# 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 maḻaiyilvāṉammīṉaṇintaṉṉa

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”: [
                "kaṭantuniṉṟāṉkamalammalarāti 1”,
                "kaṭantuniṉṟāṉkaṭalvaṇṇamemmāyaṉ 2”,
                "kaṭantuniṉṟāṉavarkkuappuṟamīcaṉ 3”,
                "kaṭantuniṉṟāṉeṅkumkaṇṭuniṉṟāṉē. 4"

]       “poem_parsed”: [
                “கடந்துநின்றான்கமலம்மலர்ஆதி 1”,
                “கடந்துநின்றான்கடல்வண்ணம்எம்மாயன் 2”,
                “கடந்துநின்றான்அவர்க்குஅப்புறம்ஈசன் 3”,
                “கடந்துநின்றான்எங்கும்கண்டுநின்றானே. 4”
        ],
        “poem_translit_p”: [
                "kaṭantuniṉṟāṉkamalam malar āti 1”,
                "kaṭantuniṉṟāṉkaṭalvaṇṇamemmāyaṉ 2”,
                "kaṭantuniṉṟāṉavarkkuappuṟamīcaṉ 3”,
                "kaṭantuniṉṟāṉeṅkumkaṇṭuniṉṟāṉē. 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).