்டாமிடம் வகிக்கும் நிலையில், மத்திய அரசின் சமீபத்திய அறிக்கையொன்றின் படி, இந்திய அளவில் தமிழகம் 2.68 கோடி பயன்பாட்டாளர்களுடன் இரண்டாமிடத்தை பெற்றுள்ளது. சென்ற ஆண்டு நடந்த சட்டப்பேரவை தேர்தலில் மொத்தம் 5.79 கோடிபேர் வாக்களிக்கும் தகுதியை பெற்றிருந்தனர். அதில் ஒரு கோடிக்கும் மேற்பட்டோர் முதல்முறையாக வாக்களிக்கும் இளைஞர்கள் என்பது குறிப்பிடத் தக்கது.

மேற்கோள்கட்டுரைகள்(Review of Literature)

இவ்வாய்வுக் கட்டுரையின் தலைப்போடு ஒத்த கருத்துடைய சில ஆய்வுக்கட்டுரைகள் குறித்த குறிப்புகள் கீழே அளிக்கப்பட்டுள்ளது. மேலும், இக்கட்டுரைகளானது, தேர்தலும் – சமூக இணைய தளங்களின் தாக்கமும் என்பதைமையக் கருத்தாககொண்டு, அமெரிக்க அதிபர் தேர்தலின் போதும் மற்றும் இந்திய நாடாளுமன்ற தேர்தலின்போதும் மேற்கொள்ளப்பட்டவைகளாகும். ஆனால், இவ்வாய்வானது முதல்முறையாக தமிழக சட்டமன்ற தேர்தலின்போது இளைஞர்களின் சமூக இணையதளப் பயன்பாட்டையும் அதன் தாக்கத்தையும் கண்டறிவதற்காகவும் மேற்கொள்ளப்பட்டுள்ளது.

காயத்ரி வாணி மற்றும் நிலேஷ் அலோன் (2014) ஆகியோர் “A Survey on Impact of Social Media on Election System” என்ற தலைப்பில் செய்த ஆய்வில், கடந்த 2014ம் பாராளுமன்றத் தேர்தலின்போது சமூக இணைதளங்களில் அரசியல் கட்சிகளின் பக்கங்களால் வாக்காளர்களிடையே ஏற்படும் தாக்கம் மற்றும் அதன் காரணமாக

தேர்தல் முடிவுகளில் ஏற்பட்டுள்ள மாற்றங்கள் குறித்து ஆய்வு மேற்கொள்ளப் பட்டது.

"Use of New Media in Election Campaigning (Lok Sabha Elections 2014)" என்ற தலைப்பில் சுபாகத்த பட்டாச்சார்யா (2014) என்பவர் செய்த ஆய்வின் மூலம், "மொத்தமுள்ள 815 மில்லியன் வாக்காளர்களில் வெறும் 12 சதவீத பேர் மட்டும்தான் புதிய தொழில்நுட்பங்களை பயன்படுத்துபவர்களாக உள்ளனர் என்றும், எனவே தற்போதைக்கு சமூக இணைய தளங்கள் அதிகளவில் தாக்கத்தை ஏற்படுத்தவில்லை என்றாலும் வருங்காலத்தில் அதன்தாக்கம் மிகப்பெரிய அளவில் அதிகரிக்கும் என்று குறிப்பிடப் பட்டுள்ளது.

மீடியா அஜீர் (2014) என்பவர் "The Effects of Internet Usage on Voter Choice in the 2012 United States Presidential Elections" என்ற தலைப்பில் செய்த ஆய்வின்படி, "அமெரிக்க குடியரசு தலைவர் தேர்தலில் குறிப்பிட்ட கட்சியினுடைய வேட்பாளரின் இணையதளத்தை பார்வையிட்டவர் பெரும்பாலும் அவருக்கே வாக்களித்ததாகவும், கடந்த நான்கு ஆண்டுகளாக சமூக இணைய தளங்களை பயன்படுத்துபவர்கள் சனநாயக கட்சிக்கு வாக்களித்ததாகவும்" கண்டறியப் பட்டுள்ளது.

கருத்தியல் செயற்திட்டம் (Theoretical Framework)

இந்த ஆய்வானது "Uses and Gratification Theory" என்னும் கோட்பாட்டை அடிப்படையாக கொண்டு மேற்கொள்ளப் பட்டுள்ளது. இந்த கோட்பாட்டை நிறுவியவரான எலிஹூகாட்ஸ் என்னும் அமெரிக்க சமூகவியல் அறிஞரின் கூற்றுப்படி, "மக்கள் சமூக மற்றும் அறிவு சார்ந்த தேவைக்காக ஊடகத்தை பயன்படுத்துகின்றனர். அது தனிப்பட்ட நபரைபொறுத்து, பொழுபோக்கு சார்ந்த தேவைக்காகவோஅல்லது தகவல் சார்ந்த தேவைக்காகவோ, தங்களுக்கு தேவையான குறிப்பிட்ட ஊடகவகையை தேர்ந்தெடுத்து தங்களின் தேவையை நிறைவேற்றிக் கொள்கின்றனர்". மேலும், குறிப்பாக இந்த கோட்பாடானது மக்கள் எதற்காக ஊடகத்தை பயன்படுகின்றனர் என்ற கேள்விக்கும், அதன் மூலம் கிடைக்கும் தகவல்கள் ஒருவரின் முடிவெடுக்கும் செயல்முறையில் எவ்வகையான பங்கை வகிக்கின்றது என்ற கேள்விக்கும் பதிலளிக்கிறது.

எனவே, இந்த கோட்பாட்டை அடிப்படையாக கொண்டு 2016ம் ஆண்டு நடந்த தமிழக சட்டப்பேரவை தேர்தலின்போது சமூக இணையத் தளங்களில் மேற்கொள்ளப்பட்ட பிரச்சாரமும், இளைஞர்களின் பயன்பாடும் எந்தளவிற்கு தாக்கத்தை ஏற்படுத்தி உள்ளதுஎன்பதை கண்டறிவதில் இந்த ஆய்வு கவனம் செலுத்துகிறது.

ஆய்வுமுறை (Methodology)

இந்த ஆய்விற்கு தேவையான தரவுகளை பெறுவதற்காக, கேள்வித்தாள் ஒன்று தயார் செய்யப்பட்டு சென்னையில் வசிக்கும் 18 வயதிற்கு மேற்பட்ட 500 நபர்களிடம் பதில்கள் பெறப்பட்டன. ஆண், பெண் என இருபாலினரும் பங்கேற்ற இந்த ஆய்வில் கேட்கப்பட்ட கேள்விகளனைத்தும், அக்குறிப்பிட்ட நபர்களின் சமூக இணையதளப் பயன்பாடு மற்றும் சமூக இணைய தளங்களின் மூலம் அரசியல் கட்சிகளால் மேற்கொள்ளப்பட்ட பிரச்சாரங்கள் குறித்த அவர்களின் கருத்தை அறியும் வகையில் அமைக்கப் பட்டிருந்தது.

கேள்வித்தாள் மூலம் பெறப்பட்ட தரவுகளை பகுப்பாய்வு செய்வதற்காக SPSS என்னும் மென்பொருளும், Microsoft Excel 2016 என்ற மென்பொருளும் பயன்படுத்தப் பட்டன. கீழ்க்காணும் மூன்று குறிக்கோள்களைமையப் படுத்தி இந்த ஆய்வு மேற்கொள்ளப்பட்டது.

1) 2016ம் ஆண்டு தமிழக சட்டப்பேரவை தேர்தலின் போது சமூக இணையதளங்களின் பரவல்.
2) அரசியல் சார்ந்த தகவல்களை சமூக இணையத்தளங்கள் மூலம் பெற்றதில் மக்களுக்கு கிடைத்த விழிப்புணர்வும், அது வாக்களிப்பதில் ஆற்றிய பங்கும்.
3) சமூக இணையதளங்கள் மூலம் மேற்கொள்ளப்படும் தேர்தல் பிரச்சாரம் குறித்த மக்களின் மனநிலையை தெரிந்துக்கொள்ள.

ஆய்வுமுடிவுகள் (Findings)

பங்கேற்பாளர்கள்:

இந்தஆய்வில் பங்கேற்ற 500 பேரில் 279 பேர்ஆண்கள், 221 பேர்பெண்கள். மேலும், 500 பங்கேற்பாளர்களில் 240 பேர் 18-21 வயதுப்பிரிவையும், 170 பேர் 21-25 வயதுப் பிரிவையும் மற்றும் மீதமுள்ள 90 பேர் 25 மற்றும் அதற்கு மேலுள்ள வயதுப் பிரிவையும் சேர்ந்தவர்களாவர். மேலும், ஆய்வில் பங்கேற்ற 500 பேர்களில் அதிகபட்சமாக 310 பேர் சமூக இணையதளங்களை பயன்படுத்த விருப்பமான மின்னணு கருவி திறன்பேசிகள் (ஸ்மார்ட்போன்) என்றும், 105 பேர்மேசை/ மடிக்கணினி என்றும், மீதமுள்ள 85 பேரின் விருப்பமாக கையடக்கக் கணினியும் உள்ளது.

பங்கேற்பாளர்கள் சமூக இணைய தளங்களில் செலவிடும் நேரம்:

அட்டவணை (1) கேள்விஎண் 1ன்படி ஆய்வில் பங்கேற்ற 500 பேர்களில் அதிகபட்சமாக 350 பேர்சமூக இணையதளங்களை மூன்று வருடங்களுக்கு மேலாகவும், 110 பேர் 2-3 வருடங்களாகவும், மீதமுள்ள 40 பேர் கடந்த ஒரு வருடமாக இணைய தளங்களை பயன்படுத்தி வருவதாக தெரிவித்துள்ளனர்.

அட்டவணை 1: பங்கேற்பாளர்களின்சமூகஇணையதளப்பயன்பாடு

தெரிவு	**எண்ணிக்கை**
1. எத்தனை வருடங்களாக சமூக இணைய தளங்களை பயன்படுத்திவருகிறீர்கள்? (உதா. பேஸ்புக், ட்விட்டர்)	
சென்ற ஒரு வருடமாக	40
2-3 வருடங்களாக	110
3 வருடங்களுக்கு மேலாக	350
2. செய்திகளை அறிந்துகொள்வதற்காக சராசரியாக ஒரு நாளில் எத்தனை மணி நேரத்தை சமூக இணையத் தளங்களில் செலவிடுகிறீர்கள்?	
ஒருமணி நேரத்திற்கும் குறைவாக	85
1-3 மணிநேரங்கள்	190
3-5 மணிநேரங்கள்	160
5 மணிநேரத்திற்கும் மேலாக	65

செய்திகளை அறிந்து கொள்வதற்காக சமூக இணைய தளங்கள்:

அட்டவணை (1) கேள்விஎண் 2ன்படி ஆய்வில் பங்கேற்ற 500 பேர்களில் அதிகபட்சமாக 190 பேர்ஒரு நாளைக்கு 1-3 மணி நேரத்தை செய்திகளை அறிந்து கொள்வதற்காக சமூக இணையதளங்களை பயன்படுத்துவதாகவும், 3-5 மணி நேரம் என்று 160 பேரும், ஒருமணி நேரத்திற்கு குறைவாக என்று 85 பேரும் மற்றும் மீதமுள்ள 65 பேர்ஒரு நாளைக்கு 5 மணி நேரத்திற்கும் மேலாக செய்திகளை அறிந்து கொள்வதற்காக சமூக இணைய தளங்களை நாடுவதாக தெரிவித்துள்ளனர்.

அரசியல் குறித்த செய்திக்கு விருப்பமான ஊடகவகை:

அட்டவணை (2) கேள்விஎண் 1ன்படி ஆய்வில் பங்கேற்ற 500 பேர்களில் அதிகபட்சமாக 165 பேர் சமமான எண்ணிக்கையில், அதாவது அரசியல் குறித்த செய்திகளை அறிந்து கொள்வதற்காக சமூக இணையதளங்களை நாடுவதாக 165

பேரும், செய்தித்தாள்களைஅணுகுவதாக 165 பேரும், தொலைக்காட்சிகள்என்று 135 பேரும், இதழ்கள்என்று 30 பேரும், மீதமுள்ள 5 பேர் நண்பர்கள் மூலமாக தெரிந்துகொள்வதாகவும் பதிலளித்துள்ளனர். மேலும், இதன்மூலம் குடும்பத்தினரிடமிருந்து ஒருவர் கூட அரசியல் சார்ந்த செய்திகளை அறிந்துகொள்ள பயன்படுத்துவதில்லை என்பது தெரியவருகிறது.

அட்டவணை 2: அரசியல் சார்ந்த செய்திகளும், சமூக இணைய தளங்களும்

தெரிவு	**எண்ணிக்கை**
1. அரசியல் குறித்த செய்திகளை அறிந்துகொள்ள எவ்வகையான ஊடகத்தை பயன்படுத்துகிறீர்கள்?	
சமூக இணைய தளங்கள்	165
செய்தித்தாள்கள்	165
இதழ்கள்	30
தொலைக்காட்சி	135
நண்பர்கள்	5
குடும்பம்	0
2. அரசியல் மற்றும் தேர்தல்கள் குறித்து விவாதிக்கும் தளமாக சமூக இணையத் தளங்களை நீங்கள் ஏற்றுக் கொள்கிறீர்களா?	
நிச்சயமாக	235
சமூக இணைய தளங்கள் தற்போதுதான் வளரத் துவங்கியுள்ளன	210
இன்னும் இல்லை	55

மாற்று ஊடகமா சமூக இணைய தளங்கள்?:

அட்டவணை (2) கேள்விஎண் 2ன்படி ஆய்வில் பங்கேற்ற 500 பேர்களிடம் அரசியல் மற்றும் தேர்தல்கள் குறித்து விவாதிக்கும் மாற்று ஊடகமாக சமூக

இணையதளங்களை எடுத்துக் கொள்ளலாமா என்று கேட்கப்பட்டது. அதற்கு அதிக பட்சமாக 235 பேர் நிச்சயமாக என்றும், 210 பேர் சமூக இணையதளங்கள் தற்போதுதான் வளரத் துவங்கியுள்ளன என்றும், மீதமுள்ள 55 இன்னும் இல்லையென்றும் பதிலளித்துள்ளனர்.

அட்டவணை 3: சமூக இணைய தளங்களும், தேர்தல் பிரச்சாரமும்

தெரிவு	**எண்ணிக்கை**
1. அரசியல் தலைவர்கள் மற்றும் கட்சியினரிடையே அதிகரித்துவரும் சமூக இணையதள பயன்பாடு உங்களுக்கு தற்போதைய அரசியல் சூழ்நிலையை புரிந்துக்கொள்ள பயன்படுகிறதா?	
ஆம், இது பயன்படுகிறது	240
ஏதோ ஒரு வகையில் பயன்படுகிறது	210
இல்லை, குழப்பத்தைத்தான் ஏற்படுத்துகிறது	50
2. அரசியல் கட்சிகள் சமூக இணையதளங்களில் விளம்பர பிரச்சாரம் செய்வது குறித்து உங்கள் கருத்தென்ன?	
பயன்படுகிறது	100
வெறுப்பேற்றுகிறது	320
நேரவிரயம்	80
3. தற்போதைய அரசியல் சூழ்நிலை குறித்து ஏதாவது சமூக இணையதளங்கள் மூலம் அறிந்துள்ளீர்களா?	
ஆம், மிகவும் பயன்படுகிறது	170
ஆம், ஆனால் பெரும்பாலும் வேடிக்கையானவைகளே	290
எதையுமே கற்கவில்லை	40

அரசியல் கட்சிகளும், சமூக இணையதள பயன்பாடும்:

அட்டவணை (3) கேள்விஎண் 1ன்படி ஆய்வில் பங்கேற்ற 500 பேர்களில் பெரும்பாலானோர், அதாவது 240 பேர் அரசியல் கட்சிகள் மற்றும் அதன் தலைவர்களின் அதிகரித்துவரும் சமூக இணையதள பயன்பாடு தற்போதைய அரசியல் போக்கை அறிந்து கொள்ள பயன்படுகிறதென்றும், 210 பேர் அவை ஏதோ ஒருவகையில் பயன்படுவதாகவும், மீதமுள்ள 50 பேர் அத்தகைய பதிவுகள் குழப்பத்தைத்தான் விளைவிப்பதாகவும் கருத்து தெரிவித்துள்ளனர்.

அரசியல் கட்சிகளும், சமூக இணையதள விளம்பர பிரச்சாரமும்:

அட்டவணை (3) கேள்விஎண் 2ன்படி ஆய்வில் பங்கேற்ற 500 பேர்களிடம் அரசியல் கட்சியினர் தேர்தல் குறித்து சமூக இணைய தளங்களில் மேற்கொள்ளும் விளம்பரங்கள் குறித்து கேட்டபோது, பெரும்பான்மையானவர்கள் அதாவது 320 பேர், அவை தங்களுக்கு வெறுப்பேற்றும் வகையிலுள்ளதாகவும், 100 பேர்அவை பயன்படுவதாகவும், மீதமுள்ள 80 பேர் அவை நேரத்தை விரயம் செய்வதாகவும் கூறியுள்ளனர்.

சமூக இணையதளங்களும், பயன்பாடும்:

அட்டவணை (3) கேள்விஎண் 3ன்படி ஆய்வில் பங்கேற்ற 500 பேர்களில் 290 பேர் சமூக இணைய தளங்களில் தாங்கள் காணும் விடயங்கள் பெரும்பாலும் வேடிக்கையானவைகளாகவே இருந்ததாகவும், 170 பேர் தற்போதைய அரசியல் சூழ்நிலையை அறிந்து கொள்ள பயன்பட்டதாகவும், மீதமுள்ள 40 பேர் தாங்கள் எதையுமே கற்கவில்லை யென்றும் தெரிவித்துள்ளனர்.

சமூக இணையதள விவாதங்களில் பங்கேற்பு:

அட்டவணை (4) கேள்விஎண் 1ன்படி ஆய்வில் பங்கேற்ற 500 பேர்களில் பெரும்பான்மையானோர்அதாவது 345 பேர் சமூக இணையதளங்களில் நடைபெற்ற விவாதங்களில் தங்கள் பங்கேற்றதில்லை என்றும், மீதமிருக்கும் 155 பேர் தாங்கள் விவாதங்களில் பங்குக்கொண்டதாகவும் தெரித்துள்ளனர்.

வாக்களிக்க உதவிய காரணி:

அட்டவணை (4) கேள்வி எண் 2ன்படி ஆய்வில் பங்கேற்ற 500 பேரில் 180 பேர்தங்களின் நண்பர்கள் மற்றும் சமூக இணையதளமே தேர்தலில் ஒரு குறிப்பிட்ட கட்சிக்கு வாக்களிக்க உதவியதாகவும், 100 பேர் குடும்பம் மற்றும் சமூக இணையதளம் என இரண்டுமே உதவியதாகவும், குடும்பத்தினரின் கருத்தே

உதவியதாக 80 பேரும், குடும்பம் மற்றும் நண்பர்கள்என்று 70 பேரும், சமூக இணைய தளங்கள்என்று 50 பேரும் மற்றும் மீதமிருக்கும் 20 பேர்நண்பர்கள் என்றும் கருத்துத் தெரிவித்திருந்தனர்.

அட்டவணை 4: சமூக இணையதளங்களில் தேர்தல் பிரச்சாரமும், முடிவெடுக்கும் திறனும்

தெரிவு	**எண்ணிக்கை**
1. சமீபத்திய தேர்தல் குறித்த சமூக இணையதளங்களில் நடைபெற்ற விவாதங்களில் பங்கேற்றீர்களா?	
ஆம்	155
இல்லை	345
2. கீழே உள்ள எந்த காரணி உங்களை ஒரு குறிப்பிட்ட ஒரு கட்சிக்கு வாக்களிக்க உதவியது?	
சமூக இணைய தளம்	50
குடும்பம்	80
நண்பர்கள்	20
குடும்பம் மற்றும் நண்பர்கள்	70
நண்பர்கள் மற்றும் சமூகஇணையதங்கள்	180
குடும்பம் மற்றும் சமூக இணையதளங்கள்	100

முடிவுரை:

மேற்கண்ட ஆய்வுமுடிவுகளை பகுப்பாய்வு செய்யும்போது பேஸ்புக் மற்றும் ட்விட்டர் போன்ற சமூக இணையதளங்கள் அரசியல் மற்றும் தேர்தல் உள்ளிட்ட பல்வேறு வகையான செய்திகள் மற்றும் தகவல்களை அறிந்து கொள்ள உதவும் முக்கிய ஊடகமாக உருவெடுத்துள்ளதாக தெரிய வருகிறது. எனவே, அரசியல் கட்சிகள் பாரம்பரிய தேர்தல் பிரச்சார வழிமுறைகளை மட்டுமே கையாண்டு வாக்காளர்களை கவர இயலாது என்றும், ஒரு அரசாங்கத்தை அமைப்பதில் முக்கிய

பங்கை வகிக்கும் நிலையை நோக்கி சமூக இணையதளங்கள் முன்னேறி வருவதையும் இதன் மூலம் அறிவியலுகிறது. மேலும், இந்த ஆய்வானது அரசியல் கட்சியினர் சமூக இணையதளங்களை பயன்படுத்துவதை மக்கள் வரவேற்றாலும், அதன் மூலம் செய்யப்படும் விளம்பர பிரச்சார அணுகுமுறையை அவர்கள் விரும்பவில்லை என்பது தெரிய வருகிறது. வருங்காலங்களில் மக்களிடையே, குறிப்பாக இளைஞர்களிடையே சமூக இணைய தளங்களின் பயன்பாடு மேலும் அதிகரித்து அவர்களின் முடிவெடுக்கும் திறனில் முக்கிய பங்கையும், சமூக பிரச்சனைகளை விவாதிக்கும் தளமாகவும் சமூக இணைய தளங்கள் மாறும் என்பதில் ஐயம் ஏதுமில்லை.

குறிப்புகள்:

1. Gayatri Wani , Nilesh Alone (2014) 'A Survey on Impact of Social Media on Election System', *(IJCSIT) International Journal of Computer Science and Information Technologies,* Vol. 5 (6), pp. 73637366, http://www.ijcsit.com/docs/Volume%205/vol5issue06/ijcsit20140506100.pdf
2. Media Ajir (2014) 'The Effects of Internet Usage on Voter Choice in the 2012 United States Presidential Elections', *Creighton University*, pp. 1-13 [Online]. Available at: *https://www.creighton.edu/fileadmin/user/CCAS/departments/PoliticalScience/Journal_of_Political_Research__JPR_/2014_JSP_papers/Media_A.pdf*
3. Subhagata Bhattacharya (2014) *Use of New Media in Election Campaigning (Lok Sabha Elections 2014)*, Delhi University: Available at : https://www.academia.edu/7486078/Use_of_New_Media_in_Election_Campaigning_Lok_Sabha_Elections_2014_

எழில் - பொது பயன்பாட்டிற்கும், வெளியீடு நோக்கிய சவால்களும்¨

கருணாகரன் கணேசன், கிரேச்குமார் ராமராசு,

அருண்ராம் ஆத்மசரன், மற்றும் முத்து அண்ணாமலை.

எழில் மொழி அறக்கட்டளை, சான் ஓசே, காலிஃபோர்னியா.

சுருக்கம்

எழில் ஒரு தமிழ் கணினி மொழி [1-3]- இந்த தமிழ் கணினி மொழி என்னும் பட்டியலே ஆரோக்கியமாக Clj-Thamil போன்ற திட்டங்களினால் [4], வளரும் ஒரு நோக்கில் உள்ளது. பல ஆண்டுகள் தொடர் முயற்சியாலும், அனுபவத்தினாலும் தற்போது எழில் மொழி ஒரு பொது வெளியீட்டை அடையும் நிலையில் உள்ளது. இந்த தருணத்தில் எழில் கடந்த பாதையையும், சந்தித்த வெற்றி தோல்விகளையும், சவால்களையும் ஒரு வரலாற்று பார்வைக்கும், பொறியாளர் திறனாய்வுக்கும் ஆவணப்படுத்தும் நோக்கில் இந்த கட்டுரை அமையும்.

எழில் மொழி இடைமுகம்

பல வழிகளாக எழில் மொழியை இணையம் வழி வழங்க முற்பட்டோம். இதில் முதல் முயற்சி [1] சுயமாக மடிக்கணினியில் ஒரு web server நிறுவி; அடுத்து ezhillang.org என்ற தளத்தில் இணையம் வழி எழில் மொழியை வழங்கி, மேலும் இந்த [2-3] இணையவழி இடைமுகத்தை syntax highlighting மூலமும் மேம்படுத்தி கொடுத்தோம். ஆனால் கிடைத்த தாக்கமோ மிக குறைவு; எங்களுக்கும் server மேற்பார்வைக்கு வேலை அதிகம். முக்கியமாக எழில் மொழி சென்றடைய வேண்டிய சிற்றூர் மாணவர்களுக்கு கணினியும் கிடையாது, இணையமும் கிடையாது. இதனை உணர்ந்த பின் எழில் மொழியை கணினியில் மட்டும் தேவைப்படும் அளவுக்கு இடைமுகத்தை கொடுக்கும் படி வடிவமைக்க முற்பட்டோம். இதுவே "**எழுதி**" (படம்:3).

மென்பொருள் வெளியீடு

ஒரு மென்பொருள் வடிவமைப்பை தாண்டி, இயக்கும் நிலையில் இருக்கும் என்றாலும் அதனை பொது வெளியீடு என்று வரும் பொழுது ஒரு பட்டியலிட்டு அதனில் உள்ள குணாதிசயங்களையும், சிறப்பம்சங்களையும் சரிவர வேலைசெய்கிறதா என்றும் பரிசோதித்து வெளியிடுவது மிக முக்கியமான கணினி பொறியியல் வளர்ச்சி நிலை. இந்த நிலையை எட்ட எழில் [படம் : 1,2] மொழியில் பல படிகள் இந்த ஆண்டு

எடுக்கப்பட்டது. முக்கியமாக Windows, Linux இயங்கு தளங்களில் எழில் திரட்டி - “எழுதி” என்ற இடைமுகம் [படம்: 3, 5] - தயார் நிலையில் பயன்டுத்தும் வகை சேய்த்தோம். மேலும் Android தளத்தில் ஒரு “எழில் கையேடு” [படம் 4] என்ற குறுஞ்செயலி (mobile app) ஒன்றையும் படைத்துள்ளோம். இந்த வேலையில் நாங்கள் சந்தித்த சவால்களையும் பற்றி இந்த கட்டுரையில் பதிவு செய்கிறோம். எழில் மொழியை வெளியீடு செய்ய பங்களித்த அனைவருக்கும் எங்களது நன்றிகள் - எங்களது முன்னேற்றத்தை பட்டியல் 1-இல் பார்க்கலாம்:

Project	Stars	Language	closed issues	open issues	pull requests	Fork	Contributors
Ezhil-Lang	78	Python	100	93	59	35	11

பட்டியல் 1:Github-இல் எழில் மொழி பங்களிப்பு தளத்தின் புள்ளிவிவர பட்டியலில்

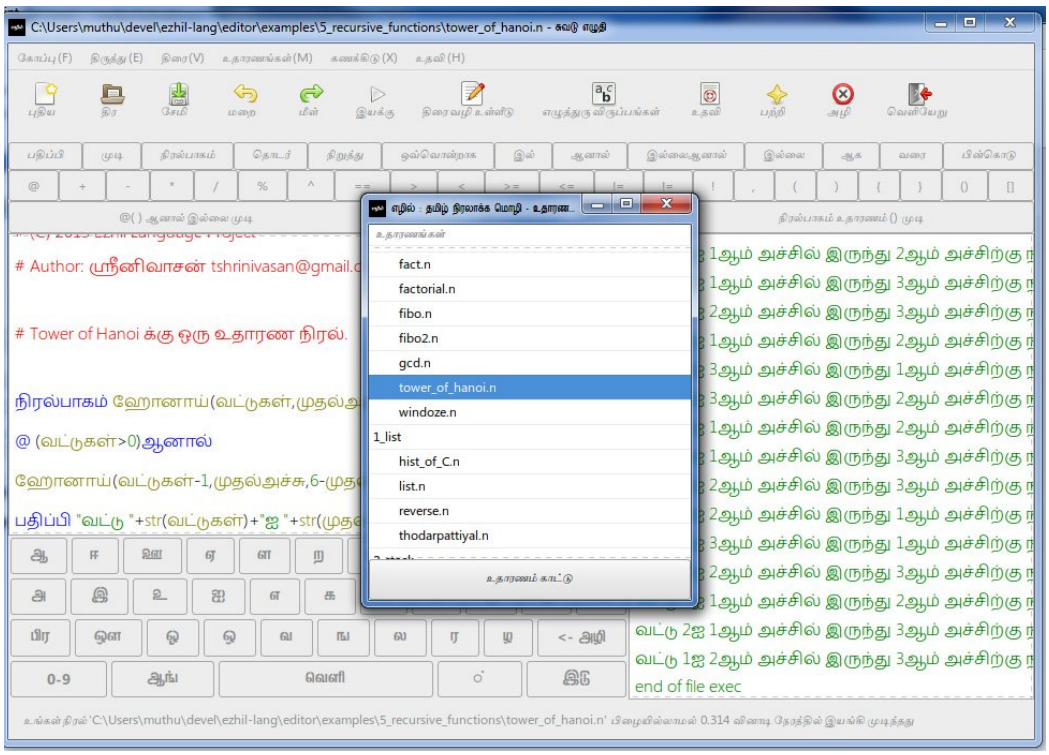
எழில் மொழி பல ஆண்டுகளாக ஆமை வேகத்தில் இதன் வளர்ச்சி தொடர்ந்து வந்திருக்கிறது; இருந்தாலும் வளர்ச்சியை நிரல் வரிகளில் எண்ணமுடியாது - மொழியின் தாக்கத்திலும், கடைசி பயனாளர்களிடத்திலும் மட்டுமே காண முடியும்.

படம் 1: எழுதி செயலி தொடக்க படம்.

படம் 2: எழில் மொழி சின்னம்.

எழில் மொழியை இப்படி சிறுவர் சிறுமியர் பொது பயன்பாட்டிற்கும், அவர்களுக்கு கற்றுவிக்கும் ஆசிரியர்களுக்கு தேவைப்படும் எழில் மொழி விளக்கம், கணிமை பாடங்கள், விளக்க உரை, போன்றவற்றை தயார் செய்து கல்வி நுகர்வோர்க்கு இணங்க செய்வதே முக்கியமான வேலை. எழில் மொழியையே நிரல்படுத்தி, பரிசோதிக்கும் வேலையை தாண்டி, இந்த எழில் மொழி கல்வி பள்ளிக்கூடங்களில்

இலகுவாக எழிலை புகட்ட தேவையான அம்சங்கள் அனைத்தையும் மென்பொருள் வெளியீட்டு ரீதியில் அளிக்க இங்கு முயல்கிறோம்.



படம் 3: எழுதி - எழில் மொழி செயலி உத்தாரணங்களுடன் பயனர் இடைமுகம்

எழில் கற்க பயிற்சி பாடங்கள்

தமிழில் நிரல் எழுது புத்தகம் [5.1], யூடியூப் காணொளிகள் [5.2] போன்றவற்றை தயாரித்து வெளியிட்டும் எழில் மொழியை பயில வழிவகுத்துள்ளோம். காணொளியில் கிட்டதிட்ட 35min நிமிடங்கள் பயிற்சி [5.2]-இல் இலவசமாக உள்ளது. இதனை எழுதி என்கிற மென்பொருளில் இயக்கி காணொளியை கண்டபடியே படிக்கலாம். இதையும் தொடர்ந்து வழங்கி வருவது எங்கள் குறிக்கோள். படம் 3. -இல் காட்டியபடி எழுதி செயலியில் கிட்டத்தட்ட 175 இக்கும் கூடுதலாக நிரல் உதாரணங்கள் உள்ளன. படம் 5. இல் எழில் மொழியில் LOGO போன்ற படங்கள் வரைவதை எடுத்துக்காட்டாக செய்துள்ளோம்.

இணைய தளத்தில் சிக்கல்கள்

எழில் மொழியில் இணைய தளம் http://ezhillang.org மற்றும் http://urbantamil.com என்ற இரு தமிழ் மொழி சார்ந்த தளங்கள் AWS Amazon மேக கணினியில் இயங்கி வந்தன. ஆனால் எழில் மொழி “கூடம்” இணையம் வழி எழில் [3] என்றபடி உங்கள் உலவியில் இருந்தே எழில் மொழியை கற்று கொள்ளலாம் என்றபடி உள்ள சலுகை வழியாக தீயவர்கள் எழில் மொழி இணையத்தை அக்டொபேர் 2016 தொடங்கி

தரகர்த்தனர். இதன்பின் மூன்று மாதங்களுக்கு இந்த இணையதளங்கள் இரண்டும் செயலற்று கிடந்தன. இன்றும், இதன்பின், கூடம் என்ற இணையம் வழி எழில் என்ற சேவையை நிறுத்தினோம்; urbantamil.com என்ற இணையதளம் முழுதாக தாக்கப்பட்டு முடங்கியது.

இணையத்தில் மின்னுவதெல்லாம் பொன்னல்ல என்றும் பாதுகாப்பாக செயலிகளை நிறுவவேண்டும், செயலிகளை பாதுகாக்கவேண்டும் என்பது ஒரு கடினமான பாடத்தை கற்றோம்.

திறன்பேசியில் எழில்

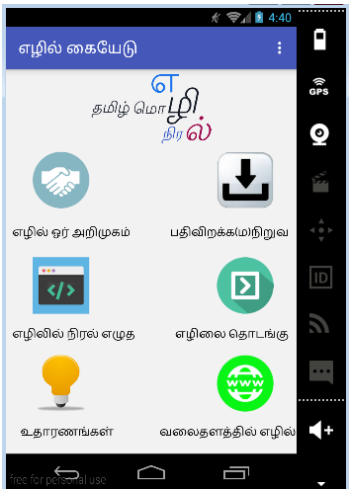
தற்போதய காலத்தில் எங்கும் கைபேசிகள் திறன்பேசியால் மாற்றம் அடைவதனால் இந்த ஒரு இயங்கு தளத்திலும் மாணவர்களை சென்றடைய எழில் மொழி கையேடு என்ற ஒரு செயலியை உருவாக்கினோம். இது இணையதளத்தின் தாக்குதலுக்கு முன்பே இணையத்தில் server-களை மேம்படுத்துவதும் சவாலாக உள்ளதை கண்டு நாங்கள் திட்டமிட்ட ஒன்று; இணைய தளம் செயலுற்றபின் இந்த முயற்சியில் மிகவும் கவனம் செலுத்துகிறோம்.

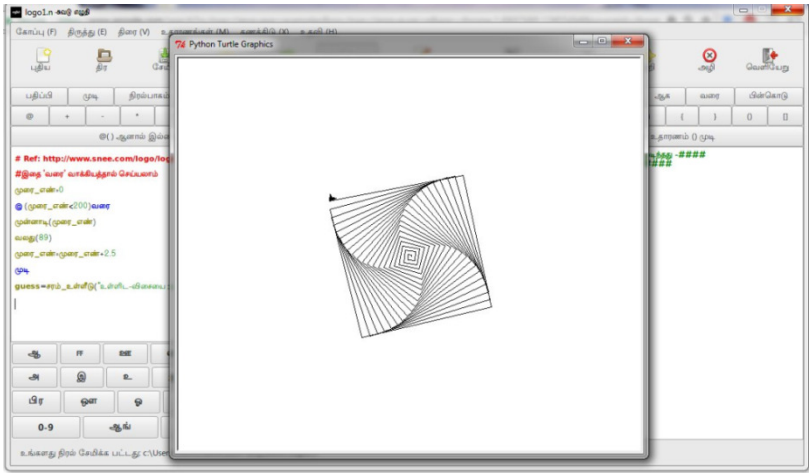
தற்போது “எழில் கையேடு” என்ற செயலியை நீங்கள் பரிசோதனை முறையில் Google Play Store-இல் இருந்து பெறலாம் [6]. இது December, 2016 அன்று வெளியிடப்பட்டு 40-50 நபர்களால் தங்கள் கைபேசியிலேயே எழில் மொழியை பயின்று, இணையம் வழி ezhillang.org “கூடம்” வழி நிரல்களை இயக்கியும் எழில் மொழியை பழகலாம். தற்போது “கூடம்” இயக்கத்தில் சவால்கள் உள்ளதால் இந்த செயலியை முழு வீச்சில் வெளியிடவில்லை. இதற்கும் தீர்வு காண்போம் என்பதில் எங்களுக்கு நம்பிக்கை உண்டு. எழில் கையேடு செயலியை முதல் படியாக கருணாகரன் கணேசன் உருவாக்கி, எங்களுடன் இதனை வெளியீட்டு நிலைக்கு கொண்டுவந்தார். இவர் வருங்கால பொறியாளர், பழகும் மாணவர்.

எழுதி இடைமுகம் பரிசோதனை, தரப்படுத்தல்

வெளியீடு செய்யும் முன் ஒரு மென்பொருளில் சில வழிமுறைகளை பின்பற்றுவது பொறியியல் நடைமுறை.

1. தனி பரிசோதனை (unit tests)
2. கூட்டு பரிசோதனை (integration tests)
3. இயங்குதளம் நிறுவல் பரிசோதனை (operation system installation tests)
4. பயனர் இடைமுகம் பரிசோதனை (user interface testing)





படம் 4 (இடது): எழில் மொழி "எழில் கையேடு" ஆண்ட்ராய்டு கைபேசி செயலி.

படம் 5 (வலது): எழுதி செயலியில் சதுரத்தில் இருந்து சுழற்சி வரைவது நிரல்.

இவை எல்லாவற்றையும் செய்தால் ஒரு நல்ல மென்பொருளை தரமாக உருவாக்கலாம் என்பது கணினியியலில் நடைமுறை புரிதல். இவற்றை எழில் மொழியில் படி 1, 2, போன்றவற்றை நேரடியாகவே எழில் source code இல் continuous integration வழி [7] செய்துள்ளோம்; அடுத்த படி 3, 4 பரிசோதனைகளை கைவழி செய்தோம்.

எங்களுக்கு தெரிந்தவரை "எழுதி" மென்பொருளை பரிசோதனை செய்தும், பின்னூட்டங்களை அணுகியும் ஒரு தரமான மென்பொருளை உருவாக்கியுள்ளோம் என்பது நம்பிக்கை. மென்பொருளில் பிழை/வழு இல்லாத மென்பொருள் என்பதே இல்லை. இதில் உள்ள பொறியியல் சிக்கல் (complexity) கையாள்வதற்கு ஒரு மனிதர் தேவை – இதனை முழுதுமே செயற்கை நுண்ணறிவால் (A.I.) தானியங்கி படுத்தமுடியுமா என்பது காலத்தால் மட்டுமே சொல்லக்கூடிய கேள்வி.

அடுத்த கட்டம் - முடிவுரை

எழில் எங்கே போகிறது என்பதை பற்றி யோசித்துள்ளோம் [8]; இதன்படி ஒரு தரமான எழுதியை உருவாக்கியுள்ளோம். இதனை சரிவர விளம்பரம் செய்து, மாணவர்களுக்கு பயிற்சி அளிப்பது அடுத்த கட்டம். எழில் மொழி இப்போது பயிற்சிக்கு தயார், ஆனால் மாணவர்களும், ஆசிரியர்களும், பள்ளிகளும் தயாரா ? அவர்கள் தேவைகளை எழுதி மட்டும் எழில் அணி சந்திக்குமா என்பது களத்தில் அறிந்துகொள்ள வேண்டிய உண்மை. அதன்பின் ஒருநாள் எழில் மொழியும் தமிழகமெங்கும் தரப்படுத்தப்படும் நாள் கூட வரும் என்பதும் எங்கள் குறிக்கோள்.

மேற்கோள்கள்

1. M. Annamalai, “Ezhil (எழில) : A Tamil Programming Language,” ArXiV/0907.4960 (2008).
2. M. Annamalai, et-al “An Introduction to Ezhil - Programming the Computer in Tamil,”Tamil Internet Conference(INFITT), Kuala Lampur, Malaysia (2013).
3. M. Annamalai, et-al “Learning Ezhil Language via Web,” Tamil Internet Conference(INFITT), Puducherry, India (2014).
4. Elango Cheran, “Exploring Programming Clojure in Other Human Languages,” Clojure/West, Portland, Oregon (2015)
5. (1) Write Code in Tamil - Ezhil Programming Language, paperback (2013), available https://www.amazon.com/Write-Code-Tamil-Programming-Language/dp/1547233915

   (2) எழுதி - எழில் கணினி மொழி (காணொளி பட்டியல்) https://www.youtube.com/playlist?list=PLo6YZ6ilTE53mRVQelij9lXg6xiOonhXM
6. Ezhil Language Handbook (எழில் கையேடு), Google Play Store at https://play.google.com/store/apps/details?id=com.ezhil.handbook
7. Travis C.I. Ezhil Language Continuous Integration and Unit Testing https://travis-ci.org/Ezhil-Language-Foundation/Ezhil-Lang
8. M. Annamalai, “எழில் எங்கே போகிறது”, http://ezhillang.wordpress.com/2017/03/02/ வலை பதிவு.

தமிழின் பெருந்தரவகத் தரவுகள் தேவையும், பயன்பாடும்

செல்வமுரளி

விசுவல்மீடியா டெக்னாலஜிஸ், Email : murali@visualmediatech.com

நாள்தோறும் வளர்ந்துவரும் தமிழ்க் கணிமையின் வளர்ச்சி தற்போது உரையிலிருந்து ஒலிக்கு மாறியிருப்பது குறிப்பிடத்தக்கது என்றாலும் நாள்தோறும் வளர்ந்து ஒவ்வொருத்துறையிலும் தமிழை நிலைநிறுத்துவது தமிழ் சமூகத்திற்கு தேவையான ஒன்று.

தமிழில் ஆதியிலிருந்து இதுவரை கிடைத்த தகவல்களை தொகுத்து நம்மிடையே நம் பயன்பாட்டுக்கு வைத்திருக்கிறோமா என்றால் இல்லை. தமிழை ஒரு கருவியாக பயன்படுத்துவதை விட தமிழை ஒரு சேவையாக கொண்டுவருவதே அடுத்த தலைமுறைக்கு நாம் செய்யும் சிறந்த சேவை என்றும் கூறலாம். அதாவது தமிழை வர்த்தகமொழியாக மாற்றுவது. தமிழ் மொழிக்கான வர்த்தக சேவை ஆண்டு பல கோடி ரூபாய்க்கு மேல் இருக்கிறது. ஆனால் ஆங்கில மொழிக்கான வர்த்தக சேவை பல லட்சம் கோடிகளில் உள்ளது.எல்லா கருவியிலும் ஆங்கிலம் உள்ளது. ஏனெனில் ஆங்கிலம் ஒரு வர்த்தக மொழி.

நம் தமிழை வர்த்தக மொழியாக மாற்றத் தேவையான பல தேவைகளில் தமிழுக்கான பெருந்தரவகம் (BigData Database) ஒன்று. பெருந்தரவரக தரவுத்தளத்தினை உருவாக்கிவிட்டால் தரவு அறிவியல் (டேட்டா சைன்ஸ்) கொண்டு எல்லா துறைகளிலும் சமூக பயன்பாட்டுக்குத்தேவையான பயன்பாடுகளை எளிதாக உருவாக்கிட முடியும். இதன் மூலம் தமிழையும், தமிழோடு சமூகத்தினையும் உயர்நிலைக்கு கொண்டு செல்ல முடியும். இங்கே விவசாயம், மருத்துவம் ஆகிய இரண்டு துறைகளை மட்டும் நான் என் ஆய்வுக்கு எடுத்துக்கொண்டிருக்கிறேன்

விவசாயம்

சங்கக்காலத்தில் வகைப்படுத்திய ஐந்திணை நிலங்களில் உள்ள மக்கள் தம் நிலத்தில் என்னென்ன பயிர்களை விளைத்திருக்கிறார்கள் என்பதை நாம் பாடப்புத்தகங்கள் வழியே தெரிந்துகொண்டாலும் விவசாயம் சார்ந்த பருவநிலைகள், பருவநிலை மாற்றத்தால் ஏற்பட்ட சீரழிவுகள் , பேரழிவு போன்றவற்றையும் கண்டறிந்து தரவு அறிவியல் வழியே எதிர்காலத்தில் ஏதேனும் சிக்கல் நேர்ந்தால் சமாளிக்க ஏதுவாக அதன் வழிமுறைகளை உருவாக்கலாம்.

ஐந்திணைகளில் கொடுக்கப்பட்டுள்ள பருவ நிலை

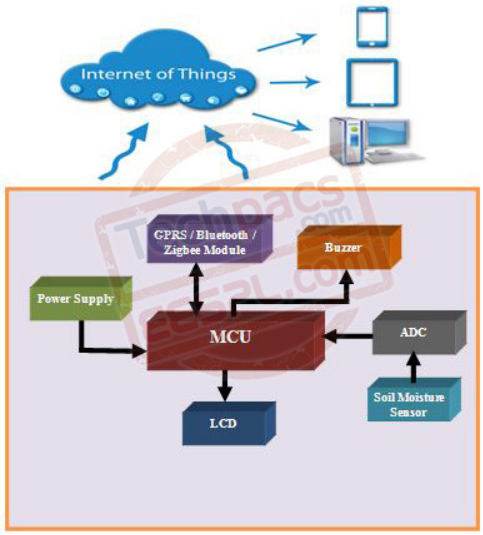
இளவேனில், முதுவேனில் , கார், கூதிர் (குளிர்), முன்பனி, பின்பனி என்ற பருவநிலைகளில் பிரித்து வைத்தாலும் இளவேனில் என்பதற்கான வெப்பநிலை, ஈரப்பதம், மழைப்பொழிவு, காற்றின் வேகம், காற்றின் திசை போன்றவற்றை

வளரக்கூடிய பயிர்கள், இச்சூழ்நிலைக்கு பொருந்தா பயிர்கள் என எல்லாவற்றையும் தொகுக்கவேண்டியும் அவசியம்.

இவைகள் ஒருபுறம் இருக்க பொருட்களின் இணையம் (IoT) மூலம் பல ஏக்கர் அளவிலான விவசாயத்தினை வெகுஎளிதாக செய்யலாம் என்ற அளவிற்கு தொழில்நுட்பங்கள் வந்துவிட்டாலும் இன்னமும் நாம் முறையாக விவசாயத்திற்கான தரவு அறிவியலை உருவாக்கவில்லை என்பதே உண்மை.

Soil Sensor System - மண் உணர்வி

மண் உணர்வியை விவசாயம் செய்யும் மண்ணுக்குள் புதைத்து வைத்துக்கொண்டால் கம்பியில்லா இணைப்பின் மூலம் மண்ணில் ஈரப்பதம் இருக்கிறதா? இல்லையா என்பதை நம் செல்பேசியின் வழியாக மேலாண்மை செய்யலாம். மண்ணின் ஈரப்பதம் குறைவாக இருந்தாலும் நமக்கு தகவல் தெரிவிக்கப்படும். இவற்றை தானாக செயல்படுத்த வேண்டுமெனில் எல்லா பயிர்களுக்கும் தேவையான ஈரப்பதத்தின் அளவு பற்றிய தகவல் தரவுத்தளமாக நம்மிடையே இருக்கவேண்டும். உதாரணத்திற்கு நெல்லிற்கான ஈரப்பதம் அதிகமாக இருக்கவேண்டும்.

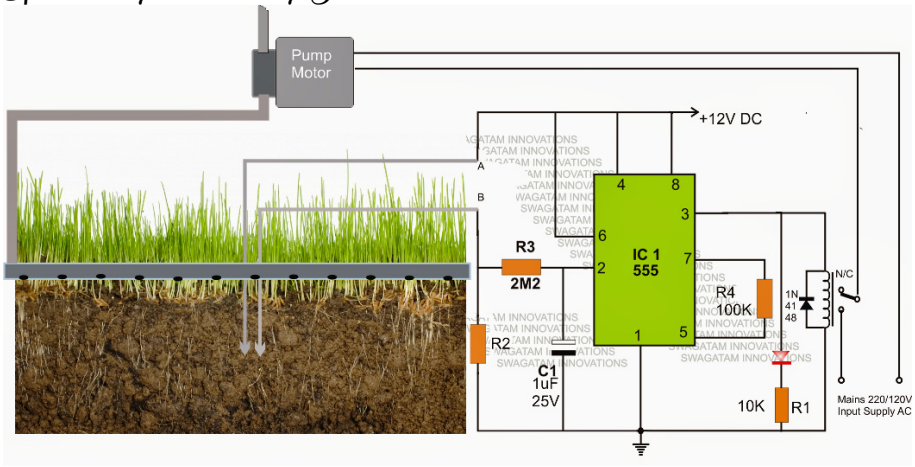


ஆனால் மற்ற பயிர்களுக்கான ஈரப்பதம் குறைவாக இருக்கவேண்டும். எனவே அனைத்து வகை பயர்களின் ஈரப்பத விபரங்களை உருவாக்கவேண்டியது அவசியமாகிறது. இத்தரவுத்தளத்தினை உருவாக்கிவிட்டால் **humidifier** என்ற ஈரப்பதமூட்டியைக் கொண்டு ஈரப்பதத்தினை தேவையான அளவு கொண்டுவந்துவிட முடியும்.

தானியங்கு நீர் மேலாண்மை அமைப்பு

விவசாய நிலங்களுக்குத் தேவையான அளவு தண்ணீரை நாமே வழங்கிட பொருட்களின் இணையம் தொழில்நுட்பம் பயன்படுகிறது. இந்த தொழில்நுட்பங்களை மேலாண்மை செய்யவேண்டுமெனில் பயிர்களுக்குத் தேவையான ஈரப்பதம் ,

தண்ணீரின் அளவு போன்வற்றிக்கான தகவல்களை திரட்டி தரவகமாக மாற்றுவது அவசிய தேவையாகிறது.



DIY an Automatic Plant Watering Device

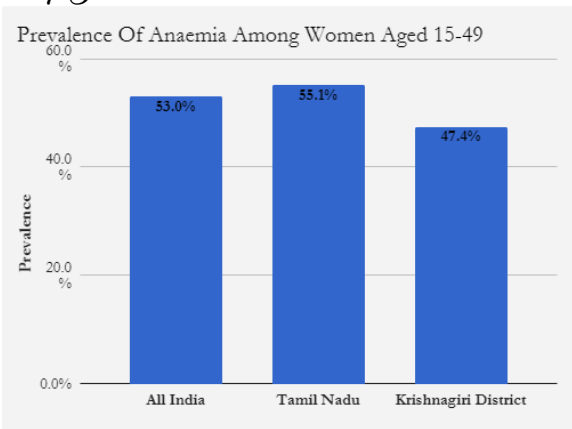
நம் வீட்டுத்தோட்டங்களுக்கு நாமில்லாத சமயங்களிலும் தண்ணீர் விட இதோ இந்தக் கருவி பயன்படுகிறது. DIY an Automatic Plant Watering Device, இதற்குத்தேவை மேலே சொன்ன மண் ஈரப்பத உணர்வியும், ஒரு சிறிய கட்டுப்பாட்டகமும் வை-பை இணைப்பும் இருந்தாலே போதும்.

பூச்சிகள் மேலாண்மை

அசைவு உணர்விகள் மூலம் பயிர்களைச் சுற்றியுள்ள பூச்சிகளின் எண்ணிக்கையை தெரிந்துகொள்ளலாம், அதோடு எந்த வெப்பநிலையில் அவைகள் பயிர்களுக்கு வருகின்றன. எந்த வெப்பநிலையில் அவவைகள் பயிரை விட்டு விலகுகின்றன போன்ற எல்லா விபரங்களும் இதனோடு பொருந்தும்.

மருத்துவம்

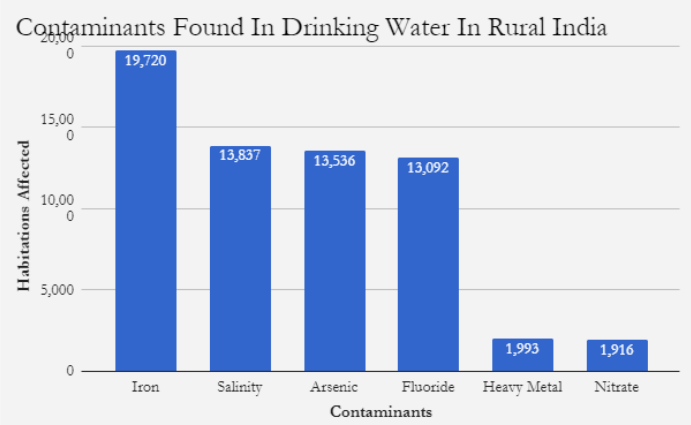
மருத்துவத்துறையில் பெருந்தரவகத்தின் தரவுகம் மிகவும் அத்தியாவசதியத்தேவையாகிறது.



Source: National Family Health Survey, 2015-16

தமிழகத்தில் தற்போது உள்ள பெண்களுக்கு ஹீமோகுளோபின் எனப்படும் ரத்தசிவப்பணுக்கள் பற்றாக்குறை 50%க்கும் மேல் உள்ளது என்று இந்திய மருத்துவ ஆய்வுக்குழு அறிவித்துள்ளது. இது ஒரு மோசமான விளைவு எனக்கொள்ளலாம்,

ஆனால் இது எதனால் ஏற்பட்டது என்பதை அறிய நமக்கு தமிழகத்தின் கடந்த 50 வருடங்களில் ஏற்பட்ட உணவு மாற்றங்கள், பருவ நிலை, உண்ணும் உணவு, வேலை, மன உளைச்சல் என பல காரணிகளை கண்டறிய வேண்டிய அவசியமும் உள்ளது.



Source: Lok Sabha (As of March 2017)

ஆனால் நம்மிடையே பொதுவான தரவுகள் மட்டுமே உள்ள உள்ளனவே தவிர மேற்கண்ட காரணிகள் நம்மிடையே இல்லை. அதோடு தமிழகத்தில் உள்ள அனைத்துப்பள்ளிகளிலும் வருடந்தோறும் மாணவ/மாணவியர்களுக்கு மருத்துவபரிசோதனைகளை பெருந்தரவகமாக மாற்றுவது மிகஅவசியமானது. ஏனெனில் பள்ளி மாணவ/மாணவிர்களிடையே வருடந்தோறும் கிடைக்கும் மாணவர்களின் தகவல்களின் அடிப்படையில் ஏதேனும் உடல் ரீதியான சிக்கல்கள் இருந்தால் அவற்றினை கண்டறிந்து அவர்களுக்கு ஆரம்பக்கட்டத்திலேய குணப்படுத்திட முடியும். இதன் மூலம் ஏதேனும் சிக்கல்கள் ஏற்பட்டால் எதனால் இந்த சிக்கல் ஏற்படுகிறது என்பதை மேலே விவசாயத்தில் கொடுத்த தகவல்களை கொண்டு ஆராய்ந்திட முடியும். இதுபோன்று எல்லா துறைகளுக்கும் தமிழகத்திற்கான, தமிழக மக்களுக்கான, தமிழகத்தில் உள்ள நிலங்கள், மலைகள், காடுகள் என எல்லாத்தரவுகளையும் பெருந்தரவகத்தில் சேர்க்கவேண்டியது அவசியம்.

இப்போது இவையெல்லாம் சில விழுக்காடுகள் கிடைத்தாலும் தொகுக்கப்படாமல் இருக்கின்றன. ஏற்கனவே கிடைத்த தகவல்களின் அடிப்படையில் எனவே இவற்றினை முழுமையாக தொகுக்கவேண்டியது அவசியம் இதை உருவாக்குவதில் உள்ள தொழில்நுட்ப சிக்கல்கள், நடைமுறை சிக்கல்களை ஆராய்ந்து இதனை செயல்படுத்த என்ன தேவைகள் என்பதை வெளி கொணர்வதே இந்த ஆய்வின் நோக்கம்

தீர்வுகள்

விவசாயம் மற்றும் மருத்துவத்துறை சார்ந்த சிக்கல்களை பெருந்தரவகம் மற்றும் தரவு அறிவியலைக் கொண்டு எளிதாக எதிர்கால பிரச்னைகளை தீர்க்கலாம். ஆனால் அதற்கு நமக்கு தேவையான தகவல்களை திரட்டி ஓரிடத்தில் சேமிக்கவேண்டியது அவசியமாகிறது. அதற்கு கீழ்க்காணும் தொழில்நுட்பங்கள் பெரும் உதவி புரியும்.

விவசாயத்திற்கு தேவையான காரணிகளை அடையாளம் காண பொருட்களின் இணையம் (IoT), அவற்றினை சேமிக்க மேகக்கணிமை (Cloud Computing) போன்ற

தொழில்நுட்பங்கள் நம்மிடையே எளிதே கிடைக்கின்றன. இதன் மூலம் விதையின் தரம், மண்ணின் வளம், மழை தண்ணீரின் pH மதிப்பு, காற்றின் வேகம், காற்றின் மாசு அளவு, பூச்சிகள் மேலாண்மை என எல்லா விபரங்களையும் சேர்த்து விவசாயிகளுக்கான மையத்தரவுத்தளத்தினை உருவாக்கிடலாம். மேலே உள்ள அனைத்து தகவல்களையும் கட்டுச்சர தொழில்நுட்பம் (Blockchain Technology) மூலம் ஒருங்கிணைத்தால் எந்த விதைகள் எல்லாம் சிறந்த முறையில் விளைந்திருக்கின்றன, எந்த நிலங்கள் சிறந்த விளைச்சலை கொண்டு இருக்கின்றது என்பது வரை எல்லா விசயங்களையும் நாம் முன் கூட்டியே கணித்திடமுடியும்

நாங்கள் நடத்திவரும் விவசாயம் தளத்தின் மூலம் சமீபத்தில் எடுத்தக் கருத்துகணிப்பில் ஒவ்வொரு வீட்டிலும் ஒரு நாளைக்கு 50 லிட்டர் தண்ணீர் துணி துவைப்பதற்கு செலவிடுடப்படுகிறது. இந்த தண்ணீரை மீண்டும் நாம் மறுசுழற்சிக்கு பயன்படுத்த இயலாது. ஏனெனில் துணி துவைக்க டிட்டர்ஜெண்ட் பவுடரை பயன்படுத்ததுவதால். ஒரு வீட்டுக்கு 50 லிட்டர் தண்ணீர் எனும்போது தமிழகம் முழுக்க குறைந்த பட்சம் 3 கோடி லிட்டர் தண்ணீர் வீணாகிறது. ஒரு நாளைக்கு இவ்வளவு தண்ணீர் வீணாகும்போது ஒரு மாதத்திற்கு, ஒரு வருடத்திற்கு என பார்த்தால் நமக்குத் தெரிந்தே தண்ணீர் வீணாகிறது. இவைகள் அனைத்தையும் தரவு அறிவியல் துணைக்கொண்டு மக்களிடையே கொண்டு சேர்க்க வேண்டியது அவசியம்.

குறிப்பு

இந்த ஆய்வை மேற்கொள்ள ஜெர்மனியில் இருந்து சில உணர்விகளை தருவித்திருந்தோம். அந்த உபகரணங்கள் சரியான நேரத்தில் வந்து சேராததால் நாங்கள் திட்டமிட்டிருந்த திட்டத்தினை மட்டும் இங்கே கொடுத்திருக்கிறோம். உபகரணங்கள் வந்தபின்னர் இந்த ஆய்வு முழுவீச்சுடன் தொடரப்பட்டு இதன் முடிவுகள் அடுத்த மாநாட்டில் சமர்பிக்கப்படும்

Resources

1. http://www.homemade-circuits.com/2014/03/simple-automatic-plant-watering-circuit.html
2. http://www.instructables.com/id/DIY-an-Automatic-Plant-Watering-Device/ http://www.indiaspend.com/cover-story/63-million-indians-without-clean-drinking-water-population-of-australia-sweden-sri-lanka-bulgaria-13088

பிராந்திய மொழியை பயன்படுத்தி இலத்திரணியல் வணிகத்தினை மேற்கொள்வதில் செல்வாக்கு செலுத்தும் காரணிகள். இலங்கையின் மட்டக்களப்பு மாவட்டம் தொடர்பான ஆய்வு.

செ. ஜெயபாலன் [1], இ. ரோகினி [2]

[1]. உயர்தொழில்நுட்பக் கல்விநிறுவகம், கோவில்குளம் ஆரயம்பதி மட்டக்களப்பு

(Advanced Technological Institute, Kovilkulam, Arayampathi, Batticaloa)

[2]. புனித சிசிலியாஸ் தேசிய பெண்கள் பாடசாலை, மட்டக்களப்பு

(BT/St/Cecilia's Girls National School, Batticaloa)

jeyapalanps@yahoo.com, @ rasaratnamrohini@gmail.com

ஆய்வுச் சுருக்கம்

இணைய வணிகத்தின் ஊடாக புதிய மாற்றம் ஒன்றினை ஏற்படுத்தல் என்பதனை பிரதான இலக்காக கொண்டு பிராந்திய மொழியை பயன்படுத்தி இலத்திரணியல் வணிகத்தினை மேற்கொள்வதில் செல்வாக்கு செலுத்தும் காரணிகள். எனும் இவ் வாய்வானது மட்டக்களப்பு மாவட்டத்தினை மையப்படுத்தி மேற்கொள்ளப்பட்டுள்ளது. இவ்வாய்விற்கான மாறிகளாக தனிப்பட்ட காரணிகள், மனப்பான்மையுடன் தொடர்புடைய மாறிகள்,தொழிநுட்பத்துடன் தொடர்புடை மாறிகள்,கட்டமைப்பு மாற்றத்துடனான மாறிகள் என நான்கு சுதந்திர மாறிகளாகக் கொண்டு நிலைத்துநிற்கும் தமிழ் மொழிரீதியிலான இணைய வணிக சேவைமையங்களின் வெற்றி என்ற சாந்த மாறியினையும் ஆய்வு மாதிரியாக கொண்டு. இவ்வாய்வானது மேற்கொள்ளப் பட்டுள்ளது. இலங்கையின் மட்டக்களப்பு மாவட்டதின் கிராமப்புற உற்பத்தியாளர்கள், பிரதேச செயலகப்பிரிவுகளில் தொழிற்படும் பொருளாதார ஈடுபாடுகளுடன் தொடர்புடைய உத்தியோகத்தர்கள், அரச அதிகாரிகள் சாதாரண நுகர்வோர் போன்றோரிடமிருந்து இவ்வாய்வுக்கு தேவையான பச்சைத் தரவுகள் வினாக்கொத்து, மற்றும் நேர்முகப் பரீட்சை அட்டவணைகள் மற்றும் அவதானிப்பு அட்டவணைகள் என்பவற்றைப் பயன்படுத்தி பெற்றுக்கொள்ளப்பட்டுள்ளது. இதற்காக 14 பிரதேச செயலகப் பிரிவுகளிலுமிருந்து படையாக்கப்பட்ட மாதிரியினைப் பயன்படுத்தி100 பேர் ஆய்வுற்கு உட்படுத்தப்பட்டனர். சேகரிக்கப்பட்ட தரவுகள் தொகைரீதியாகவும், பண்புரீதியாகவும் பகுப்பாய்வுக்கு உட்படுத்தப்பட்டது

இவ்வாய்வானது பின்வரும் முடிவுகள் கண்டறியப்பட்டதுவணிக அபிவிருத்தி நோக்கம் கருதி பிரதேச செயலகங்ளில் நியமிக்கப்பட்டுள்ள பட்டதா ாிகளுள் 85 வீதமான பட்டதா ாிகள் கலைப்பாிவு பட்டதாிகள் ஆவர் இவர்களிடத்தில் வியாபாரம் சம்மந்தமான அறிவு குறைவாகக் காணப்படுகின்றது.

இணைய வசதிகள் ஏற்படுத்திக் கொடுப்பதில் 55 வீதமான பங்களிப்பு காணப்படுகிறது, இணைய வணிகம் தொடர்பான விழிப்புணர்வு மிகவும் குறைந்த மட்டத்தில் காணப்படுகின்றது. தமிழ் மொழியில் இணைய வணிகத்தளத்தில் தரவேற்றப்படுவதனை 79 வீதமானவர்கள் ஆதரிக்கின்றனர். நேரலைக் கட்டளைகளுக்கு சௌகரிகமாக பொருள் வழங்கீடு செய்யக்கூடிய வசதிகள் இருப்பதாக 29 வீதமானவர்களே உடன்படுகின்றனர். குறுஞ் செய்திச் சேவையில் மேம்படுத்தல் தகவல்கள் வழங்கப்படுவதனை 35 வீதமானவர்களும் சமுகவலைத்தளங்களில் மேம்படுத்தல்கள் மேற்கொள்ளப்படுவதனை 67 வீதமானவர்களும் உடன்பட்டு ஏற்றுக்கொள்கின்றனர். இலவசமாக இணைய வணிக செயற்பாடுகளை தமிழ் மொழியில் கிராமப்புற உற்பத்தியாளர்கள் சார்பாக மேற்கொள்ளும் பொருளாதார ஈடுபாடுகளுடன் தொடர்புடைய ஊழியர்களின் மனப்பான்மை மாறிகளின் தொடர்புறு குணகம் 0.4 ஆக காணப்படுவதனை ஆய்வு வெளிப்படுத்துகின்றது.இது தமிழ் மொழி மூலமான இணைய வணிக பொருளாதார சேவைகள் மையங்களின் வெற்றிக்கும் ஊழியர்களின் அர்ப்பணிப்பு ரீதியிலான மனப்பான்மைக்குமிடையில் நேரான உறவு உள்ளதனை வெளிப்படுத்துகின்றது. தொழில் நுட்ப மாறிகளின் ஏற்றுக்கொள்ளலுக்கும் மையத்தின் வெற்றிக்குமிடையில் 0.25 இணைவுக் குணகம் காணப்படுவதுடன் கட்டமைப்பு ரீதியியாலான மாறிகளின் ஏற்றுக்கொள்ளலுக்கும் மையத்தின் நிலைத்துநிற்கும் வெற்றிக்குமிடையில் 0.12 எனும் மிகவும் நலிந்த ஆனால் சாதகமான இணைவுக் குணகம் காணப்படுவதனை ஆய்வு வெளிப்படுத்துகின்றது. வெளிநாட்டு இணைய வணிக இணையத்தளங்களின் ஊடாக மேற்கொள்ளப்படும் சந்தைப்படுத்தலில் மோகம் கொண்டு அந்நியச் செலாவணிகளை இழக்கும் எமது இன்றைய சமுகம் பணத்தினை மாத்திரமல்ல தமிழையும் பொருளாதாரத்தினையும் இழப்பதனைத் தடுப்பதற்கு இவ்வாறான இலவச இணைய வணிக பொருளாதார சேவைகள் மையங்கள் கிராமப்புறம் சார்ந்ததாக நலிவுற்ற கிராம உற்பத்தி மற்றும் விவசாயிகளுக்கு அருகிலிருந்து தொழிற்படவேண்டியதும் அச்சேவைகளில் ஈடுபடும் ஊழியர்களின் அர்ப்பணிப்பினை உயர்த்துவதற்கு சிறந்த ஊக்குவிப்புக்களை அரச மற்றம் அரசசார்பற்ற நிறுவனங்கள் வழங்க முன்வரவேண்டும் என்பதனையும், அரசு ஒரு அலகாக இந்த இணைய வணிக மையத்தினை தமது அரச கட்டமைப்புக்குள் ஏற்றுக்கொண்டு கிராம மட்ட உற்பத்தியாளர்களுக்கு தொடர்ச்சியான விழிப்புணர்வு பயிற்சிகள் தொழில்நுட்ப உதவிகள் கணினிமயப்படுத்தப்பட்ட ஒரு மெய்யிலி நிறுவன அமைப்பினை உருவாக்குவதில் மிகுந்த ஈடுபாடு காட்டவேண்டும் என்பதனையும் சரியான வேறுபடுத்தல் தந்திரோபாயங்களையும், பணபரிமாற்றல் வழிமுறைகளிலும், உரிய பொதியமைத்தல் செயற்பாடுகளில் மாற்றங்களை கிராம மட்ட உற்பத்தியாளர்கள் கொண்டிருக்க வேண்டியதையும் இவ்வாய்வு பரிந்துரைக்கின்றது.

பிரதான சொற்கள்
எண்ணிமப் பொருளாதாரம், மனப்பான்மை, கட்டமைப்பு மாற்றம்

1.0 அறிமுகம்

பிரதேசம் ஒன்றின் அபிவிருத்தியிலும், மனிதர்களின் வாழ்க்கைத்தர உயர்விலும் வணிக முயற்சியாண்மைச் செயற்பாடுகள் பாரிய தாக்கம் செலுத்துகின்றது. புரட்சிகரமாக மாறிக்கொண்டிருக்கும் இன்றைய எண்ணிமப் பொருளாதார காலகட்டத்தில் இலங்கை போன்ற காலணித்துவத்திற்கு உட்பட்ட அபிவிருத்தி-யடைந்துவரும் நாடுகளில் செம்மொழியாம் தமிழ் மொழியை மாத்திரம் பேச்சு மற்றும் எழுத்து மொழிகளாக பயன்படுத்தும் பிரதேசங்களின் கிராமப்புற விவசாயிகள், பாரம்பரிய சிறு உற்பத்தியாளர்கள் தாரளமாயமாக்கப்பட்டுள்ள உலகசந்தையில் உள்நுளைந்து அவர்களாகவே தமது உற்பத்திப் பொருட்களை வழங்கீடு செய்வதில் ஒரு இலாபகரமான சந்தைப்படுத்தல் சேவையினை மேற்கொள்ளவில்லை.

மட்டக்களப்பு மாவட்டத்தில் 99வீதமானவர்கள் தமிழ் பேசுபவர்கள். இலங்கையின் பொருளாதாரத்திற்கு அம்பாறை,திருகோணமலை மற்றும் மட்டக்களப்பு மாவட்டங்களை உள்ளடக்கிய கிழக்கு மாகாணம் 6 வீத மத்தியவங்கி அறிக்கை 2015) பங்களிப்பினை மாத்திரம் வழங்கும் அதேவேளை மட்டக்களப்பு மாவட்டம் 1.2 வீத பங்கினையே வழங்குகிறது. மேலும் நாட்டின் வறுமையிலும் முன்னணி மாவட்டமாக இது காணப்படுகின்றது.

இவ்வாய்வின் பிரதான நோக்கமாக இலங்கையின் மட்டக்களப்பு மாவட்டத்தில் காணப்படும் 14 பிரதேச செயலகப் பி ாிவுகளிலும் பொருளாதார ஈடுபாடுகளுடன் தொடர்புடைய அரச உத்தியோகத்தர்களை தொடர்புபடுத்தி உருவாக்கப்பட்டுள்ள தமிழ் மொழி மூலமான இணைய வணிக பொருளாதார சேவைகள் மையங்கள் (www.easbatti.org, www.harisarisale.com) ஏற்படுத்தியுள்ள தாக்கத்தினையும், அவை எதிர்நோக்கும் சவால்களையும் கண்டறிந்து. உரிய விரிவுபடுத்தல் தந்திரோபாயங்களை முன்வைப்பதன் மூலம் உலக சந்தைநோக்கிய மட்டக்களப்பின் கிராம மட்ட உற்பத்திப் பொருட்களுக்கு சந்தை வாய்ப்பினை ஏற்படுத்தத் துண்டுவதுமாகும்.

ஆய்வின் துணை நோக்கமாக ஒரு புதிய வணிக ஆலோசனை சேவை மாதிரி ஒன்றினை இலங்கையின் அரச நிர்வாக கட்டமைப்புக்குள் உள்வாங்கக் கூடிய விதத்தில் அபிவிருத்தி செய்து முன்மொழிதலாகும்.

1.1 ஆய்வுப் பிரச்சனை

செம்மொழியாம் தமிழ் மொழியை மாத்திரம் பேச்சு மற்றும் எழுத்து மொழிகளாக பயன்படுத்தும் பிரதேசங்களின் கிராமப்புற விவசாயிகள், பாரம்பாரிய சிறு உற்பத்தியாளர்கள் தாரளமயமாக்கப்பட்டுள்ள எண்ணிமப் பொருளாதாரத்தில் இலத்திரனியல் வணிகத்தின் ஊடாக உலகசந்தையில் உள்நுளைந்து அவர்களாகவே

தமது உற்பத்திப் பொருட்களை வழங்கீடு செய்வதிலும் இலாபகரமான சந்தைப்படுத்தல் சேவையினை வழங்குவதிலும் பின்நிக்கின்றனர் என்பது அவதானிக்கப்பட்டுள்ளது. அத்துடன் இது தொடாடர்பில் போதுமான ஆய்வுகள் மேற்கொள்ளப்படாமல் இருப்பதும் இதுதொடர்பான தகவல் இடைவெளியினை ஏற்படுத்தியுள்ளது இவ்வாய்வானது இத் தகவல் இடைவெளிப்பிரச்சினைக்கு ஒரு பாலமான நோக்கங்களை நோக்கி நகர்கின்றது.

1.2 ஆய்வின் முக்கியத்துவம்

பிராந்திய மொழிகளை இணைய வணிகத்தில் பயன்படுத்துவதானது அம்மொழி வணிக மொழியாக வளர்வதற்கும் மொழிச் சொந்தக்காரர்கள் புதிய கண்டுபிடிப்புக்களை மேற்கொள்ளவும் வாய்ப்பாக அமையும். இல்லையெனில் தமது பிரதேசங்களுக்குள் மாத்திரம் தமது உற்பத்திகளை வழங்கும் ஒரு குறும்பார்வைச் சந்தைப்படுத்துணர்களாக இவர்கள் அடையாளப்படுத்தப்படுவதுடன் உலக சந்தையில் இவர்கள் நுழையாது இடைத்தரகர்களுக்கு வெறும் மூலப்பொருள் வழங்குணர்களாகவும் எந்தவித பெறுமதி சேர்நடவடிக்கையிலும் ஈடுபடாதவர்களாகவும் காணப்படுவர்.இது ஒரு நலிவடையும் கூலி சமுதாயமாக இவர்களை மாற்றிவிடும். இவ்வாய்வானது கிராம மட்டத்தில் வேறு எந்த மொழி அறிவுமற்ற ஒரு மொழிச்சொந்தக்காரர்களின் வணிக உற்பத்திப் பொருட்களை அதே மொழிபேசும் உறவுகளுக்கு இணைய வணிகத்தில் இணைத்து உலகம் முழுவதிலும் சந்தைப்படுத்துவதற்கான வாய்ப்புக்களை இவ்வாய்வு ஆய்வு செய்ய முனைவது மிகுந்த முக்கியத்துவம் வாய்ந்ததாகும்.

ஆய்வின் துணை நோக்கமாக ஒரு புதிய வணிக ஆலோசனை சேவை மாதிரி ஒன்றினை இலங்கையின் அரச நிர்வாக கட்டமைப்புக்குள் உள்வாங்கச் செய்யவேண்டியதன் அவசியத்தினை உணரச்செய்து அதற்கான மாதிரி ஒன்றினை எல்லோரும் ஏற்றுக் கொள்ளக்கூடிய விதத்தில் அபிவிருத்தி செய்து முன்மொழிதலாகும்.

எனவே இவ்வாய்வானது எதிர்காலத்தில் இது தொடர்பான ஆய்வுகளை மேற்கொள்வர்களுக்கு ஒரு அடிப்படையினை ஏற்படுத்திக்கொடுப்பதுடன், பல்வேறுபட்ட வணிக பங்காளர்கள் மற்றும் கொள்கைவகுப்பவர்கள், மாணவர்கள் நுகர்வோருக்கு மிகுந்த பயனுள்ளதாகவும் அமைவதால் கலத்தின் தேவையான ஒரு ஆய்வாக இது அமைகின்றது.

1.3 ஆய்வின் பொதுவான நோக்கம்

மாற்றத்தின மூலக்கூறுகளான தனிப்பட்ட காரணிகள், மனப்பாங்கு,கட்டமைப்பு, மற்றும் தொழில்நுட்பம், போன்ற காரணிகளுக்கும் நிலைத்திருக்கும் இணைய வணிக அலகுகளின் வெற்றிக்கும் இடையிலான உறவினை கண்டுகொள்ளல்.

1.3.2 ஆய்வின் குறிப்பிடத்தக்க நோக்கங்கள்

1. பொருளாதாரஆலோசனை சேவை அலகு வெற்றியை நோக்கி அபிவிருத்தி உத்தியோகத்தர்களின் மனநிலையை கண்டறிதல்
2. பொருளாதார ஆலோசனை சேவைஅலகு வெற்றியைசெல்வாக்கு செலுத்தும் கட்டமைப்பு காரணிகளை கண்டறிதல்
3. பொருளாதார சேவை அலகு வெற்றியை நோக்கி தொழில்நுட்ப காரணிகளின் அளவை கண்டறிதல்
4. பொருத்தமான சிபா ா சுகளை முன்வைத்தல்

2.0 வரலாற்று முன்னாய்வுகள்

அன்றாடம் மாற்றமடையும் வணிக உலகமயமாதலில் சரியான பிரயோகங்களை வணிகத்தில் மேற்கொள்வதற்கு மொழித்திறன் அவசியமாகும். எந்த தனித்துவமான தாய் மொழியும் உலகத்தின் பொதுவான மொழியாக கொள்ளப்படுவது கிடையாது. டுக்கன்(2007) நிறுவனங்களுக்கிடையில் சாதகமான விளைவு மாற்றங்களுக்கும் வணிக வெற்றிகளும் எற்படுவதற்கு பெறுமதிசார் மொழித்திறனும் கலாச்சார ஊக்குவிப்புக்களும் மிகவும் அவசியமானது என சுட்டிக்காட்டுகின்றார்ஹெர்மான். (2007), ஒரு மேம்பட்ட மொழித்திறனானது "

வெற்றிகரமானஉள்ளூர்தயாரிப்புகளைஉருவாக்கஅல்லது மேம்படுத்துவதற்கு, வாடிக்கையாளர்கள், உள்ளூர் சமூகங்கள் மற்றும் பங்காளிகளிடையே புரிந்துணர்வின் முழு சுற்றுச்சூழல்தேவை. வாடிக்கையாளர்களுடனும், சமூகங்களுடனும், கூட்டாளிகளுடனும் உள்ள நம்பகமான உறவுகளுக்கு வழிவகையாக அமையும் எனக் குறிப்பிடுகின்றார்" எலிசபெத். (2007) "வெளிநாட்டு சந்தைகளில் அமெரிக்க பங்கு அதிகரிப்பதற்கு அமெரிக்க வணிகர்களிடையே மொழிதிறமை இல்லாதது ஒரு பெரிய தடை ஆகும்" மேம்படுத்தப்பட்ட பொதுவான மொழிகள் தொழில்சூழலைப் பற்றியும், உற்பத்தி, மூலப்பொருட்களின் மற்றும் சந்தைப்படுத்தல் மற்றும் வர்த்தகம் பற்றிய புதிய யோசனையையும் இயல்பான எண்ணங்களை உருவாக்கி சிறந்த தகவலை பெற உதவுகின்றன. சேனல்கள் ஒரு மொழிமூலோபாயம் கொண்டிருப்பதோடு, உள்ளூர் மொழி பேசும்மொழி, திறமை வாய்ந்த பணியாளர்கள் மற்றும் நிபுணத்துவ மொழிபெயர்ப்பாளர்களின் கலவையைப் பயன்படுத்தி வணிகவெற்றியை கணிசமான அளவில் அதிகரிக்கும்.

ஹார்வர்ட்வணிகஆய்வு (2013) நாடுகளின் தேசிய வருமானத்திற்கும் ஆங்கில மொழித்தினுக்கு-மிடையில் ஒரு வலுவான இணைவுக்குணகம் காணப்படுகின்றது என சுட்டிக்காட்டுகின்றது. ஆய்வுப் பிரதேசத்தில் 90 வீதத்திற்கும் மேற்பட்டவர்கள தமிழர்கள் அத்துடன் கிராமட்ட உற்பத்தியாளர்களில் 80 வீதத்திற்கும்

மேற்பட்டவர்களுக்கு தமிழைத்தவிர எந்தவிதமான மொழியாற்றலும் இல்லை என்பது சுட்டிக்காட்டத்தக்கது.

2.1 மனப்பாங்குமாற்றம்

ஆட்ஜூமன் (2014: 187) மின்வணிகத் தொழில்நுட்பம் மற்றும் அது தொடர்பான தொழில்நுட்பங்களை நோக்கி ஊழியரக்ளின் மனோபாவம் தொடர்புடையதாக இருப்பதைக் கண்டறிந்தார், ஜெயலனி மற்றும் பலர், 2009: 54) போன்ற காரணிகளால் பாதிக்கப் படலாம். தொழில்நுட்பத்தைப் பயன்படுத்துவது, பணியாளர்களின் (மஹாதா&யங்லெசன், 2013: 75) ஒரு நல்ல அறிவுத்தளத்தை அடிப்படையாகக் கொண்டது. தொழில் நுட்பத்தில் தொழில் நுட்ப நிபுணத்துவம் தொழில்நுட்பத்தை ஏற்றுக் கொள்வதற்கும், திறமையுடன் நிறுவனங்களுக்கிடையில் பயன்படுத்தப்படுவதற்கும் நீண்ட காலமாக செல்லலாம். கபோக்லோ மற்றும் பலர். (2011: 56) தொழில்நுட்ப முன்னேற்றங்களைப் பயன்படுத்தி பணியாளர்களுக்கு தொழில்நுட்ப ரீதியிலான திறன்கள் முக்கியம் என்பதையும் நிறுவனங்களின் குறிக்கோள்களை சந்திப்பதில் உயர்மட்ட முகாமைத்துவத்தின் முயற்சியை ஆதரிப்பதும் முக்கியமானதாகும்.

தொழில்நுட்ப திறன்களைக் கொண்டுள்ள பணியாளர்கள் எப்போதும் வணிக வளர்ச்சிக்கான ஆசைகளை வெளிப்படுத்தும் நிறுவனங்களை (புதுடோகன் & பேங்கோல், 2016, கௌகாக்லூ மற்றும் பலர், 57: 57) விரும்புகின்றனர்..

2.2 கட்டமைப்பு ரீதியான மாற்றம்

சிகர் மற்றும் பலர். (2003) 1978 முதல் 1995 வரை தரவுகளைப் பயன்படுத்தி மேற்கொண்ட ஆய்வில் இக்காலப்பகுதியில் 17 வீதமான வளர்சிக்கு காரணம் கட்டமைப்பு மாற்றம் என சுட்டிக்காட்டுகின்றனர்.மேலும் மாக்மில்லன் மற்றும் ரோட்ரிக் (2011) சீனாவின் ஒட்டுமொத்த உழைப்பு உற்பத்தித்திறனுக்கான கட்டமைப்பு மாற்றம் மற்றும் உள்ளக உற்பத்தித்திறன் வளர்ச்சியின் பங்களிப்பை சுட்டிக் காட்டுகின்றனர்.

சில்வா (2016) தொழில்அனுபவம், தொடர்பு, ஆர்வம், திட்டமிடல், கண்டுபிடிப்பு, சந்தை மற்றும் ஆராட்சிச்செலவு, சந்தைநோக்கம், வியாபாரக் குறி, அங்கீகாரம், நம்பகத்தன்மை, வலயமைப்பு, நிதி ஆதாரங்கள் மற்றும் தகவல் தொழில்நுட்டம் என்பன இலங்கையின் தொழிற்துறை ஆரம்ப நிலைக்கான வெற்றிகரமான காரணிகள் எனக் குறிப்பிடுகின்றார்

.

2.3 தொழில்நுட்ப காரணிகள்

(டோர்னட்ஸ்கி & க்ளீன், 1982) நிறுவனமொன்று இது தமக்கு மிகப்பொருத்தமான தொழில்நுட்பந்தான் என கண்டுகொண்ட உடன் போட்டியாளருக்கு மன்னர் அதனை உள்வாங்கவேண்டும் எனக் குறிப்பிடுகின்றார். (ஜு&க்ரேமர், 2005: 65) வியாபார நடவடிக்கைகளுடன் அதிகரித்த உற்பத்தித்திறன் மற்றும் மேம்பாட்டு செயல்திறனைப் பற்றி அறிமுகப்படுத்தும் ஒரு தொழில்நுட்பத்தை தொழில்நுட்பம் தத்தெடுக்கும்போது, தற்போது இருக்கும் மற்றும் சாத்தியமான சிக்கல்கள் தீர்க்கப்பட வாய்ப்புள்ளது என சுட்டிக் காட்டுகின்றார்.

Wanjau et al. (2012: 76) கூற்றுப்படி, வளரும் பொருளாதாரங்களில் சிறிய நடுத்தரளவு நிறுவனங்கள் இலத்திரணியல் வர்த்தக தொழில்நுட்பங்களைப் பயன்படுத்துவதில் மெதுவாக இருந்து வருகின்றன. இது தொழில்நுட்பத்திலிருந்து பெறப்படும் சாத்தியமான நன்மைகளைத் தவிர. சிறிய நடுத்தரளவு நிறுவனங்கள் இலத்திரணியல் வர்த்தக தொழில்நுட்பங்களைப் பயன்படுத்துவதில் மெதுவாக செயல்படுகின்றன என்பதை பல்வேறு ஆய்வுகள் உறுதிப் படுத்துகின்றன.

3. ஆய்வு முறை

பிரதேச செயலகப்பிரிவுகளில் தொழிற்படும் பொருளாதார ஈடுபாடுகளுடன் தொடர்புடைய உத்தியோகத்தர்களிடமிருந்து அரச அதிகாரிகள் சாதாரண நுகர்வோர் போன்டமிருந்து இவ்வாய்வுக்கு தேவையான பச்சைத் தரவுகள் வினாக்கொத்து, மற்றும் நேர்முகப் பரீட்சை அட்டவணைகள் மற்றும் நேரடி அவதானிப்பு என்பவற்றைப் பயன்படுத்தி பெற்றுக்கொள்ளப்பட்டுள்ளது. இதற்காக 14 பிரதேச செயலகப் பிரிவுகளிலுமிருந்து படையாக்கப்பட்ட மாதிரியினைப் பயன்படுத்தி 100 பேர் ஆய்வுற்கு உட்படுத்தப்பட்டனர். சேகரிக்கப்பட்ட தரவுகள் தொகைரீதியாகவும், பண்புரீதியாகவும் ஒப்பீடுகள், சராசரிகள், அட்டவணைகள், சட்டவரைபுகள் விளக்கம் வியாக்கியானம், தொடர்புறுகுணகம், தொடர்புப்போக்குப்பகுப்பாய்வு கணிதரீதியிலான நுட்பங்கள் பயன்படுத்தப்பட்டுள்ளன. இவ்வாராட்சியானது பின்வரும் கண்டறிதல்களை வெளிப்படுத்துவதாக அமைந்து காணப்படுகின்றது. ஆய்விற்கான வினாக்கொத்தானது நான்கு பிரிவுகளாக 27 வினாக்களைக் கொண்டிருந்தது இவ் 37 வினாக்களில் 09 வினாக்கள் தனிப்பட்ட தகவல்களைத் திரட்டவும் 15 வினாக்கள் மனப்பாங்குமாற்றம், தொழில்நுட்ப மாற்றம் மற்றும் கட்டமைப்பு மாற்றம் போன்ற மாற்ற காரணிகளின் மாறிகளை அடிப்படையாக் கொண்டிருந்ததுடன் 3 வினாக்கள் தமிழ் மொழிமூலமான இணைய வழி பொருளாதார ஆலோசனை மையங்களின் வெற்றியினை மதிப்பிடவும் பயன்படுத்தப்பட்டது. ஆய்வானது முன்னாய்வுகளை மதிப்பிடல் மற்றும் பரிசோதனை முறையான தன்மையில் மேற்கொள்ளப்பட்டதுடன் ஆய்வுக்கு தேவையான தரவுகள் முதல்தர மற்றும் இரண்டாந்தர மூலங்களிலிருந்து பெற்றுக்கொள்ளப்பட்டது.மேலும் ஆய்வானது சமுக விஞ்ஞானத்திற்கான புள்ளிவிபரவியல் மென்பொருளினை(SPSS16.0) பயயன்படுத்தி அட்டணைகள் மற்றும் வரைபடங்கள்

மற்றும் இணைவுக்குணக மற்றும் காரணி ஆய்வுகளைப் பயன்படுத்தி ஆய்வு மாதிரிக்கு ஏற்றால் போல் மாறிகளுக்கிடையிலான தொடர்பு ஆய்வு செய்யப்பட்டது.

4.0 தரவுப் பகுப்பாய்வு

ஆய்வுக்குட்படுத்தப்பட்ட 100 போில் ஆண்கள் 54 வீதமாகவம் பெண்கள் 46 வீதமாகவும் காணப்படுகின்றனர் அத்துடன் பதிலளித்தவர்களில் 30 வயதுக்கும் குறைந்தவர்கள் 10 வீதமாகவும் 30 -39 வயதுக்கு உட்பட்டவர்கள் 60 வீதமாகவும் 40-49 வயதுக்கு உட்பட்டவர்கள் 25 வீதமாகவும் 50-60 வயதுக்கு உட்பட்டவர்கள் 3 வீதமாகவும் 60 வயதுக்கும் மேற்பட்டவர்கள் 2 வீதமாகவும் காணப்படுகின்றனர். இதிலிருந்து பொருளாதார சேவைநிலைய உத்தியோகத்தர்களின் அதிகமானவர்கள் ஆண்களாக இருப்பதுடன் இவர்களில் அதிகமானவர்கள் 30-39 வயதிலும் காணப்படுவது சுட்டிக்காட்டத்தக்கது.

85 வீதத்திற்கும் மேல் கலைப்பிரிவுப் பட்டதாரிகளே பொருளாதார ஈடுபாடுகளுடன் தொடர்புடைய உத்தியோகத்தர்களாக வணிக அபிவிருத்தி நோக்கம் கருதி பிரதேச செயலகங்களில் நியமிக்கப்ட்டுள்ளனர்.இணைய வசதிகள் ஏற்படுத்திக் கொடுப்பதில் 55 வீதமான பங்களிப்புக் காணப்படுவதாகவும், இணைய வணிகம் தொடர்பான விழிப்புணர்வு மிகவும் குறைந்த மட்டத்திலிருப்பதனை ஆய்வு சுட்டிக்காட்டுகின்றது. தமிழ் மொழியில் இணைய வணிகத்தளத்தில் தரவேற்றப்படுவதனை 79 வீதமானவர்கள் ஆதரிக்கின்றனர்.நேரலைக் கட்டளைகளுக்கு சௌகரிகமாக பொருள் வழங்கீடு செய்யக்கூடிய வசதிகள் இருப்பதாக 29 வீதமானவர்களே உடன்படுகின்றனர். குறுஞ் செய்திச் சேவையில் மேம்படுத்தல் தகவல்கள் வழங்கப்படுவதனை 35 வீதமானவர்களும் சமுகவலைத்தளங்களில் மேம்படுத்தல்கள் மேற்கொள்ளப்படுவதனை 67 வீதமானவர்களும் உடன்பட்டு ஏற்றுக்கொள்கின்றனர்.

அட்டவணை: 4.1 மனப்பாண்மை காரணிகளுக்கும் பொருளாதார சேவை மையங்களின் வெற்றிக்கும் இடையிலான தொடர்பு

மாறி		வெற்றி மட்டம்
மனப்பாங்கு	பியசன் இணைவுக் குணகம்	.40**
	முக்கியத்துவம். (2-வால்)	.000
*. 0.05 மட்டத்தில் இணைவுக்குணமக் முக்கியம் வாய்ந்தது (2-வால்).		

மூலம் ஆய்வுத் தரவுகள்

வெற்றிவிகிதம் மற்றும் மனப்பாண்மைக் காரணிகளுக்கு இடையிலான நெகிழ்வான தொடர் புகுணகம் 0.40 ஆகும் மற்றும் இது 0.01 நிலை அளவில் புள்ளியியல் முக்கியத்துவம் வாய்ந்தது.

எனவே வெற்றிவிகிதம் மற்றும் மனப்பாண்மைக காரணிகள் இடையே மிதமான உறவு உள்ளதனை ஆய்வு சுட்டிக் காட்டுகின்றது. தமிழ் மொழி மூலமான இணைய வணிக பொருளாதார சேவைகள் வழங்குகின்ற இவ்வகையான இணைய மையங்களின் வெற்றிக்கும் ஊழியர்களின் அர்ப்பணிப்பு ரீதியிலான மனப்பான்மைக்குமிடையில் நேரான உறவு உள்ளதென்பது இவ்வாறான மக்களக்கு அருகில் இலவசமாக அவர்களது கிராமத்து உற்பத்திகளை மக்கள் நலன்கருதி சந்தைப்படுத்தக்கூடிய செயற்பாட்டிற்கு ஒரு உந்து கோலாகக் காணப்படுகின்றது.

அட்டவணை: 4.2 தொழில்நுட்பக் காரணிகளுக்கும் பொருளாதார சேவை மையங்களின் வெற்றிக்கும் இடையிலான தொடர்பு

மாறி		வெற்றி மட்டம்
தொழில் நுட்பம்	பியசன் இணைவுக் குணகம்	.250**
	முக்கியத்துவம். (2-வால்)	.000
*. 0.05 மட்டத்தில் இணைவுக்குணமக் முக்கியம் வாய்ந்தது (2-வால்).		

மூலம் ஆய்வுத் தரவுகள்

வெற்றிமட்டம் மற்றும் தொழில் நுட்பக் காரணிகளுக்கு இடையிலான நெகிழ்வான தொடர்பு குணகம் 0.250 ஆகும் மற்றும்இது 0.01 நிலை அளவில் புள்ளியியல் முக்கியத்துவம் வாய்ந்தது. இங்கு காணப்படும் பலவீனமான சாதகமான உறவு இணைய பொருளாதார சேவை மையங்களில் மேற்கொள்ப்படவேண்டிய தொழில்நுட்ப ரீதியிலான மேம்படுத்தல்களை சுட்டிக்காட்டுவதாக அமைகின்றது.

அட்டவணை: 4.3 கட்டமைப்புகாரணிகளுக்கும் பொருளாதார சேவை மையங்களின் வெற்றிக்கும் இடையிலான தொடர்பு

மாறி		வெற்றி மட்டம்
கட்டமைப்பு	பியசன் இணைவுக் குணகம்	.120**
	முக்கியத்துவம். (2-வால்)	.000
*. 0.05 மட்டத்தில் இணைவுக்குணமக் முக்கியம் வாய்ந்தது (2-வால்).		

மூலம் ஆய்வுத் தரவுகள்

0.120 இணைவுக் குணகம் காணப்படுவதுடன் கட்டமைப்பு ரீதியியாலான மாறிகளின் ஏற்றுக் கொள்ளலுக்கும் மையத்தின் நிலைத்துநிற்கும் வெற்றிக்குமிடையில் 0.120 எனும் மிகவும் நலிந்த ஆனால் சாதகமான இணைவுக் குணகம் காணப்படுவதனை ஆய்வு வெளிப்படுத்துகின்றது. சரியான ஒரு கட்டமைப்பின் தேவையினை இது வெளிப்படுத்துவதுடன் கட்டமைப்பு ரீதியிலான ஒரு மாற்றத்திற்கு இவ் வணிக பொருளாதார சேவை மையங்கள் உள்வாங்கப்பட வேண்டிய தேவையினை இது சுட்டிக்காட்டுகின்றது.

5.0 ஆய்வுப் பரிந்துரை மற்றம் முடிவு

வெளிநாட்டு இணைய வணிக இணையத்தளங்களின் ஊடாக மேற்கொள்ளப்படும் சந்தைப்படுத்தலில்தந்திரோபாயங்களில் மோகம் கொண்டு அந்நியச் செலாவணிகளை இழக்கும் எமது இன்றைய சமுகம் பணத்தினை மாத்திரமல்ல தமிழையும் பொருளாதாரத்தினையும் இழப்பதனைத் தடுப்பதற்கு இவ்வாறான இலவச இணைய வணிக பொருளாதார சேவைகள் மையங்கள் கிராமப்புறம் சார்ந்ததாக நலிவுற்ற கிராம உற்பத்தி மற்றும் விவசாயிகளுக்கு அருகிலிருந்து தொழிற்படவேண்டியதது காலத்தின் கட்டாயமாகும். என்பதனை ஆய்வின் சாரா மாறிகளான மனப்பாங்கு,தொழில்நுட்பம் மற்றும் கட்டமைப்பு ரீதியிலான காரணிகளுக்கும் சார்ந்த மாறியான தமிழ் மொழி மூலம் மட்டக்களப்பு மாவட்டத்தில் பிரதேச செயலகப்பிரிவுகளில் தொழிற்படும் இணைய வணிக பொருளாதார ஆலோசனை சேவை மையங்களுக்கும் இடையில் நேரான இணைவுக்குணகங்களின் சேர்வைகள் சுட்டிக்காட்டுகின்றது அச்சேவைகளில் ஈடுபடும் ஊழியர்களின் அர்ப்பணிப்பினை உயர்த்துவதற்கு சிறந்த ஊக்குவிப்புக்களை அரச மற்றம் அரசசார்பற்ற நிறுவனங்கள் வழங்க முன்வரவேண்டும் என்பதனையும், அரசு ஒரு அலகாக இந்த இணைய வணிக மையத்தினை தமது அரச கட்டமைப்புக்குள் ஏற்றுக்கொண்டு கிராம மட்ட உற்பத்தியாளர்களுக்கு தொடர்ச்சியான விழிப்புணர்வு பயிற்சிகள் தொழில்நுட்ப உதவிகள் கணினிமயப்படுத்தப்பட்ட ஒரு மெய்யிலி நிறுவன அமைப்பினை உருவாக்குவதில் மிகுந்த ஈடுபாடு காட்டவேண்டும் என்பதனையும் சரியான வேறுபடுத்தல் தந்திரோபாயங்களையும், பணபரிமாற்றல் வழிமுறைகளிலும், உரிய பொதியமைத்தல் செயற்பாடுகளில் மாற்றங்களை கிராம மட்ட உற்பத்தியாளர்கள் கொண்டிருக்க வேண்டியதையும் இவ்வாய்வு பரிந்துரைக்கின்றது.

Bibliography

1. Ali, M. & Kurnia, S. 2011. Interorganizational Systems (IOS) adoption in the Arabian Gulf region: The case of Bahraini grocery industry. Journal of Information Technology and Development, 17 (4) (2011), pp. 253–267
2. Ardjouman, D. 2014. Factors influencing small and medium enterprises (SMEs) in adoption and use of technology in Cote d'Ivoire. International Journal of Business and Management, 9(8):179-190.

3. Annenberg Center for Communication, University ofSouthern California, Los Angeles, California, (1999), Conference on Issues in Global Electronic Commerce.
4. Arpaci, I., Yardimci, Y.C., Ozkan, S. & Turetken, O. 2012. Organizational adoption of information technologies: A literature review. International Journal of eBusiness and e-Government Studies,4(2):37
5. Chan, S. & Lu, M. 2004. Understanding internet banking adoption and use behaviour: A Hong Kong perspective. Journal of Global Information Management, 12(3):21-43.
6. Curry, J.P., Wakefield, S.D., Price, J.L. & Mueller, C.W. 1986. On the causal ordering of job satisfaction and organisational commitment source. The Academy of Management Journal, 29(4):847-858.
7. Dahnil, M.I., Marzuki, K.M., Langgat, J. Fabeil, N.F. 2011. Factors influencing SMEs adoption of social media marketing. Procedia- Social and Behavioural Sciences, 148(1):119-126.
8. Dakora, E.A.N, Bytheway, A.J. & Slabbert, A. 2010. The Africanisation of South African retailing: A review. African Journal of Business Management, 4(4).
9. Dauda, Y.A. & Akingbade, W.A. 2011. Technology change and employee performance in selected manufacturing industries in Lagos state of Nigeria. Australian Journal of Business and Management Research, 1(5):32-43.
10. Dlodlo, N. & Dhurup, M. 2010. Barriers to e-marketing adoption among small and medium enterprises (SMEs) in the Vaal Triangle. Acta Commercii, 164-180.
11. Hough, J., Thompson, A.A., Strickland, A.J. & Gamble, J.E. 2011. Crafting and executing strategy: Creating sustainable high performance in South African businesses. Berkshire: McGraw-Hill.
12. Mahadea, D. & Youngleson, J. 2013. (Eds). Entrepreneurship and small business management. Cape Town: Pearson Education.
13. Maholtra, N.K. 2010.Marketing research: An applied orientation. 6th edition. New York: Pearson Education.
14. Wanjau, K., Macharia, N.R. & Ayodo, E.M.A. 2012. Factors affecting adoption of electronic commerce among small-medium enterprises in Kenya: Survey of tour and travel firms in Nairobi. International Journal of Business Humanities and Technology, 2(4):76-91.

15. Yaghoubi, N.M. & Bahmani, E. 2010. Factors affecting the adoption of online banking. International Journal of Business and Banking, 5(9):150-165; Srilanka Central Bank report 2012,2013, 2014,2015,2016; Batticaloa distric Development Plan report, Eas unit reports.

தமிழ்ச் சூழலில் திறந்த இணைப்புத் தரவுக்கான மெய்ப்பொருளிய உருவாக்கம் நோக்கி

இ. நற்கீரன்
நூலக நிறுவனம்

கட்டுரைச் சுருக்கம்
தமிழ்ச் சூழலில் நினைவு நிறுவனங்கள் தமது தரவுகளை ஒரு பொது மெய்ப் பொருளியத்தைப் பயன்படுத்தி இணைப்புத் தரவாக வெளியிடும் தேவை உள்ளது. அந்த மெய்ப்பொருளியம் ரிம் பேர்னேர்ஸ்-லீ 2006 இல் முன்வைத்த கொள்கைகளுக்கு ஏற்பதாகவும், அனைத்துல சீர்தரங்களையும், ஏற்கனவே பயன்பாட்டில் உள்ள மெய்ப்பொருளியங்களைப் பயன்படுத்தியும் வடிவமைக்கப்பட வேண்டும். இந்த ஆய்வு அனைத்துலக அருங்காட்சியகங்கள் மன்றத்தின் கருத்துரு குறிப்பு மாதிரியை (CIDOC Conceptual Reference Model - சி.ஆர்.எம்) அடிப்படையாகக் கொண்டு, பரந்த பயன்பாட்டில் உள்ள டப்பிளின் கருவகம் (Dublin Core), இசுகீமா (Schema) போன்ற சொற்தொகுதிகளைப் பயன்படுத்தி எளிமைப்படுத்தப்பட்ட ஒரு மெய்ப்பொருளியத்தை முன்வைக்கிறது. இந்த மெய்ப்பொருளியம் சேகரிப்பு, படைப்பு, கருத்துருப் பொருள், பௌதீகப் பொருள், நபர், குழு, இடம், நிகழ்வு, நேரம்-காலம் ஆகிய எட்டு முதன்மை வகுப்புகளைக் கொண்டது. மேலும், இந்த மெய்ப்பொருளியத்தை ஐலண்டோரா குளோ (Islander CLAW) என்ற இணைப்புத் தரவு தளத்தில் நிறுவுவதற்கான வழிமுறையையும் விபரிக்கிறது. இவ்வாறு தகவல் வளங்களையும், அவை சுட்டும் தரவுகளையும் இணைத்துப் பகிர்வதன் மூலம் அவற்றை இலகுவாக அணுக, தேட, வினவ, பகுத்தறிய முடியும்.

குறிசொற்கள்: இணைப்புத் தரவு, மெய்ப்பொருளியம், வள விபரிப்புச் சட்டகம், எண்ணிமக் களஞ்சியம், மும்மை, மும்மைத்தரவுத்தளம்

1. முன்னுரை

உலகளாவிய வலை (World Wide Web) பாரிய தொழில்நுட்ப, சமூக, பொருளாதார புரட்சியை ஏற்படுத்தியது. ஆவணங்களை, அவற்றுக்கு இடையேயான இணைப்புக்களை உரலிகள் அல்லது வலை முகவரிகளைப் (URL) பயன்படுத்தி இணையம் ஊடாக அணுகுவதற்கான நுட்பக் கட்டமைப்பே உலகளாவிய வலை ஆகும். இதன் நீட்சியாக, அடுத்த தலைமுறைத் தொழில்நுட்பமாக பொருளுணர் வலை (Semantic Web) அல்லது இணைப்புத் தரவு (Linked Data) தொழில்நுட்பங்கள் விளங்குகின்றன. இணைப்புத் தரவு தொழில்நுட்பங்கள் கடந்த சில ஆண்டுகளாக

முதிர்ச்சி அடைந்துள்ளன, பரவலான பயன்பாட்டுக்கு வந்துள்ளன. நாம் எண்ணிம வளங்களை அல்லது தரவுகளை வெளியிடும் (publishing), பரிமாறும் (data exchange), ஒருங்கிணைக்கும் (integration), கண்டுபிடிக்கும் (discovery), பயன்படுத்தும் முறைகளில் பாரிய மாற்றங்களையும் வாய்ப்புக்களையும் இது கொண்டுவருகின்றது.[1] பெருந்தரவை ஒழுங்குபடுத்தவும் (structuring big data), தரவுகளைப் பகுத்தறியவும் (reasoning with data), தரவுகளைப் பற்றி துல்லியமான கேள்விகளைக் கேக்கவும், முடிவுகளை எடுக்கவும் இணைப்புத் தரவு நுட்பங்கள் உதவுகின்றன.

வலை ஆவணங்கள் உலகளாவிய வலைக்கு அடிப்படையாக அமைந்தன என்றால், இணைப்புத் தரவுக்குப் பொருட்கள் (things) அல்லது பொருட்களைப் பற்றிய விபரிப்புக்கள் அடிப்படையாக அமைகின்றன.[2] பொருட்கள் ஒரு படைப்பாக, நபராக, இடமாக, கருத்தாக, நிகழ்வாக, எதுவாகவும் அமையலாம். இந்தப் பொருட்களைப் பற்றியும், அவற்றுக்கு இடையேயான தொடர்புகளை விபரிக்கவும் பயன்படும் அடிப்படைத் தொழில்நுட்பமே வள விபரிப்புச் சட்டகம் (Resource Description Framework - RDF - ஆர்.டி.எப்) ஆகும். வள விபரிப்புச் சட்டகம் ஒரு பொருளை எழுவாய் - பயனிலை - செயற்படுபொருள் (subject-predicate-object) என்ற இயற்கை மொழி வசனத்தின் அமைப்பைக் கொண்ட கூற்றுக்களால் (statements) அல்லது மும்மைகளால் (triples) விபரிக்கிறது. ஒவ்வொரு பொருளும் ஒரு தனித்துவமான உரலியால் அடையாளம் காணப்படுகின்றது. இந்த உரலியே அவற்றை இணைக்கப் பயன்படுத்தப்படுகின்றது. வள விபரிப்பு மும்மைகள் மும்மைத்தரவுத்தளம் (Triplestore) ஒன்றில் சேமிக்கப்பட்டு, எசுபார்க்கிள் (SPARQL - SPARQL Protocol and RDF Query Language) போன்ற மொழிகள் ஊடாக வினவப்படலாம்.

வள விபரிப்புச் சட்டகம் எளிமையானது. ஒரு பொருளை பலர் வெவ்வேறான முறைகளில் உருவகிக்க முடியும் (represent/model) விபரிக்க (describe) முடியும். இதனால் தரவுகளைப் பகிர்வதில், பயன்படுத்துவதில் தடைகள் ஏற்பட்டன. ஒரு பொதுவான, பகிரப்படக் கூடிய அணுகுமுறை அல்லது கருத்தோற்ற முறைமை (schema) தேவை என்பது நன்கு உணரப்பட்டது. இவ்வாறு ஒரு குறிப்பிட்ட துறை பற்றிய தரவுகளை அல்லது அறிவை உருமாதிரியாக்கப் பயன்படும் வகுப்புகள் (classes/types/sets), பண்புகள் (properties/attributes) மற்றும் உறவுகளைக் (relationships) கொண்ட சட்டகமே மெய்ப்பொருளியம் ஆகும். இணைப்புத் தரவுக்கான மெய்ப்பொருளியங்களை வள விபரிப்புச் சட்டக கருத்தேற்ற முறைமை (RDF Schema), வலை மெய்ப்பொருளிய மொழி (OWL), எளிய அறிவு ஒழுங்கமைப்பு முறைமை (SKOS) போன்ற மொழிகளைப் பயன்படுத்தி உருவாக்க முடியும். இந்த

மொழிகளைப் பயன்படுத்தி மெய்ப்பொருளியத்தையும் இணைப்புத் தரவாகவே வெளிப்படுத்த முடியும் என்பது குறிப்பிடத்தக்கது.

தமிழ்ச் சூழலில் எண்ணிமப்படுத்தல், எண்ணிமப் பாதுகாப்பு, எண்ணிம வளங்களை உருவாக்கும் செயற்திட்டங்கள் கடந்த இருபது ஆண்டுகளுக்கு மேலாக முன்னெடுக்கப்பட்டு வருகின்றன. மதுரைத் திட்டம், நூலக நிறுவனச் செயற்திட்டங்கள், படிப்பகம், தமிழ்க் கல்விக் கழகம், தமிழ்நாடு வேளாண்மைப் பல்கலைக்கழகம், இந்திய எண்ணிம நூலகம், சிங்கப்பூர் தமிழ் மின்மரபுடைமைத் திட்டம், தமிழ் விக்கியூடகங்கள் என்று பல செயற்திட்டங்கள் தமிழ் எண்ணிம வளங்களை அணுக்கப்படுத்துகின்றன. பிற நிறுவனங்களுக்கும் அணுக்கப்படுத்தல் பணிகளைத் தொடங்கியுள்ளன. இந்த எண்ணிம வளங்களை, தரவுகளை வளர்ந்துவரும் இணைப்புத் தரவுக் கட்டமைப்புக்கு ஏற்ற வகையில் வெளியிடுவது, கட்டமைப்பது அவசியமாகும். அவ்வாறு செய்தாலே கூகிள் தேடுபெறியில் துல்லியமாக இடம்பெறச் செய்வதில் இருந்து சிக்கலான கேள்விகளுக்கு பதிலளிக்க கூடிய வரைக்குமான பயன்களைப் பெற முடியும். இவ்வாறு பல செயற்திட்டங்கள் தரவுகளை இணைப்புத் தரவாக வெளியிட, பகிர முன்வரும் போது மெய்ப்பொருளியங்களின் தேவை எழுகிறது. மேற்சுட்டப்பட்ட நிறுவனங்கள் பாதுகாத்து அணுக்கப்படுத்தும் நூல்கள், இதழ்கள், நாளிதழ்கள், பல்லூடகங்கள், வலைத்தளங்கள், தரவுகள் உட்பட்ட எண்ணிம வளங்களையும், அவற்றோடு தொடர்புடைய பண்பாட்டு, வகைப்படுத்தல் மற்றும் பொது அறிவுத் தரவுகளையும் (classification/knowledge organization, general knowledge) ஆக்க, பகிர, பராமரிக்கத் தேவையான ஒரு மெய்ப்பொருளியத்தை உருவாக்குவதை நோக்கி இந்த ஆய்வுக் கட்டுரை அமைகிறது.

முதலில் இக் கட்டுரை எண்ணிம நூலகம், ஆவணகம், நினைவு நிறுவனங்களுக்குத் தேவையான மெய்ப்பொருளியத்தின் நோக்கங்களை, அதை வடிவமைப்பதில் பின்பற்ற வேண்டிய கொள்கைகளை விபரிக்கும். இரண்டாவது இது மெய்பொருளியத்தை வடிவமைப்பதற்குப் பயன்படும் முறையியலை (methodology) விபரிக்கும். மூன்றாவது ஒர் அடிப்படை மெய்ப்பொருளியத்தை இது விபரிக்கும். இறுதியாக இந்த மெய்பொருளியத்தை அடுத்த தலைமுறைக் களஞ்சிய மென்பொருளான ஐலண்டோரா குளோவில் (Islandora CLAW) எப்படிப் பயன்படுத்தலாம் என்று எடுத்துரைக்கும்.

2. மெய்பொருளியத்தின் தேவைகளும் நோக்கங்களும்

மெய்ப்பொருளியம் ஒரு குறிப்பிட்ட ஆய்வுத் துறையின் கருத்துருக்களையும் (concepts) அவற்றுக்கு இடையேயான உறவுகளையும் (relationships) உருவகப்படுத்த, விபரிக்க, ஒழுங்குபடுத்த உதவுகின்றது.[3] நாம் விபரிக்க முனையும் துறைகள் நூலகம்,

ஆவணகம் மற்றும் அருங்காட்சியகங்களின் அறிவுத் துறைகள் ஆகும். இவை முறையே நூல்கள், ஆவணங்கள், அரும்பொருட்களோடு தொடர்புடையவை. நாம் உருவாக்க முற்படும் மெய்ப்பொருளியம் இத் துறைகளில் பயன்படும் சிக்கலான தகவல் வளங்களை (எ.கா தொடர்கள், வலைத்தளம், கலைப்பொருள்) விபரிக்கக் கூடியதாக இருக்க வேண்டும். இந்த வளங்களை இலகுவாகத் தேட, கண்டுபிடிக்க உதவ வேண்டும்.

மேற்குறிப்பிட்ட துறைகள் மொத்த அறிவுத்துறைகள் தொடர்பான தகவல் மூலங்களை வெளிப்படுத்தும், வகைப்படுத்தும், ஒழுங்குபடுத்தும், பகிரும் மேநிலைப் பணியில் ஈடுபட்டுள்ளன. அந்த நோக்கில் மொத்த அறிவுத் துறைகளை வகைப்படுத்தவும் ஒழுங்குபடுத்தவும் மெய்ப்பொருளியம் உதவ வேண்டும். இதனை நூலகவியலின் மரபுசார்ந்த வகைப்படுத்தல் சிக்கல் என்று குறிக்கலாம்.

தனியே ஒரு நிறுவனத்திடம் மட்டும் இருக்கும் தகவல்களை விபரிக்க, ஒழுங்கமைக்க மட்டும் அல்லாமல் பல்வேறு நிறுவனங்களுக்கு இடையே தகவல் பரிமாற்றம் (data exchange) செய்யவும், பலபடித்தான தகவல்களை (hetrogenous information) ஒருங்கிணைக்கவும் மெய்ப்பொருளியம் உதவ வேண்டும். இத் தகவல்களை கணினி மூலம் பகுத்தறிய (reasoning) செய்யக் கூடியதாக இருக்க வேண்டும். மையப்படுத்தப்பட்ட கட்டுப்பாடுகள் இன்றி (decentralized), ஒப்பீட்டளவில் இலகுவாக நிறைவேற்றக் கூடியதாகவும் (implementable), நீட்டப்படக் (extendable) கூடியதாகவும் இருக்க வேண்டும்.[4]

இன்னுமொரு வகையில் நோக்குவதானால் இந்த மெய்ப்பொருளியத்தை அடிப்படையாகக் கொண்ட ஒரு நுட்ப கட்டமைப்பு குறிப்பான கேள்விகளுக்கு நேரடியாக, இலகுவாகப் பதில் தரக் கூடியதாக இருக்க வேண்டும். எ.கா பெரியார் எழுதிய நூல்கள் எந்தத் துறைகள் பற்றி பெரிதும் பேசுகின்றன? இந்த ஊரோடு தொடர்புடைய அரும்பொருட்கள் எவை? இந்த ஊரில் போரில் ஈழப் போரில் இறந்த மக்கள் யார்? இந்தக் கருத்துப் பற்றி தொடர்புடைய ஆக்கங்கள் எவை? இந்தத் கருத்துரு/துறைசார்ந்த எழுத்தாளர்கள் யார்? இந்த எடுத்துக்காட்டுக்கள் போன்று படைப்பு, அமைப்பு, நிகழ்வு, பொருள் என்று பல்வேறு விடயங்கள் பற்றிய கேள்விகளுக்கு பதில் தரக்கூடியதாக இந்த மெய்ப்பொருளியம் அமைய வேண்டும். ஒரு குறிப்பிட்ட துறைசார்ந்த கேள்விகளுக்கு (எ.கா உயிரணுவின் பாகங்கள் எவை) இது நேரடியாகப் பதில் தர முடியாமல் இருக்கலாம். ஆனால் அந்தத் துறைசார்ந்த மெய்ப்பொருளியத்தோடு மேல்நிலையில் தொடர்புறக் கூடியதாக அமைய வேண்டும்.

3. மெய்ப்பொருளிய வடிவமைப்புக்கான கொள்கைகள்

இணைப்புத் தரவுக்கான நான்கு கொள்கைகளை உலகளாவிய வலையின் கண்டுபிடிப்பாளாரான ரிம் பேர்னேர்ஸ்-லீ 2006 இல் பின்வருமாறு முன்வைத்தார்:[5]

- தகவல் வளங்களையும் (எ.கா வலைப்பக்கம், கோப்பு, படிமம்), தகவல் அல்லா வளங்களையும் (எ.கா பொருள், கருத்து) யு.ஆர்.ஐ (URI) பெயர் கொண்டு இனங்காட்டுதல்.
- எச்.ரி.ரி.பி யு.ஆ.ஐ களைப் பயன்படுத்தல் மூலம் இந்த வளங்களைப் பற்றிய தகவல்களைக் கண்டறிய உதவுதல் (dereferencing using http)
- ஒருவர் யு.ஆர்.ஐ அணுகும் போது, பயன்படக்கூடிய தகவல்களைத் திறந்த சீர்தரங்களைப் பயன்படுத்தி வழங்குதல் (எ.கா ஆர்.டி.எப் (RDF), எசுபார்க்குவல் (SPARQL))
- பிற வளங்களுக்கு இணைப்புத் தருதல். இதன் ஊடாக மேலதக வளங்களைக் கண்டறிய உதவுதல்.

நாம் கட்டமைக்க முனையும் எந்தவொரு இணைப்புத் தரவு மெய்ப்பொருளியமும் இந்த அடிப்படைக் கொள்கைகளை மதித்து அமைய வேண்டும்.

மெய்பொருளியங்களை வடிவமைக்கும் போது எதிர்நோக்கப்படும் ஒரு பொதுச் சிக்கல் நேர்ப்படுத்தல் (Mapping and Alignment) சிக்கல் ஆகும். அதாவது ஒரே துறையைச் சார்ந்த மெய்ப்பொருளியங்கள் அத் துறையை சிறிய வேறுபாடுகளுடானா வகுப்புக்களாலும் பண்புகளாலும் (language-level mismatches) வரையறை செய்ய முற்படும். அல்லது அத் துறையை விபரிப்பதில் பொருள் சார்ந்த வேறுபாடு (semantics-level mismatches) காணப்படும். இந்தச் சிக்கலைத் தீர்க்க பொது மேநிலை மெய்ப்பொருளியங்கள் (Upper Ontologies) அல்லது குறிப்பு மெய்ப்பொருளியங்கள் (Reference Ontologies) உதவுகின்றன.[6]

இணைப்புத் தரவின் அடிப்படை விழுமியமே ஏற்கனவே உள்ளவற்றை பயன்படுத்தல், உள்ளவற்றுக்கு இணைப்புத் தருதல் ஆகும். அந்த நோக்கில் ஏற்கனவே பரவலான பயன்பாட்டில் உள்ள மெய்ப்பொருளியங்களைப் இயன்றவரை நாம் பயன்படுத்த வேண்டும்.[7] மேலும், மெய்பொருளியம் அடிப்படைப் பண்புகளை வரையறை செய்து, பயனர்கள் தங்கள் தேவைகளுக்கு ஏற்றவாறு தேர்ந்து பயன்படுத்தவும், நீட்டக் கூடியவாறும் மெய்ப்பொருளியம் அமைய வேண்டும்.

4. மெய்ப்பொருளியத்தை உருவாக்குவதற்கான முறையியல்

மெய்ப்பொருளியங்களை விபரிக்க பல முறையியல்கள் உண்டு. பொதுவாக பின்பற்றப்படும் ஒரு முறையியலை பின்வருமாறு விபரிக்கலாம்:[8]

* செயற்பரப்பை தெளிவாக வரையறை செய்தல் - Define the scope

* கருத்துருக்களை வரையறை செய்தல், எ.கா வகுப்புக்கள் - Define the concepts

* கருத்துருக்களை படிநிலைகள் கொண்ட ஒரு பாகுபாட்டில் ஒழுங்குபடுத்தல் - Organize the concepts/classes into a taxonomy

* வகுப்புக்களுக்கு இடையேயான உறவுகளை வரையறை செய்தல் - Define the relations

* வகுப்புக்களின் பண்புகளையும், அந்தப் பண்புகள் எடுக்கக் கூடிய மதிப்புக்களையும் வரையறை செய்தல் - Define properties, their domains and ranges

* குறிப்பான பொருட்களை வரையறை செய்தல் - Defining the instances

* அடிகோள்களை (axioms), விதிகளை(rules), செயற்கூறுகளை (functions) விபரித்தல்

எமது நோக்கம் ஒரு அடிப்படையான மேநிலை மெய்ப்பொருளியத்தின் மீது, ஏற்கனவே பரந்த பயன்பாட்டில் உள்ள மெய்ப்பொருளியங்களைப் பயன்படுத்தி நீட்டி வடிவமைப்பது ஆகும். அந்த வகையில் இத் துறையில் பயன்படக்கூடிய மேநிலை மெய்ப்பொருளியங்களாக (Upper Ontologies) அடிப்படை முறைசார் மெய்ப்பொருளியம் (Basic Formal Ontology -பி.எப்.ஓ) மற்றும் அனைத்துலக அருங்காட்சியகங்கள் மன்றத்தின் கருத்துரு குறிப்பு மாதிரி (CIDOC Conceptual Reference Model - சி.ஆர்.எம்) அமைகின்றன. துறைசார் மெய்பொருளியங்களாக (Domain Ontologies) பிப்ஃபேரம் (BIBFRAME), போர்ட்லன்ட் பொது தரவு மாதிரி (Portland Common Data Model), திறந்த ஆவணக தரவு மாதிரி (Open Archive Data Model) போன்றவை அமைகின்றன. மேலும், டப்பிளின் கருவகம் (Dublin Core), எசுகீமா (Schema.org), FOAF போன்றவை குறிப்பான கருத்துருக்களை வெளிப்படுத்த பயன்படக் கூடியன.

நாம் முன்வைக்கும் மெய்ப்பொருளியம் அனைத்துலக அருங்காட்சியகங்கள் மன்றத்தின் கருத்துரு குறிப்பு மாதிரியை (CIDOC Conceptual Reference Model - சி.ஆர்.எம்) அடிப்படையாகக் கொள்ளும். இந்தக் குறிப்பு மாதிரி பண்பாட்டு மரபுரிமை தொடர்பான தகவல்களை கட்டுப்பாடான வழியில் உருவாக்குவதற்கும், பகிர்வதற்கும் பயன்படும் ஒர் அனைத்துலகச் சீர்தரம் (ISO 21127:2014) ஆகும்.[9] இதில் வரையறை செய்யப்படும் வகுப்புகளை உறவுகளை பண்புகளை தேவைக்கேற்ப தேர்ந்து பயன்படுத்த இது அனுமதிக்கிறது. எனினும் எமது பயன்பாடு ஒரு குறிப்பிட்ட முறையில் இந்த குறிப்பு மாதிரியில் இருந்து வேறுபடும். சீர்.ஆர்.எம் இன் நிகழ்வை

மையப்படுத்திய அணுகுமுறையைப் பின்பற்றாது, நடைமுறையில் பயன்பாட்டில் இருக்கும் பொது மெய்ப்பொளியங்களுக்கு நெருக்கமாக அமையும். குறிப்பாக விக்கித்தரவு[10], இசுகீமா (schema.org), பிபிசி மெய்ப்பொருளியங்களுக்கு[11] நெருக்கமாக அமையும்.

பன்மொழியாக்கம் (Multilinquality) மெய்ப்பொருளிய வடிவமைப்பில் ஒரு முக்கிய கூறு ஆகும். தற்போது ஒரு மதிப்பு (value) எந்த மொழியில் உள்ளது என்பதைக் குறிப்பதற்கான முறைகள் உள்ளன. ஆனால் மதிப்புகள் மட்டும் அல்லாமல் வகுப்புகள், பண்புகள் உட்ப்பட்ட முழுமையான சூழலும் பன்மொழிகளில் இருந்தால்தான் ஒரு விடயத்தை வினவி முழுமையான தகவல்களை தமிழிலேயே பெற முடியும். அதற்கு முதற் கட்டமாக ஒரு மெய்ப்பொருளியத்தை வடிவமைத்து மொழிபெயர்க்க வேண்டும்.

பதிப்பு மேலாண்மை (Versioning) மெய்ப்பொருளிய வடிவமைப்பில் கருத்தில் கொள்ள வேண்டிய இன்னுமொரு முக்கிய விடயம் ஆகும். ஒரு மெய்ப்பொருளியம் வடிவமைக்கப்பட்ட பின்னர் அது பல காரணங்களுக்கா இற்றைப்படுத்தப்பட வேண்டி வரும். அதை முறைப்படி பராமரிக்கவும், இற்றைப்படுத்தவும் தேவையான உள்கட்டமைப்பு அவசியம் ஆகும். இந்த விடயங்கள் தமிழ்ச் சூழலுக்குத் தேவையான ஒரு நிலையான மெய்ப்பொருளியம் உருவாக்கப்படும் போது கவனிக்கப்பட வேண்டியவை.

5. மெய்ப்பொருளிய விபரிப்பு

பி.எப்.ஓ (BFO) மற்றும் சி.ஆர்.எம் (CRM) ஆகிய இரண்டும் உலகில் உள்ள அனைத்தையும் நிலையானவை (continuant - persistent), காலச்சார்புடையவை அல்லது நிகழ்பவை (occurants - temporal) என்று வகைப்படுத்துகின்றன.[12] சி.ஆர்.எம் பல படிநிலைக் கட்டமைபைக் (multi level hierarchy) கொண்டது. இதில் இருந்து மாறுபட்டு, நாம் தட்டையான எளிய வடிவமைப்பை முன்வைக்கிறோம். அந்த வகையில் அண்டத்தில் உள்ள அனைத்தையும் (E1 CRM Entity) உருபொருள் (Entity) என்று தொடங்கி, அதன் கீழேயே முதன்மை வகுப்புகளை பிரிக்கிறோம். இந்த தட்டையான அணுகுமுறை இசுகீமா, விக்கித்தரவு, பிபிசி மெய்ப்பொருளிய வடிவமைப்பை ஒத்தது. ஆனால் இந்த வகுப்புகள் சி.ஆர்.எம் வகுப்புகளில் இருந்து அதே பொருளோடு இங்கு பயன்படுத்தப்படுகின்றன.

E73 Information Object:

- E78 Collection
- F1 Work

E28 Conceptual Object
E18 Physical Thing
E21 Person
E74 Group
E53 Place
E5 Event
E52 Time-Span

E73 Information Object/E73 தகவல் பொருள் என்ற உருபொருள் அல்லது வகுப்பினைப் பயன்படுத்தி நூலகம், ஆவணம், அரும்பொருட்கள் மற்றும் மரபுரிமைகள் பற்றிய பதிவுகளைக் குறிக்க நாம் பயன்படுத்த முடியும். இவற்றின் சேகரிப்புக்களையும் படைப்புகளை சிறப்பாக விபரிக்க போர்ட்லன்ட் தரவு மாதிரி (Portland Data Model) பயன்படுகிறது. போர்ட்லன்ட் தரவு மாதிரி சேகரிப்பு (pcdm:Collection), படைப்பு (pcdm:Object), மற்றும் கோப்பு (pcdm:File) ஆகியவற்றை விபரிக்கிறது. ஓவ்வொரு சேகரிப்பையும் படைப்பையும் டப்பிளின் கருவகம் (Dublin Core) போன்ற விபரிப்பு மீதரவுப் பண்புகளைப் பயன்படுத்தி விபரிக்க முடியும்.

E28 Conceptual Object/E28 கருத்துருப் பொருள் முதன்மையாக பொருள் அல்லதான நுண்புலமான விடயங்களைக் குறிக்கப் பயன்படும். வகைப்படுதலில் பயன்படும் தலைப்புகளையும் (Subjects/Topics) இதன் உதவியுடன் உருவகிக்க முடியும். பெரும்பாலும் இத்தகைய வகைப்படுத்தல்கள் அல்லது கலைச்சொற் தொகுப்புகள் எளிய அறிவு ஒழுங்கமைப்பு முறைமையைப் (Simple Knowledge Organization System (SKOS)) பயன்படுத்தி உருவாக்கப்படுகின்றன. ஓர் எடுத்துக்காட்டு அனைத்துலக தசம வகைப்படுத்தல் முறைமையின் (Universal Decimal Classification) இணைப்புத் தரவு வெளியீடு ஆகும்.

E18 Physical Thing/E18 பௌதீகப் பொருள் அண்டத்தில் உள்ள தகவல் அல்லது கருத்துரு அல்லாத எல்லாப் பொருட்களையும் அடங்கியது. இது மேலும், தனிப் பொருளா (E19 Physical Object) அல்லது இன்னொன்றின் கூறா (E26 Physical Feature), மனிதர் உருவாக்கியதா (E22 Man-Made Object, E25 Man-Made Feature) அல்லது இயற்கையானதா (E19 Physical Object, E26 Physical Feature), உயிர் உள்ளதா (E20 Biological Object) என்று கூர்மையாக வரையறை செய்யப்படுகின்றன. E21 Person/E21 நபர், E74 Group/E74 குழு அல்லது அமைப்பு, E53 Place/இடம் ஆகியன E18 பௌதீகப் பொருளின் உப வகுப்புகள்தான். எனினும் நூலக, ஆவணக, அருங்காட்சியகப் படைப்புகளை விபரிக்க, தேட, பயன்படுத்த இவை முக்கியம் ஆகும்.

E18 பௌதீகப் பொருள் என்பதற்குள் வரமால் காலச்சார்புடையவற்றை அல்லது நிகழ்பவையை (occurants - temporal) E5 Event/ நிகழ்வு மற்றும் E52 Time-Span/நேரம்-காலம் வகுப்புகள் கொண்டு விபரிக்க முடியும். E5 Event/E5 நிகழ்வு வரலாற்றின் நிகழ்வுகளையும், ஒரு படைப்புக்கு நிகழக் கூடிய நிகழ்வுகளையும் ஒருங்கேயே குறுக்கிறது. E52 Time-Span/நேரம்-காலம் ஒரு குறிப்பிட்ட நேர பரிமாணத்தைப் பதிவுசெய்ய பயன்படுகிறது (temporal extent). எ.கா 2009, மே 1981, சங்க காலம். ஒரு குறிப்பிட்ட நேரம் அல்லது கணத்தையும் (instant of time), கால இடைவெளியையும் (interval of time) வேறுபடுத்தியே அணுக வேண்டும்.

மேலே, நாம் முன்வைக்கும் மெய்ப்பொளியத்தின் முதன்மை வகுப்புகளை நோக்கினோம். அடுத்து அவைக்கு இடையேயான உறவுகளை நாம் வரையறை செய்ய வேண்டும். இங்கு உறவுகளுக்கும் (Relations) பண்புகளுக்கும் (Properties) இடையே உள்ள வேறுபாட்டை சுட்டுவது பயன்மிக்கது. உறவுகள் பொதுவாக இன்னுமொரு பொருளை சுட்டும். பண்புகள் நிலையுருகளைப் (literal) பயன்படுத்தி அந்தப் பொருளை நேரடியாக விபரிக்கும்.

ஒரு படைப்பை கருத்துரு, இடம், நிகழ்வு அல்லது காலம், அமைப்பு, நபர் போன்றவற்றோடு பின்வரும் உறவுகளைப் பயன்படுத்தி தொடர்புபடுத்த முடியும்: dcterms:subject, dcterms:spatial, dcterms:temporal, schema:organization, schema:person. இதே போன்று ஒரு நபர் dbp:knownFor, dbo:birthPlace, schema:birthDate, schema:memberOf, schema:relatedTo இனை தொடர்புபடுத்த முடியும். இவ்வாறு ஒவ்வொரு வகுப்புக்கும் இடையேயான உறவுகளை வரையறை செய்து செழுமையான இணைப்புத் தரவுக் கட்டமைப்பாக ஒரு மெய்ப்பொருளியத்தை உருவாக்க முடியும். இவ்வாறு உருவாக்கப்பட்ட ஒரு மெய்ப்பொருளியத்தின் வரைவுனை இங்கு காணலாம்: https://goo.gl/CwXs42

6. ஐலண்டோரா குளோவில் (Islandora CLAW) நிறுவுதல்

ஐலண்டோரா (islandora.ca) என்பது எண்ணிம வளங்களை பாதுகாக்க, மேலாண்மை செய்ய, அணுக்கப்படுத்த பயன்படும் ஒரு கட்டற்ற மென்பொருள் தளம் ஆகும். இந்த மென்பொருளை 150 க்கும் மேற்பட்ட பல்கலைக்கழகங்கள், நூலகங்கள், ஆவணகங்கள், அருங்காட்சியகங்கள் பயன்படுத்துகின்றன. இதன் அடுத்த தலைமுறை பதிப்பு ஐலண்டோரா குளோ (Islandora-CLAW), இணைப்புத் தரவுச் செயலிகளுக்கான தளமாக பல நிறுவனங்களால் கூட்டாக கிட்கப்பில் (https://github.com/Islandora-CLAW/CLAW) உருவாக்கப்பட்டு வருகிறது.

ஐலண்டோரா குளோ டுரூப்பல் 8.3 (drupal.org) உள்ளடக்க மேலாண்மை ஒருங்கியம் (Content Management System), ஃபெடோரா 4.7 (fedorarepository.org) இணைப்புத் தரவு தளம் (Linked Data Platform), பிளேசுகிராப் 2.4 (blazegraph.com) மும்மைத் தரவுத்தளம் (Triplestore) ஆகியவற்றினைப் பயன்படுத்தி உருவாக்கப்படுகிறது. டுரூப்பல் ஆர்.டி.எப் (RDF) இக்கு அடிப்படையான ஆதரவினைத் தருகிறது. குளோ டூர்ப்பலில் ஆர்.டி.எப் வளங்கள் (RDF Resources), ஆர்.டி.எப் இல்லாத வளங்கள் (Non RDF Resources) என்ற இரு உள்ளடக்க மாதிரிகளை (Content Types) அமைத்துத் தருகின்றது. இந்த அடிப்படை உள்ளடக்க மாதிரிகளைப் பயன்படுத்தி நாம் எந்தவொரு உள்ளடக்க மாதிரியையும் உருவாக்க முடியும். எ.கா சேகரம், படைப்பு, நபர், அமைப்பு, இடம். இந்த உள்ளடக்க மாதிரிகளில் எமக்குத் தேவையான RDF Mapping ஐ செய்ய முடியும். இவ்வாறு RDF Mapping செய்வதன் ஊடாக டுரூப்பல் உள்ளடக்கங்களை இணைப்புத் தரவாக வெளியிட முடியும். ஐலண்டோரா குளோ இவ்வாறு வெளியிடப்படும் ஆர்.டி.எப் மும்மைகளை ஃபெடோராவில் சேமிக்கிறது. மேலும், அவற்றை பிளேசுகிராப் மும்மைத் தரவுத்தளத்தில் குறியீடு செய்து (index) தருகின்றது. இந்த மும்மைத் தரவுத்தளத்தை எசுபார்க்கிள் (SPARQL) முனை ஊடாக வினவ முடியும். ஆர்.டி.எப் ஆக வெளியிடப்படும் தகவல்களையும் யாரும் அணுக முடியும், சுட்ட முடியும், தமது மும்மைத் தரவுத்தளத்தில் சேமிக்க முடியும்.

7. உரையாடலும் முடிவுகளும்

தமிழ்ச் சூழலில் நினைவு நிறுவனங்கள் தமது தரவுகளை ஒரு பொது மெய்ப்பொருளியத்தைப் பயன்படுத்தி இணைப்புத் தரவாக வெளியிடும் தேவை உள்ளது. இந்த ஆய்வு அனைத்துலக அருங்காட்சியகங்கள் மன்றத்தின் கருத்துரு குறிப்பு மாதிரி (CIDOC Conceptual Reference Model - சி.ஆர்.எம்) அடிப்படையிலானா எளிமைப்படுத்தப்பட்ட மெய்ப்பொருளியம் ஒன்றை விபரித்துள்ளது. மேலும், அதை நிறுவக்கூடிய ஒரு கட்டற்ற நுட்பக் கட்டமைப்பு ஒன்றையும் முன்வைத்துள்ளது. தற்போது மேல்நிலையில் விபரிக்கப்பட்டுள்ள இந்த மெய்ப்பொருளியத்தை துல்லியமாக வரையறை செய்துகொள்வது அடுத்த கட்டப் பணியாக அமையும். அந்தச் சந்தர்ப்பத்தில் பன்மொழியாக்கம் தொடர்பாக சிக்கல்களும் கூடிதலாக ஆயப்படவேண்டும்.

8. மேற்கோள்கள்

[1] Fenández, J., & Kiesling, E. et al. (2016) Propelling the Potential of Enterprise Linked Data in Austria (Tech.). Leobersdorf: Edition mono/monochrom. doi:https://www.linked-data.at/wp-content/uploads/2016/12/propel_book_web.pdf

[2] Fenández, J., & Kiesling, E. et al. (2016) Propelling the Potential of Enterprise Linked Data in Austria (Tech.). Leobersdorf: Edition mono/monochrom. doi:https://www.linked-data.at/wpcontent/uploads/2016/12/propel_book_web.pdf

[3] What is a Vocabulary? (2015). W3C. Retrieved July 15, 2017, from https://www.w3.org/standards/semanticweb/ontology

[4] Holland, S. V., & Verborgh, R. (2016). 2 Modelling. In Linked Data for Libraries, Archives and Museums. London: Facet Publishing. doi:http://book.freeyourmetadata.org/chapters/1/modelling.pdf

[5] Tim Berners-Lee (2006-07-27). "Linked Data". Design Issues. W3C. Retrieved July 15, 2017 from https://www.w3.org/DesignIssues/LinkedData.html

[6] Diallo, Papa Fary, et al. "Ontologies-Based Platform for Sociocultural Knowledge Management." Journal on Data Semantics 5.3 (2016): 117-139. Doi: https://hal.inria.fr/hal-01342912/document

[7] Bontas, E. P., Mochol, M., & Tolksdorf, R. (2005, June). Case studies on ontology reuse. In Proc. of the 5th International Conference on Knowledge Management. Retrieved July 15, 2017, from https://pdfs.semanticscholar.org/6c29/58820daf02afeb29d92765cb1cfa2e8c459f.pdf

[8] Bermejo, J. (2007). A simplified guide to create an ontology. Madrid University. Retrieved July 15, 2017, from https://www.researchgate.net/profile/MJ_Healy/publication/2405846_Ontology_Reuse_and_Application/links/5452af590cf2cf51647a48b0.pdf

[9] ISO - International Organization for Standardization. (2014, October 16). Retrieved July 15, 2017, from https://www.iso.org/standard/57832.html

[10] Wikidata:List of properties. (n.d.). Retrieved July 15, 2017, from https://www.wikidata.org/wiki/Wikidata:List_of_properties

[11] Ontologies - Core Concepts Ontology. (n.d.). Retrieved July 15, 2017, from http://www.bbc.co.uk/ontologies/coreconcepts

[12] Kurtz, D. (2014, July 01). The CIDOC Conceptual Reference Model (CIDOC-CRM): PRIMER. Retrieved July 16, 2017, from http://www.cidoc-crm.org/Resources/the-cidoc-conceptual-reference-model-cidoc-crm-primer

CROSS LINGUAL PERSONALIZED TRAVEL RECOMMENDATION USING LOCATION BASED SOCIAL NETWORKS

R.Kalaiselvan[1] Dr.T.Mala[2] Shri Vindhya[3]

[1]PG Student, Department of Information Science and Technology,
Anna University, Chennai, Tamil Nadu; rkalaiselvan6@hotmail.com

[2]Associate Professor, Department of Information Science and Technology,
Anna University, Chennai, Tamil Nadu: malanehru@annauniv.edu

[3] Research Scholar, Department of Information Science and Technology,
Anna University, Chennai, Tamil Nadu; space.safia@gmail.com

ABSTRACT

The Social Networks are becoming one of the modern platforms to connect with other people to share the information such as texts, video, audio, images, etc. There is a special type social network available which allows us to share the locations in the form of check-ins, geotags to others called Location Based Social Networks that uses geolocation attributes for sharing. The proposed approach makes use of available large amount of social data from both travel based services (such as Foursquare) and community contributed photos with added heterogeneous data (such as Flickr), to provide efficient recommendation for single POIs (Point Of Interest) and for the Travel Sequence based on the user interests. Thus by using the unstructured social media data like images, geo-tags, check-ins with advanced machine learning algorithms are used to provide efficient recommendations. This proposed work in this paper follows sequential approach to provide the travel recommendation to the users, namely, collection of data from multiple sources and preprocessing those data, selecting venue features to provide recommendations, applying venue recommendation algorithm which predicts venues similar to user queries by using Content Based recommendation algorithm, to analyze the sentiment on each locations by using NLTK packages which further used to provide recommendations in Tamil language and to provide travel sequence to the users by predicting set of places as a travel sequence to the interacting people.

1. Introduction:

Personalization in LBSNs allows users to modify their searches as needed and recommendations has to be provides as per the user's interests. The proposed system takes the input as the location name which has to be geo coded to obtain the geo location information for the location such as Latitude and Longitude. These geo location information helps us to collect the data belong to those locations using APIs from social network service providers. These Application Programming Interfaces allows us to use the

given input and acts like medium to crawl data from the service providers in the major formats like JSON and XML. Location Based Social Networks (LBSN) like Foursquare and Photo sharing services like Flickr allows users to collect their data by obtaining the credentials for accessing the publicly shared information on their services. The data collected are in the format of JSON. The data collected using the API contains noises such as missing data, symbols and short words, emoticons, data with no geotags and irrelevant data to the query that are preprocessed and preprocessed data is stored in the database.

2. Methodology

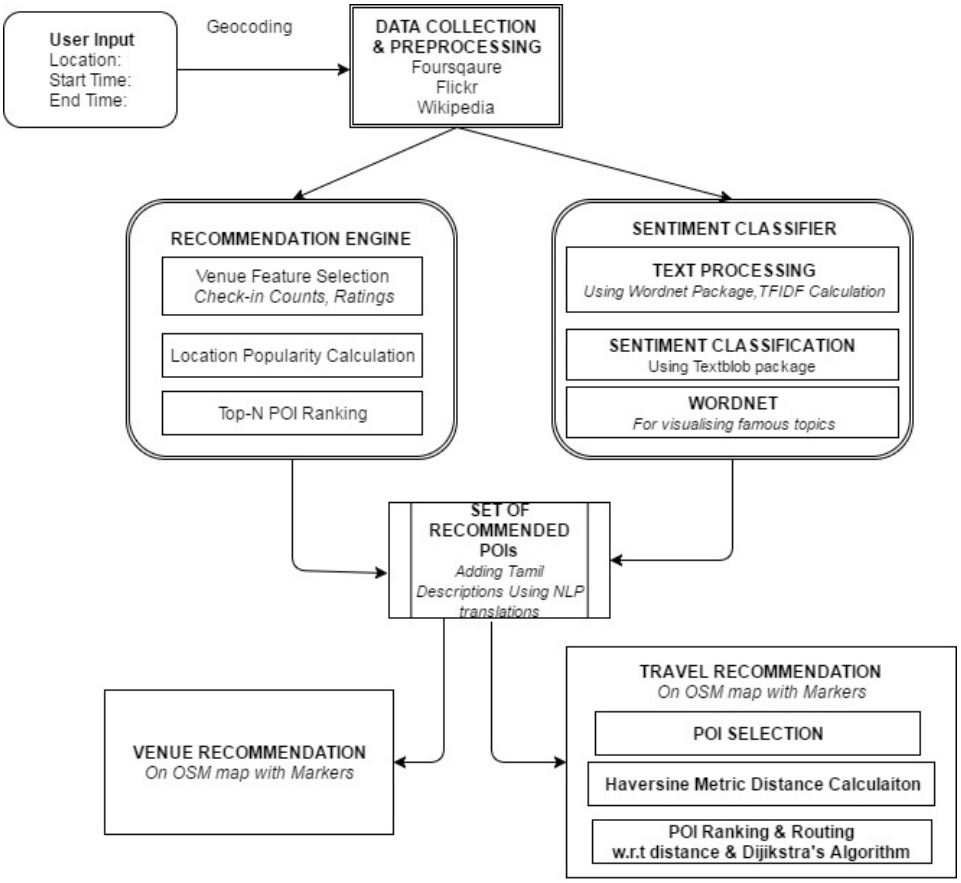
The proposed system mainly consists of two parts such as Sentiment analysis and Travel recommendation. The content based recommendation engine uses venue features like check-in counts, ratings to provide the recommendations from which the Top-N places are recommended. The proposed system predicts the set of POIs which the users might like using the venue features and popularity ranking methodology approach to sort predicted places based on maximum ratings to each locations. This work allows users to select for both the venue recommendation as well as the travel recommendation. Venue recommendations are made based upon category and places of interest by the users; the proposed system allows users to select categories like Food, Entertainment, Drinks, Cafe, Outdoor Activities and Venues to obtain better personalized recommendations.

3. System Architecture

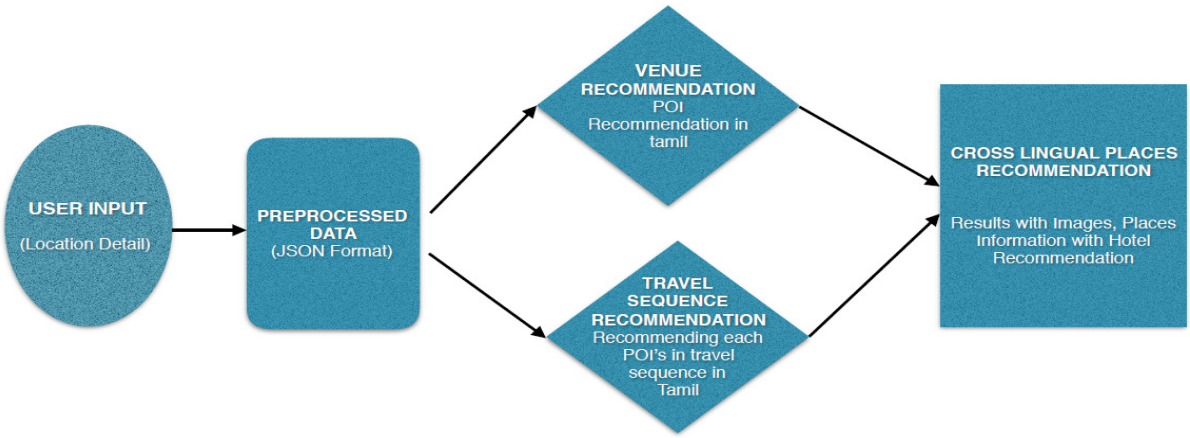
The detailed description of proposed travel recommendation process has been given as follows.

Travel recommendation methodology takes series of POIs and sorts them based on the distances to each location by using the Haversine metric to compute distances using Great Circle distance formula. These distances are used to select the POIs and to create the Travel Sequence for Top-N recommendations. This travel sequence recommendation consists of attributes like next POI to be visited, estimated travel time to the next venue and venue features like check-in counts and ratings of each POIs. Sentiment analysis is done on each venues based on the semantic data obtained from the Flickr and Foursquare tips and applied to NLP techniques to preprocess the unwanted data, then the sentiment is checked using the popular machine learning library Textblob. The sentiments on each location are displayed to the users where user can sense the opinion of each venue by previously visited users. Finally, the proposed work provides the recommendations in the form of Maps on which each location information is pointed with a marker, the location descriptions are provided in both the Tamil and English. Tamil description for each places are added by applying NLP techniques which involves translation on famous tips and description to the location.

Architecture of Travel Recommendation System



ARCHITECTURE OF TRAVEL RECOMMENDATION SYSTEM



Architecture of Tamil Descriptions to the Recommendations

4. Conclusion

Including Tamil Description to the Recommendation

Tamil Description is added to all the recommendation including particular Point Of Interest recommendation as well as Travel Sequence Recommendation with use of Google

Translate and Wikipedia. Tamil description to each places added if there is any description available for given locations.

This inclusion of Tamil Description can provide better readability to Native Tamil Users. Only specification of places will be given with using Tamil Descriptions and Packages will be containing Alpha-numeric characters.

5. REFERENCES

[1] H. Liu, T. Mei, J. Luo, H. Li, and S. Li, "Finding perfect rendezvous on the go: Accurate mobile visual localization and its applications to routing," in Proc. 20th ACM Int. Conf. Multimedia, 2012, pp. 9–18.

[2] J. Bao, Y. Zheng, and M. F. Mokbel, ―Location-based and preference aware recommendation using sparse geo-social networking data,‖ in Proc. 20th Int. Conf. Adv. Geographic Inf. Syst., 2012, pp. 199–208.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, ―Latent Dirichlet allocation,‖J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.

[4] A. Cheng, Y. Chen, Y. Huang, W. Hsu, and H. Liao, ―Personalized travel recommendation by mining people attributes from community contributed photos,‖ in Proc. 19th ACM Int. Conf. Multimedia, 2011,pp. 83–92.

[5] M. Clements, P. Serdyukov, A. de Vries, and M. Reinders, ―Using flickr geo-tags to predict user travel behaviour, in Proc. 33rd Int. ACMSIGIR Conf. Res. Develop. Inf. Retrieval, 2010, pp. 851– 852

[6] Y. Gao, J. Tang, R. Hong, Q. Dai, T. Chua, and R. Jain, ―W2go: A travel guidance system by automatic landmark ranking, in Proc. Int.Conf. Multimedia, 2010, pp. 123–132.

[7] J. Li, X. Qian, Y. Y. Tang, L. Yang, and T. Mei, "GPS estimation for places of interest from social users' uploaded photos," IEEE Trans. Multimedia, vol. 15, no. 8, pp. 2058–2071, Dec. 2013.

தமிழ் APIகள் மூலம் இணைப்பில் இல்லா இணையதளங்களை (Offline Websites) தமிழில் உருவாக்குதல், அவற்றின் முக்கியத்துவம் மற்றும் பயன்கள் - ஓர் ஆய்வு

ரா.சு. சிவ சுப்பிரமணியம், ப. செந்தில் குமார்

கட்டுரைச் சுருக்கம்

இந்த ஆய்வுக்கட்டுரை இணையத்தில் இருக்கக்கூடிய தமிழ் இலக்கியங்களை செயலி நிரலி இடைமுகம் (API) மூலம் எவ்வாறு ஒருங்கிணைப்பது என்றும், அவ்வாறு ஒருங்கிணைக்கப்பட்ட உள்ளடக்கங்களை இணைய இணைப்பு இல்லாவிடிலும் பார்க்கக்கூடிய இணையதளமாக (முன்னேற்ற இணைய செயலி - Progressive Web App) வடிவமைப்பது எவ்வாறு என்றும் விளக்குகிறது. இந்த ஆய்வுக்கட்டுரைக்கென குறுந்தொகைக்கான செயலி நிரலி இடைமுகம் (API) ஒன்று வடிவமைக்கப்பட்டுள்ளது. இந்த குறுந்தொகை API உதவியுடன் இணைய இணைப்பில் இல்லாவிடிலும் இயங்கக்கூடிய இணையதளம் (முன்னேற்ற இணைய செயலி) ஒன்றும் வடிவமைக்கப்பட்டுள்ளது. தமிழ் இலக்கிய உள்ளடக்கங்களை கணினியில் மாங்கோ தரவுத்தளம் (database) மூலம் ஒழுங்குபடுத்தி சேமித்து, அவ்வாறு சேமித்த தரவுகளை ஜாவா நிரலி உதவியுடன் ஒரு செயலி நிரலி இடைமுகமாக உருவாக்கி உள்ளோம். அவ்வாறு உருவாக்கப்பட்ட API யை கொண்டு ஒரு இணையதளத்தை உருவாக்கி உள்ளோம். உலாவியில் (browser) உள்ள சேவை பணியாள் (Service Worker) மற்றும் சேமிப்பகம் (Local Storage) உதவியுடன் இணையதளம் இயங்குவதற்கு தேவையான உள்ளடக்கங்கள் பதுக்ககத்தில் (cache) சேமிக்கப்படுகின்றன. மறுமுறை பயனர் இந்த இணையதளத்தைப் பார்க்கும் பொழுது பயனரிடம் இணைய இணைப்பு இல்லை என்றால் அவரின் உலாவி பதுக்ககத்தில் இருந்து பயனருக்கு இணையதளம் இயங்குவதற்கான உள்ளடக்கங்களைக் கொடுக்கிறது. இணைப்பில் இல்லா இணையதளம் என்பது தமிழ் API களின் ஒரு பயன்பாடே ஆகும். இத்தகைய API களை கொண்டு ஒரு நிரலர் எந்த வகையான இணையதளமாகவோ அல்லது ஏற்கனவே உள்ள இணையதளத்தில் ஒரு நிரல் பலகையாகவோ (Wizard) வடிவமைக்க முடியும். தமிழில் தற்போது உள்ள Project Madurai போன்ற இணையதளங்களை கொண்டு இது சாத்தியப்படாது. மேலும் இணைப்பில் இல்லா இணையதளங்கள் ஆன்ட்ராய்ட் போன்ற செயலிகளைப் போல சேமிப்பகத்தை எடுத்துக் கொள்ளாததால், பயனருக்கும் இந்த தொழில்நுட்பம் உதவிகரமானதாக இருக்கும்.

முன்னுரை

தற்போது இணையத்தில் தமிழ் இலக்கியங்களை அணுகுவதில் பின்வரும் சவால்கள் உள்ளன:

1) தமிழ் இலக்கிய படைப்புகளை ஒரே இடத்தில் அணுக முடிவதில்லை.

தமிழ் இலக்கிய படைப்புகள் Project Madurai, தமிழ் விக்கி போன்ற இணையதளங்களில் பல்வேறு வடிவங்களில் கிடைக்கின்றன. உதாரணமாக குறுந்தொகையை எடுத்துக் கொள்வோம். இதை Project Madurai இணையதளம் சென்றால் பி.டி.எப் அல்லது HTML வடிவத்தில் மட்டுமே அணுக முடியும். குறுந்தொகையில் பாடல்களுக்கான பொருள் வேண்டும் என்றால் தமிழ் விக்கி தளத்தை அணுக வேண்டி இருக்கிறது. இவ்வாறு நமக்கு தேவையான விவரங்களை அணுக வெவ்வேறு தளங்கள் செல்ல வேண்டி இருக்கிறது

2) தமிழ் இலக்கிய படைப்புகளை நமக்கு விரும்பிய வடிவத்தில் அணுக முடிவதில்லை

இணையத்தில் கிடைக்கும் தமிழ் இலக்கிய படைப்புகளை அந்த தளத்தில் கொடுக்கப்பட்டுள்ள வடிவத்தில் மட்டுமே அணுக முடிகிறது. உதாரணமாக குறுந்தொகைக்கென ஒரு நிரலர் ஒரு இணையதளத்தை உருவாக்க விரும்பினால், மேலே குறிப்பிட்டுள்ள எதோவொரு இணையதளத்திற்கு சென்று, தேவையான விவரங்களை படி எடுத்து, தன்னுடைய இணையதளத்தை உருவாக்கலாம். ஆனால் இதில் மனித ஆற்றல் அதிகம் செலவாகும். மேலும் இது விரிவாக்கக் கூடிய (scalable) தீர்வும் அல்ல. நிரலி மூலமாக தேவையான விவரங்களை எடுப்பதே தீர்வாகும். (திருக்குறள் போன்ற வெகு சில APIகள் தவிர தமிழில் வேறு APIகள் கிடையாது.) எனவே மேலே குறிப்பிட்டுள்ள இணையதளங்களில் நிரலி மூலமாக விவரங்களை சேகரிப்பது மிக மிக கடினமாக காரியம்.

3) இலக்கிய படைப்புகளைப் பற்றி ஆராய்வது கடினமாக உள்ளது.

உதாராணமாக தென் தமிழ்நாட்டின் சில பகுதிகளில் 'பைய' என்கிற சொல்லை, ‘மெதுவாக’ என்ற பொருளில் இன்றும் பயன்படுத்துகிறார்கள். இதே சொல் 2000 வருடங்களுக்கு முன் இயற்றபட்ட திருக்குறளிலும் வருகிறது. இந்த சொல் சங்க இலக்கியங்களில் வேறெந்த இடத்தில் பயன்படுத்தப்படுகிறது என ஒருவர் ஆராய விரும்பினால் தற்போது ஒவ்வொரு தளமாக சென்று அங்குள்ள படைப்புகளில் தேட வேண்டும். இது நேர விரயம். விரிவாக்கவும் இயலாது. தற்போது உள்ள இணையதளங்களில் நிரலி கொண்டு இவ்வாறு ஆராய முடியாது.

4) இணைய இணைப்பு இல்லாவிட்டாலோ, வேகம் குறைவாக இருக்கும் போதோ தமிழ் இலக்கிய படைப்புகளை அணுக முடிவதில்லை. சமீபத்தில் அகமாய் நிறுவனம் நடத்திய ஆய்வு ஒன்றில், இந்தியாவின் சராசரி இணைய வேகம் 4.5Mbps என குறிப்பிடப்பட்டுள்ளது. இந்தியா உலக அளவில் இணைய வேகத்தில் 105 ஆவது இடத்தில் உள்ளது. மேலும், இந்தியாவில் கைப்பேசி வைத்திருப்பவர் எல்லா நேரங்களிலும் இணையத்தோடு தொடர்பில் இருப்பதில்லை. கைப்பேசி சமிக்ஞைகளில் பிரச்சனை, விலையேறி வரும் இணைய சிப்பம் (Internet package) போன்ற பல காரணங்கள் உள்ளன. இம்மாதிரியான சமயங்களில் அல்லது இணைய வேகம் மிக குறைவாக இருக்கும் சமயங்களில் ஒருவர் தமிழ் இணையதளங்கள் மற்றும் அவற்றின் உள்ளடக்கங்களை அணுகுவதில் சிக்கல் ஏற்படுகிறது.

5) தமிழ் இலக்கியத்திற்கென இருக்கும் ஆண்ட்ராய்ட் செயலிகளில் உள்ள சிக்கல்கள்
ஒருவர் ஆண்ட்ராய்ட் கைப்பேசி பயனராக இருக்கும்பட்சத்தில் ஏற்கனவே பதிவிறக்கம் செய்யப்பட்ட செயலிகள் மூலம் அவர் தமிழ் உள்ளடக்கங்களை இணைய வசதி இல்லாத நேரத்திலும் பயன்படுத்த முடியும். உதாரணம்: பாரதியார் பாடல்கள் ஆண்ட்ராய்ட் செயலி. ஆனால், இது போன்ற செயலிகளின் சிக்கல் என்னவென்றால் இவை கைப்பேசிகளில் இடத்தை எடுத்துக் கொள்கின்றன. சராசரி இந்தியர்கள் பயன்படுத்தும் ஆண்ட்ராய்ட் கைப்பேசிகளில் நிறைய செயலிகளை தரவிறக்கம் செய்து கொள்வதற்கான வசதிகள் இல்லை. ஓர் அளவுக்கு மேல் செயலிகளை தரவிறக்கம் செய்து பயன்படுத்தும் போது கைப்பேசியின் செயல் திறன் குறைவதை ஒருவர் அறியலாம்.

தமிழ் செயலி நிரலி இடைமுகம் (API), மற்றும் அவற்றைப் பயன்படுத்தி இணைப்பில் இல்லா இணையதளத்தை உருவாக்கி மேலே குறிப்பிட்டுள்ள சவால்களை எவ்வாறு தீர்க்கலாம் என இந்த ஆய்வுக்கட்டுரையில் பார்க்கலாம்.

முன்மொழியும் தீர்வு

முதலில் தமிழ் இலக்கியங்களை எதேனும் ஒரு தரவுத்தளத்தில் (database) சேமிக்க வேண்டும். அவ்வாறு சேமித்த தரவுகளை ஜாவா போன்ற நிரலியின் உதவியுடன் செயலி நிரலி இடைமுகம் (API) ஏற்படுத்த வேண்டும். இந்த APIயை பயன்படுத்தி இணைப்பில் இல்லா இணையதளம் ஒன்றை உருவாக்க வேண்டும்.

தமிழ் இலக்கிய படைப்புகளுக்கான தரவுத்தளம்

தமிழ் இலக்கிய படைப்புகளை எங்கிருந்தாலும் அணுகுவதற்கான முதல் படி அனைத்து இலக்கியங்களுக்கும் ஒரு தரவுத்தளம் (database) அமைப்பதாகும். இதற்கு ஆரக்கிள், எஸ்.க்யூ.எல் போன்ற தீர்வுகள் இருக்கின்றன. ஆனால் ஒரே மாதிரியான கட்டமைப்பு உள்ள தரவுகளை மட்டுமே இவற்றில் சேமிக்க முடியும். ஆனால் தமிழ் இலக்கியங்கள் அனைத்தும் ஒரே கட்டமைப்பு உடையவை அல்ல. உதாரணமாக திருக்குறள் மூன்று வகையான பால்களைக் கொண்டு, 133 அதிகாரங்களுடன் 1330 பாக்கள் உள்ளன. ஆனால் குறுந்தொகையோ 5 வகையான நிலங்களை உள்ளடக்கிய பாடல்களைக் கொண்டுள்ளது. மேலும் குறுந்தொகையின் ஒவ்வொரு பாடலுக்கும் ஒரு ஆசிரியர் உண்டு, ஆனால் குறளுக்கோ ஒரே ஆசிரியர் தான். இதே போல வெவ்வேறு இலக்கிய படைப்புகளும் வெவ்வேறு வடிவங்களைக் கொண்டுள்ளது. எனவே அவற்றை ஆரக்கிள், எஸ்.க்யூ.எல் போன்றவற்றில் சேமிக்க முடியாது. இதற்கு நாம் வடிவமற்ற தரவுகளை சேமிக்கக் கூடிய மாங்கோ, கசாண்ட்ரா போன்ற தரவு தளங்களை பயன்படுத்தலாம். இந்த ஆய்வுக்கட்டுரையில் நாங்கள் மாங்கோ தரவுத்தளத்தைப் பரிந்துரைக்கிறோம்.

குறுந்தொகைக்கான மாங்கோ வடிவத்தை பின்வருமாறு அமைத்துள்ளோம். இதே போல் மற்ற படைப்புகளுக்கும் அமைக்க முடியும்.

```
{
    id: <பாடலின் தனித்துவ எண்>,
    name: <பாடலின் பெயர்>,
    content: <குறுந்தொகை மூல பாடல்>,
    explanation: <பாடலின் விளக்கம்>,
    explanationBy: <பாடல் விளக்கம் எழுதியவர்>,
    situation: <பாடல் இடம்பெற்ற தருணம்>,
    genre: <பாடலின் வகைமை>,
    author: <பாடலின் ஆசிரியர்>,
    tags: [
    {
        id: <வகைகளின் தனித்துவ எண்>,
        tagName: <திணை அல்லது மற்ற வகை>,
        tagValue: <திணை>,
    }
```

]
}

செயலி நிரலி இடைமுகம்

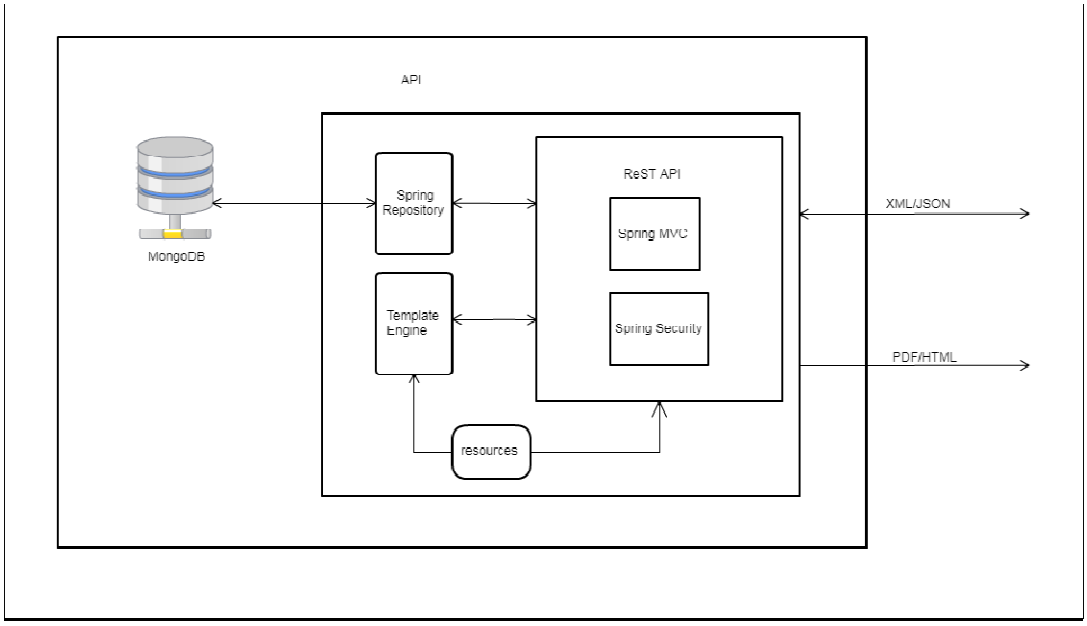
மேலே குறிப்பிட்டபடி நம்மிடம் இப்போது தமிழ் இலக்கிய படைப்புகள் மாங்கோ தரவுத்தளத்தில் உள்ளது. மாங்கோ தரவுத்தளத்தில் இருந்து எந்த நிரலி மொழியையும் பயன்படுத்தி ஒரு API உருவாக்கலாம். எனினும் நாங்கள் ஜாவா மொழியைத் தேர்ந்தெடுத்துள்ளோம்.

முதலில் இந்த APIயின் வடிவத்தைப் பற்றி பார்க்கலாம்.

<இணையதளமுகவரி>/குறுந்தொகை/<பாடல் எண்>

உதாரணமாக குறுந்தொகையின் 14 ஆவது பாடல் நமக்கு தேவைப்படுகிறது என்றால், பின்வரும் சுட்டியைப் பயன்படுத்தலாம்.

https://hidden-reef-62795.herokuapp.com/public/item/குறுந்தொகை/14.json



இந்த சுட்டியில் இருந்து நமக்கு கிடைக்கும் தரவில், குறுந்தொகையின் 14 ஆவது பாடல், அதை எழுதிய ஆசிரியர், அந்த பாடலுக்கான பொருள் JSON வடிவில் கிடைக்கிறது. இதே போல மற்ற பாடல்களையும் பெறலாம். இதை ஒரு நிரலர் தான் விரும்பிய வடிவத்தில் இணையதளமாகவோ அல்லது நிரல் பலகையாகவோ பயன்படுத்த முடியும்.

மேலே குறிப்பிட்ட API ஆனது Java, Spring, JPA, Hiberate உதவியுடன் உருவாக்கப்பட்டது. இந்த தீர்வைப் பின்வருமாறு விளக்கலாம். முன்னுரையில் குறிப்பிட்ட சில சவால்களுக்கு இந்த API எவ்வாறு தீர்வாகும் என பார்க்கலாம்.

1) தமிழ் இலக்கிய படைப்புகள் ஒரே இடத்தில் அணுக முடிவதில்லை.
நாங்கள் இப்போது குறுந்தொகைக்கு மட்டுமே API உருவாக்கி உள்ளோம். இதே போல மற்ற இலக்கிய படைப்புகளுக்கும் இதே இணைய முகவரியில், ஒரே தரவுதளத்தைப் பயன்படுத்தி API தயார் செய்யலாம். அவ்வாறு செய்யும் பொழுது உலகில் உள்ள எந்த ஒரு நிரலரும் இந்த சுட்டியைப் பயன்படுத்திக் கொள்ளலாம்.

2) தமிழ் இலக்கிய படைப்புகளை நமக்கு விரும்பிய வடிவத்தில் அணுக முடிவதில்லை.
இது JSON வடிவமாக மட்டுமல்லாமல், பிடிஎப் (PDF), எக்ஸ்.எம்.எல் (XML) ஆகவோ அல்லது html வடிவமாகவோ பெற்றுக்கொள்ளலாம். உதாரணம்: https://hidden-reef-62795.herokuapp.com/public/item/குறுந்தொகை/14.pdf
https://hidden-reef-62795.herokuapp.com/public/item/குறுந்தொகை/14.xml
https://hidden-reef-62795.herokuapp.com/public/item/குறுந்தொகை/14.html

மேலே குறிப்பிட்டபடி ஒரு நிரலர் இந்த API பயன்படுத்தி அவருடைய விருப்பத்திற்கு ஏற்றவாறு இணையதளமோ அல்லது கைப்பேசி செயலியோ தயாரிக்க முடியும்.

3) இலக்கிய படைப்புகளைப் பற்றி ஆராய்வது கடினமாக உள்ளது.
API என்பது நெகிழ்வான தன்மை உடையது. உதாரணமாக குறுந்தொகையில் குறிஞ்சி திணைப்பாடல்கள் அனைத்தும் பெறுவதற்கு பின்வருமாறு ஒரு சுட்டி உருவாக்கலாம் https://hidden-reef-62795.herokuapp.com/public/item/குறுந்தொகை?திணை=குறிஞ்சி. இதே போல பைய எனும் ஒரு வார்த்தை உள்ள பாடல்கள் அனைத்தும் பெற வேண்டுமானால் https://hidden-reef-62795.herokuapp.com/public/item/குறுந்தொகை?சொல்=பைய என்ற சுட்டியை பயன்படுத்தலாம். மேலே குறிப்பிட்டவை ஒரு உதாரணம் மட்டுமே. இதே போல நமக்கு தேவையான வகையில் தரவுகளை அணுக நிரலி எழுத முடியும்

இணைப்பில் இல்லா இணையதளம்

தற்போது நம்மிடம் மாங்கோ தரவுதளத்தில் இலக்கிய படைப்புகள் உள்ளன. ஜாவா நிரலியின் உதவியோடு இந்த மாங்கோ தரவுத்தளத்தில் இருந்து API மூலமாக தரவைப்

பெறும் சுட்டியும் உள்ளது. இதைப் பயன்படுத்தி இணைப்பில் இல்லாமலும் இயங்கக் கூடிய இணையதளம் ஒன்று உருவாக்கி உள்ளோம் - https://kurunthogai.herokuapp.com இந்த இணையதளத்தை பயனர் ஒரு முறை இணையதள உதவியோடு பார்வையிட்டால் மறுமுறை இந்த தளத்தை பார்வையிடும் போது இணைய இணைப்பு இல்லாவிடிலும் இந்த தளம் வேலை செய்யும்.

தற்போது நாம் பயன்படுத்தும் குரோம், ஃபயர்பாக்ஸ் போன்ற உலவிகளில் சேவை பணியாள் (Service Worker) என்று ஒரு அம்சம் உள்ளது. இது தான் இணையதளங்கள் இணைப்பில் இல்லாமலும் வேலை செய்ய வழி வகை செய்கிறது. இதற்கு முதலில் இந்த வசதி வேண்டுகிற எந்த ஒரு இணையதளமும் உலவியின் சேவை வேலையாளிடம் தனது தளத்தின் எந்தெந்த உள்ளடக்கங்களை பதுக்ககத்தில் (cache) சேமித்து வைக்க வேண்டும் என்று பதிய வேண்டும். (இதற்காக முதல் முறையேனும் இந்த தளத்தை இணைய வசதியோடு பார்வையிட வேண்டும்) பிறகு பயனர் அதே தளத்தை இணைய வசதி இல்லாமல் பார்வையிடும் போது உலவியின் சேவை பணியாள் தனது பதுக்ககத்தில் இருந்து உள்ளடக்கங்களை எடுத்துக் கொடுக்கும்.

குறுந்தொகைக்கான இணையதளம் ReactJS, Redux, Webpack தொழில்நுட்பங்களோடு உருவாக்கப்பட்டுள்ளது. இந்த இணையதளம் சேவை வேலையாளிடம் அனைத்து HTML, CSS, JS கோப்புகளையும் தனது பதுக்ககத்தில் சேமிக்குமாறு கேட்டுக்கொள்கிறது. இந்த HTML, JS, CSS கோப்புகள் அனைத்தும் இணையதளத்தின் வார்ப்புருவாக (template) செயல்படும். ஆனால் API யை இவ்வாறு பதுக்ககத்தில் சேமிக்க முடியாது. எனவே இதற்காக உலவியில் இடம்சார்ந்த சேமிப்பகத்தில் (Local Storage) API இல் இருந்து வரும் தரவை சேமித்து வைக்கிறோம். எனவே இணைய இணைப்பு இல்லாத போது தளத்தின் வார்ப்புருவை (template) சேவை பணியாள் அளிக்கும். குறுந்தொகை பாடலை இடம்சார்ந்த சேமிப்பகத்தில் இருந்து எடுத்துக் கொள்கிறோம்.

இப்போது இந்த இணைப்பில் இல்லா இணையதளம் தீர்க்கும் சவால்களை பார்க்கலாம்

1) மிகக் குறைந்த இணையதள வேகம் அல்லது இணைப்பில் இல்லாவிட்டால் எவ்விதமான படைப்புகளையும் தற்போது அணுக முடியவில்லை
இணைப்பில் இல்லா இணையதளம் மூலம் பயனர் இணைப்பில் இல்லாவிட்டாலும், சேவை பணியாள் மற்றும் உலவியின் இடம்சார்ந்த சேமிப்பகம் உதவியுடன் உருவாக்கிய குறுந்தொகை தளத்தைப் பார்த்தோம்.

2) தமிழ் இலக்கியத்திற்கென இருக்கும் ஆண்ட்ராய்ட் செயலிகளில் உள்ள சிக்கல்கள்
இந்த தளத்தைப் பார்ப்பதற்கு பயனரின் கைப்பேசியில் ஏற்கனவே உள்ள குரோம் போன்ற உலவி ஒன்றே போதுமானது. இதனால் மேலும் பல செயலிகளை கைப்பேசியில் தரவிறக்கத் தேவையில்லை.

முடிவுரை

இந்தத் தீர்வை விரிவுபடுத்துதல் மற்றும் முன் உள்ள சவால்களை நாம் காணலாம்.

இப்போது நாங்கள் குறுந்தொகைக்கு மட்டுமே API மற்றும் தளத்தை உருவாக்கி உள்ளோம். இதே போல மற்ற இலக்கிய படைப்புகளுக்கும் இந்த தீர்வை விரிவாக்க வேண்டும். அப்போது ஏற்கனவே கூறியபடி, இலக்கிய படைப்புகளை மாங்கோ தரவுத்தளத்தில் சேமிக்க மனித ஆற்றல் செலவாகும். நிரலி மூலம் சேமிப்பது மிக கடினமான பணி. ஆனால் ஒருமுறை மாங்கோவில் நாம் சேமித்து விட்டால், அதன் பிறகு மனித ஆற்றல் வீணாகாது. நமக்கு தேவையான தரவுகளை நிரலி மூலம் பெற்றுக்கொள்ளலாம்.

மேலும் இணைப்பில் இல்லா தளத்தை வடிவமைக்கும் போது நாம் இடம்சார்ந்த சேமிப்பகத்தில் (Local Storage) குறுந்தொகை பாடலை சேமிப்பதாக கூறினோம். உலவியில் இருக்கும் இடம்சார்ந்த சேமிப்பகத்தின் அதிகபட்ச கொள்ளளவு 5 எம்.பி மட்டுமே. எனவே நமக்கு தேவையான மொத்த குறுந்தொகை பாடல்களின் மொத்த அளவு 5 எம்.பி யைத் தாண்டுமானால் நம்மால் எல்லா பாடல்களையும் சேமிக்க முடியாது. இதற்கு நாம் வேறு பல உத்திகளைக் கையாள வேண்டி வரும். உதாரணமாக பயனரிடம் அடுத்த முறை இணைய இணைப்பு கிடைக்கும் போது நாம் ஏற்கனவே சேமித்த தகவல்களை அப்புறப்படுத்தி விட்டு புதிய பாடல்களை சேமித்து வைக்கலாம்.

மேற்கோள்கள்

[1] https://developers.google.com/web/progressive-web-apps/

[2] https://developer.mozilla.org/en-US/docs/Web/API/Service_Worker_API/_Using_Service_Workers

[3] https://projects.spring.io/spring-framework/

[4] https://www.html5rocks.com/en/tutorials/offline/storage/

Tamil Open-Source Landscape – Opportunities and Challenges

Muthiah Annamalai*, T. Shrinivasan+
* - ezhillang@gmail.com, + tshrinivasan@gmail.com

Introduction

General tenet of Open-Source and FreeSoftware originally founded by pioneers like Richard Stallman and Eric. S. Raymond [1a,b] continues to rely on collective good of developing software in open and creatively monetizing it outside of closed-source traditional models. Tamil open-source software has its roots from various GNU/Linux user-groups across Tamilnadu [2a], and individuals motivated by Tamil support on Internet [2b,c]. More recently a rise in awareness created by Tamilnadu branch of FSF [2d].

Tamil open-source software (TOSS) continues to grow with over a hundred repositories in github that contribute code for libraries, applications, speech synthesis tools, Android/iOS apps, OCR tools, web-utilities, translations and font faces. While Tamil open-source work truly may have started with translation efforts and localization of KDE (tamillinux in early 2000s) and GNOME (especially GTK) in same period of late 90s and early 2000s, today it continues to grow in an organic, if unorganized way. In the meanwhile many projects, have come and gone and morphed in their existence. (Please note this is not, by any means complete, representative, inclusive history of Tamil open-source development or Tamil computing – just a partial glimpse).

The various challenges are identified below, will be elaborated with appropriate examples,

1. Limited market for Tamil software
2. Marketing efforts for digital Tamil products
3. Lack of reusable, ready s/w components for Tamil software delays development and increases costs of development, production and post-production are limiting future projects
4. Tamil origin Tech workers are indifferent to TOSS causes

Barriers to entry into the Tamil open-source software (TOSS) space are identified and removal of these issues could spurt a growth phase in TOSS

1. Software not addressing the market – teaching developers to address real market needs
2. Address demographic needs:
   What about other adults, young-adults, teenagers, boys and girls usage of software?
3. Sometimes CS challenges are hard – CS education and continuous growth are recommended for developers
4. Multiple roles required for software development, graphic art, testing, documentation, packaging and release, which are lesser known to potential Tamil contributors

We also like to highlight several open-source Tamil projects which helped create newer projects and grow the ecosystem. We also found Tamil project “OCRWikisource” by one of authors inspired an Oriya language application. So there is tremendous potential for growth in Indian-language computing by learning from one-another.

Further issues in TOSS include leadership and fund-raising abilities, continuity of the developer community sustenance and growth, promotion of open-source way of contributing to digital Tamil. For security of the Tamil open-source community we realize that a non-profit model with Wikipedia style rapid-grants for 6month periods to grow the community, certify young engineers contributions, and provide leadership will be ideal. Currently Thamizha and Ezhil Language Foundation have provided some aspect discussed here. Malayalam engineers have provided leadership for their efforts through **Swatantra Computing** [3a] (SMC) and maintaining frameworks like SILPA and 11-Indian language TTS system called Dhvani [3b].

Image 1

We provide a list of representative Tamil open-source projects and number of contributors, and bugs reported/fixed and the pull requests from GitHub data in Table. 1. Our full review of TOSS landscape is expected to serve as a milestone of present day Tamil adoption and growth aspects.

Project	Stars	Language	closed issues	open issues	pull requests	Fork	Contributors
Ezhil-Lang	78	Python	100	93	59	35	11
open-tamil	37	Python, Java,	61	54	8	26	6
OCR4wikisource	26	Python	54	32	9	17	5
EkType/Mukta	46	Font	24	2	5	16	3
ratreya/lipika-ime	21	Objective-C	17	3	9	13	2
VanillaandCream/Catamaran-Tamil	26	Font	7	10	4	7	3
FreeTamilEbooks (Android client)	6	Android/Java	3	10	0	11	1
thamizha/eKalappai	11	C++	3	13	4	8	3
velsubra/Tamil	6	Java	2	0	5	3	2
ashokr/TamilNLP	9	Python	1	0	1	4	1
mayooresan/Android-TamilUtil	15	Java	0	0	0	15	1
thamizha/android-tamilvisai	14	Java	0	3	3	14	6
godlytalias/Bible-Database	18	PL/pgSQL	0	0	1	9	2
vasurenganathan/tamil-tts	24	php	0	0	0	6	1
rdamodharan/tamil-stemmer	22	C	0	0	0	4	1
echeran/CLJ-Thamil	36	Clojure	0	0	0	3	1
psankar/Korkai	10	Go	0	0	0	3	1
rprabhu/Tamil Dictionary	8	Javascript	0	0	0	2	1

Table 1: Representative List of Tamil Open-Source projects by Git-Hub community

GitHub collaboration space

GitHub is an open-source project development space that is popular and many Tamil origin developers are present and contributing to various open-source efforts – within and outside of Tamil computing.

Typically the project founders create a GitHub repository and choose one of the open-source licenses like MIT, Apache, GPL etc. and start uploading their code and setting up unit-tests and continuous-integration tests via Travis-CI. Other developers may join in by forking the git repository and working on one or more issues (bugs) or features and sending a pull-request to the original (founder's) repository.

The founder can have one or more comments and after resolving any outstanding disagreements the software pull-request is committed to the source repository and merged into one. Rinse-and-repeat of this flow is usual Git-Hub development

Collectives in GitHub

Thamziha group [4] in GitHub is the single largest pool of volunteer developers followed behind by Ezhil-Language-Foundation. Thamizha group has 42 volunteer developers, 24 source projects, 2 forks, and contains sources for important projects like eKalappai, tamil-fonts, visaineri, Peyar, and Thamizha.org website. Major coding languages used in 26 repositories of Thamizha are JavaScript, Python, HTM, PHP. and Java. Thamizha is a open-source volunteer group with a supporting mailing list at freetamilcomputing [5], with -the logo in **Image. 2.**

Image 2: Thamizha group at Git Hub

Repositories by Language

To gain a measure of interest and popularity of language based projects by developers in Github, we collected data by running searches on Github [7] and found the surprising result in **Image. 3.** Tamil language repositories outnumber Hindi or Malayalam, Kannada and Telugu efforts together.

While our data collection methods are by Github search and admittedly coarse grained, we still see on the 760 Tamil projects hosted on Github, even if only 20% of them are active, a resounding interest in Tamil computing. We need all of these 760 projects and developers to come on to mainstream and become active, thriving contributors. This data is a call for general Tamil community to support these budding open-source developers and channelize their energies to create useful, inventive software to further improve Tamil computing landscape. The more active among these 760 projects are dictionaries, Android and iOS applications, keyboards, transliteration software, programming languages in Tamil, encoders/converters for Tamil, Tamil font collections, book translations etc.

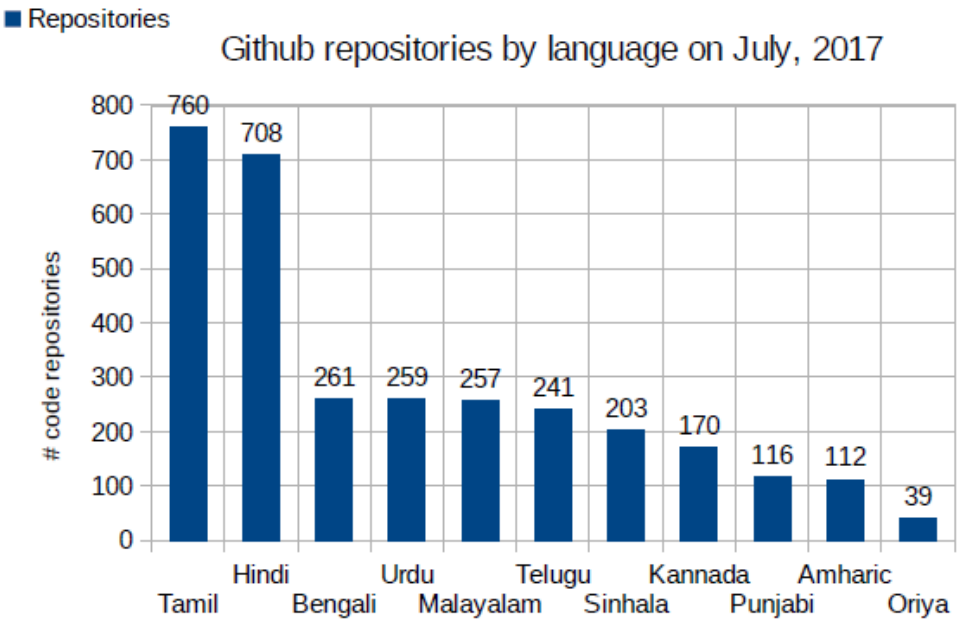


Image 3: Chart of repositories in Github by language in July, 2017.

Why TOSS does not reach end-users?

With such a large body of software available within reach of developers it is natural to wonder why Tamil software usage has not "taken-off" in a big way. In this section we present some detailed analysis of why Tamil open-source software has missed the big-impact on general Tamil users and continues to under-perform its reach.

We present this analysis naming several valuable, popular projects with intention of highlighting their missing pieces and improving their impact – with complete understanding of how strapped for support, funding and manpower many of these projects can be. Hence authors request creative license to critique:

1. Software is not packaged well

Most of the software are in development stage always; except a handful of TOSS projects rest of them binary executable packages for Linux/Windows is available. Not everyone has access to a development environment. This batteries excluded approach fails us.

Examples:

(a) Until recently the Ezhil-Language software was not available for downloads to Windows and Linux platform for general users to try out.

(b) A popular software component often requested by developers but not available as a component is the tamil-stemmer by R. Damodharan [8]

2. Lack of online demos

Even for software libraries, there no online demonstrations to have a taste of them, before installing. Example - open-tamil, has a non-unicode to unicode convertor, with many font combinations then any other available tool. since, it has no web interface or site, nobody knows and uses it.

3. No windows version

As windows is still used by regular users, we need to ship packagesfor windows too. Example : OCR4WikiSource is a connector for Google OCR and Wikisource, but works only for Linux. New users found it difficult use with Linux. So, they not using it.

4. No showcase site for all the web applications or fonts

There are many web applications for Tamil. But there is no showcase site to list them all, install and let the users to play around it. Few apps have their own sites. But they work sometime and die without maintenance.

We have TamilTTS, avalokitam, Velmurugan's Tamil tools [9], anunaadam, Pallanguzhi, Ezhil Lang etc as source code in github. Few sites work and many are not. If we have a common portal with samples installed, will be easy to explore. Need a portal like http://www.opensourcecms.com/. Need a section to list all fonts with sample images.

5. Cant use with major software

Few Tools work as standalone. But the real need would to be used as plugin to any other major system. Example: Spellchecker by Thamizha. It works with Mozilla browser. But the needs is to work with MS Office/LibreOffice/PageMaker/InDesign. There may be tech issues like proprietary software may not allow plugins to be developed by others.

6. Lack of Graphical User Interfaces (GUI)

UI should be nice and easy to easy adoption. Most developers are not good UI designers, and the applications are not looking good. Example - Tamil tools by Velmurugan Subramanian. This is good in linguistic features, but not good at UI.

7. Lack of user/developer documentation

With no or less documentation, users find it is difficult on using a software further. When a developer is left with no developer documentation, he/she looks for another project with nice developer doc. Example : Velmurugan Subramanian's Tamil application is good. But its documentation is missing with the source code. The site for doc is not working now.

8. Lack of marketing done by developers

The applications are developed and source is pushed to github. There no announcements in public mailing lists like FreeTamilComputing or similar. There may be few Facebook posts or tweets. But to get more people's attention, the project releases should be announced to wider audience.

9. Lack of Offline events

The offline events like intro talks on Linux users groups, hackathons will inspire local people to contribute. These kind of events are not happening.

10. Not mobile friendly

Most of the web apps wont work well with mobile browsers. Cant make mobile apps with webview. They should have mobile friendly CSS styles. Possible applications should be converted as mobile apps too. Like tamil games, spell checkers, TTS, OCR etc, should be available as mobile apps. Example: Thamizha's spellcheckers wont work with mobile browsers.

11. No lobbying

Cant convince big giants to install the TOSS as inbuilt. Example: e-kalappai can be a part of windows by default. Indic-keyboard can be a part of android. Still these are not happened. We don't know how to do this and whom to contact.

12. No funding.

There are no private/govt agencies for funding. No events like GSOC. No events/prizes/goodies for developers.

13. Transliteration is cheap alternative

Institutional users of Tamil – mainly Tamil cinema industry – does not strongly patronize Tamil software in script writing, re-recording, dubbing, music production/playback singing etc. and chooses to use transliterated Tamil. However there are problems in Tamil prosody, diction which is appalling in Tamil cinema industry today, but more importantly they fail to be a market for Tamil software.

14. Lack of reusing

Most of the features are buried with the applications itself. No libraries/API are provided.

15. Other basic issues for end-user:

a) He/She don't know tamil typing.

b) Tamil keyboards (physical) are not available; most of them are onscreen keyboards. keyboard covers with tamil letters embossed are available

c) Too many keyboard layouts like Tamil Typewriter, Tamil99 cause confusion to user

d) Too many fonts/engines. Still unicode is not a standard. Introduction of TACE/16 makes it complex especially without open-source tools for TACE encoding.

e) Fear of typing wrongly as no default spellcheckers are available across the OS – Tamil writing is more self-critical and acts negatively to suppress Tamil usage online.

We think addressing one or more of these issues highlighted above will create a positive feedback growth-cycle for the Tamil Open-Source landscape.

Recommendations

We feel the following recommendations are necessary to groom young talents into Tamil software generally and open-source Tamil computing in particular:

1. A central FAQ about accessing Tamil script/text functions in various programming languages to address developer training
2. There is a need for institutional effort to have a Tamil coding school
3. There is a need for Tamil marketplace for software

Conclusions

We report in this paper, Tamil open-source software community is a vibrant place with software developers, font designers, translators, voice-over artists, and general user testers, who come together for love of their language, and promotion of critical thinking, and modern language usage in Tamil. We identify a need for institutional support at various stages from

grooming software developers in Tamil, to marketing platform for Tamil software. There is bright future for tamil software if we will meet challenges it brings with it.

References

1. Richard Stallman, “GNU Manifesto” (1985), Eric. S. Raymond, “The Cathedral & the Bazaar” (1999).
2. (a) India Linux User Group Chennai, (b) GNU Linux User Group Trichy (2003-2008), (c) Thagadoor Gopi of http://higopi.com; Suratha Yarlvanan of http://suratha.com, (d) TamilNadu branch of FreeSoftwareFoundation, https://fsftn.org/
3. Dhvani TTS, https://github.com/dhvani-tts/dhvani-tts
4. Swatantra Malayalam Computing, https://smc.org.in/
5. Thamizha group in GitHub, https://github.com/thamizha
6. Thamizha Free Tamil Computing mailing list, https://groups.google.com/forum/#!forum/freetamilcomputing
7. Github, www.github.com
8. R. Damodharan, “A Rule Based Iterative Affix Stripping Stemming Algorithm For Tamil ” Tamil Internet Conference, Malaysia, (2013).
9. Velmurugan Subramanian, Java project 'Tamil' with tools for processing Tamil language, https://github.com/velsubra/Tamil

The role of VLE Frog in assisting students, teacher and parents in M - learning and usage of ICT tool such as smartphones and computational devices in school curriculum.

Shanti Ramalinggam ***<shanti.usm@gmail.com>***
SJK(T) West Country Barat, Selangor, Malaysia

Abstract

The latest teaching-learning approach M-learning is known as learning through mobile computational devices (Quinn, 2000). Network-based learning content (Malin en, Kari, & Tiusanen, 2003). Wireless network-learning (Boerner, 2002) or technology-based curriculum (Anderson, 2001). Emergence of a new pedagogical approach which promotes student-centred learning experience using technologies based on M-learning will offer opportunities, convenience, advantages and dynamic environment enabling students to succeed in their studies.

Mobile learning is a way of learning where students and teachers don't need to be concerned regarding the time and place restrictions. Small sized technologically advanced gadgets such as PDA(Personal Digital Assistant), Compaq iPaq, laptop, wireless screen phone, smartphone and many other are used to send and receive information in regards to learning and teaching. Study note, homework, and assignments are synchronized through group chats and websites and received by students and parents.

Students from Year four till Year six at SJK (T) West Country Barat were used as the sample for this research and questionnaire is used as instrument. All the data were processed using SPSS software to identify the mean score percentage frequency, Pearson correlation and One Way ANOVA.

Introduction

In Malaysia, YTL Communications has collaborated with Education Department in providing free smartphones to teachers to incorporated Mlearning in their curriculum. VLE Frog, a branding under YTL's 1Bestari Network has been creating and managing online and smartphone based applications for students, teachers and parents to actively involve in teaching-learning activities using the latest technological advancement.This article highlights the concept and benefits of M-learning through VLE Frog applications, and discusses the prospects of M-learning in the future curriculum.

Among the modules are:

- **PIE** (Participation, Interaction, and Engagement) :

This module will guide participants to attract attention and increase students interactivein learning and teaching process through creation of a new website, using the 3 widget Text, Media and Wall in the Frog VLE. The workshop will also provide space for teachers to share ideas in classroom management.

- **CAKE** (Connect and Keep Engaging) :

The module aims to empower teachers and students with a means of creative and innovative learning through video conferencing technology with the use of the Frog Connected Classroom learning concept.

- **PUFFS** (Project Using Frog for Students) :

This module is designed for teachers who wish to engage students in 21st century learning through project based learning using Frog VLE. This module will also provide a space for teachers to share their experiences and designing digital projects for students with Frog widget.

- **GULAI** (Game Up Learning And Involvement) :

This module will introduce FrogPlay as reviewing applications for students with pre-existing quizzes, mini-games and a robust performance reporting.

This workshop will also include a way for teachers to guide students' learning through ongoing feedback on the performance of students with the use of application FrogPlay.

- **NASI LEMAK** (Smartphone based Frog applications) :

NasiLemak module is a module that includes all the modules in YES Altitude smartphones and Frog mobile application. Teachers can choose to start the workshop using pieces of any module in question.

Methodology

Questionnaire was used as the instrument in this research. According to Cohen, Manion & Morrison (2007) questionnaire is suitable because it is able to directly collect data from a large size sample whereby it will improve the accuracy of the statistic sample to estimate to population parameter thus decreasing the sample differences, it isalso believed to increase the accuracy and true reaction given by respondents without being influence by the researcher' behaviour (Mohd Majid Konting, 2005). Level twostudents were the population. In order to avoid biasness, random sampling was used because it is easier to control. Especially for this research, the selection of sample will be based on the students' attendance sheet (table). The final sample of the research consists of 85 standard six students from West Country Barat Primary School (Tamil), Kajang, Selangor. The sample includes male and female students from various family backgrounds. 22 samples were selected randomly through selection of odd numbers for the reason of pilot study. L.G Gay, G.E. Mills and P. Airasian (2006) and Sekaran (2000) stated that selection of sample from population must be similar to the real sample.

The validity of the questionnaire was also proven. According to MohdMajidKonting (2005) testing the level of validity of questionnaire is vital to determine whether the item formed are suitable with the respondent. Internal Consistency Reliability in the Alpha

Cronbach Test indicated the value of Alpha Cronbach for all aspects in Part B are 0.6 till 0.9, which indicate a high value of reliability.

In conducting the research, approval from the District Education Office was required. The duration of time in conducting the research was after school session (curriculum period) to ensure that it will not impede with the students' learning process. All questionnaire collected will be analysed by using SPSS software. Inferential statistic is used to assist in analysing the data. Pearson correlation and One Way ANOVA was used to analyse research question 1, 3, 4, 5 & 6.

The findings of the research indicated positive significant improvements among students, who use the Mlearning software to assist in their learning. Comparison among students who use the software and those who don't has given a clear indication of the impacts of the ICT usage in learning which could help improve students creative thinking.

Conclusions

The research has helped the school management identify more fun and constructive learning system through the usage of smart phone and other computational devices and able to improve the students thinking ability and creative learning. The usage of Frogasia website and modules has helped students, teachers, and parents to interact in a more personal way through video conferencing and social messaging services. The study also helped teachers to construct a thinking based learning system rather than exam orientated learning through the use of smart devices. The drawback to this is to provide the students with enough tools such as computers and smartphones for them to use the applications. Although YES has been marketing their YES ALTITUDE smartphone for RM 180 but still it's a burden to provide the device to all students and parents. Apart from that education institutions must invest in more facilities in order to make this learning method available to all students.

References

1. Anderson, 1. (2002). Revealing the hidden curriculum of eLeaming. Dalam C. Vrasidas, & G.Y. Glass (Eds.), *Current perspectives in applied information technologies.* Greenwich, CT: Information Age.
2. Boerner, G. L. (2002). The brave new world of wireless technologies: A primer for educators. *Syllabus Technology for Higher Education.*Dimuatturunpada Jun 20, 2004, daripada http://www.campus-technology.com/article.
3. Malinen, 1, Kari, H., &Tiusanen, M. (2003). Wireless networks and their impact on network-based learning content. *Enable Network-Learning.* DimuatturunpadaJulai, 2004, daripada http://www/enable.evitech.fi!enable99/papers/ malinen/ melinen.html
4. MohdMajidKonting. (2005). *Kaedahpenyelidikanpendidikan*. Kuala Lumpur: Dewan BahasadanPustaka.
5. Sekaran, U. (2002). *Research Methods for Business: A Skill Building Approach.* (3rd ed.).

New York: John Wiley and Sons, Inc.

6. Quinn, C. (2000). *Mobile, wireless, in-your-pocket learning.*Dimuatturunpada Julai 1, 2004, daripadahttp://www.1inezine.com/2.l/features/cqmmwiyp.htm

Acknowledgement

1. SJK (T) West Country Barat
2. FrogasiaSdn. Bhd.
3. Ms. Elizabeth Lopez ((FrogAsia @ Head of Transformation Management
4. YTL Corporation.

கற்பித்தலில் தரவகமொழியியலின் பங்கு

Dr A Ra Sivakumaran

Head, Tamil Language & Cultural Division, National Institute of Education, Singapore

---

உலகமொழிகளில் செம்மொழியாகக் கருதப்படும் சில மொழிகளில் காலத்தால் முந்தைய மொழியாக விளங்குவது தமிழ்மொழி. அம்மொழி உலகில் பலபாகங்களிலும் கற்பிக்கப்-படுகிறது - கற்கப்படுகிறது. மொழிக்கல்வியின் நோக்கம், கற்பவருக்குக் குறிப்பிட்ட மொழியைப் பயன்பாட்டு நோக்கில் கற்பிக்கவேண்டும் என்பதே ஆகும். மொழியின் இலக்கணத்தையும் சொற்களஞ்சியத்தையும் அவற்றின் இயல்பான கருத்துப் புலப்பாட்டுப் பயன்பாட்டிலிருந்து (communicative function) தனிமைப் படுத்திக் கற்றுக்கொடுப்பது மொழிக்கல்வி (teaching the language) ஆகாது. அவ்வாறு கற்றுக் கொடுப்பது மொழியைப்பற்றிக் (teaching about the language) கற்றுக் கொடுப்பதாகவே அமையும்.

கருத்துப் புலப்பாட்டுப் பயன்பாட்டு நோக்கில் (Communicative approach) மொழியைக் கற்றுக்கொடுக்க முதல்தேவை அந்நோக்கத்தில் மொழிப்பாடங்களை அமைப்பதே ஆகும். மொழிப்பாடங்களில் இடம்பெறும் பனுவல் (lessons/texts) செயற்கையாக உருவாக்கப்பட்ட ஒன்றா (synthetic) அல்லது அம்மொழியைப் பயன்படுத்துவோர் இயல்பாக நேரடியான கருத்துப் புலப்பாட்டில் பயன்படுத்திய தொடர்களை உள்ளடக்கியதா (authentic) என்பது கவனத்தில் கொள்ளவேண்டிய ஒரு முக்கியக்கூறாகும்.

இயல்பான கருத்துப் புலப்படுத்தத்தில் பயன்படுத்தியதொடர்களைக் கொண்ட தரவுத்தளத்தை (Corpus) அடிப்படையாகக் கொண்டு அமைக்கப்படுகிற மொழிப்பாடங்களே சிறந்தது என்ற கருத்து தற்போது மேலோங்கி வருகிறது. ஒருமொழிக்கு உருவாக்கப்படுகிற தரவுத்தளங்கள் அத்தளங்களின் பயன்பாட்டிற்-கேற்பத் தமது இயல்பில் மாறுபடும். இக்கட்டுரையானது தமிழ்மொழிக் கல்விக்குப் பயன்படுகிறமின் தரவுத்தளத்தை (electronic learning corpus) அடிப்படையாகக் கொண்டது.

இந்நோக்கில் அமைக்கப்பட்ட தமிழ்மொழி மின்தரவுதளத்தை எவ்வாறெல்லாம் பயன்படுத்தி, மாணவர்களுக்குத் தமிழைக் கற்றுக்கொடுக்கலாம் என்பதே இக்கட்டுரையின் நோக்கம்.

தமிழ்மொழி மின்தரவுதளத்தை மொழிக் கல்வியில் இரண்டு வகைகளில் பயன்படுத்தலாம். ஒன்று, பாடங்களைஉருவாக்குவதற்குப் (curriculum / syllabus design, text book preparation) பலவகைகளில் பயன்படுத்துவது. மற்றொன்று, மாணவர்களுக்குத் தமிழைப் பயன்பாட்டு நோக்கில் பயன்படுத்தப் பயிற்றுவிக்கும் நேரடி நடவடிக்கைகளுக்குப் (teaching / learning activities) பயன்படுத்துவது.

தமிழ்மொழியைக் கற்றுக்கொள்ள வேண்டும் என்றால் முதலில் அம்மொழியில் வழங்கப்படும் சொற்களின் பொருளை அறிந்து கொள்ள வேண்டும். பொருளை அறிந்து கொள்வதற்கு அகராதி துணையாக இருக்கும் என்பதில் கருத்து வேறுபாடு இல்லை. ஆனால் மொழியில் இடம்பெறும் அத்தனை சொற்களும் பயன்பாட்டில் இருக்கும்போது அகராதிப் பொருளை மட்டுமே தருவதில்லை, மேலும் ஒருசொல் ஒருபொருளில் மட்டுமே வருவதில்லை. பயன்படுத்துவோரின் ஆட்சிக்குட்பட்டு இடத்திற்கு ஏற்பப் பலபொருளில் (usage contexts) அச்சொல்வருவதுண்டு. இடத்திற்கேற்ப வரும்பொருளைக் கற்பவர் அறிந்து கொள்ளும்போதே மொழியைக் கைவரப்பெற இயலும். ஒருசொல் எந்தெந்த இடத்தில் எந்தப்பொருளில் எல்லாம் ஒருபனுவலில் வந்துள்ளது என்பதை நாம் ஒட்டுமொத்தமாகத் தெரிந்து கொள்வதற்கு இத்தரவுதளம் மிகவும் உதவிபுரிகிறது. இதில் நேரடியாக மாணவர்களை ஈடுபடுத்திப் பயிற்சியளிப்பதற்குச் சொற்சூழல் அடைவி (Concordancer) மிகவும் பயன்படும். இச்செயலை மேற்கொள்வதற்குப் பயன்படும் சில தரவுதள மொழி ஆய்வுக்கருவிகள் பற்றி (corpus analysis tools) இங்குப் பேசப்படுகிறது. ஒருசொல் பொதுவாக வேறு எந்தச்சொல்லுக்கு அருகில்வருகிறது அல்லது இணைந்து வருகிறது (collocation) என்பதைப் பொருத்து அதன்பொருள் மாறுபடுகிறது. எந்தச்சொல் எதன் அடிப்படையில் மாறுகிறது என்பதைப் பயன்பாட்டின் அடிப்படையிலேயே மாணவர்கள் எளிதில் தெரிந்து கொள்ளமுடிகிறது. எடுத்துக் காட்டுக்குப் “முயல்” என்ற சொல்லை எடுத்துக் கொள்வோம்.

"அங்கே முயல் ஓடுகிறது" என்ற தொடரில் அது விலங்கினத்தை(பெயர்ச்சொல்) குறிக்கிறது; "நீ தேர்வில் அதிக மதிப்பெண்கள் எடுப்பதற்கு இன்னும் நன்றாக முயல்" என்ற தொடரில் மாணவர் மேற்கொள்ளவேண்டிய செயலை (வினைச்சொல்) குறித்து நிற்கிறது. "கடலை"என்ற சொல், " கடலைப் பார்த்துவிட்டுச் செல்லலாம்" என்ற தொடரில் ஒரு பொருளையும் (கடல்+ஐ), "கடலை சாப்பிடுவோமா" என்ற தொடரில் வேறு பொருளையும் (கடலை) குறித்து நிற்கிறது. இங்கெல்லாம் குறிப்பிட்ட சொல் அமைகிற மொழிச் சூழலே சரியான பொருளைப் பெற உதவுகிறது.

ஒரு குறிப்பிட்ட தொடரில் ஒரு குறிப்பிட்ட சொல்லின் பொருளை அறிந்து கொள்ள அகராதிகள் பயன்படலாம்.. அகராதியானது அதே குறிப்பிட்ட சொல்லுக்குப் பல பொருள்கள் இருந்தாலும் அத்தனை பொருள்களையும் பட்டியலிட்டுத் தரும். ஆனால் குறிப்பிட்ட தொடரில் அச்சொல் பெறுகிற பொருளைத் தெரிந்துகொள்ள

மனிதமூளையின் உலகியல் அறிவே பயன்படுகிறது. மேலும் ஒரு சொல்லின் பொருளைத் தெரிந்துகொள்வதில் அச்சொல்லின் இலக்கண வகைப்பாடு மிகவும் உதவுகிறது. குறிப்பிட்ட சொல்லானது பெயரா, வினையா என்பதைப் பொறுத்தே பொருள் அமைகிறது. ஒரு அடிச்சொல்லே பெயராகவும் வினையாகவும் அமையும்போது, அந்த அறிவை அகராதி அளித்துவிடும். ஆனால் சில தொடர்கள் வடிவத்தில் ஒரே மாதிரியான விகுதிகளைப் பெற்று பொருள் குழப்பத்தை ஏற்படுத்தும். எடுத்துக்காட்டுக்கு “ஓடுவது நல்ல குதிரை”, “ஓடுவது நல்ல பழக்கம்” என்னும் இருதொடர்களில் "ஓடுவது" என்பது முதல் தொடரில் வினையாலணையும் பெயராகவும், இரண்டாவது தொடரில் தொழிற்பெயராகவும் அமைகிறது. முதல் தொடரில் ‘அது’என்பது வினையாலணையும் பெயர் விகுதி. இரண்டாவது தொடரில் அதே ‘அது’என்பது தொழிற்பெயர் விகுதி. அது என்னும் விகுதி இரு பொருளிலும் வரும் என்பதை அறிந்தவர்கள் பொருள் குழப்பத்திற்கு ஆளாக மாட்டார்கள் மற்றவர்கள் பொருள் குழப்பத்திற்கு ஆளாகுவார்கள். இத்தகு குழப்பத்தை அகற்றக் குறிப்பிட்ட சொல்லுக்குப் பின்னால் வருகிற சொற்கள் உதவும்.

அதற்குக் கணினிமொழியியலில் என்-கிராம் (n-gram) அடிப்படையில் அமைகிற சொல் பட்டியல் உதவுகிறது. என்-கிராம் சொல் பட்டியலில் குறிப்பிட்ட சொல்லுக்கு முன்னால் வருகிற சொற்களும் பின்னால் வருகிற சொற்களும் கொடுக்கப்படும். ஆனால் அவ்வாறு குறிப்பிட்ட சொல்லுக்கு முன்னாலும் பின்னாலும் இடம்பெறுகின்ற எத்தனை சொற்களின் விவரங்கள் தேவைப்படும் என்பதைக் கவனத்தில் கொள்ளவேண்டும். மேலே கொடுக்கப்பட்ட எடுத்துக்காட்டுகளில் "ஓடுவது" என்ற சொல்லுக்கு அடுத்து "நல்ல" என்ற சொல்லே இடம்பெறுகிறது. நல்ல என்னும் சொல்லைக்கொண்டு இங்கு ஓடுவது என்னும் சொல்லின் பொருளை அறிய இயலாது. அதற்கு மாறாக நல்ல என்னும் சொல்லுக்கு அடுத்து வருகின்ற ‘குதிரை’ ‘பழக்கம்’என்ற சொற்களே பொருள் குழப்பத்தைத் தெளிவுபடுத்த உதவுகிறது. அதாவது குறிப்பிட்ட சொல்லுக்கு அடுத்த ஒரு சொல் மட்டுமல்லாமல், அதற்கும் அடுத்த சொல்லும் தேவைப்படுகிறது. இவ்வாறு என்-கிராம் அடிப்படையில் குறிப்பிட்ட தொடர்களின் சொற்களைத், தரவக அடிப்படையில் சுட்டிக்காட்டப்படுகின்ற பொழுதே மாணவர்கள் தாங்களாகவே தங்களது மொழியில் வடிவ ஒற்றுமையினால் அமைந்த சொற்களின் பொருளை அறிந்து கொள்ள இயலும். எடுத்துக்காட்டுக்குக் கீழ்க்கண்ட வாக்கியங்களைப் பார்ப்போம்.

‘ஒவ்வொன்றையும் முறையாகப் **பார்ப்பது** அவசியம்’. அங்கிருந்து முறைத்துப் **பார்ப்பது** யார்? ‘அவர் **அமர்ந்தது** சரியான செயலே’! ‘அங்கு **அமர்ந்தது** அந்தப் பெரிய யானை’. **‘படிப்பது** நல்ல பழக்கம்’. ‘அங்கு **படிப்பது** நல்ல பிள்ளை’.

மேற்கண்ட வாக்கியங்களில் முதல் வாக்கியத்தில் இடம்பெறுவது தொழிற்பெயராகவும் இரண்டாம் வாக்கியத்தில் இடம்பெறுவது வினையாலணையும் பெயராகவும் உள்ளன.

‘காலையில் குழந்தை **பிறந்தது’. ‘பிறந்தது** நல்ல குழந்தை’.
‘சேக்கிழார் நற்குடியில் பிறந்தார்’. ‘நற்குடியில் பிறந்தார் நல்லது செய்வார்’
‘மாடு தண்ணீர் **குடித்தது’. ‘**பால் **குடித்தது** மாறன் அல்ல; பூனை’.
‘பாகற்காய் பந்தலில் **தொங்கும்’. ‘**மரத்தில் **தொங்கும்** குரங்கு சில வேளைகளில் தாவும்’.
மேற்கண்ட வாக்கியங்களில் முதல் வாக்கியத்தில் அடிக்கோடிடப்பட்ட சொல் வினைமுற்றாகவும் இரண்டாம் வாக்கியத்தில் அதே சொல் வினையாலணையும் பெயராகவும் வருகிறது. சொற்கள் ஒன்றுபோலவே இருந்தாலும் வெவ்வேறு இலக்கணப்பொருளில் அவை இடம் பெறுகின்றன. இவ்வாக்கியங்களில் அடிக்கோடிட்ட சொற்களின் பொருளை அறிந்து கொள்வதற்கு அந்தச்சொல்லுக்கு முன்னாலும் பின்னாலும் அமைகிற சொற்கள் உதவுகின்றன. இதனை நமக்குக் காட்டவல்லது தரவுமொழியில் அமைந்துள்ள N-Gram என்ற மொழிஆய்வுக்க ருவியாகும். ஒரு சொல்லுக்குப் பல்வேறு பொருள்கள் (Word Senses) இருக்கின்றபோது அச்சொல் பயின்று வருகிற சூழலைப் பொறுத்தும், அதற்குமுன்பு அல்லதுபின்பு அமைகிற சொல்லைப் பொறுத்தும், அச்சொல் எந்தப்பொருளில் வந்துள்ளது என்பதை அறியச்செய்வதே N-Gram என்ற மொழிஆய்வுக் கருவிஆகும். மேற்கண்ட வாக்கியங்களில் அடிக்கோட்டிட்ட சொற்களின் பொருளை தரவகங்களின் மூலம் நன்கு அறிந்து கொள்ள இயலும் (இக்கூற்றை தரவகத்தின் மூலம் கணினியில் காட்டுகின்ற பொழுது தெளிவாக விரைவாக ஐயமின்றி தெரிந்துகொள்ள இயலும்)

மேலும் மாணவர்களே தாங்கள்கற்கும் புதியசொற்களைக் கொண்டு ஒரு அகராதியை அமைத்துக் கொள்ளவும் வாய்ப்பு இருக்கின்றது. தாங்கள் உருவாக்கும் அகராதியில் தாங்கள் கற்கும் புதிய சொற்களுக்கு என்னென்ன பொருள் என்பதை அவர்களே குறித்துக்கொண்டு ஒரு அகராதியை உருவாக்கிக் கொள்ளலாம்.

இத்தரவுதளத்தை வைத்துக்கொண்டு கற்பிக்கும்போதும் கற்கும்போதும் மொழியை எளிதாவும் தெளிவாகவும் கற்க, கற்பிக்க வாய்ப்பு மிக அதிகம். மேலும் பாடப்புத்தகங்கள் எழுதுவோர்க்கு இத்தரவுத்தளம் பலவகைகளிலும் பயனுள்ளதாக அமையும். இக்கட்டுரையில் அமைந்துள்ள செய்திகளைக் கணினியின்வழி அறிகின்றபோதே முழுமையாகஅறிய முடியும்.

*** தரவுத்தளத்தின் பல்வேறு பயன்பாடுகளை இக்கட்டுரையில் விளக்குவதற்கு எனக்குப் பயன்பட்ட தமிழ்மென்பொருள் “என்டிஎஸ் லிங்க்சாப்ட் சொலூஷன்ஸ் தயாரித்துள்ள மென்தமிழ் – ஆய்வுத்துணைவன்” என்பதாகும்.

COLLABORATIVE AND INTERACTIVE VIDEO QUIZ IN TAMIL USING COMPUTATIONAL OFFLOADING

Shalini Lakshmi A J [1] and Vijayalakshmi M[2]
[1] Research Scholar <ajshalini2020@gmail.com>
[2] Assistant Professor (Sr. Gr.) <vijim@annauniv.edu>
Department of IST, Anna University, Guindy, Chennai, Tamilnadu

---

Abstract

An enhanced mobile application for Tamil learners named ”Interactive Video Quiz” is developed to create a collaborative environment among Tamil students in a classroom. This work tries to ensure that all Tamil students and educators benefit from a learning environment that provides adequate access to high quality applications in Tamil. In order to support Next Generation Learning in Tamil (NGL-T), an effective formative assessment practices including planning, implementing and refining student’s formative assessment practices has been considered. The large computations in the applications that require higher bandwidth are computationally offloaded to resource rich cloud. By doing so, QoS/QoE parameters are satisfied and the quality of the application is improved.

1 Introduction

NGL-T equipped with state-of-the-art audio, visual technology aims to enhance the collaborative classroom level engagement among Tamil learners and instructors both onsite and at remote locations. NGL-T transforms the way of acquiring education through latest technologies focusing on competency among Tamil learners. Next Generation Classrooms include multiple pieces of equipment such as Tablets, Laptop Computers and Smartphones like Micromax, Motorola, iPhone, Microsoft phone, Samsung etc that offer a variety of options for instruction.

A lot of mobile gaming applications have become popular currently. Such applications are in need of high resource mobile devices that can handle heavy computational tasks just like a personal computer [6]. Computational Offloading is a Mobile Cloud Computing (MCC) technique to take care of those tasks. The basic principle behind this technique lies on identifying the computationally low and heavy tasks and shifting the heavy tasks to a rich server in the vicinity [7].

There are two ways in which a larger task can be made to execute in another resource rich device. One is Clonecloud; the other one is Cloudlet. Clonecloud [11] possesses some serious limitations like machine dependent applications and mobility which are not possible in it, whereas in cloudlet, the machine dependent tasks alone can be executed locally while executing the rest in another device. It is also sufficient to handle mobile situations [9].

In order to achieve an interactive Tamil learning environment using Bring Your Own Device (BYOD) method, network bandwidth of the mobile phones plays a vital role [8]. With the help of cloud technology, the possibility rate of collaboration learning in a classroom can be achieved up to an optimum level. The learning content is screen shared between Tamil

educators and learners in the classrooms for achieving cooperative and interactive engagement learning as shown in the prototype below.

Fig 1: Screen Sharing in Classroom

The main objectives of this work are:

- To deliver a mobile application for Next Generation Learning (NGL-T) to the Tamil learners and educators to benefit from a learning environment that provides adequate access to high quality resources using cloud.
- To support the learning with the effective formative assessment practices, including planning, implementing and refining their formative assessment practices.
- To satisfy the QoS/QoE standards by applying computational offloading technique for using resource rich cloud to carry out large computations effectively.

2 System Model

The following diagram Fig 2. shows the overall system architecture of NGCL using cloud. The process of screen sharing the interactive video session among the classroom learners is classified into the following major tasks:

- Identifying the nearby compute cloud (Cloudlets)

  To carry out the large computations faster, nearby cloud is searched.
- Computational offloading of huge processes

  The learning video content is a highly resource intensive application that will be migrated to the nearer cloudlet for processing.
- QoS/QoE based Video Analytics
  - Pre-fetching the video from Video Service Provider
    - Admission control
    - Packet Classification
    - Packet Scheduling
    - Traffic policing and Shaping
  - Overlaying questionnaires on the video
  - Buffering video to the mobile devices

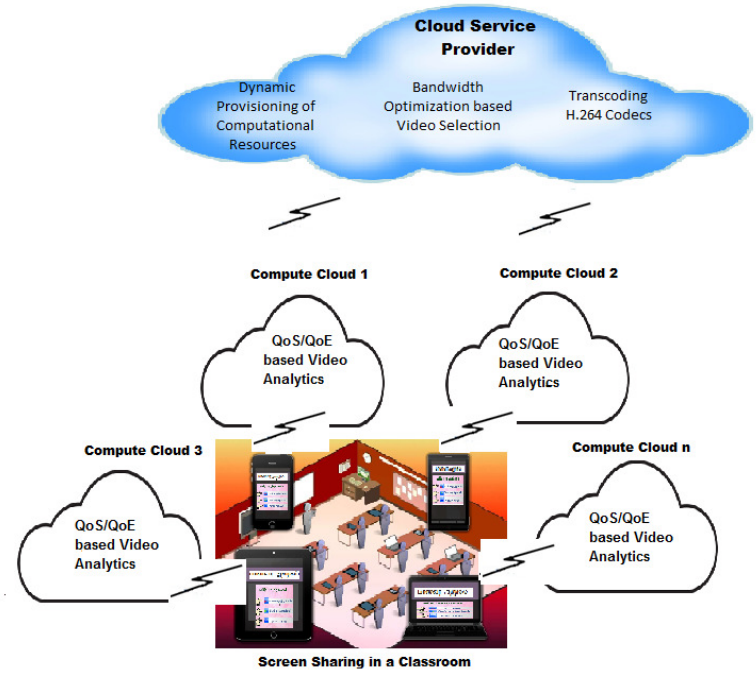


Fig 2. Overall System Architecture of NGCL

Features of Mobile App

This classroom interaction learning app constitutes of two learning components. One is Interactive Quiz session; the other one is Group Collaboration.

A. Interactive Quiz Session

The usage of online video as a primary resource in Tamil education is comparatively rare. The construction and instrumentation of an Interactive Video Lecture Platform measures the student engagement with interactive video quizzes in Tamil. Here quiz questions are overlaid on the frames of the video.

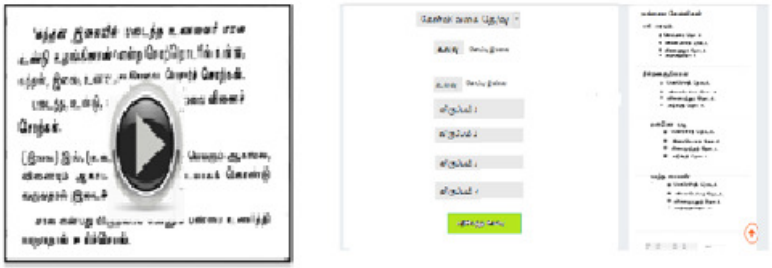


Fig 3. Embedding Questions in video

The dynamic radio environment with fluctuating network bandwidth and wide range of device capabilities are managed by the mobile cloud [10] that provides offloading of the streaming process in the cloud.

B. Group Collaboration

Using a shared display of mobile devices within the same physical space allows teachers and students to share the same information, so that teachers can detect any problems and clarify specific concepts if necessary. This may increase the investment costs of computers to each student and the educators. A much cheaper way to build an Interpersonal Computer with a shared display is to follow the Bring Your Own Device (BYOD) method.

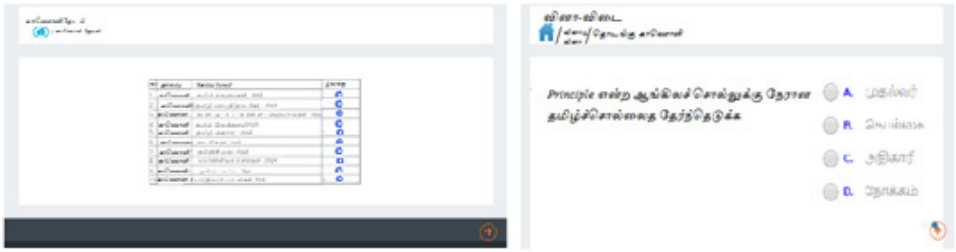
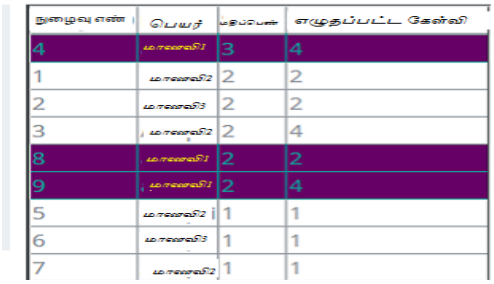


Fig 4. Video with Quiz overlaid

C. Monitoring Learning Behaviour of the Students

The purpose of these questionnaires is to understand Tamil learner perceptions of the technology as well as their behaviours towards answering them. The questions are inspected in terms of cognitive complexity to determine whether complexity affects the students' engagement.



நுழைவு எண்	பெயர்	பதில்கள்	எழுதப்பட்ட கேள்வி
4	[illegible]	3	4
1	[illegible]	2	2
2	[illegible]	2	2
3	[illegible]	2	4
8	[illegible]	2	2
9	[illegible]	2	4
5	[illegible]	1	1
6	[illegible]	1	1
7	[illegible]	1	1

Fig 5. Student Scoreboard

3 Performance Measures

The application is tested with different set of mobile users and with variation in number of users in the learning discussion that happened within a classroom. The resultant video in the application is shared among the Tamil learners who participated in the discussion. The streaming time taken by the video has been tabulated by varying the number of users from 1 up to 10. The tested video is of 5 minutes in length.

It is observed that the streaming time among the users increases gradually when the number of mobile devices in the discussion increases. The scores of each learner is also calculated based on their outcome in the quiz overlaid in the video.

Number of users	Streaming time (in s)
1	2
2	2
3	2
4	5
5	6
6	6
7	9
8	9
9	11
10	11

Table 1. Number of users Vs Streaming time

4 Conclusion

NGL-T targets at affording emerging learning technologies in Tamil, collecting and sharing evidence of what works, and fostering a community of innovators and adopters resulting in a robust pool of solutions and greater institutional adoption which, in turn, will dramatically improve the quality of learning experiences among Tamil learners. The purpose of this work is to encourage Tamil learners to use the latest technology applications as English learners do. The quizzes are embedded into the video lectures to allow Tamil students to follow along at their own pace and clarify the results with their educators all within the lecture.

Our future work has been planned to decrease the hike in the streaming time even when the number of mobile devices in the discussion increases.

References

[1] Jian Wu, Chau Yuen, Ngai-Man Cheung, Junliang Chen & Chang Wen Chen 2015, Enabling Adaptive High-Frame-Rate Video Streaming in Mobile Cloud Gaming Applications, IEEE Transactions on Circuits and Systems for Video Technology, vol. 25, no. 12, pp. 1988-200.

[2] Stephen Cummins, Alastair R. Beresford & Andrew Rice 2015, Investigating Engagement with In-Video Quiz Questions in a Programming Course, IEEE Transactions on Learning Technologies, vol. 9, no. 1, pp. 57-66.

[3] Shuangshuang Ji, Huaping Liu, BinWang & Fuchun Sun 2013, Key-Frame Extraction for Video Captured by Smart Phones, IEEE International Conference on Robotics and Biomimetics, Shenzhen, pp. 2154-2159.

[4] Tal Rosen, Miguel Nussbaum, Carlos Alario-Hoyos, Francisca Readi & Josefina Hernandez 2014, Silent Collaboration with Large Groups in the Classroom, IEEE Transactions on Learning Technologies, vol. 7, no. 2, pp. 197-203.

[5] Guozhu Liu & Junming Zhao 2010, Key-Frame Extraction from MPEG Video Stream, Third International Symposium on Information Processing, Qingdao, china, pp. 423-427.

[6] Fernando, N., Loke, S. W., & Rahayu, W. (2016). Computing with nearby mobile devices: a work sharing algorithm for mobile edge-clouds. IEEE Transactions on Cloud Computing.

[7] Fan, W., Liu, Y. A., Tang, B., Wu, F., & Zhang, H. (2016). TerminalBooster: Collaborative Computation Offloading and Data Caching via Smart Basestations. IEEE Wireless Communications Letters, 5(6), 612-615.

[8] Zhou, B., Dastjerdi, A. V., Calheiros, R., Srirama, S., & Buyya, R. (2015). mCloud: A Context-aware offloading framework for heterogeneous mobile cloud. IEEE Transactions on Services Computing.

[9] Ahmed, E., Gani, A., Sookhak, M., Ab Hamid, S. H., & Xia, F. (2015). Application optimization in mobile cloud computing: Motivation, taxonomies, and open challenges. Journal of Network and Computer Applications, 52, 52-68.

[10] Zhang, S., Wu, J., & Lu, S. (2016). Distributed workload dissemination for makespan minimization in disruption tolerant networks. IEEE Transactions on Mobile Computing, 15(7), 1661-1673.

[11] Zhang, Z., & Li, S. (2016, March). A Survey of Computational Offloading in Mobile Cloud Computing. In Mobile Cloud Computing, Services, and Engineering (MobileCloud), 2016 4th IEEE International Conference on (pp. 81-82). IEEE.

Engaging Augmented Reality and Collaborating With Learners To Inspire and Maximize Learning of Tamil Language

Shahul Hameed M M (Shah)

Abstract

Learners retain a very small amount of the information that they hear and a slightly larger percentage of what is shown to them. However, when they become actively involved in an experience, they retain much more information that is presented to them. In particular, the visual triggering of Augmented Reality sparks enthusiasm in learners, maximizing the opportunity for interaction, encouraging critical response and the adoption of new perspectives and positions. This paper describes how Augmented Reality has been used to promote the learning of Tamil Language. Abstract concepts or ideas in Tamil that might be difficult for learners to comprehend can be presented through an enhanced learning environment. Augmented Reality can harness both asynchronous and synchronous learning, allowing learners to acquire the ability to make links across different areas of knowledge and to generate, develop and evaluate ideas and information. Augmented Reality is indeed a powerful way of promoting active language learning.

1. Introduction

In Singapore, every student takes up Mother Tongue Language as a mandatory second language subject. Students dread learning Tamil Languagebecause they find it difficult to comprehend. Furthermore, most consider Tamil Language as having very little or no relevance to their future. Hence, many learnershavelow commitment towards learning Tamil Language.Most learners are able to read texts in Tamil, but many are unable to comprehend what they read. They are also unable to interpret what they hear. In addition, many are not able to express themselves clearly when speaking or writingTamil.As a result of all these, many learnersresort to learning Tamil through English. For instance, they take notes in English of what is being taught in Tamil Language.

This scenario did not dampen my belief that students will learnTamil better if their language learning interest is kindled. I decided to inspire their learning through engaging them in a meaningful and active learning environment.To do this, I decided to use Augmented Reality Applications to provide a more authentic learning environment that engages learners in ways that were never possible before. With Augmented Reality, all learners can have their own unique discovery path through real-life immersive simulations, with no time pressure and no real consequences if mistakes are made during learning. This technology thus promotes active learning and encourageslearners to take on diverse thinking perspectives.EngagingAugmented Realityboth stimulates the learning interest of students and equips them with the desire to explore further avenues of developing the various Tamil Language skills.

2. Methodology

The following objectives and strands explain how and why Iuse Augmented Reality to enhance the joy of learning.

2.1 Learning Objectives

1. Learners are first introduced to Augmented Reality. They discover its uses, especially how it can act as a vital trigger for learning. Learners then find out about the diverse range and

characteristics of Augmented Reality and describe and explain their extensive, applied uses. They will get to choose the type of Augmented Reality that most fascinates them and demonstrate how it affects their learning.

2. Learners also have the opportunity to analyze and compare learning through Augmented Reality withthat involving traditional pedagogies. In doing this, they evaluate the impact of Augmented Reality and portray their understanding and learning of Tamil Language.

2.2. Strand 1: Understanding learning and every learner

I have been teaching since 1997 and pursued a Master Degree in 2013. Subsequently, after a hiatus of two years, Ireturned to teaching in 2015. After teaching Tamil Language for one term, I developed abetter understanding ofhow a learner-centric approach can be effective in promoting learning. This involves the teacher understanding the needs and strengths of learners, the process of learning, and how learners' minds are engaged in the learning process. As the progress of growing technology affects the educational ambience, I explored how it can be tapped to supportthe learning of Tamil Language. With my conviction that every child can learn, I am also learning from my day-to-day experiences of teaching Tamil Language the underlying principles of learning for every learner.

2.3. Strand 2: Nurturing and inspiring learners

While I passionately support the conventional pedagogies, I am keen to harness Computer and online-assisted teaching in the learning of Tamil to provide efficacious lessons. Motivating my students to learn the language effortlesslyhas been one of my goals. This approach establishes anadvantageous learning environment that stimulates and empowers learners. What I have gathered from my experience in teaching for nearly two decadesis that very little learning can occur unless learners are motivated on a consistent basis through the five key ingredients impacting students' motivation: student, teacher, content, method or process, and environment. Bearing in mind that in general, students are complex creatures with complex needs and desires, I amalgamated these five ingredients to motivate them.

2.4. Evidence of Impact on Student Learning

In Singapore, students studying Tamil Language in secondary schools are streamed according to their academic abilities. The students in Express and Normal Academic streams sit for papers with higher weightingfor written assessments. Those in the Normal Technical stream do Basic Tamil where the focus is on Oracy. I teach Tamil Language to all three streams. The students from these three streams are of mixed ability, but the common factor found among them was that they were not keen to learn Tamil. My observations suggested that adopting a range of teaching techniques would aid in effectively engaging these students in their learning. Keeping my lessons well-paced, I used various teaching techniques and online toolsto engage my students of differing abilities.

Students started to display traces of interest, especially when I designed my lessons utilisingonline tools such as Educreations, Kahoot, Plickers,Socrative, Padlet, Media Broadcasts and social platforms like Skype and Facebook. I got wind of what they wanted and used these to spur their interest further. Students' interest were evident from their keenness to attend Tamil Lessons, timely attendance to Tamil Lessons, on time submission of assignments, and improving results in both formative and summative assessments.

Informal techniques such as Checks For Understanding (during teaching), Written Reflections (End of Lesson), Polls (End of Module) and formal techniques such as Quizzes, Online Learning Modes along with the usual in-class activities and deliverables to evaluate individual students'

learning, provided me with the evidence that this approach has impacted students' learning. Embarking on the use of Augmented Reality was a natural progression for me in my continued quest to ignite students' interest in learning. I found out that Augmented Reality motivated and engagedmy students to learnwith genuine interest and commitment. Augmented Reality also helped to foster greater students' attention and satisfaction. In short, engaging Augmented Reality in the teaching and learning of Tamil Language truly inspired my learners.

2.5. How will students experience lessons involvingAugmented Reality?

Educators can usethe "Explain, Elaborate and Experience Approach" (EEEA)to generate a meaningful involvement of the learners.

2.5.1. Explain

Firstly, teachers can do a presentation to demonstrate the basic idea of Augmented Reality which is to superimpose graphics, audio and other sense enhancements over a real-world environment in real-time. They can further explain how Augmented Reality adds graphics and sound to the natural world and how it augments the real world scene in such a way that the user can maintain a sense of presence in that world. The user interacts with the real world and, at the same time, see both the real and virtual world co-existing – the user is not cut off from reality.

2.5.2. Elaboration

Studentsmay then be provided opportunities to explore and harness Augmented Reality from widely available tools online. After gaining exposure and trying out these online tools, students can be asked to share how Augmented Reality has truly changed the way they view the world, how the technology and its components loomed informative graphics and audio to coincide with whatever they see, and how it has enhanced their learning. Students will then share how Augmented Reality apps have become transportable and generally available on various devices. They will share their findingson how Augmented Reality is beginning to occupy its place in audio-visual media and used in various fields in our life in tangible and exciting ways such as news and sports. Above all, they will show how Augmented Reality is being used to facilitate the learning of Tamil Language.

2.5.3. Experience

This involves lessons where the students are taken through hands-on activities to better understand how Augmented Reality creates immersive, computer-generated environments for them to gain more experience. They will comprehend how using Augmented Reality creates the interface for learners to use these systems to learn more about a certain historical event. The hands-on activities will enable students to walk into a Civil War battlefield and see a recreation of historical events in panoramic view.From their own classroom, students can see The Majestic Taj Mahal 'appearing' right in front of them and experience a panoramic tour of it. It would immerse the learners in the learning session and they would alsoestablish better collaboration among one another.

3. Conclusion

It can be established that students' learning interest will certainly be ignited when using Augmented Reality. However, educators must be careful not to let the wide availability of Augmented Reality apps, which are mainly in English, to distract from the main aim of its use as a tool for learning of Tamil Language. There are so many Augmented Reality apps available online and, if not monitored, students may shift their interest to try out every one of them, for all these apps carry intriguing demonstrations and games. We need to alertstudents to be mindful of this. We need to partner with parents to create an awareness of how Augmented Reality can create a useful learning

environment. Once the students are au fait with the rudiments of the Augmented Reality, they can then progress to the next stage, where they will learn how to create their own Augmented Reality and this can be used to assign projects.

We can liaise with some of the leading Augmented Reality app owners to showcase quality student projects. If successful, thesestudents'projects will be available in Tamil Language for many to use and benefit. Moving on, students with entrepreneurial abilities may even develop and market their own Tamil Augmented Reality apps, thus maximizing the use of Augmented Reality in Tamil Language.

References

[1] Dede, C. "Immersive Interfaces for engagement and learning". (Jan 2, 2009) Science Vol 323 (591), 66-69.

[2] Dias, A. "Technology enhanced learning and augmented reality: An application on multimedia Interactive books. International Business & Economics Review", (2009). Vol 1(1).

[3] Stewart-Smith, H. "Education with augmented reality: AR Textbooks released in Japan", (2012) Znet.com.

[4] Wirzesien, M., &Alcanis Raya, M. "Learning in serious virtual worlds: Evaluation of learning effectiveness and appeal to students in the E-Junior project," (Jan 20, 2010) "Science Digest".

[5] http://educause.edu/ir/library/pdf/ELI7007.pdf :"7 Things you should know about Augmented Reality" (Oct 15, 2005). "Educause Learning Initiative".ID:ED17007)

[6] http://www.esquire.com/the-side/feature/augmented-reality-technology-110909 : "Augmented Reality technology-Esquire augmented reality practical guide." (Nov 9, 2009).

இலக்கணப் பிழைகளின்றி தமிழ் எழுதிட எட்மோடோ (Edmodo) வழிமெய்நிகர் கற்றல் கற்பித்தல் அணுகுமுறை

சு. புஷ்பராணி & இரா. மோகனதாஸ்

இந்திய ஆய்வியல் துறை, மலாயாப் பல்கலைக்கழகம், கோலாலம்பூர்

ஆய்வுச் சுருக்கம்

மொழியின் அடிப்படைக் கூறுகளில் இலக்கணம் மிகப் பெரிய பணியைச் செய்யவல்லது. முறையான இலக்கணமே ஒரு மொழியைச் சரியாகப் பேசிட, எழுதிட வழிவகுக்கும். தமிழ் மொழியினைத் தாய்மொழியாகக் கொண்டு தமிழ்ப் பயிலும் மாணவர்களிடையே இலக்கணப் பிழைகள் இன்றி தமிழ் எழுதிடும் ஆற்றல் மிகக் குறைவாகவே உள்ளது. இச்சிக்கலைக் களைய மெய்நிகர் கற்றல் கற்பித்தல் அணுகுமுறை கையாளப்பட்டுத் தீர்வுக்கான இவ்வாய்வு மேற்கொள்ளப்படுகின்றது. இவ்வாய்வுக்கு**எட்மோடோ (Edmodo)** வழி இணையத்தளம் வாயிலான மெய்நிகர் கற்றல் கற்பித்தல் முறை அணுகப்பட்டுள்ளது. இவ்வாய்வில்Online Collaborative Learning Theory (OCL)கோட்பாடுபயன்படுத்தப்பட்டுக் கற்றல் கற்பித்தல் முறை மேற்கொள்ளப்பட்டது.இவ்வாய்விற்காக மலேசியக் கல்வி அமைச்சின் ஆரம்பக் கல்விக்கான பாடநூல்கள் படிநிலை 1 மற்றும் படிநிலை 2 பயன்படுத்தப்பட்டு அதிலுள்ள இலக்கணப் பாடக் கல்வி மட்டுமே பயன்படுத்தப்பட்டுள்ளன. ஆய்விற்கு உட்படுத்தப்பட்ட மாணவர்களுக்கு மெய்நிகர் கற்றல் கற்பித்தல் 4 வாரங்கள் தொடர்ந்து எட்மோடோ வழி இருவழி தொடர்பில் போதிக்கப்பட்டது. பின்னர், மாணவர்களின் கட்டுரையில் காணப்படும் எழுத்துப்பிழைகள் **எட்மோடோ (Edmodo)** வழி மெய்நிகர் கற்றல் கற்பித்தல் ஆய்வுக்கு முன்னும் பின்னும் ஆராயப்பட்டுக்கலந்துரையாடப்பட்டுள்ளது. 10 மாணவர்களிடம் ஆய்வுக்கு முன் 5 கட்டுரைகளும் மெய்நிகர் கற்றல் கற்பித்தலுக்குப் பின் 5கட்டுரைகளும் வழங்கப்பட்டுத் திருத்தப்பட்டன. அக்கட்டுரைகளில் காணப்படும் எழுத்துப்பிழைகளின் சராசரி கண்டறியப்பட்டு ஆராய்விற்கு உட்புகுத்தப்பட்டன.

குறிச்சொல் :எட்மோடோ, இலக்கணப் பிழைகள், மெய்நிகர் கற்றல் கற்பித்தல், மொழி, இலக்கணம்

Keywords :Edmodo, Grammar mistakes, Virtual Learning, Language, Linguistic

1.0அறிமுகம்

இலக்கணப் பிழைகளின்றி தமிழ் எழுதும் ஆற்றல் மாணவர்களுக்குச் சிறு வயது முதற்கொண்டே கற்றுத் தரப்பட வேண்டிய ஒரு முக்கியக் கூறாகும். இலக்கணப் பிழைகளின்றி எழுதும் ஆர்வத்தைத் தூண்ட மெய்நிகர் கற்றல் கற்பித்தல் அணுகுமுறை சிறந்த களமாகப் பயன்படுகின்றது. மெய்நிகர் கற்றல் கற்பித்தல் முறை கல்வித் துறையில் பரவலாகக் காணப்படும் தற்கால போதனா முறை என்பது அனைவரும் அறிந்த ஒன்றே. www.internetworldstats.com/stats.htm -இன் கணக்கெடுப்பின்வழி 31.03.2017 வரை கணினியைப் பயன்படுத்துவோரின் எண்ணிக்கை உலக மக்கள் தொகையில் சராசரி 3 பில்லியன் என ஆய்வு காட்டுகின்றது. பரவலாகப் பயன்படுத்தப்பட்டு வரும் இணைய உலகில் பல்வேறு மாற்றங்களை உள்ளடக்கிய Web 2.0 தொழில்நுட்பம் கல்வித் துறைக்குப் பெரிதும் பங்காற்றி வருகின்றது.(Al-Kathiri, 2014)இத்தொழில்நுட்பத்தில் ஒன்றான எட்மோடோ எனப்படுவது இலவசமாகவும் பாதுகாப்பாகவும் பயன்படுத்தும் தளமாகும். இத்தளத்தின் வாயிலாக ஆசிரியர்கள் மாணவர்கள் என இரு தரப்பினரும் கற்றல் கற்பித்தல் நடவடிக்கையை மேற்கொள்ளலாம்.அதோடு, சமூக வலைத்தளங்களின் பங்கு அனைத்துத் துறைகளிலும் பெரும் அளவில் காணப்படுவதனால்அதனைக் கல்வியோடு தொடர்புக் கொண்டு மாணவர்கள் பயன்பெறும் வகையில் மாணவர்களின் எதிர்ப்பார்ப்புகளைப் பூர்த்தி செய்வது கற்றல் கற்பித்தல் முறையில் அவசியமான ஒன்றாகும்.(Foster & Neal, 2012)அதோடு, இன்றைய கல்வியாளர்கள் சமூக வலைத்தளங்களாகிய முகநூல், டிவிட்டர், யூடியூப், வலைப்பூக்கள், தெலிகிராம் போன்ற சமூக வலைத்தளங்களைக் கல்வியோடு இணைத்துவிட்டனர். (Balakrishnan & Gan, 2016)ஆகவே, இன்றைய காலத்தில் உடனிணைந்து பணியாற்றுதல், நுண்ணாய்வுச் சிந்தனை, படைப்பாக்கம், கருத்துப்பரிமாற்றம் ஆகியவனவற்றிற்கு ஏதுவாக இத்தளம் பயன்வழங்கி மாணவர்கள் இலக்கணப்பிழைகளை இன்றி தமிழ் எழுதிட உறுதுணையாகப்பயன் வழங்கியுள்ளது.

2.0செயல்முறை

2.1 Online Collaborative Learning Theory (OCL)

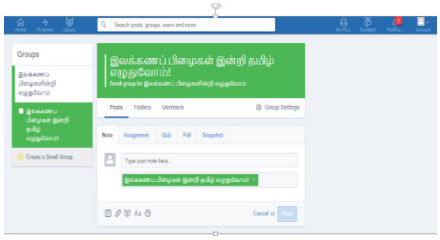
கற்றல் கற்பித்தலில் அணுகுமுறைகளைக் கல்வியல் உலகம் நான்கு பெரும் குழுமத்தில் இணைத்துள்ளன. அவை Behaviorism, Cognitivism, Humanism மற்றும் Contructivism ஆகும். அவற்றுள்புதிய அணுகுமுறையான **OCL**–பரவலாகக் கல்வித் துறையில் பயன்படுத்தப்பட்டு வருகிறது. இந்தக் கோட்பாடு மாணவர்கள் கற்றல் சவாலை உடனிணைந்து களைவதற்கு முக்கியத்துவம் வழங்குகின்றது. இந்த அணுகுமுறை 2012 ஆம் ஆண்டு லிண்டா ஹரசிம் என்பவரால் அறிமுகப்படுத்தப்பட்டது. இந்த அணுகுமுறை முற்றிலும் கணினி வழி தொடர்பு மற்றும் சமூக தொடர்பு ஒன்றிணைந்த கற்றல் கற்பித்தலை வழியுறுத்தும் கோட்பாடாகும். இக்கோட்பாடு உடனிணைந்து பணியாற்றுகையில் மூன்று வகையாக அறிவு மேம்படுகிறது என்பதனை விவரிக்கின்றது. அவையானவை, தகவலை உருவாக்குதல், தகவலை ஒருங்கிணைத்தல் மற்றும் மூன்றாவது கட்டமாக முன்னறிவைத் திரட்டிப் புதிய தகவலைப் படைத்தல் ஆகியன ஆகும்.இக்கோட்பாட்டின் நன்மையானது கூட்டிணைக் கற்றல்நடவடிக்கையோடு, நுண்ணாய்வு சிந்தனை ஆற்றல், பகுப்பாய்வு, மதிப்பிடல் ஆகிய உயர்நிலை சிந்தனையாற்றலும் வளர உதவுகின்றது.

2.2 வடிவமைப்பு

இவ்வாய்விற்கு உட்படுத்தப்பட்ட 10 மாணவர்களுக்கு 4 வாரங்களுக்கு மெய்நிகர் கற்றல் கற்பித்தல் முறை போதிக்கப்பட்டது. இம்மாணவர்களுக்கு இணையம் வழி எட்மோடோவைப் பயன்படுத்தி கல்வி கற்கும் முறை முதற்கட்டமாகப் போதிக்கப்பட்டது. மெய்நிகர் கற்றல் கற்பித்தலுக்கு முன்னரும் பின்னரும் ஒரே தலைப்பிலான கட்டுரைகள் வழங்கப்பட்டன. மெய்நிகர் கற்றல் கற்பித்தலுக்கு முன் திருத்தப்பட்ட கட்டுரைகள் மாணவர்களிடம் மீண்டும் வழங்கப்படவில்லை. இதன் மூலம் மெய்நிகர் கற்றல் கற்பித்தலுக்குப் பின் மாணவர்களின் இலக்கணப் பிழைகளின் மாற்றங்களை எளிதில் அடையாளம் காண முடிந்தது.

2.3எட்மோடோ அமைப்பு முறை

மாணவர்கள் சுயமாக இத்தளத்தில் பதிய இயலாது. ஆசிரியர் வழங்கும் கடவுச்சொல்லைப் பயன்படுத்தி மட்டுமே இத்தளத்தில் உறுப்பினராகப் பதிய முடியும். இத்தளத்தில் மாணவர்கள் உறுப்பினராகப் பதிய மின்னஞ்சல் முகவரி தேவையில்லை.



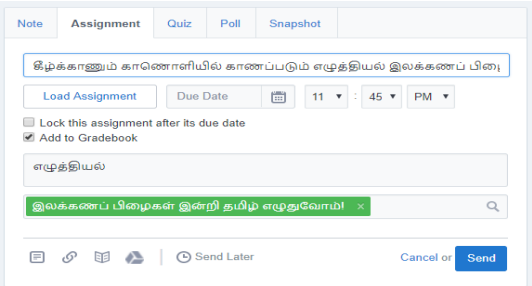
படம் 1 : எட்மோடோ இணையத்தளம்

I. Note

எளிய இலக்கணப் பாட முறை மாணவர்களைக் கவரும் வண்ணப் படக்காட்சிகள், காணொளிகள், மின்னூல் ஆகிய வகையில் தயாரிக்கப்பட்டு மாணவர்களுக்கு வழங்கப்பட்டன. மாணவர்கள் பாடத்தை எட்மோடோ வழிச் சுயமாகக் கற்றனர். அவர்கள் கற்ற இலக்கணப் பாடங்களை எட்மோடோ குழுவில் பதிவேற்றம் செய்தனர். இலக்கணப் பாடத்தில் ஏற்படும் ஐயங்களைக் களையமாணவர்களுக்கு ஆசிரியர் மற்றும் சக நண்பர்கள்விளக்கம் அளித்தனர்.ஆசிரியர் மாணவர்களுக்கிடையிலான உரையாடல் மாணவர்களுக்குத் தூண்டுகோலாகவும் சக நண்பர்களின் பாட சம்மந்தமான சிக்கல்களைக் களைய ஏதுவாகவும் பயன்பட்டது. உடனிணைந்து செயலாற்றுதல் மற்றும் கருத்துப்பரிமாற்றம் செய்திட இப்பக்கம் உறுதுணையாக இருந்தது.

II. Assignment

ஒவ்வொரு இலக்கணக் கூறுகளைக் கற்றப்பின் மாணவர்கள் ஆசிரியரால் வழங்கப்பட்ட பணியை முறையாகச் செய்து பதிவேற்றம் செய்தனர். உதாரணத்திற்கு மின்நாளிதழ்களில் காணப்படும் இலக்கணக் பிழைகளை அடையாளம் காணுதல். பாடல், கதைகள், வானொலி தொலைக்காட்சி நிகழ்ச்சிகள், அறிவிப்புகள் ஆகிய காட்சிகள் எழுத்து வடிவில் இணைக்கப்பட்டு அதில் காணப்படும் இலக்கணக் கூறுகள் சரிவர அடையாளம் கண்டு பட்டியலிட்டுப் பதிவேற்றம் செய்யக்கூடிய பயிற்சிகளும் வழங்கப்பட்டன. அதோடு,வழங்கப்பட்ட படக்காட்சிகளுக்கு ஏற்றகதைகளை உருவாக்கிப் பதிவேற்றம் செய்தனர். சக மாணவர்கள்மற்றவர் படைப்பில் காணப்படும் இலக்கணப் பிழைகளைக் களைய உதவினர். இந்தப் பணிகளை ஆசிரியர் தனிநபர் முறை அல்லது குழு முறை எனக் குறிப்பிட்ட கால அவகாசத்துடன் நிர்ணயித்து வழங்க இப்பக்கம் பயன் வழங்கியது.



படம் 2 : எட்மோடோ வழி பணி உருவாக்கம்

III. Quiz

ஆசிரியர் மாணவர்களுக்கு உயர்நிலை சிந்தனை கேள்விகள் வழங்கிக் குறிப்பிட்ட மணி நேரத்திற்குள் கேள்விகளுக்கான பதிலை அளிக்கும் திறனைச் சோதித்தல். மாணவர்கள் உயர்நிலை சிந்தனை கேள்விகளுக்கான பதில்களைத் தனிநபர் முறையில் அளித்தல். பல்வகை தேர்வு, சரிபிழை, குறுகிய விடை, கோடிட்ட இடத்தை நிரப்புதல், இணைத்தல் போன்ற கேள்வி அணுகுமுறைகள் உருவாக்கிட இப்பக்கம் உதவிற்று.

IV. Poll (கருத்துக் கணிப்பு)

மாணவர்கள் கேள்விகளுக்கான தங்கள் பதில்களைப் பகிர்ந்து கொள்ள ஏதுவான களமாக இது பயன்பட்டது. இதனால், மொத்தம் எத்தனை மாணவர்கள் சரியான பதிலை அளித்துள்ளனர், எத்தனை மாணவர்கள் தவறான பதில் அளித்துள்ளனர் எனக் கண்டறிந்து மாணவர்களுக்கு மேலும் விளக்கமளிக்க இத்தளம் பயன் வழங்கிற்று. மாணவர்களின் அடைவு நிலையை உறுதி செய்ய உதவியது.

Reward

சிறந்த அடைவுநிலையை அடைந்த மாணவர்களுக்கு வெகுமதி வழங்கும் நோக்கில், சிறந்த மாணவர் வாரம் ஒரு முறை தேர்ந்தெடுக்கப்பட்டுப் பாராட்டைப் பெற்றது மாணவர்கள் கற்றல் கற்பித்தல் நடவடிக்கைகளில் மேலும் ஆர்வத்தைத் தூண்ட துணை புரிந்தது.

3.0 ஆய்வு முடிவு

அட்டவணை 1, அட்டவணை 2, அட்டவணை 3

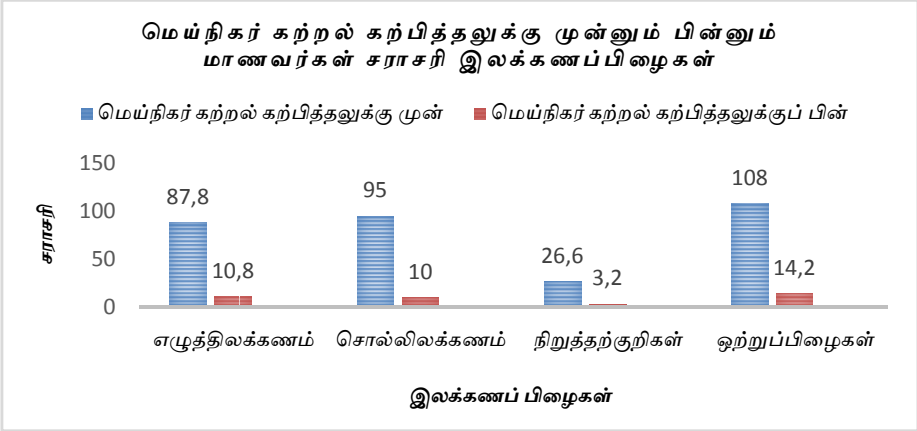
அட்டவணை 3 : எட்மோடோ (Edmodo) வழி மெய்நிகர் கற்றல் கற்பித்தலுக்கு முன்னும் பின்னும் மாணவர்களின் சராசரி இலக்கணப் பிழைகள்.

இலக்கணப்பிழைகள்	மெய்நிகர் கற்றல் கற்பித்தலுக்கு முன்னும் பின்னும் மாணவர்களின் சராசரி இலக்கணப் பிழைகள்	
	மெய்நிகர் கற்றல் கற்பித்தலுக்கு முன்	மெய்நிகர் கற்றல் கற்பித்தலுக்குப் பின்
எழுத்திலக்கணம்	87.8	10.8
சொல்லிலக்கணம்	95	10
நிறுத்தற்குறிகள்	26.6	3.2
ஒற்றுப்பிழைகள்	108	14.2

அட்டவணை 3

4.0 கலந்துரையாடல்

வகுப்பறை சூழல் போதனா முறையினைக் காட்டிலும் மெய்நிகர் கற்றல் கற்பித்தல் அணுகுமுறை மாணவர்களுக்குச் சிறந்ததொரு கற்றல் அனுபவத்தைத் தந்ததோடு, மாணவர்களின் கற்றல் ஆர்வத்தை மேலும் அதிகரித்துள்ளது. தனிநபர் முறையில் மாணவர்களின் கற்றல் ஐயங்களைக் களைய மேற்கொண்ட அணுகுமுறையும் சக நண்பர்களின் கலந்துரையாடல் ஆகியன மாணவர்களின் இலக்கணப் பிழைகளை நிவர்த்தி செய்ய வழிவகுத்தன.



பட்டைக் குறிவரைவு 1

5.0 முடிவு

மெய்நிகர் கற்றல் கற்பித்தல் மாணவர்களின் இலக்கணப் பிழைகளைக் களைவதில் பெரும் பங்காற்றியுள்ளது. இலக்கணப் பிழைகள் இன்றி எழுத மேற்கொண்ட ஆய்வின் நோக்கம் நிறைவேறியது. இவ்வாய்வின் இறுதியில் எட்மோடோ வழி கற்றல் கல்வி பெற்ற மாணவர்களின் கட்டுரைகளில் எழுத்துப்பிழைகள் இன்றிஎழுதும் ஆற்றல் ஆழக் காணப்பட்டது. அதோடு, மாணவர்களிடையே மேலும் மெய்நிகர் கற்றல் கற்பித்தல் அணுகுமுறையைப் பின்பற்றிக் கல்வி கற்கும் ஆர்வம், தமிழ் மொழியினைப் பிழைகளின்றி எழுதும் ஆற்றலை வளர்த்துள்ளது.

References

Al-Kathiri, F. (2014). Beyond the Classroom Walls: Edmodo in Saudi Secondary School EFL Instruction, Attitudes and Challenges.

Balakrishnan, V., & Gan, C. L. (2016). Students' learning styles and their effects on the use of social media technology for learning. *Telematics and Informatics, 33*(3), 808-821. doi: http://dx.doi.org/10.1016/j.tele.2015.12.004

Balakrishnan, V., Liew, T. K., & Pourgholaminejad, S. (2015). Fun learning with Edooware – A social media enabled tool. *Computers & Education, 80*, 39-47. doi: http://dx.doi.org/10.1016/j.compedu.2014.08.008

Balasubramanian, K., Jaykumar, V., & Fukey, L. N. (2014). A Study on "Student Preference towards the Use of Edmodo as a Learning Platform to Create Responsible Learning Environment". *Procedia - Social and Behavioral Sciences, 144*, 416-422. doi: http://dx.doi.org/10.1016/j.sbspro.2014.07.311

Berns, A., Gonzalez-Pardo, A., & Camacho, D. (2013). Game-like language learning in 3-D virtual environments. *Computers & Education, 60*(1), 210-220. doi: http://dx.doi.org/10.1016/j.compedu.2012.07.001

Falloon, G. (2015). What's the difference? Learning collaboratively using iPads in conventional classrooms. *Computers & Education, 84*, 62-77. doi: https://doi.org/10.1016/j.compedu.2015.01.010

Foster, R., & Neal, D. R. (2012). 12 - Learning social media: student and instructor perspectives *Social Media for Academics* (pp. 211-226): Chandos Publishing.

Gan, B., Menkhoff, T., & Smith, R. (2015). Enhancing students' learning process through interactive digital media: New opportunities for collaborative learning. *Computers in Human Behavior, 51, Part B*, 652-663. doi: https://doi.org/10.1016/j.chb.2014.12.048

Grosseck, G., & Holotescu, C. (2010). Microblogging multimedia-based teaching methods best practices with Cirip.eu. *Procedia - Social and Behavioral Sciences, 2*(2), 2151-2155. doi: http://dx.doi.org/10.1016/j.sbspro.2010.03.297

Holotescu, C., Grosseck, G., & Danciu, E. (2014). Educational Digital Stories in 140 Characters: Towards a Typology of Micro-blog Storytelling in Academic Courses. *Procedia - Social and Behavioral Sciences, 116*, 4301-4305. doi: http://dx.doi.org/10.1016/j.sbspro.2014.01.936

Ng, W. (2012). Can we teach digital natives digital literacy? *Computers & Education, 59*(3), 1065-1078. doi: http://dx.doi.org/10.1016/j.compedu.2012.04.016

Phungsuk, R., Viriyavejakul, C., & Ratanaolarn, T. Development of a problem-based learning model via a virtual learning environment. *Kasetsart Journal of Social Sciences*. doi: https://doi.org/10.1016/j.kjss.2017.01.001

Seralidou, E., & Douligeris, C. (2015). Identification and Classification of Educational Collaborative Learning Environments. *Procedia Computer Science, 65*, 249-258. doi: http://dx.doi.org/10.1016/j.procs.2015.09.073

Won, S. G. L., Evans, M. A., Carey, C., & Schnittka, C. G. (2015). Youth appropriation of social media for collaborative and facilitated design-based learning. *Computers in Human Behavior, 50*, 385-391. doi: https://doi.org/10.1016/j.chb.2015.04.017

Zhang, M. (2014). Who are interested in online science simulations? Tracking a trend of digital divide in Internet use. *Computers & Education, 76*, 205-214. doi: http://dx.doi.org/10.1016/j.compedu.2014.04.001

Websites

https://www.learning-theories.com/online-collaborative-learning-theory-harasim.html

www. edmodo.com

அட்டவணை 1 : மெய்நிகர் கற்றல் கற்பித்தல் முறைக்கு முன்னர் மாணவர்களின் கட்டுரைகளில் காணப்பட்ட எழுத்துப்பிழைகளின் எண்ணிக்கை

	மெய்நிகர் கற்றல் கற்பித்தலுக்கு முன் மாணவர்களின் கட்டுரைகளில் காணப்படும் எழுத்துப்பிழைகளின் எண்ணிக்கை																			
	தமிழ் மொழியின் சிறப்பு				காடுகளின் பயன்கள்				வாசிப்பின் பயன்				செம்மொழி சான்றோர்கள்				நீர்த் தூய்மைக்கேடு			
மாணவர்கள்	எ.பி	சொ.பி	தி.கு	ஒ.பி	எ.பி	சொ.பி	தி.கு	ஒ.பி	எ.பி	சொ.பி	தி.கு	ஒ.பி	எ.பி	சொ.பி	தி.கு	ஒ.பி	எ.பி	சொ.பி	தி.கு	ஒ.பி
மாணவர் 1	5	15	4	9	4	14	0	5	6	6	2	8	7	9	5	16	8	9	4	14
மாணவர் 2	5	14	5	4	5	15	1	4	9	6	2	5	13	20	4	5	6	14	0	5
மாணவர் 3	7	8	5	15	4	8	4	19	2	7	5	15	7	9	2	18	6	8	1	16
மாணவர் 4	8	9	4	10	8	1	2	14	21	6	2	12	7	15	2	18	19	7	5	17
மாணவர் 5	5	11	4	14	11	16	5	16	15	7	5	8	10	9	4	15	7	10	4	15
மாணவர் 6	9	21	5	11	9	6	4	14	2	8	2	5	8	17	4	9	6	7	1	8
மாணவர் 7	9	6	0	7	7	9	0	8	14	5	0	7	9	12	5	7	14	5	1	15
மாணவர் 8	11	14	5	18	6	9	4	15	12	7	2	14	17	14	4	16	13	12	5	13
மாணவர் 9	15	8	5	9	15	8	4	15	10	2	2	5	8	15	4	15	24	10	2	10
மாணவர் 10	0	5	0	2	6	8	5	7	4	6	0	8	8	6	2	7	6	8	0	8

அட்டவணை 2 : மெய்நிகர்|கற்றல் கற்பித்தல் முறைக்குப் பின்னர் மாணவர்களின் கட்டுரைகளில் காணப்பட்ட எழுத்துப்பிழைகளின் எண்ணிக்கை

	மெய்நிகர் கற்றல் கற்பித்தலுக்குப் பின் மாணவர்களின் கட்டுரைகளில் காணப்படும் எழுத்துப்பிழைகளின் எண்ணிக்கை																			
	தமிழ் மொழியின் சிறப்பு				காடுகளின் பயன்கள்				வாசிப்பின் பயன்				செம்மொழி சான்றோர்கள்				நீர்த் தூய்மைக்கேடு			
மாணவர்கள்	எ.பி	சொ.பி	தி.கு	ஒ.பி	எ.பி	சொ.பி	தி.கு	ஒ.பி	எ.பி	சொ.பி	தி.கு	ஒ.பி	எ.பி	சொ.பி	தி.கு	ஒ.பி	எ.பி	சொ.பி	தி.கு	ஒ.பி
மாணவர் 1	1	0	0	1	1	0	0	3	1	1	0	2	0	2	0	2	2	2	2	1
மாணவர் 2	2	1	0	0	0	1	0	1	1	0	1	0	0	1	0	1	1	1	0	4
மாணவர் 3	1	2	0	3	2	2	0	2	0	2	1	1	0	0	0	1	1	3	0	1
மாணவர் 4	1	1	0	2	1	1	1	4	1	1	1	1	1	1	1	0	0	2	0	2
மாணவர் 5	3	2	0	1	1	3	0	1	0	0	0	3	2	0	0	0	1	1	0	0
மாணவர் 6	1	1	0	1	3	1	0	2	0	1	1	1	1	0	0	3	2	1	0	0
மாணவர் 7	2	0	0	2	1	2	0	1	1	0	1	2	4	0	0	2	1	0	0	1
மாணவர் 8	0	0	0	1	1	1	2	1	0	0	2	0	1	1	0	3	2	1	1	1
மாணவர் 9	1	2	0	0	0	0	0	1	1	1	0	0	2	1	0	3	1	0	0	3
மாணவர் 10	0	1	0	1	1	2	0	2	0	1	2	0	3	2	0	2	1	1	0	1

தமிழ் கற்றல் கற்பித்தலில் 21ஆம்நூற்றாண்டுத் தகவல் தொழில்நுட்ப மதிப்பீடு : *குயிசிஸ்* (Quizizz)

சிவபாலன் திருச்செல்வம்[1] & மோனேஸ்ரூபிணி தியாகராஜன்[2]

[1] தேசியவகைஇந்துவாலிபசங்கத்தமிழ்ப்பள்ளி, பேரா, மலேசியா

[2] சுங்கைபோகாத்தோட்டத்தமிழ்ப்பள்ளி, பேரா, மலேசியா

[1] shivabalanthiruchelvam@gmail.com [2]monesrubini@gmail.com

---

1.0 முன்னுரை

உலக நகர்ச்சிக்கும் சமூக வளர்ச்சிக்கும் ஏற்ப அவ்வப்பொழுது கற்றல் கற்பித்தலில் மாற்றம் ஏற்படுவதும், புத்தம் புதிய சிந்தனைகள் உட்புகுத்தப்படுவதும் அவசியமாகின்றது. உலகோடு ஒட்ட வாழ்தல் மட்டுமன்றி, உலகத் தேவைக்கேற்பவும் வாழச்செய்யும் ஆற்றலைக்கொடுக்கும் கல்வியே வாழும் களத்திற்கும் காலத்திற்கும் ஏற்புடையதாக அமையும் (நாராயணசாமி, 2012). இத்தகைய கல்வியே தனிமனிதனுக்கும், குடும்பத்திற்கும் சமூகத்திற்கும், நாட்டிற்கும் பயன்மிக்க பங்கினை ஆற்றதுணை நிற்கும். இந்தக் கல்வியின் முழுமையான விளைபயனை அறிய அதற்கான மதிப்பீடு மிக முக்கியஇடத்தினை வகிக்கின்றது. இந்த மதிப்பீடானது மிகத்துள்ளியமாகவும் உடனுக்குடனும் நடத்தப்படுவது அவசியமாகின்றது. அதனை அவ்வாறு செய்ய இயலாத ஒரு சூழ்நிலை ஏற்படுகின்றபோது, 21ஆம் நூற்றாண்டுத் திறன்களில் ஒன்றான தகவல் தொடர்புத் தொழில்நுட்பத்தைக் கொண்டு எவ்வாறு அதனைச் சாத்தியப்படச் செய்வது என்பதனை விளக்கவல்லதே இவ்வாய்வு.

2.0 ‘குயிசிஸ்’ (Quizizz)

ஒரு பாடத்தின் இறுதியில் அதன் விளைபயனை அறிவதற்காக மேற்கொள்ளப்படும் மதிப்பீட்டை மாணவர்களுக்கு மனமகிழ்வாகவும் ஆசிரியர்களுக்கு எளிமையாகவும், துள்ளியமாகவும் தொழில்நுட்பஉதவியுடன் மேற்கொள்ள வழிவகுப்பதுதான் இந்த ‘குயிசிஸ்’ புதிர்வகை மதிப்பீடு. முற்றிலும் இலவயமாகப் பயன்படுத்தும் வகையில் அமைக்கப்பட்டிருக்கும் இந்த மென்பொருள், திறன்பேசிகளில் பயன்படுத்தும் வண்ணம் ஆன்டிராய்டு, ஆப்பிள் அங்காடிகளிலும் செயலியாக இருக்கின்றது. இந்த மென்பொருளில் பயனராக மட்டும் செயல்படாது, நமது தேவைக்கேற்ப உள்ளீடு செய்பவராகவும் பங்காற்றலாம். இது ஆங்கிலத்தில் உருவாக்கப்பட்ட மென்பொருள் என்றாலும், தமிழ்க் கல்வி மதிப்பீட்டிற்கு இது முழுமையான பங்கினை ஆற்றவல்லது.

3.0 ஆய்வுக்குரியசிக்கல் :

3.1 மாணவர்களின் தனியாள் அடைவு நிலையைத் துள்ளியமாக அறிவதில் ஆசிரியர்களுக்குச் சிக்கல் ஏற்படுகின்றது.

3.2 மாணவர்களின் தனியாள் அடைவுநிலையை உடனுக்குடன் அறிவதில் ஆசிரியர்களுக்குச் சிக்கல் ஏற்படுகின்றது.

4.0 ஆய்வின்நோக்கம் :

4.1 மாணவர்களின் தனியாள் அடைவுநிலையைக் 'குயிசிஸ்' வழி துள்ளியமாகக் கண்டறிதல்.

4.2 மாணவர்களின் தனியாள் அடைவுநிலையைக் 'குயிசிஸ்' வழி உடனுக்குடன் கண்டறிதல்.

5.0 ஆய்வின்வினா :

5.1 மாணவர்களின் தனியாள் அடைவுநிலையைக் 'குயிசிஸ்' வழி துள்ளியமாகக் கண்டறிய இயலுமா?

5.2 மாணவர்களின் தனியாள் அடைவுநிலையைக் 'குயிசிஸ்' வழி உடனுக்குடன் கண்டறிய இயலுமா?

6.0 ஆய்வுக்குட்பட்டோர்

மலேசியாவின் பேரா மாநிலத்தின் லாருட்மாத்தாங் செலாமா, கிரியான் ஆகிய இரு மாவட்டங்களைச் சார்ந்த 34 தமிழ்ப்பள்ளி ஆசிரியர்கள்.

7.0 ஆய்வின்முறைமை :

இந்த ஆய்வுபண்புசார், அளவுசார் ஆகிய இரு முறைமைகளையும் உள்ளடக்கி நடத்தப்பட்டது. 'குயிசிஸ்' பயன்பாட்டிற்கு முன்பும் 'குயிசிஸ்' பயன்பாட்டிற்குப் பின்பும் தமிழ் மொழி கற்றல் கற்பித்தல் மதிப்பீடு மீது ஆசிரியர்களின் பார்வையையும் பயன்பாட்டையும் இந்த இரு முறைமைகளும் கூறி நிற்கின்றன.

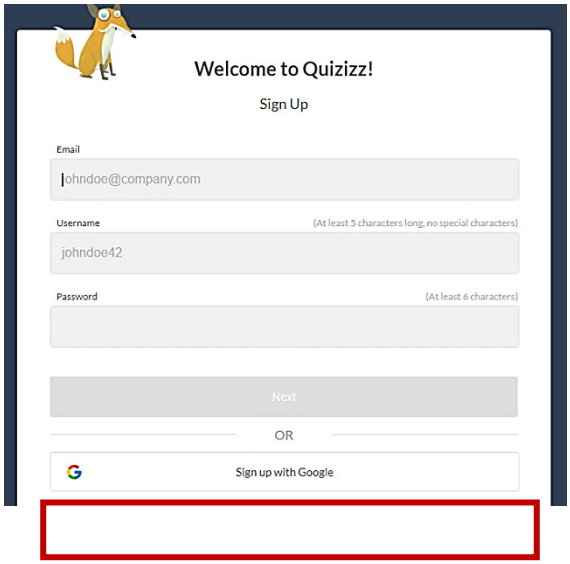
8.0 ஆய்வின்அமலாக்கம் (திட்டப்பணி) :

இந்தப் புதிர்வகை மதிப்பீடான 'குயிசிஸ்'ல் இணைய https://quizizz.com/admin என்ற இணையமுகவரியைப் பயன்படுத்துதல் வேண்டும். பின் இதன் முழுமையான பயனைப்

பெற சரியான செயல்முறைகளைக் கையாள்வது அவசியமாகின்றது. தொடர்ந்து அதன் செயல்முறை விளக்கத்தினைக் காண்போம்.

8.1 சுயக்கணக்கைஉருவாக்குதல்

இந்த மென்பொருளைப் பயன்படுத்த முதலில் அதன் பக்கத்தில் நமது சுயக்கணக்கை உருவாக்குதல் வேண்டும்.

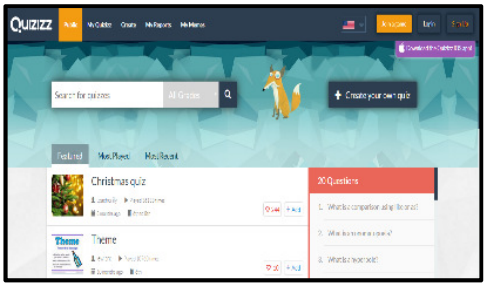


படி1 :தங்கள்சுயவிவரங்களைக்கொண்டகணக்கைஉருவாக்குதல்.

கூகுள் கணக்கு வைத்திருப்பவர்கள் தனியாக ஒரு கணக்கை உருவாக்க வேண்டிய அவசியமன்று. நமது கூகுள் கணக்கை இந்த மென்பொருளுடன் ஒருங்கிணைத்தால் போதுமானது.

8.2 வினாக்களைஉருவாக்குதல் / தேர்ந்தெடுத்தல்

இந்த மென்பொருளை ஏற்கனவே பயன்படுத்திய பயனர்கள்தங்களுக்குத் தேவையான வினாக்களைத் தயார்செய்து இங்கு உள்ளீடு செய்திருப்பார்கள். நமது கற்றல் கற்பித்தலுக்கு ஏற்ற வினாக்களாக அது இருக்கும் பட்சத்தில் நாம் அதையே நமது மதிப்பீட்டிற்குப் பயன்படுத்தலாம். இல்லையேல், நமது தலைப்பிற்கு ஏற்பவும், மாணவர்களின் நிலைக்கு ஏற்பவும் நமக்கான வினாக்களை நாமே மிக எளிமையாக உருவாக்கலாம்.முன்பே கூறியது போன்று, இந்த மென்பொருளில் பயனராக மட்டும் அல்லாது நாம் உள்ளீடு செய்பவராகவும் பங்காற்றலாம். எனவே, நாம் உருவாக்கிய இந்த வினாக்கள் பின்னாளில் மற்ற பயனர்களால் பயன்படுத்தப்படும்.

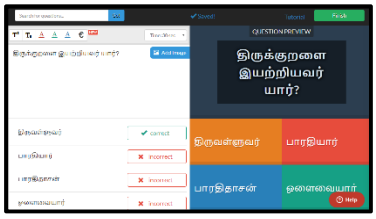


படி2 :

உங்களுக்கான வினாக்களை உருவாக்குதல் அல்லது தேர்ந்தெடுத்தல். தேர்ந்தெடுக்க :*search my quizziz* உருவாக்க :*create own*

படி3 :

வினா உருவாக்க முதலில் வினா தலைப்பு, வினா முகப்புப்படம், மொழி ஆகியவற்றை உள்ளிடுதல்

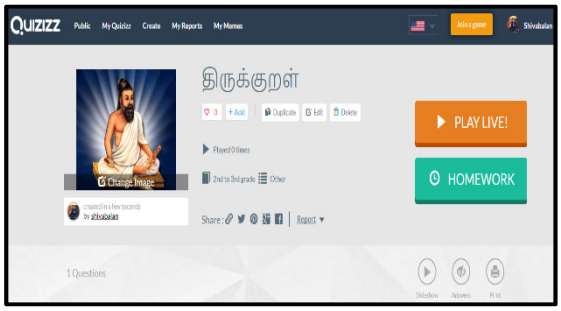


படி4 :

உங்கள் வினாக்களை எண்களுக்கேற்ப உருவாக்குதல். அதற்கான விடைகளை உள்ளிடுதல் (ஒரு சரியான, மூன்று பிழையான விடைகள்).

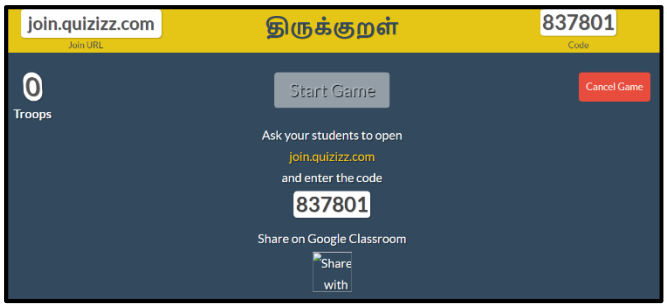
8.2 மாணவர்களை இணைத்தல்

வினாக்களை உருவாக்கியப்பின் மாணவர்களை ஒரே நேரத்தில் தனியாள் முறையில் இந்தப் பக்கத்திற்குள் நுழையச் செய்தல் வேண்டும். நாம் உருவாக்கிய வினாப் புதிருக்கு ஒரு கடவுஎண் வழங்கப்படும். அதைக்கொண்டு மாணவர்கள் இந்தத் தளத்தில் நமது கணக்கோடு இணையலாம்.



படி5 :

வினாப்புதிர் தயார் ஆனதும் கடவுஎண்பெற *play live* குறியைச் சொடுக்குதல்.



படி6 :

கடவுஎண் ஆசிரியரின் முகப்பில் தோன்றியதும் மாணவர்கள் *join game* வழிபுதிரில் இணைதல். அனைத்து மாணவர்களும் இணைந்தப்பின் புதிரைத் தொடங்க *Start Game* சொடுக்குதல்.

அனைத்து மாணவர்களும் இணைந்தபின், ஆசிரியர் புதிரைத் தொடங்கலாம். மாணவர்கள் ஒவ்வொரு வினாவிற்கும் பதிலளிக்கும் நேரத்தையும் ஆசிரியர் கட்டுப்படுத்தலாம். மாணவர்கள் வினாக்களுக்கு விடையளித்து முடித்த உடனேயே மாணவர்களின் புள்ளிகள் ஆசிரியரின் முகப்பில் தெரியவரும். இதைக்கொண்டு அன்றையக் கற்றல் கற்பித்தலின் நோக்கத்தை அடைந்த / அடையா மாணவர்களை ஆசிரியர் மிகவிரைவாகவும் துள்ளியமாகவும் கண்டறிய இயலும்.

9.0 ஆய்வின்பகுப்பாய்வு :

ஆய்வின் பகுப்பாய்வு ஆசிரியர்களிடம் மேற்கொண்ட நேர்காணல் மூலம் பண்புசார் அளவிலும், வினாநிரல் மூலம் அளவுசார் அளவிலும் மேற்கொள்ளப்பட்டது.

	நோக்கங்கள்	முன்		பின்	
		ஆம்	இல்லை	ஆம்	இல்லை
1	குயிசிஸ்வழி துள்ளியமாகக் கண்டறிய இயல்கிறது	3	31	34	0
2	குயிசிஸ்வழி உடனுக்குடன் கண்டறிய இயல்கிறது	5	29	34	0

10.0 ஆய்வின்முடிவு :

10.1 மாணவர்களின் தனியாள் அடைவுநிலையைக் ‘குயிசிஸ்’ வழி துள்ளியமாகக் கண்டறிய இயலும்.

10.2 மாணவர்களின் தனியாள் அடைவுநிலையைக் ‘குயிசிஸ்’ வழி உடனுக்குடன் கண்டறிய இயலும்.

11.0 முடிவுரை

இந்த 'குயிசிஸ்' புதிர்வகை மென்பொருள் எந்த நேரத்திலும் எந்த இடத்திலும் எந்தக் கருவியைக் கொண்டும் பயன்படுத்தும் வகையிலும் பயனடையும் வகையும் உருவாக்கப் பட்டுள்ளது. இந்த 'குயிசிஸ்' போன்றே இன்னும் பல அரிய மென்பொருள்கள், குறிப்பாகத் தமிழ்க் கல்விக்கு ஏதுவான மென்பொருள்கள் இன்று அதிக அளவில் காணப்படுகின்றன. இந்த மென்பொருள்களை ஆசிரியப் பெருமக்கள் சரியாகப் பயன்படுத்தி நிறைவான பயனை அடைதல் வேண்டும். இந்த வகை 21ஆம் நூற்றாண்டுத் தகவல் தொழில்நுட்பத்தைக் கற்றல் கற்பித்தலில் ஒருங்கிணைப்பதானது ஆசிரியர்கள், மாணவர்களைத் தாண்டி தமிழுக்கும் தமிழ்க் கல்விக்கும் மிகப்பெரிய நன்மையைப் பயப்பதாய் அமைகின்றது என்பதில் மாற்றுக் கருத்தில்லை.

12.0 மேற்கோள்பட்டியல்

1. சிவபாலன், தி.(2016, அக்டோபர் 18). *தமிழ்க் கல்வியும் 21ஆம் நூற்றாண்டுத் திறன்களும்*: ஒருபார்வை. எடுத்தாளப்பட்டது மார்ச் 5 2017, மூலம்
2. சுப்புரெட்டியார், ந) .2002). *தமிழ் பயிற்றும் முறைகள்*. சிதம்பரம் மெய்யப்பன் : தமிழாய்வகம்.
3. மலேசியக்கல்வி அமைச்சு. (2016). *தமிழ்ப் பள்ளிக்கான சீரமைக்கப்பட்ட தமிழ் மொழிக் கலைத்திட்டத் தர மற்றும் மதீப்பீட்டு ஆவணம் :ஆண்டு 1*. மலேசியக் கல்வி அமைச்சு.
4. Jamalludin Harun. (2003). *Multimedia dalam Pendidikan.* Kuala Lumpur: PTS Publications.
5. https://quizizz.com

ஊடாடல், நகர்ப்படங்கள் கலந்த மின்னூல்கள்வழிக் குழந்தைகளுக்கானத் தமிழ்க்கல்வி

கஸ்தூரி இராமலிங்கம்
khasturiam@gmail.com

முன்னுரை

மழலையர்களும் தொடக்கப் பள்ளி மாணவர்களும், தமிழ்க்கதைகளை எளிதாகவாசிப்பதற்கும் அவற்றோடு ஒன்றுவதற்கும், மின்னூல்களை எவ்வாறு உருவாக்கலாம் என்பதை ஆராய்வதே இக்கட்டுரையின் நோக்கம். வாசிப்பதோடு நின்று விடாமல், கதைகளில் வரும் பெயர்களையும் சொற்களையும் நினைவில் வைத்துக்கொள்ளும் வகையில், நூலிலேயே துணை நடவடிக்கைகளைச் சேர்ப்பது பற்றியும் இக்கட்டுரை ஆராய்கிறது.

2011-ஆம் ஆண்டு மலாயாப் பல்கலைக்கழகத்தில் நடந்த"கற்றல் கற்பித்தலில் புதிய சிந்தனைகள்" பன்னாட்டு மாநாட்டில் கணிஞர் முத்து நெடுமாறன் படைத்த"கையடக்கக் கருவிகளில் தமிழ் மின்னூல் உருவாக்கம்'என்ற தலைப்பிலான கட்டுரை, இந்த ஆய்வினை மேற்கொள்ள வித்தாக அமைந்தது. இக்கட்டுரையில்,ஐ-பேட்-இல் இ-பப் வடிவிலான மின்னூல்களை வாசிப்பதற்கு ஐ-புக் (iBook) எனும் செயலி உள்ளதைச் சுட்டிக்காட்டினார். இந்தச் செயலியின் வழி வாசிக்கப்படும் நூல்களில்,'உரக்க வாசிக்கும்' (read aloud) ஏந்து (வசதி) உள்ளது. மின்னூலில் உள்ள வரிகளை ஒலிப்பதிவு செய்து நூலோடு சேர்க்கலாம். நூலைத் திறந்து ஐ-புக் செயலியை அதில் உள்ள வரிகளை வாசிக்கச் சொல்லலாம். தமிழ் உச்சரிப்பைச் சரிவரக் கற்பிப்பதற்கு இந்த ஏந்து பேருதவியாக இருக்கும் என்று கூறிய அவர், மின்னூலின் ஈர்ப்புத் தன்மையைக் கூட்டுவதன் வழி, மாணவர்கள் நூலின் மேல் வைத்துள்ள ஈடுபாட்டையும் கூட்டுகிறது என்றும் கூறுகிறார்.

இதனைத் தொடர்ந்து, சிங்கப்பூரில் உள்ள தமிழ்க் குழந்தைகளுக்கு, அவர்களின் குரலிலேயே 'வாசிக்க வைக்கும்' மின்னூல்களை உருவாக்குவதற்கான செயலியையும் உருவாக்கினார்.

இந்த மேம்பாட்டின் தொடர்ச்சியாக, வாசிக்கும் தன்மையோடு,நகர்ப்படங்களும் அவற்றோடு ஊடாடும் வசதியும், கதைகளுக்கேற்றத் துணை நடவடிக்கைகளும் ஐ-புக் செயலியின் உதவி இல்லாமலேயே தமிழ் மின்னூல்களுக்கான சிறப்புச் செயலி ஒன்றனில் சேர்க்கப்படவுள்ளது. இந்த மின்னூல்கள், எளிய சொற்களை அறிமுகப்படுத்துவதோடு, அவற்றைக் கேட்டல் முறையில் முறையாக உச்சரிக்கும் வழிமுறைகளையும் பயிற்றுவிக்கவுள்ளது.

இவை யாவும் மாணவர்களுக்கு வாசிக்கும் ஆர்வத்தைத் தூண்டுவதுமட்டுமல்லாமல் சொற்களை, முறையாக உச்சரிக்கும் வழிமுறைகளையும் உடன் கற்கச் செய்கின்றது. மாணவர்கள் கதைகளை வாசிக்கும் போது சரியான ஏற்றம், தொனி உச்சரிப்பு, நயம், ஆகியவற்றுடன் வாசிக்க உதவுகிறது.

இம்முயற்சி மழலையர்களுக்கும் மாணவர்களுக்கும் சிறு வயது முதலே தமிழ் மொழியின் மீது ஆர்வத்தினையும் ஆற்றலையும் வளர்க்கும் கருவியாக அமையும் என்பதே எங்கள் நம்பிக்கை.

இன்றைய சூழல்

தமிழ் மொழிப் புலமையோடு குழந்தை இலக்கியத்தில் ஆழ்ந்த ஈடுபாட்டைக் கொண்டவர்களால் குழந்தைகளுக்கு ஏற்பச்சிறந்தப் படைப்புகளை உருவாக்க இயலும். இந்த உருவாக்கங்களை, நோக்கம் குலையாமல், மின்னுட்பக் கருவிகளுக்குக் கொண்டு செல்ல வேண்டும். இதனைச் செய்யும் நுட்பவல்லுனர்கள், குழந்தைகள் உலகை நன்கு புரிந்துகொண்டுள்ளவர்களாகவும், தமிழின் அழகையும் சுவையையும் உணர்ந்தவர்களாகவும் இருந்தால்,உருவாக்கங்கள் மேலும் சிறப்படையும். இந்த இரண்டு துறையும் இணைந்து உருவாக்கப் படும் மின்னூல்களே சிறந்த முறையில் பயனர்களைச் சென்றடைகின்றன.

குழந்தைகளுக்காக உருவாக்கப் பெற்ற தமிழ் மின்னூல்களை ஒரு பார்வையிட்டோம். அவை ஏற்ற அமைப்பில் உள்ளனவா என ஆராய்ந்தோம். கட்டுரையின் முன்னுரையில் கூறப்பட்ட நோங்கங்களை இலக்காகக் கொண்ட மின்னூல்கள்,எண்ணிக்கையில் குறைவாகவே இருந்தன. கூகுள், ஆப்பிள் ஆகிய செயலிக் கூடங்களில் மின்னூல்கள் பல உருவாக்கப்பட்டிருப்பதைக் கண்டறிந்-துள்ளோம். குறிப்பாகக் குழந்தைகளுக்காக உருவாக்கப்பட்ட மின்னூல்களின் பயன்பாட்டினை ஒப்பாய்வு செய்ததில்,கீழ்க்காண்பவை தெளிவாகத் தென்பட்டன:

பெரும்பாலான மின்னூல்களின் உள்ளடக்கங்களை அச்சில் வழங்குவது போலவே மின்வடிவிலும் வழங்குகின்றன. இவற்றில் பல இலவசமாகவே பதிப்பிக்கப் படுகின்றன. நூல்களைப் பெறுவதற்கு இந்த முயற்சிகள் உதவினாலும், ஒலி, நுகர்நுட்பம், ஊடாடல் போன்ற தன்மைகளுக்கு இவை முதன்மையான இடத்தைத் தரவில்லை.

சற்று அதிக நுட்பங்களைச் சேர்க்கும் மின்னூல்கள் காணொளிகளைச் சேர்க்கின்றன. காணொளிகள் ஒரு நூலோடு குழந்தைகள் ஈடுபடும் பட்டறிவைச்சற்றுக் குலைக்கின்றன என்பது எங்கள் கருத்து. எனவே, கதைகளுக்கு ஏற்ற பின்னணி ஒலி, நகர்படங்கள், கதைகளில் வரும் பாத்திரங்களோடும், பக்கங்களில் தோன்றும்

பொருள்களோடும் குழந்தைகள் ஊடாடுவதற்கான வாய்ப்பு, படித்தக் கதைகளை ஒட்டிய விளையாட்டு நடவடிக்கைகள் முதலியக் கூறுகளைக் கொண்ட மின்னூல்களோ செயலிகளோ தமிழில் இல்லை என்பதே எங்களின் தேடலின் இறுதியில் நாங்கள் கண்ட முடிவாகும்.

பல மின்னூல்கள் வெற்றிகரமாக உருவாக்கம் கண்டிருந்தாலும் அவற்றின் பயன்பாடுகுறைவாகவே உள்ளது. சிங்கப்பூரில்2015-ஆம் ஆண்டு நடைபெற்ற 14-வது தமிழ் இணைய மாநாட்டில், ஆய்வுக் கட்டுரையாளர் திரு வாசுதேவன் இலச்சுமணன்மேற்கொண்ட ஆய்வில், 137 தமிழ்ப்பள்ளி ஆசிரியர்களுள் 18.9 விழுக்காட்டினர் மட்டுமே தமிழில் மின்னூல் வாசிப்போராக அடையாளம் காணப்பட்டுள்ளனர் என்றும்; தமிழ் மின்னூல் வாசிக்காதோர் 63.5 விழுக்காட்டினர், என்றும் கூறினார்.

இச்சூழலின் விழைவாக,மின்னூல்களை உருவாக்கும் ஆர்வலர்கள், ஊக்கமும்,உற்ற அங்கீகாரமும் இன்றி அவரவர் முயற்சிகளில் இருந்து பின் வாங்கும் நிலையே ஏற்படுகின்றது. தமிழ் மின்னூல்களின் எண்ணிக்கையோடு ஆங்கில மின்னூல்களின் எண்ணிக்கையை ஒப்பிட்டால், தமிழ் மின்னூல்களின் எண்ணிக்கை விழுக்காட்டளவிலும் மிகப்பெரிய வேறுபாட்டைக் காட்டுகின்றது.

ஆங்கில மின்னூல்களைப் பயன்படுத்தும் முனைப்பு, குழந்தைகளிடையே பெரிய அளவில் காணப்படுவதற்குப் பல காரணங்கள் உள்ளன. பல்வகை துறை சார்ந்த மின்னூல்கள் ஆங்கில மொழியில் பெருமளவில் வெளியீடு செய்யப்பட்டுள்ளன. மேலும், இம்மின்னூல்கள் நிறங்கள் நிறைந்த காட்சிகளோடும், நகர்ப்படங்களோடும் மழலையர்களுக்காகவே சிறப்பாக உருவாக்கப்படுவதை நம்மால் காண முடிகிறது. குழந்தைகள் இவ்வாறு அமைக்கப்பட்டுள்ள மின்னூல்களைத் தான் தொடர்ந்து மகிழ்வுடன் பயன்படுத்தி மேலும் படிக்க முனைப்பு காட்டுகின்றனர். அதுமட்டுமல்லாமல், இம்மின்னூல்களில் பலவகையான பயிற்சிகளும் ஈர்க்கும் வண்ணம் அமைக்கப்பட்டுள்ளன. இப்படிப்பட்ட மின்னூல்கள் தமிழில் காணப்படுவது அறிதாக உள்ளது. இச்சூழல் பெற்றோர்களுக்கு மட்டுமல்லாமல் குழந்தைகளுக்கும் ஏமாற்றத்தையே தருகின்றது.

வளர்ந்து வரும் தொழில்நுட்ப உலகில் தமிழ் மின்னூல்கள் ஆங்கில மின்னூல்களின் தன்மைகளைக் காட்டிலும், ஈர்ப்புத்தன்மை, ஊடாடல் ஆகிய அடிப்படையில் வளர்ச்சி காணாவிடில் தமிழ்க் குழந்தைகள் தமிழ் மொழிக்காக தொடர்ச்சியான நேரம் ஒதுக்க முனைப்பு காட்ட மாட்டார்கள் என்பது வெள்ளிடை மலை.

அணுகுமுறை

இந்த ஆய்வுக்காக பலவகை வடிவத்திலான ஊடாடல் வசதிகொண்டதமிழ் மின்னூல்களை ஒப்பீடு செய்து, அவற்றில் இடம் பெற்றுள்ள நகர்படங்களையும் தொடர்ப்படங்களையும் அடையாளம் கண்டோம். வெவ்வேறு அணுகுமுறைகள் இருப்பினும் அவை அனைத்தையும் குழந்தைகளுக்குப் பொருந்தும் வகையில் அமையாததையும் கண்டறிந்தோம். கீழ்க்காணும் சில தன்மைகள் தெளிவாகத் தென்பட்டன:

அ. காணொளி சேர்க்கை

பல்லூடக இணைப்புகளைச் சேர்ப்பது, மின்னூல்களில் உள்ள பல வசதிகளில் ஒன்று. படங்கள், ஒலிப் பதிவுகள், நகர்படங்கள், காணொளிகள் போன்றவற்றை விரும்பும் பக்கங்களில் சேர்க்கலாம். இருப்பினும், மற்ற ஊடகப் பதிவுகளைவிட, காணொளிப் பதிவுகள் ஒரு நூலை வாசிக்கும்போது கிடைக்கும் அனுபவத்துக்கு மாறுபட்ட ஓர் அனுபவத்தைத் தருகின்றன.

நூல்களை வாசித்தலும், காணொலிகளைக் காண்பதும், வெவ்வேறு பயனர் அனுபவங்களாகும். நூல்களில், அடுத்தடுத்தக் கட்டங்களுக்குச் செல்லும் கட்டுப்பாடு, வாசகரின் கையிலேயே உண்டு. வாசிப்பினை மேற்கொள்ளும்போது வாசிப்பவர்களால் புத்தகத்தில் உள்ள தகவல்களையும் உள்ளடக்கத்தையும் அவரவர் விரும்பும் வேகத்திற்கேற்ப உள்வாங்க முடிகின்றது. எதிர்மாறாக, ஒரு காணொலியில் உள்ள செய்திகளும் உள்ளடக்கமும் வாசகர்களுக்குத் 'தள்ளப்படுகின்றன'. ஏனெனில், காணொளிகள் அமைக்கப்பட்ட வேகத்தோடு, அக்காணொளியின் செய்தியையும் உள்ளடக்கத்தையும் வாசகர்கள் பின்பற்ற வேண்டும். எனவே, மின்னூல்களில் காணொளிகளைச் சேர்ப்பது வாசகர்களுக்குஒருதொடர்ச்சியான அனுபவத்தை வழங்க உதவவில்லைஎன்னும் கருத்தை இங்கே முன்வைக்கிறோம்.

ஆ. கதைகளின் பின்னனி

தமிழ் மொழியைக் கற்பிக்கும் நோக்கில் மின்னூல்கள் உருவாக்கப்பட்டிருந்தாலும், கதைகளின் அமைப்பு, ஓர் அனைத்துலகத் தன்மையைக் கொண்டுள்ளனவாகத் தென்படவில்லை. மலேசியா, சிங்கப்பூர் போன்ற நாடுகளில் வாழும் குழந்தைகள் பல்லின மக்கள் வாழும் சூழலில் வளர்கின்றனர். பல்வேறு பண்பாட்டு விழாக்களையும் கொண்டாடுகின்றனர். இவற்றோடு ஒட்டி இருக்கும் கதைகளைக் குழந்தைகள் ஈடுபாட்டோடு படிக்கவும், அவர்கள் வாழும் சூழலோடு தங்கள் தாய்மொழியை இணைக்கவும் வாய்ப்பளிக்கும்.

இந்தத் திட்டத்தின் வழி உருவாக்கப்படும் நூள்களில், தமிழர்கள் அதிகம் வாழும் நாடுகளில் உள்ள கல்விக் கொள்கைகளைக் கருத்தில் கொண்டுக் கதைகள் எழுதப்படும். வாசிக்கும் இடங்களுக்கேற்ப படங்களையும் சூழல்களையும்

மாற்றியமைக்கும் நுட்பம், நூல்களை வாசிக்கப் பயன்படும் செயலிகளிலேயே சேர்க்கப்படும்.

இ. விளையாட்டு நடவடிக்கைகள்

பெரும்பாலானத் தமிழ் மின்னூல்கள் வாசிப்போடு நின்று விடுகின்றன. வாசிக்கப்பட்ட கதைகளை ஒட்டியத் தொடர் நடவடிக்கைகள் சேர்க்கப்படவில்லை.
இந்தத் திட்டத்தின் கீழ் வரும் நூல்களில், வாசிப்பின் ஈடுபாட்டை அதிகரிக்க மொழி விளையாட்டுகள் இணைக்கப்படும். விளையாடும் மாணவர்களுக்கு ஊக்குவிப்புப்பரிசுகளும்நூலிலேயே வழங்கப்படும். இப்படிச் செய்வதன் வழி மாணவர்களிடையே மேலும் வாசிக்க வேண்டும் என்னும் ஆர்வம் கூடுகின்றது. அதுமட்டுமின்றி இம்மின்னூலில் உருவாக்கப்பட்ட நடவடிக்கைகள் யாவும் கதையின் உள்ளடக்கம், கதைமாந்தர்கள், சொற்கள் ஆகியவற்றைக் கொண்டு உருவாக்கப்-பட்டவை. மாணவர்களிடையே தொடர்பு கொள்ளும் வாய்ப்பை ஏற்படுத்தி அக்கதைகளை மீட்டுணர்ந்து உள்வாங்கச் செய்யும் ஆற்றலை ஏற்படுத்துவதற்காகவும் இந்த மின்னூல்கள்உருவாக்கப்படுகின்றன.

ஈ. மின்னுட்பப் பயன்பாடு

மின்னுட்பத்தைப் பயன்படுத்திப் பற்பல செயல்களை மேற்கொள்ள முடியும். ஆய்வு செய்யப்பட்டமின்னூல்கள் சிலவற்றில், தொழில்நுட்பம் இருக்கும் காரணத்தால் மட்டுமே சில கூறுகள் சேர்க்கப்பட்டிருப்பதைக் காணமுடிகின்றது. இது மிகச் சிறந்த அணுகுமுறை என்று கூறுவதற்கு அல்ல.

தொழில்நுட்பத்தால் இந்தப் பயன்பாட்டை வழங்க இயலும் என்பதை விட, இந்தப் பயன்பாடு குழந்தைகளுக்குப் பயனுள்ளதாக இருக்க வேண்டும் என்பதே முதன்மையானது.

ஆகவே, நாங்கள் குழந்தைகளுக்காகவே சிறப்பான உள்ளடக்கத்தைத் திட்டமிட உறுதி செய்தோம். அதன்பிறகே, இவ்வுள்ளடக்கங்களை மிகத் துல்லியமான வழியில் குழந்தைகளுக்குச் சென்றடைய தகுந்த தொழில்நுட்பத்தைத் தெரிவு செய்தோம். தற்போதைய நிலையில், ஊடாடல், நகர்ப்படங்களை உட்புகுத்தி மின்னூல்களை உருவாக்க பள்ளிகளிலும் வீட்டிலும் பயன்பாட்டில் உள்ள மின்னுட்பக் கருவிகளை, குறிப்பாகக் கையடக்கக் கருவிகளைக் கண்ணோட்டமிட்டோம். பொதுவாக, மூன்று வகையான கருவிகளே உள்ளன. பள்ளிகளில் ஐ-பேட் கருவிகள் பரவலான பயன்பாட்டில் உள்ளன. வீட்டில் ஆண்டிராய்டுகருவிகளை அதிகப் பயன்பாட்டில் உள்ளன. இருப்பினும் இச்சூழல் நாட்டுக்கு நாடு வேறுபட்டு இருப்பதையும்கண்டோம். எடுத்துக்காட்டாக, அமெரிக்க நாடுகளில் ஐ-பேட்டுகள்வீடுகளிலும் அதிகமாகப் பயன்படுத்தப்படுகின்றன.

முழுமைபெற்ற எங்கள் மின்னூல்கள் அதிகப் பயன்பாட்டிலுள்ள ஐ-பேட், ஆண்டிராய்டு, விண்டோசு ஆகிய மின்னுட்பக் கருவிகளில் ஒரே தரத்தில் செயல்படுவதை உறுதி செய்துள்ளோம்.

முதன்மைக் கூறுகளில் ஒன்றான ஊடாடும் வசதி, நகர்ப்படம், அசைவூட்டம், குரல் பதிவு இம்மின்னூலுக்குத் தேவைப்பட்டன. பின்னனிக் குரல் பதிவைச் சேர்ப்பது, குழந்தைகளுக்குச் சொற்களை முறையாக உச்சரிக்கும் வழிமுறைகளையும் உடன் கற்கச் செய்கின்றது. மாணவர்கள் கதைகளை வாசிக்கும் போது சரியான உச்சரிப்பு, நிறுத்தம், நயம், தொனி ஆகியவற்றுடன் வாசிக்கும் வழிமுறைகளையும் கற்பிக்கின்றது. இப்படிச் செய்வது, குழந்தைகளுக்குக் கதைகளையும் சொற்களையும் விரைவாக நினைவில் கொள்ளும் ஆற்றலையும் வளர்க்கின்றது. மின்னூல், தானாகவே வாசித்துக் காட்டும் சூழல், பின்னனிக் குரல் இல்லாமல் குழந்தைகளே வாசிக்கும் சூழல் என இரு சூழல்கள் சேர்க்கப்பட்டுள்ளன.

நடவடிக்கை வகைகள்

நூல்களோடு மாணவர்களின் ஈடுபாட்டைக் கூட்ட, மொழி விளையாட்டுகள் இணைக்கப்பெற்றுள்ளன. “வாசித்துக் காட்டு”, “நானே வாசிக்கிறேன்” எனும் வாசிப்பு நடவடிக்கைகளை மேற்கொண்ட பின்னரே மாணவர்கள் இந்த விளையாட்டுகளை விளையாடுவர். முதல் கட்டத்தில் 4 வகையான விளையாட்டுகள் இணைக்கப் பட்டுள்ளன:

1. படக்குவியல்

படக்குவியலில் மாணவர்கள் வெட்டப்பட்ட படங்களை முறையாக அடுக்குவர். ஒவ்வொரு முறை விளையாட்டை மேற்கொள்ளும் போதும் மாணவர்களுக்கு வெவ்வேறு படங்கள் வழங்கப்படும். காண்பவர் கண்களைக் கவர்ந்திழுத்து மேலும் மேலும் பயிற்சியைச் செய்யத் தூண்டும் இப்பயிற்சிகள், மாணவர்களுக்கு சலிப்பினைத் தராது.

2. படத்தோடு சொல்லை இணைக்கும் விளையாட்டு

கதையை வாசித்த பின்னர், கதையில் இடம்பெற்ற படங்களும் சொற்களும் இவ்விளையாட்டில் புகுத்தப்படும். மாணவர்கள் படங்களைத் தகுந்த சொல்லுடன் இணைப்பர். மிகவும் எளிமையான இந்த விளையாட்டு மாணர்களுக்கு மகிழ்ச்சியையே தரும்.

3. நிறம் தீட்டுதல்

கதையில் கண்ட காட்சிகள் இந்தப் பயிற்சியில் இணைக்கப்படும். மாணவர்கள் தங்களுக்குப் பிடித்த வகையில் ஏற்ற நிறங்களைப் பயன்படுத்தி நிறமற்றப் படங்களுக்கு நிறமூட்டுவர். மாணவர்களின் சிந்தனை ஆற்றலைத் தூண்டுவதற்கு நிறங்களும் துணைப்பொருள்களும் படங்களைச் சுற்றி வைக்கப்படும்.

4. விடுபட்ட எழுத்து விளையாட்டு

இந்தப் பயிற்சியில் மாணவர்களுக்குப் படமும் விடுபட்ட எழுத்து கட்டங்களும் வழங்கப்படும். மாணவர்கள் படங்களுக்கேற்ப எழுத்து கட்டங்களில் விடுபட்ட எழுத்தினை நிரப்புவர்.

இவ்வகையான மொழி விளையாட்டுகளை மேற்கொள்வதனால், குழந்தைகளுக்கு நினைவு கொள்ளும் ஆற்றலும் சொற்களஞ்சியமும் வளர்க்கின்றன. மேலும் ஒவ்வொரு விளையாட்டிலும் மாணவர்களுக்கு ஊக்குவிப்புப்பரிசு வழங்கப்படுகின்றது. இவ்வகையான நடவடிக்கைகள் மாணவர்களை மேன்மேலும் மின்னூல்களைப் பயன்படுத்த ஆர்வமூட்டி, மாணவர்களின் வாசிப்புத் திறனை வளர்க்க மிகவும் துணைபுரிகின்றன.

முடிவுரை

தமிழில் இன்னும் வளர்ச்சிப்பெறாத துறை இது. இம் முதல் முயற்சி பல்லாற்றானும் சிறப்பாய் அமையப் பல்வேறு கூறுகள் ஆராயப்பெற்றன. கதைவடிவமும் பயிற்சி முறைகளும் ஆய்ந்து தேர்ந்து இணைக்கப் பெற்றன. தகுந்த இடத்தில் தக்க அளவில் தொழில்நுட்பம் பயன்படுத்தப் பட்டது. இம் முயற்சி முத்தாய்ப்பாய் அமையப் பட்டறிவுமிக்க கல்வியாளர் பெருமக்கள்தம் கருத்துகளும் பெறப்பட்டன. இப்பெருந்திட்டத்தின் வெற்றி இது போன்ற நுட்பங்கள் அடங்கிய நூட்களை வெளியிட பலருக்கும் உந்துதல் அளிக்கும் என நம்புகிறோம்!

கட்டுரை எழுதத் தகவுரை கூறி ஆற்றுப்படுத்திய,முனைவர் முரசு நெடுமாறன், கணிஞர் முத்து நெடுமாறன், ஆகியோருக்கு என் நன்றியதலைப் புலப்படுத்தி அமைகிறேன்.

கலைச்சொற்கள்:

பல்லூடகம் - multimedia; ஊடாடல் - interactive; பயனர்பட்டறிவு - user experience

குறிப்புகள்:

1. http://ta.wikipedia.org/wiki/மின்னூல்
2. முத்து நெடுமாறன் (2011) கையடக்கக் கருவிகளில் தமிழ் மின்னூல் உருவாக்கம்', 12.8.2011 முதல் 14.8.2017, மலாயாப் பல்கலைக்கழகம், மலேசியா.http://anjal.net/ebooks/eBooks-in-Tamil.epub
3. வாசுதேவன் இலட்சுமணன் (2015) 21ஆம் நூற்றாண்டு கற்றல் திறன்களில் அட்டைக் கணினி வழித் தமிழ் மின்னூல் உருவாக்கம்: ஓர் ஆய்வு, 14வது தமிழ் இணைய மாநாடு, சிங்கை.

PADAM INAITHAL-WORDS MATCHING WITH IMAGES GAME

Gokul Kumar, M [1] Keerthana, V[2] Hemanandhini, S[3] Dr. Mala, T[4] Yesodha, K[5]

[1 2 3] UG Scholars, [4] Associate Professor, [5] Research Scholar,

Department of Information Science and Technology,

Anna University, Chennai, Tamil Nadu

[4] malanehru@annauniv.edu [5]yeshoda.ammu@gmail.com

ABSTRACT:

Recent digital games have been developed not only for entertainment purposes, but also to promote learning. In this research work, Kids AR Memory Matching game in Tamil has been proposed which is an Augmented Reality (AR) game for learning words in English and Tamil language. First person game controllers have been used as user interfaces to increase immersion in playing games. Game controllers allow a player to explore and move, reload and destroy then activate secondary functions such as zooming within a forest environment. The goodness of the features extracted from the instances and the number of training instances are key components for building an effective model. This work is completely implemented in Unity 3D and has been compared with ABC3D Augmented Reality game. This work consists of three components – i. Matching Images with Word, ii. First Person Controller and iii. Third Person Controller.

1. INTRODUCTION

The objective of the research work is to develop an Augmented Reality based game which can help children to learn words and their usages. The outcome of the research work is the Words matching with images game (Padam Inaithal) which is a specially designed game. The implementation of Padam Inaithal is divided into three main components. The first part includes matching the images with words. This is achieved using the vuforia packages from vuforia developer portal which is a database for right and wrong answer. The second part is the first person controller which is present in the forest environment. The traditional Doom-style first person controls are not physically realistic. The solution is the specialized Character Controller. It is simply a capsule shaped Collider which can be told to move in some direction from a script. The third part is basically interfacing the third person controller in the forest environment. Third-person controller games almost always incorporate an aim-assist feature, since aiming from a third-person camera is difficult.

2. RELATED WORK :

R.Ballagas et al. [1] have proposed the world of mobile technologies is in a constant and growing popularity, becoming more ubiquitous every day. Mobile devices have increasingly

demonstrated their usefulness and applicability in a daily basis, to assist users in their work, to be used in a familiar environment or to support forms of entertainment. These devices are equipped with high resolution cameras and other resources such as GPS and accelerometer. A. Mulloni et al. [2] have proposed the interest in implementing augmented reality applications on mobile systems has increased significantly. These systems, which integrate reality with virtual elements, provide the user with an easy and safe interaction, without prior knowledge of this technology. Augmented reality can be useful in any application that needs to display information not available. P. Skalski et al. [4] proposed game controllers that are used as user interfaces to increase immersion in video games. Research on the subject has focused on studies of how interactivity can affect enjoyment leading to find positive relation between the two.

3. SYSTEM ARCHITECTURE

The following **figure 1** depicts the System Architecture Diagram for Kids AR Memory Matching game in Tamil namely the Padam Inaithal. It shows the overall schematic view of the process involved in gaming. The user interface module provides bilingual languages (both Tamil and English). The image part contains the image target and corresponding meta-data which is stored in the cloud database. The user places the marker on the webcam and using vuforia SDK, image detection and tracking is done for correctness of matching.

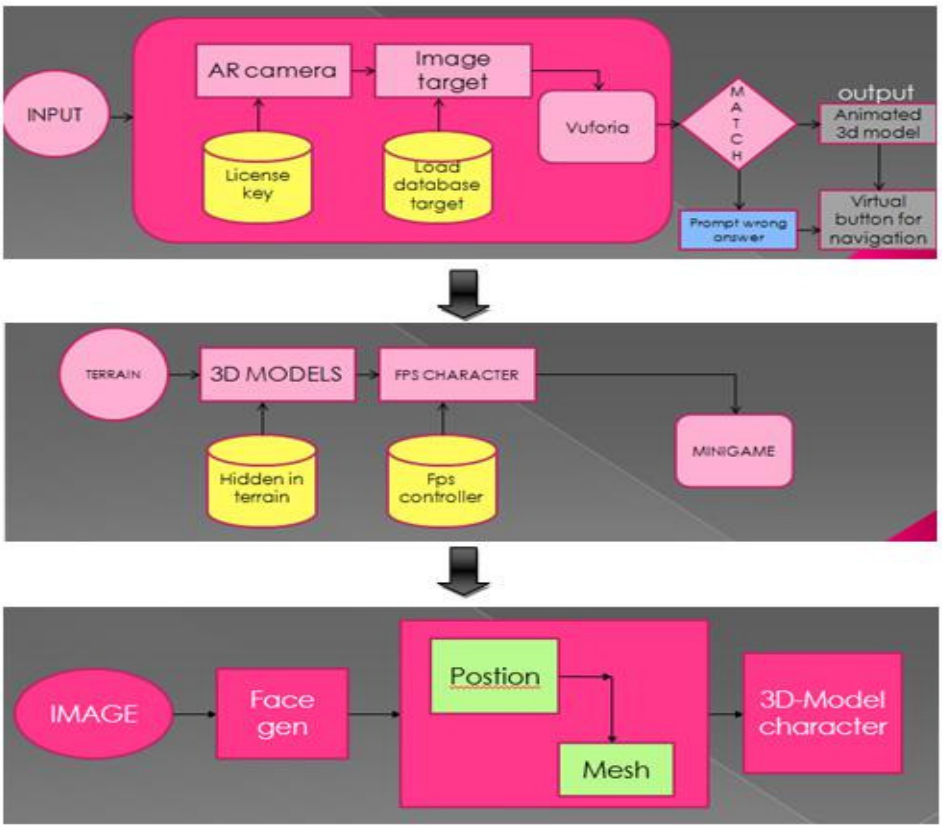


Figure. 1. System Architecture Diagram for Kids AR Memory Matching game in Tamil

For the correct match, the game is processed based on scores such as AR game, virtual play, First person controller and Third person controller. For the third person controller users their face is recognized and introduced as a 3D model in the game to provide immersive AR gaming.

The main menu of Padam Inaithal consists of free mode game, settings and exit user interfaces. The settings consist of two language interfaces English and Tamil. The Tamil language interface is used to specify buttons, textual contents and immersive gaming experiences. The words are constructed using online Tamil keyboard and used as a user interface in Kids AR Memory Matching game in Tamil. The required Tamil word cards are created to play the Kids AR Memory Matching game in Tamil and the image recognition technique using Vuforia identifies feature selection points to extract the unique feature points for correct recognition. The interface between Tamil and English is easily switchable using settings options.

To facilitate the children to play, both Tamil and English language interface were developed. Tamil is a consistently head-final language. The verb comes at the end of the clause, with a typical word order of subject-object-verb. However, word order in Tamil is also flexible, so that surface permutations of the SOV order are possible with different pragmatic effects. Tamil has postpositions rather than prepositions. Demonstratives and modifiers precede the noun within the noun phrase. Subordinate clauses precede the verb of the matrix clause. The Tamil language interface is also implemented in certain coding of the paper using mono develop unity. The main menu, free mode game and its scoring, virtual butterfly-squash gaming, first person controller and third person controller gaming are developed for both Tamil and English language interfaces.

4. RESULTS

The words and the image when matched correctly results in correct animal 3D model and if the words and images are not matched correctly, then it results in wrong answer 3D model. For the correct answer test case, the score is increased by one, and proceeds to the next levels. For the wrong answer test case, the score is decreased by one and goes to preceding levels. The experience of Padam Inaithal is shown in **figure** 2.

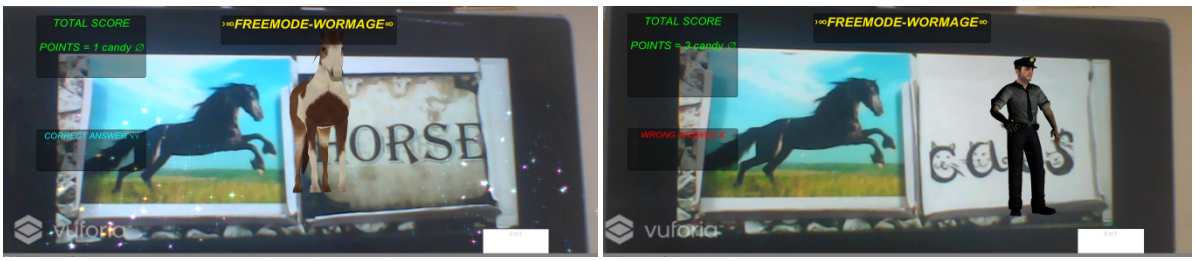


Figure 2. Correct case “Horse”and Wrong case “Cats”

4.1 Evaluation

The PADAM INAITHAL game is compared with ABC3D mobile game, in order to show the various analysis parameters like Augmented Reality Immersive, gaming levels, user rating, image recognition, user interface and the results are given in **table. 1**

Analysis	Wormage AR game	ABC3D AR game
Image Recogniton	25(out of 26 image content)	26(out of 26 textual image content)
AR Immersive	2(out of 4 levels)	2(out of 4 levels)
Gaming Levels	4(out of 5 levels)	3(out of 5 levels)
User Rating(best)	40(out of 50 people)	38(out of 50 people)
User Interface(Scene alignment methodology)	3(out of 4 UI's)	2(out of 4 UI's)

Table 1 Comparison of PADAM INAITHAL with ABC3D mobile game

In the image recognition technique, there are total 26 cards available and out of the 26 cards PADAM INAITHAL game identifies 25 image content cards and ABC3D game identifies 26 textual image content cards. In the PADAM INAITHAL AR game out of 4 levels, 2 are Augmented Reality oriented and in ABC3D mobile game out of 4 levels, 2 are Augmented Reality oriented. From the survey of 50 people, 40 of them rated PADAM INAITHAL AR game as best, 5 of them rated the game as fair, 3 of them rated it as average and 2 of them rated the game as poor. Also, for the ABC3D mobile game 38 of them rated PADAM INAITHAL AR game as best, 4 of them rated the game as fair, 2 of them rated it as average and 4 of them rated the game as poor. In the PADAM INAITHAL AR game out of 5 levels, 4 levels are completed (ARgame, FPC, TPC, Virtual play) and story environment is not working perfectly. In the ABC3D game out of 5 levels, 3 levels are completed (Textual recognition AR game, Story environment, Virtual play) FPC and TPC is not working perfectly. In the context of user interface Padam Inaithal uses both English and Tamil interfaces whereas ABC3D mobile game uses only English interface.

5. CONCLUSION

Augmented Reality game PADAM INAITHAL (words matching with images game) thought of learning of English language, focused on specific words, in a simple, interesting and interactive way. The results indicate that children who used the Augmented Reality game had a superior English learning progress than those who only used traditional methods. This work indicates that the use of Augmented Reality games has a positive pedagogical impact in the learning process concerning young children, more exactly in the progressive domain of oral recognition of words and concepts and their corresponding written form. Accordingly, we strongly believe that AR will be in a short term, an important tool in the class activities in some areas of education. As a future work, the PADAM INAITHAL (words matching with images game) can be well extended to the teaching and learning processes with students of other ages and in the teaching of other languages. Since PADAM INAITHAL (words

matching with images game) runs in a simple and cheap hardware that requires only a PC equipped with a webcam, it can be used for teaching aids in most schools.

6. REFERENCES

1. A. Mulloni, A. Dunser, and D. Schmalstieg, Zooming interfaces for augmented reality browsers . The Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services, MobileHCI, vol 10, 2010, pp.161 170.
2. Fungus in unity, https://fungusgames.com/.
3. He Zhang, Peng Ren, "Game theoretic hypergraph matching for multi-source image correspondences" Pattern Recognition Letters, Volume 87, 1 February 2017, Pages 87-95.
4. I.A. Essa, "Analysis, Interpretation and Synthesis of Facial Expressions", Perceptual Computing Technical Report, MIT Media Laboratory, 1995, pp.303-308.
5. P. Skalski, R. Tamborini, A. Shelton, M. Buncher, and P. Lindmark, Mapping, the road to fun: Natural video game controllers, presence, and game enjoyment , New Media & Society, 2010, pp.1-15.
6. R. Ballagas, J. Borchers, M. Rohs, and J. G. Sheridan, The smart phone:A ubiquitous input device . IEEE Pervasive Computing Technique, vol 5, 2006, pp.70 77.
7. Tao Wang, Quansen Sun n , Zexuan Ji n , Qiang Chen, Peng Fu 'Multi-layer graph constraints for interactive image segmentation via game theory' Pattern Recognition, Volume 55, July 2016, Pages 28-44.
8. Unity guide, https://docs.unity3d.com/ScriptReference/GUI.html.

Digitization, Distribution and Synthesizing Tamil Texts: Challenges of taking Madurai Project to its next step

Ku. Kalyanasundaram[1] and Vasu Renganathan[2]
[1] kalyan.geo@yahoo.com, [2] vasur@sas.upenn.edu

Abstract

Digitizing texts from Tamil literatures of three different genres have been a challenging task, especially in the context of crowd resourcing and distribution. "Project Madurai" (http://www.projectmadurai.org/) has been a very successful attempt ever since it was implemented about twenty years ago. The crowd resourcing method was exploited in a very sophisticated fashion to collect, digitize, proof-read, store and finally distributing to the world. Multiple methods of distribution namely in plain html format, distributable pdf format and in multiple encoding formats namely ascii, eight bit and unicode texts. With the experience we gained on crowd resourcing, we would like to layout a novel way of enriching the existing texts with extendable infrastructure, especially for its effective use in research and learning. This paper attempts to present a method of human assisted machine learning, instead of implementing any algorithm with a full-fledged learning technique. Our goal will be to convert the existing texts into other formats namely JSON, relational database, word net and others. We demonstrate in this paper with suitable interfaces for how human's effort can be included in a supplementary fashion from a crowd resourcing technique, while at the same time machine is trained with suitable data to further manipulate toward outputting them in a comprehensive form. We demonstrate in this paper how the url

http://www.thetamillanguage.com/url_gloss.php?url=

can be used to feed into any Tamil etext page and get a machine learning algorithm built with Crowd-Sourcing technique.

Introduction

The important tasks we would focus on in this paper are a) conversion of HTML texts to a relational database as well as JSON structure, b) identify a plausible and optimum structure for the relational database and JSON formats and c) identify the meaningful ways of performing the stemming practices using a Crow-Sourcing technique. While the first two tasks do not require as much effort as possible, but the success of the most important last task would fully depend upon the former. Stemming and tagging Tamil texts have been the most significant aspect of digitization and distribution of Tamil texts for the reason that the inflected Tamil words of both Sangam, medieval and modern Texts pose a daunting task for information storage as well as retrieval. Many attempts have been made so far to stem and tag Tamil texts, including the one as demonstrated in http://www.thetamillanguage.com/tamilnlp/tagit.html. The aim of this paper is mainly to extend this automatic method of stemming and tagging process into a more manageable method by implementing the crowd resourcing techniques as we demonstrated by the Madurai Project earlier. However, the technique we demonstrate in this paper extends the above automatic method and implement a crowd resourcing method by employing suitable

interfaces. Further, this method will be intended to work with texts from Sangam, medieval and modern period particularly working with the corpus that is presented in the Madurai Project. In essence, this paper attempts to illustrate as well as demonstrate a machine learning technique exclusively exploiting the crowd resourcing process with suitable interfaces for human interference.

Project Madurai

Project Madurai was started twenty years ago with volunteers across the globe performing both collection as well as digitization of Tamil data ranging from old Tamil to modern Tamil. In order to maintain the authenticity, the volunteers were divided into many categories including those collect rare materials, xerox them and send to those who are assigned the task of typing. The other category of people are those who can proof read the text to make sure the digitized data are error free in all respects. When the project was started, there was no any standardized font, as we have in the form of Unicode. With the meticulous involvement of members of INFITT and the Tamil Nadu government, a standard font called TSCII and the volunteers were trained to use this font throughout the process to maintain standard. Later, when Unicode was introduced, we had to convert all of the texts entered in TSCII to Unicode compliant format and it was done using many available convertors.

Significance of Digitization

In general, it is believed that any digitization process is meant for preservation of archaic text. In fact, advantages of digitized text should be more than preservation as far of any linguistic data is concerned. Cross-reference, searching, statistical analysis, conducting historical research etc., are some of the most significant aspects of digitized data. These scopes can not be accomplished with the digitized linguistic data either in HTML or PDF format. Instead, what requires to be done is making the stored data to be accessible in a number of different flexible formats. Storing them in any relational database, for example, would enable anyone to retrieve the data in a multiple number of formats. This paper intends to describe some of the issues involved while converting the existing Tamil literature data that is stored widely in the internet, including that of what is made available in Madurai project.

Stemming Old and Medieval Tamil Data and Limitations of Crowd Resourcing:

One of the immediate requirements for any Tamil literature data is to perform the stemming of inflected forms, as the machine can notbe trained to perform any high level processes unless suitable algorithm is written to make available morphological and syntactic knowledge of the text. Stemming Tamil words into their corresponding morphological units, in general, is a complex task especially for the reasons of its agglutinative nature. But, it is more complicated in the case of old Tamil data as most of the texts in old Tamil are in poetic forms and the words are separated in many odd ways to meet meter and other poem types.

To cite one example, consider the following line from akanāṉūṟu (264):

maḻaiyilvāṉammīṉaṇintaṉṉa

This sentence when attempted to segment using the tagger: http://www.thetamillanguage.com/tamilnlp/tagit.html(See Renganathan 2016 fore details of this tagger) we get the output as below:

[["loc","mazhai","noun"],["nom","vaanam","tr"],["nul","miinaNin"],["nul","tanna"],["nom",".","period"]]

The first two words are segmented as expected, but the other two words did not get the result as desired. This is for the important reason that these words are segmented differently from source words in the poem to maintain meter. When this sentence is realigned as ma̲laiyilvā̲nammīn̲aṇintan̲n̲a

We can get the desired output as:

[["loc","mazhai","noun"],["nom","vaanam","tr"],["nom","miin","tr"],["adv","aNi", "anna","adv_m"],["nom",".","period"]]

Thus, in order to get this desired output, one would need to train the tagger with more rules, but accounting for this type of segmented form adhering to meter is almost impossible. The only solution one would expect to have in this kind of situation is that the tagger needs to have a human intervention for segmenting words that are parsed for metrical purposes. In order to accomplish this type of tasks, though, one would surely depend on crowd resourcing involving people who have sufficient knowledge in old Tamil poems and the segmentation techniques. Limiting people with such knowledge naturally minimizes the power of crowd resourcing, as one can not expect to have as many people as one would expect to have such specialized knowledge.

Javascript Object Notation (JSON) Structure of Medieval and Old Tamil Data:

The important limitation of the Sangam and Medieval Tamil texts , as we have now in http://www.projectmadurai.org/, is that they are all in HTML and PDF formats, which don't have the convenience of searching and researching across different genres of literature. As for using Tamil literature texts for references in the context of writing research articles and books, one would surely need all the texts searchable by many factors including author, poem number, line number, chapters, composition, commentaries and so on. In order to accommodate all of these information as well as exploring Tamil words for their historical changes as well as development of meaning across the genres, what is supposed to be a meaningful attempt is to identify a plausible and resourceful structure of Tamil literature records and use them to build a very comprehensive database, either in JSON format or in any other relational database format. Unfortunately, no such attempts have been made so far, despite many efforts to digitize and preserve Tamil literature from ancient times. Often times, these digitization efforts are of the nature of reinventing the wheels with multiple efforts digitizing and typing the same text by many people.

In this section, we propose an ideal format of Tamil literature data in JSON format and how this can be advantageous over storing text in a rather linier format. JSON format has been one of the very popular data structures that is used by many programming languages including Javascript, PHP, JAVA, Python and so on for the main reason that it can be stored in text format and itdoes not require any relational databases such as MySQL, SQL server, Oracle and so on. Further, it is easily exportable to other database formats without too much programming efforts. What is important is to identify the "key" vs. "value" relationships to account for the data. For example, consider the following poem from Tirumantiram and how it can be presented in a JSON string.

Text:
கடந்துநின்றான்கமலம்மலராதி
கடந்துநின்றான்கடல்வண்ணம்எம்மாயன்
கடந்துநின்றான்அவர்க்குஅப்புறம்ஈசன்
கடந்துநின்றான்எங்கும்கண்டுநின்றானே. (திருமந்திரம் 14).

JSON:

```
[{
        "Number": "14",
        "poem_source": [
                "கடந்துநின்றான்கமலம்மலராதி 1",
                "கடந்துநின்றான்கடல்வண்ணம்எம்மாயன் 2",
                "கடந்துநின்றான்அவர்க்குஅப்புறம்ஈசன் 3",
                "கடந்துநின்றான்எங்கும்கண்டுநின்றானே. 4"
        ],
    "poem_translit_s": [
                "kaṭantuninṟāṉkamalammalarāti 1",
                "kaṭantuninṟāṉkaṭalvaṇṇamemmāyaṉ 2",
                "kaṭantuninṟāṉavarkkuappuṟamīcaṉ 3",
                "kaṭantuninṟāṉeṅkumkaṇṭuninṟāṉē. 4"

]       "poem_parsed": [
                "கடந்துநின்றான்கமலம்மலர்ஆதி 1",
                "கடந்துநின்றான்கடல்வண்ணம்எம்மாயன் 2",
                "கடந்துநின்றான்அவர்க்குஅப்புறம்ஈசன் 3",
                "கடந்துநின்றான்எங்கும்கண்டுநின்றானே. 4"
        ],
        "poem_translit_p": [
                "kaṭantuninṟāṉkamalam malar āti 1",
                "kaṭantuninṟāṉkaṭalvaṇṇamemmāyaṉ 2",
                "kaṭantuninṟāṉavarkkuappuṟamīcaṉ 3",
                "kaṭantuninṟāṉeṅkumkaṇṭuninṟāṉē. 4"
```

```
    ]
      "poem_translation":[
            "Excelled Him, Lotus, Flowers and all 1",
            "Excelled Him, Color of the Ocean, our mystery man 2",
            "Excelled Him, Surpassing Him is God 3",
            "Excelled Him, Every whereVisible, He is 4"
       ]
}
]
```

The keys used here include "Number", "poem_source", "poem_translit_s", "poem_parsed", "poem_translit_p" and "poem_translation". What is significant to note here is that this type of detailed structure as stored in JSON format can easily be used for a number of different purposes such as glossing, historical research, translations, advanced search, filtering and so on, which is not possible in the available HTML and PDF formats. However, converting all of the available e-texts, as given in Madurai Project as well as in other electronic resources of Tamil is humanly impossible for many reasons including the available quantity, expert knowledge to parse text, making a viable interface to input this type of specialized data and so on so forth. Unless one sets up a robust online interface that can allow experts to manually input all of thesedata structures through crowd-sourcing, this task can never be envisioned by any other means. As illustrated by Vamshi et al. (2017) one needs to implement what they call "Active Crowd Translation" method, according to which the Crowd Sourcing attempts along with the machine learning algorithm need to be integrated constantly. So, as and when any new linguistic structure is presented through Crowd sourcing, the machine learning algorithm should immediately use the new knowledge with the already available knowledge base in order to make it independent further. Along these lines, one can speed up the process of Crowd Sourcing by integrating the already developed taggers as shown in http://www.thetamillanguage.com/tamilnlp/tagit.html.This will minimize the amount of human intervention during this process.

Once this type of robust database of Tamil literature texts from Sangam to Modern Tamil is built, any available programming resources can be used to manipulate them any way one would want. One of such attempts was made to convert some of the Madurai Project texts into JSON text, and used for quick search using the VUE.JS technology, as can be viewed at: http://sangam.tamilnlp.com/mp/. With this type of electronic text, data mining is thoroughly possible (See Eickhoff 2011 for advantages of Crowd Sourcing and data mining.)

Crowd Sourcing Technique for Stemming Words from Tamil Poems:

As indicated in the above JSON structure for Tamil poems, all but the key "poem_parsed" requires human knowledge as well as a Crowd Sourcing technique to successfully exploit all of the Madurai project files to build very meaningful and novel resources. In order to use all of the Madurai Project files, we have implemented a webpage written in PHP, JQuery and Javascript as can be seen at:

http://www.thetamillanguage.com/url_gloss.php?url=. This script can be fed into any of the Madurai Project files as given below.

http://www.thetamillanguage.com/url_gloss.php?url=
http://www.projectmadurai.org/pm_etexts/utf8/pmuni0010_02.html

This allows users to read the text with suitable gloss from lexicon. As one can see, not all of the words can be glossed for the reason that many words like உய்க்காக்கால்are inflected and unless the head word உய்வுis stored in the database, or a tagger is built to identify the head word. By crowd sourcing, it is possible to segment this type of inflected words and save them as separate record in JSON format.

'[{"word":"உய்க்காக்கால்","annotation":"உய்வு"}]'

This system, thus, is capable of building its knowledge base through Crowd Sourcing on an ongoing basis. The more number of people get involved in this process, the more efficient and powerful the electronic texts will be. The contributions of users by this stemming process is stored in a MySQL database, so the glossing software can consult this it dynamically.

Conclusion:

It is attempted in this paperhow some of the uses of digitized Tamil texts from Sangam to Modern period can be optimized with both the process of Crowd Sourcing as well as by employing the already built machine learning algorithm through the morphological taggers and glossing. It is shown how the data prepared in JSON structure along with the Vue.JS technology allows one to build client side search engines, data mining as well as filtering of texts without having to rely on any relational databases from the server side. What is significant is to develop suitable web based user-friendly Crowd Source interfaces so the unsegmented and inflected Tamil texts of three genres can be parsed and stored as part of the JSON data, so meaningful linguistic applications can be built without much complexities. With the system as illustrated in this paper, it is possible to make use of the electronic texts as stored in Madurai Project Website and other sites in a more efficient manner possible.

References:

1. Eickhoff, Carsten and Arjen P. de Vries (2011). "How Crowdsourcable is Your Task?" WSDM 2011 Workshop on Crowdsourcing for Search and Data Mining (CSDM 2011), Hong Kong, China, Feb. 9, 2011. http://s3.amazonaws.com/academia.edu.documents/30680905/csdm2011_proceedings.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1494088024&Signature=mmu7g%2FPdtaDGhKb0hXxQnL2oHnw%3D&response-content-disposition=inline%3B%20filename%3DCrowdsourcing_blog_track_top_news_judgme.pdf#page=11

2. Renganathan Vasu. 2016. *Computational Approaches to Tamil Linguistics*. Cre-A Publishers: Chennai.
3. Vamshi Ambati, Stephan Vogel, Jaime Carbonell 2017. "Active Learning and Crowd-Sourcing for Machine Translation". Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA vamshi,vogel,jgc@cs.cmu.edu (https://www.cs.cmu.edu/~jgc/publication/PublicationPDF/Active_Learning_And_Crowd-Sourcing_For_Machine_Translation.pdf).

A Newly Digitalized Archive of Tamil Texts, Video, Audio and other Graphic Materials

Brenda Beck,
Department of Anthropology, University of Toronto

This paper is about my own personally collected and digitalized Tamil archive, something I have just recently finishing compiling and indexing. My intent is to share it with students and scholars interested in learning more about Tamil traditional life, as it was experienced, in the villages of Kongunadu not so long ago. For nearly two years, between December 1st, 1964 and August of 1966, I lived in a village named Olappalayam, located about 6 miles East of the town of Kangayam in Tamilnadu. At the time I was enrolled in doctoral studies at the Institute of Social Anthropology, Oxford University, UK. This stay in Tamilnadu was my personally chosen "fieldwork" project. My goal was to learn Tamil from local speakers and to immerse myself in local culture in an attempt to document and develop a understanding of local history and its many related social traditions. I have returned to this same village numerous times since then and, unwittingly, witnessed a stunning degree of transformation in the area over a period of just 52 years.

I choose Olappalayam for its centrality in what was then a tradition-focused and relatively unstudied area of Tamilnadu. Although Olappalayam has always been located close to a main "trunk" road it was still a quiet and culturally conventional place when I first lived there. The area has blossomed in terms of industrial investment and is now a booming community housing many immigrant workers who have come from other parts of India in search of work. A number of large employers have built factories and warehouses in the neighborhood. The main road has been vastly widened, the lovely shade trees on that road all cut down and the noise of bus and lorry traffic has replaced the gentle sound of the many squeaking bullock carts that the author remembers. Rents are high, water is scarce and locals have many complaints about how life has changed. What my research archive has "captured" is how life once was in this area, documented by many photographs, maps, charts and copious field notes. But more than documenting the look of the place and its geographic layout, my research attempted to capture the culture and social traditions common to this area.

The Kongu region encompasses a large, quite dry, upland plain that is surrounded by high hills. This area never suffered domination for long by any of the three great Southern kingdoms, each of which was dependent for its strong economy on a vibrant costal trade (the Chola, Chera and Pandiya empires). The Kongu area has always been distinctively different. As recently as the 11th and 12th centuries it was heavily forested and know for its skilled and resilient tribal residents, many of whom are mentioned in the famous Sangam texts. Kongu was a resource-rich inland plan where artisans flourished along with traders. It was watered by the great Kaveri River and cris-crossed by many overland trade routes that linked the area to distant places via a vast trade network. The people of this region were then, and still are, fiercely proud of their distinctive heritage, their relative independence and their resourceful

attitudes. They are a population that has been steeled by the challenges of living far from the central urban nodes and "cultured" courts of the grand rulers of peninsular life. Although there are Brahmins in the area, of course, the powerful families in this area are all peasant landowners by background. The culture of the area is characterized by a distinctly "non-Brahman" set of social values.

Looking back on my work of over fifty years in this region, I wanted to compile an archive of research materials that could be used by students to learn about a number of key Tamil cultural themes. One of the sub-project for this has been to collect together all of my audio tape recordings. This is a unique body of cultural memories captured in story form as shared with me by local villagers. By introducing me to their oral traditions they helped me improve my Tamil language skills and also gave me many new ways to build my understanding of their lives and to build by knowledge of their rich, subtle and textured heritage. I was also fortunate in finding a very kind female cook and companion. After some time her son joined us and became my dedicated research assistant. It was this man, K. Sundaram of Olappalayam (OKS) who faithfully collected so many tape recordings of local stoeies and legends for me, and then doggedly and patiently transcribe them all! His work now constitutes the largest part of the archive I have built. Sundaram was also a great help with my local census work, including household counts, family genealogies, and field ownership surveys. He also helped me to build charts of food exchange customs involving various castes and much more. Compiling this archive would not have been possible without his help.

The chart below outlines the bare bones of the B. Beck digital archival collection I describe here:

INDEX OF THE BECK ARCHIVE (Much more extensive indexes describing this archive are available to those who are interested) TOPIC	Total scanned pages (legal size)
Researcher's field notes (typed) 1964-66 - in order of date + card index +1964-66 diary-typed (from handwritten original)	822
Olappalayam and neighborhood 1965 Census +field ownership records	993
Local Caste Stories, Caste interaction charts	104
Coimbatore City 1966 house-to-house caste survey by Malaria workers	91
Regional Kongu structures – including the. four titled families of the Kongu area	33
MacKenzie Manuscripts accessed in Madras	129
Family lineage Histories- including temple rights	301
Annanmar-Ponnivala Story texts	6,006
Major stories-legends (except Annanmar	6,643

Ponnivala story)		
Shorter Stories	552	
Myths and stories about various gods, lesser divine beings and temple origins	517	
Kavadi songs plus lullaby and kummi songs	62	
Temple Maps	172	
Temple rituals -plus inscriptions and related stories	166	
Transcripts of human possession events by gods and/or by evil spirits	1,053	
Wedding, funeral and coming of age rituals - detailed descriptions	734	
Proverbs and riddles	22	
Research Assistant Answers to Questions from Brenda's letters	523	
Oxford B.Litt. Thesis	206	
Oxford D.Phil. Thesis	522	
Color Slide Index (pages in loose leaf notebook)	187	
Black and White Photo index - typed pages	5	
Audio Tapes Index - typed pages	8	
Also stored in Olappalayam: list of duplicate tape transcripts	1	
Total pages scanned		19,852
PLUS		
Color slides (all scanned as .jpg images)	3,256	
Black and White Photos (scan of negatives + contact prints) as .jpg files	1,429	
Original 5" audio tapes - (now all are accessible in a digital .wav format)	34	

*Note: The archive is still expanding and the totals reported here are expected to increase slightly in the coming months.

In addition to the archival materials discussed above there are a number of related interpretive materials which I am still using frequently, but which I intend to contribute to this core of research materials in due time. These include:

1. **A 13 hour animated video series** telling the Annanmar (Ponnivala) story. This large animation project was directed by an authentic Tamil folk artist, Ravichandran Arumugam. Th complete video series is available with two separate narrated sound tracks, one in English and the other in Tamil. It can be used in multiple new ways, for language teaching and more.

2. **An ipad version of the same Annanmar story**, intended for use in teaching reading (in Tamil or English). The ipad version (first half of the story only) reads the legend aloud in either language while presenting illustrations and also highlighting each spoken word in written form, as it is pronounced. The ipad version also includes some short video excerpts that relate to particular episodes. There are many technical development possibilities here, including converting this program for android use, programming a similar presentation based on the second half of the same story, narrating it in other Indian languages, and more.
3. **A digital game** developed as an educational tool for kids. The game is based on the traditional and very ancient Indian "Parcheesi" or taiyam contest.
4. **A 2-part set of graphic novels**, about 800 full-color graphic pages, telling the same Annanmar story in printed form. Both a Tamil and an English version have been produced and broadcast both in Canada and in Tamilnadu. Both graphic novel sets are available as digital files as well.

There are many ways for this archive to be used. Here are some of the thoughts I have. Others will very likely think of more:

RESEARCH SUGGESTIONS – by Data Set

Olappalayam and neighborhood 1965 Census +field ownership records

There is a striking opportunity here to document how fast change can occur and what specific directions and subtle contours have been involved. Follow-up field work with fresh mapping and interviewing would be required.

Local Caste Stories, Caste interaction charts

Detailed caste interaction customs and contrasting left and right-hand caste value systems constituted a major part of my original research focus. A follow-up investigation exploring how these attitudes (especially interaction customs) have both persisted and changed, in the face of monumental changes would be of great interest to a variety of scholars.

A Coimbatore City 1966 house-to-house caste survey done by Malaria workers

This is a truly unique data set on Coimbatore City, probably not equaled for its detail by records from any other city in India. This information was obtained courtesy of Malaria workers who repeatedly went door-to-door, in Coimbatore, in 1966 asking about fevers. In this context, a setting where they knew most of the families they spoke to, I asked the team to record the caste of the families they visited. This was not as sensitive a topic then as it is now. The results are a map-able data set of household-by-household caste identities, street by street, for the entire city. Although directly parallel information would not likely be collectable today, modern sampling techniques might produce a very interesting general picture of how caste distribution patterns have changed in fifty years, particularly striking might be how some neighbourhoods have changed more than others.

Regional Kongu structures – including the four titled families of the Kongu area

Kongunadu was a conceptually structured region that had a formal architecture of sub-nadus, titled families, specially identified temples and much more. The area provides a very interesting example of traditional Tamil thinking about the interweaving of sacred and profane spaces. My (unpublished) D.Phil. thesis also provides a lot of information on this topic. Developing a comparison between these patterns and the traditional conceptual structure of other parts of India would be informative. A recent Ph.D. thesis submitted to Pondichery University has focused on the history of Kongu's four titled families specifically and could usefully supplement what is to be found in this archive.

MacKenzie Manuscripts accessed in Madras (now Chennai)

These are notes on much older manuscripts housed in Chennai that contain details related to an impressive heritage of ordering themes that together describe Kongu as a unique social "space." These notes need to be further catalogued, after which they could be used to provide additional information that could help develop a deeper and wider study how local Tamil cultural identities have become interwoven with place, as well as how they have changed over time.

Family lineage Histories- including temple rights

This is a very large collection of personal interviews that I and my research assistant conducted with a wide range of family elders. It would be very interesting to track down some of these families today and explore how their sense of identity links to the traditions described fifty years ago by their relatives, as well as exploring how those same family and lineage identities have evolved in the modern context.

Annanmar-Ponnivala Story texts

This particular part of the collection is extensive and represents the core of much of my work over the years, especially my attempts to share this particular story widely through animation, graphic novels, large vinyl murals and more. This is a cultural gem of the first order that is not widely known or appreciated. There are several versions represented in this archive, both multiple tellings by the same bard and variants preserved and promoted by various practicing-bard who hail from different "lineages." There is a great deal of interesting further research to be pulled from this collection of texts that will help us to better understand how oral traditions evolve over time in relation to place and also get "adjusted" for presentation to a variety of audiences. Tamils have yet to develop pride in this very special part of their rich story heritage. It is my hope that getting to know this great story in depth will provide the essential starting point for a much-needed oral culture discovery process.

Major stories-legends (except Annanmar Ponnivala story)

The archive also contains a number of other important transcripts of locally performed, lengthy story narratives. Some of these are variants of pan-Indian legends (like the story of Sita's wedding). Others are extremely interesting variants of famous Tamil legends like Kovalan's story. And still others are unique to this region, as far as I know. One can also

uncover a variety of intriguing crossovers with European tales of various kinds. This is a hugely interesting area for further research and study.

Shorter Stories, Riddles, Songs, Lullabies etc.

There is much more in this archive than I have the space to describe here. For example, the collection includes an extensive collection of over 100 maps of temple layouts, lengthy descriptions of life-cycle rituals, and transcriptions of a number of "possession" events where either a god, a goddess, or some evil spirit speaks through an individual for a period of time. All these topics are worthy of further study. In the case of wedding ceremonies, which are described in great detail for a number of communities, an interesting comparison could be made with the data compilation provided in my B.Litt. thesis (also in the archive) where I look at ritual patterning by caste groupings using Edgar Thurston's wedding descriptions provided in his much earlier and very famous publication: The Castes and Tribes of South India.

Color Slides

The collection includes over 3,250 digitized colored slides. Many of these capture scenes in the life of Kongu residents that convey the feeling and the cultural texture of the region in the mid-sixties. Selections drawn from this archive can be used to illustrate various research works, but could also become an exhibition that would stand on its own as an expression of "life and times" of people in this area during that mid-century "era."

Black and White Photos

The scanned black and white photos contained in the archive are less numerous and more personal. Many were taken and then given to local families as a "thank you" for their research help. It would be of real interest to local people living in the area today if someone were to take these photographs back to this locale and find some of the families that were earlier documented. It would be a very nice way to do a follow-up study of some of the family groups covered in my original research. The archive's many family genealogies and lineage stories could also be shared in this way.

Audio Tapes

The huge collection of audio tapes digitalized for this archive will be of interest to Tamil linguists studying the Kongu Tamil dialect. Furthermore, there is the possibility of sharing many of the stories and legends housed here by broadcasting selected recordings, perhaps via the radio or sharing them on the internet.

B. Beck's Oxford B.Litt. and D. Phil. theses

Both these works have considerable research value of their own. They each contain a lot of unpublished data (never published in the case of my D.Phil. manuscript) that is organized and discussed at length. Because these two works are very difficult to find on line I have included a digitized copy of both as a possibly useful supplement to this broader archival collection.

IN SUM:

It is my hope that this freshly compiled collection of research data (roughly 100 gigabytes in size) will serve multiple scholarly interests, helping scholars to better understand how the people of Tamilnadu are evolving with the times. The fact that this is a digital archive will make this body of 50-year-old information about the Kongu area easily accessible to Tamil researchers around the world. For me it is a dream-come-true, something I never imagined would be possible when I first collected all this material during my own primary field work completed than fifty years ago. I am thrilled to now be able to "give back" much of what was so generously shared with me by local villagers at that time. I request that this material be respectfully used to further our understanding of the strength and the beauty of this one small part of a much wider, fuller and deeper corpus of the Tamil-speaking peoples' heritage culture. WHERE THIS UNIQUE DIGITAL ARCHIVE WILL FIND A HOME has not yet been determined.

Can technological advancements such as Digital Archiving play a critical role in preserving a classical language?

Ram Kallapiran
Academic and Collaborations Manager,
UK College of Business and Computing, London, Essex, UK IG2 6NW
Email: pudhuyugan@yahoo.com

Abstract
We live in an age of information overload as stated by futurist Toffler A. (1970). Enormous amounts of literature and resources are being digitally produced every day in every language. It is therefore imperative to use the right approach to seamlessly archive these digital resources lest they may be lost or misrepresented in the long run. The technology of semantic Web project has facilitated the hope that this could be achieved by transferring all printed, filmed data into the digital realm for public access along with the newly-produced digital resources. This paper aims to explore the role of technological developments in this preservation process. The uses and challenges of digital archiving have been taken up for investigation along with the discussions on the current efforts undertaken in the field. The digital archiving of a classical language such as Tamil has been taken up as an exemplar. This paper constructs a view on how these steps can be taken forward in the preservation of a classical language for the precise and purposeful use of future generations, mainly through the recommendation of a three-part process.

What is Digital Archiving?

According to Hodge (2000), Digital Archiving can be defined as long-term storage, preservation and access to information that is 'born digital' or for which the digital version is considered to be the primary archive.
On the other hand, Digital library is one that contains a collection of original, digitised resources that can be searched. The primary focus here is uniqueness.
As can be comprehended Digital archiving is not a conflicting concept to digital libraries but can be seen as a more sophisticated advancement within that. However unlike digital libraries, digital archives are not necessarily unique or part of a collection.

Why is it important?

The dangers of erosion and infestation inflicted by insects to ancient texts and traditional inscriptions that were maintained in proprietary & historic formats such as palm-leaves led to the dawn of the print age. We are undergoing a similar turn in history at this point in time. An ocean of short-term information are being produced today digitally such as museum records, literary creations, research articles, records of socio-cultural developments, government orders, blogs, social networking feeds, news items, web articles & other e-resources. However due to the very nature of such resources, factors such as accountability, ownership and aptness of the maintenance methods used, can be questioned. Digital material is risky and vulnerable to loss. Hence there is a need to preserve them in a systematic basis which is the primary use

of digital archiving.
In addition, as we know concepts like blended learning have come to the fore today, in an attempt to help e-learning realise its true potential.
“A high-quality education system is one that achieves quality and equity” (Woods A., Comber B., Iyer R., 2015)
Digital Archiving can be useful not just for reference and research but in learning and education as well, in providing equal, quality learning opportunities to all.

The challenges of Digital Archiving
Whilst the uses are evident, a number of acknowledged challenges can also be found which can’t be overlooked. In order to achieve optimum benefits in the process of preservation these challenges of digital archiving as articulated below should be evaluated and duly dealt with.
• How to standardise metadata used for Digital archiving?
• How to handle Copyrights issues on articles that have featured in multiple locations?
• How to categorise the searching mechanism for articles, authors and publications?
• How can we methodically address the incognizance of digital archiving in people who produce digital resources?

Why is it required for a classical language such as Tamil?
According to linguistic scholars like George L. Hart (2000), a classical language can be termed as one that is ancient with a copious body of literature and is an independent tradition which has come to being mostly on its own. This means most of the root words of the language can be found in the very same language. A few other scholars believe that it should also be a living language in addition to the above. Tamil fully qualifies as a classical language in all these counts.
Asko Parpola, who is a Professor emeritus of Indology at the University of Helsinki, Finland, has carried out a rigorous research on the script used in Indus Valley Civilisation which is said to have matured some 3000 years BC. He has concluded in his book ‘Deciphering the Indus script’ (1994) that the Indus script belonged to ancient Tamil or proto-Dravidian. The ancient literature in Tamil has been widely translated and acclaimed across the world (Zvelebil K., 1973).
A long surviving language can provide bounteous amounts of details on history, politics, arts, culture, advancements made by human civilisations, socio-cultural dynamics and so on. In many cases these can serve as tertiary, secondary or even primary sources of information for research. Hence it can be useful in comprehending living patterns and genealogy of entire nations at various points in time. It is important, therefore, not only to preserve the long-standing body of literature but also to preserve the newer set of digital literature that is being produced currently as offshoots emerging from their ancient counterparts or independent of them.
On the other hand there are numerous challenges that arise from the digitization of a language that dates back to thousands of years. The lack of a standard platform, loss of linking data, influences of marginal groups, and authenticity of information and so on can be named as examples. So such digitization measures should bring together state-of-the- art technology and scholarly linguistic expertise under the umbrella of a singular archiving infrastructure.

What resources can be used to achieve this?

It is very important to connect datasets across the semantic web so that future users can access data relevant to a subject. By connecting to global datasets and constantly upgrading them, we could ensure the concerted and continued preservation of the classical language. Therefore employing effective and highly customisable search mechanism as driven by technologies such as web3.0 are paramount.

The use of technologies such as SSI involving digital conversion where required, accessing digitally stored data and use concepts such as Deep Learning can be helpful here.

An archiving mechanism sponsored by an international federation such as Google, UNESCO can help here. They have indeed initiated some successful projects. Also such archiving should be added on to semantic web projects such as LOD Cloud (Insight, University of Manheim, 2017) which can connect to the global databases. This way the digital archiving process can be made seamlessly scalable and durable[1].

[1] The Digital Library Federation has made publications on the interesting developments happening in this field (Heterick B., 2002)

When it comes to preserving a classical language such as Tamil there have been some on-going but disintegrated attempts in archiving digital resources. Whilst many initiatives in creating Digital libraries can be found such as www.chennailibrary.com , www.tamilheritage.com, Madurai Project, they can be further enhanced in forms of clearly earmarked digital archives considering the humongous volumes of modern literature that are being produced currently, solely in electronic formats in websites, blogs and so on.

However these can be different to the classification made by Hodge discussed above. These archives aim to preserve existing and ancient publications that are scattered across various libraries in Tamilnadu by firstly capturing them as microfilms and later digitizing the reels. Examples include Roja Muthiah Research Library (RMRL) project on digitizing early publications on the history of Tamilnadu, Professor Anne Gilliland, University of California EAP 191 project on archiving publications of French India between 1800 and 1923 and so on (British Library Endangered Archives, No date).

How can it be taken forward?

The challenges of digital archiving as identified above can be effectively handled in the following ways

- Appropriate handling of Metadata is key
- Creating awareness on digital archiving
- Creating and managing international projects on digital archiving in order to preserve Tamil.

In addition to the above the use of suitable technologies, disaster recovery plans and data migration mechanisms should all be planned, executed and periodically managed under an overarching mechanism which can be termed as ‘Digital Archiving Infrastructure for Tamil’ [DAIT] or Digital Archiving Tamil Environment [DATE]

Recommendations

The challenges in the utilisation of digital archiving were discussed above and so were the

challenges that are associated with the preservation of a classical language. Both of these pools of challenges should be effectively handled. This can be done by the inception of an overarching standard that encompasses digital archiving as an integral part of it. This research recommends the use of a three-part process to achieve this silky balance in achieving a solution to the problem

The three parts are 'Digital Publishing – Digital Archiving – Digital searching'.
Digital publishing involves publishing and bringing all that have been published in a searchable library format. We already have some initiatives in Tamil Digital libraries as mentioned above [2]. All new publications should be made digitally available and connected to the Digital Archiving Infrastructure
[2] Digital libraries frequently use the protocol Open Archives Initiative Protocol for Metadata Harvesting.
Digital Archiving can involve all of what has been discussed above
Digital searching tries to create a connection between global databases, digital libraries and archives3 so as to facilitate deep and advanced searching which are customised. This can be presented in form of Digital Libraries which are all connected to the Infrastructure that has been recommended
In order to fully realise the potential of the three-part process mentioned above in practical terms, the following recommendations have been made;
• To identify a Project Owner – National governments can take ownership of the project and execute it through a university for enduring scholarly sustenance. For instance Tamilnadu government can embark on this initiative and function as the Project Owner. Instead of reinventing the wheel this project can utilise and integrate existing resources discussed above and expand on them
• To create a viable plan on Content Ownership – All copyrights and publication rights inclusive of display & search should be obtained with appropriate license agreements in place and a 'light archive4' access provided to users
• To Create a Digital Archiving Infrastructure [DAIT] – Digital archiving that is disintegrated will not offer the real benefit that has been envisioned here. Given that the resources on Tamil range from a variety of sources, it is vital to create an infrastructure that includes data migration, digital library, searching, backup & disaster recovery and connection to global datasets
• To use appropriate Data management technologies (such as cloud storage) – The use of state-of-the-art technologies is key in tackling all the technological challenges identified above and to achieve success
5
• To Work out the Economics – Terabytes of storage are available for very affordable prices these days. For example both Google and Dropbox offer cloud storage for rates as cheap as $10 per terabyte. So project cost should be suitably worked out for the creation and maintenance of DAIT

Conclusion

As explicated above from many angles it can be concluded that the systematic use of digital archiving created and managed in form of a suitable project as recommended in this paper, aptly aided by suitable technologies & architecture, is indeed a key Concepts like data chaining and data channelling can be effectively used here [3].

Archive that is always available for a large community of users ([4] Heterick B., 2002)

Key references on technologies have been discussed by many [5]. Some important references have been given included in the Bibliography for further reading

requirement in the preservation of the classical Tamil language for the benefit of future generations. This research can be further taken forward by investigating the use of newer technological advancement in the field as they come by. It is also crucial to maintain a constant vigil on the continued upkeep of a singular, universal voice for Tamil Digital Archiving so as to retain the integration process intact.

Bibliography

Breeding M., 2013, Digital Archiving in the Age of Cloud Computing, The Systems Librarian Available from https://faculty.washington.edu/rmjost/Readings/cloud_computing_and_digital_archives.pdf (accessed July 2017)

British Library, No Date, Endangered Archives supported by Arcadia, Available from http://eap.bl.uk/database/overview_project.a4d?projID=EAP183;r=10383 (accessed 01 May 2017)

Daintith J., Wight E. (Ed), 2008, Oxford Dictionary of Computing, Oxford University Press, UK

George Hart, 2010, The Uniqueness of classical Tamil, World Classical Tamil Conference, India

Grensing-Pophal, 2012, Preserving the digital past, Econtent, Pg: 8 - 10

Hartman H., 2003, Consumer Digital Snapshots Will Be The Next Revolution, The Seybold Report, Analyzing Publishing Technologies, Seybold Publications Volume 3, Number 14, Pg: 3 – 7.

Heterick B., 2002, Applying the lessons learned from retrospective archiving to the digital archiving conundrum, Information Services & Use 22 (2002) pg: 113–120

Hodge G M., 2000, Best Practices for Digital Archiving An Information Life Cycle Approach, D-Lib Magazine Vol 6 Number 1 Available from http://www.dlib.org/dlib/january00/01hodge.html (Accessed 05 May 2017)

Insight, University of Manheim, 2017, The Linking open Data cloud diagram, Available from http://lod-cloud.net/versions/2017-02-20/lod.svg (accessed 01 May 2017)

McCargar V., Risc Inc.,2007, Kiss Your Assets Goodbye: Best Practices and Digital Archiving in the Publishing Industry, The Seybold Report, Analyzing Publishing Technologies, Seybold Publications Volume 7, Number 16, Pg: 5 – 7

Dr. Nakeeran P.A., 2010, World Classical Tamil Conference, Tamilnadu Government, India

Parpola A., 1994, Deciphering the Indus script, Cambridge university press, UK

Toffler A, 1970, Future Shock, Bantam Books, USA

Up Front, 2002, A Digital Archiving Standard, EBSCO Publishing, The Information Management Journal, Pg: 14

Woods A., Comber B., Iyer R., 2015, Literacy Learning: Designing and Enacting Inclusive Pedagogical Practices in Classrooms, in Joanne M. Deppeler , Tim Loreman , Ron Smith , Lani Florian (ed.) Inclusive Pedagogy Across the Curriculum (International Perspectives on Inclusive Education, Volume 7) Emerald Group Publishing Limited, pp.45 – 71

Zvelebil K., 1973, The smile of Murugan on Tamil literature of South India, EJ Brill, Leiden, Netherlands

---

Discovering Deep Knowledge from Relational and Sequence Data

Andrew K.C. Wong
Systems Design Engineering and Centre of Pattern Analysis and Machine Intelligence
University of Waterloo

Abstract

This talk presents a novel method P2K (Pattern-to-Knowledge) with surprising findings that deep knowledge could be discovered from mixed-mode relational, biological and temporal sequence data without reliance on explicit prior knowledge. By deep knowledge, we mean the hidden and entangled associations, governed by different underlying factors, that could not be revealed with traditional methods but could be unveiled in different transformed disentangled statistical spaces by P2K. From relational mixed-mode datasets, biological and temporal sequences, it is able to discover the "what" and "where" of crucial associations, units, elements or mechanisms related explicitly to the source environments. The knowledge discovered is able to enhance predictive analysis and pattern clustering and reveal subtle associations and subgroup characteristics. Its outcomes could be validated via domain experts and knowledgebase (including data banks and literature) in Intra-Internet or through new tests/experiments. It enhances the experts' insights and efficiency by shortening the search process and improving predictive analysis. P2K has been applied to temporal sequence patterns with varying magnitude and time delays to discover a wide range of local relations alongacross mixed-mode time series without relying explicitly on prior models. Incorporated in P2K is a new method for predicting rare events in imbalanced data that has great potential for abnormal financial trend forecast and fault detection. P2K will open an avenue to discover deep knowledge for biology, drug discovery and medical research as well as engineering and business practice. It meets a new challenge in the era of big data.

A Brief Biography

Dr. Wongholds a Ph.D. from Carnegie Mellon University; B.Sc (Hons) and M.Sc. from the Hong Kong University. He is an IEEE Fellow for his contribution to machine intelligence, computer vision, and intelligent robotics areas. Currently, he is a Distinguished Professor Emeritus at UW. He was the Founding Director of the Pattern Analysis and Machine Intelligence Laboratory (PAMI Lab) at UW, now the Centre of PAMI. and a Visiting Distinguished Chair Professor at the Hong Kong Polytechnic University (0003). He has published over 300 papers and chapters, and holds seven US Patents. He served as the General Chair of the International Conferences IASTED (1996) and IROS (1998). Dr. Wong has served as consultant to many high-tech companies in the US, Canada and Hong Kong. Over the years, based on his developed core technologies, he co-founded Virtek Vision International Corporation, a leader in laser and vision technology and served as President (87-93), Chairman (93-97) and Director of the Board (97-03) of it. In 1997, he co-founded Pattern Discovery Software Technologies Inc. and served as its Chairman until 06/2013. He is now serving as the chief scientist of Knowledge Fund Limited.

ISSN 2313 -4887

Tamil Internet Conference 2017 Co-Sponsors

உத்தமம்
INFITT
உலகத் தமிழ்த் தகவல் தொழில்நுட்ப மன்றம்
International Forum for Information Technology in Tamil

UNIVERSITY OF WATERLOO

UNIVERSITY OF TORONTO

IEEE Canada

Media Sponsor:

ATN™
ASIAN TELEVISION NETWORK