# Tamil Interface to the Online Tamil Dictionary at Cologne

Anbumani Subramanian
Bradley Department of Electrical and Computer Engineering,
Virginia Tech, Blacksburg. VA – 24060. USA.
anbumani@vt.edu

## Abstract

The need for a complete Tamil-English dictionary on the internet has long been felt and with the increased use of Tamil in electronic formats, it is imperative that such a dictionary exists. While there exist few Tamil dictionaries on the internet, a simple interface to accept Tamil characters is not yet implemented. Instead, the dictionaries use various transliteration schemes to represent Tamil alphabets using English letters. Therefore an end user has to learn this transliteration scheme to use the dictionaries. In this paper, a solution to the problem is presented, using a wrapper technique with a discussion on the Tamil interface developed for the Online Tamil Lexicon (OTL) at the University of Cologne. This technique can be easily extended to other existing resources and thus enable users to comfortably interact with Tamil characters. The primary advantage of this technique is that the existing interface needs no modification and yet Tamil input/output can be implemented.

**Keywords:** Tamil dictionary, TSCII converter.

## 1. Introduction

The rising popularity of the internet in the mid-90s led to the birth of Tamil on the web. Soon many began to share ideas in Tamil, through email and websites across the world. This information exchange grew to a larger extent and thoughts on developing other Tamil resources emerged. The aim for publicly available dictionaries and literature collection initiated voluntary efforts to build online dictionaries and libraries. Among these were few attempts to build a complete Tamil dictionary.

There exist few Tamil dictionaries on the internet now, although only one of them is complete. The Online Tamil Lexicon (OTL) at University of Cologne [1] is mostly popular among the research community. This dictionary is an exhaustive collection of more than 130,000 Tamil words from the University of Madras Tamil Lexicon [1] and hence can be considered a complete Tamil dictionary. The University of Chicago, through its Tamil Dictionary Project now hosts Kriyavin Tarkalat Tamil Akarati, in its Electronic Text Services collection [2]. As a part of The Alternative Dictionaries, there is an Alternative Tamil Dictionary with a goal to build a dictionary of slang words in Tamil [3]. The Tamil Dictionary project [4] is an internet effort by Umar to build an online dictionary. Although not a dictionary, Tamil Computing Words [5] is also an internet initiative to become a reference for Tamil words equivalent to often used technical words in English.

The online dictionary at Cologne follows the transliteration scheme used in University of Madras Tamil Lexicon (TL) and does not accept Tamil character input. The research scholars well acquainted with this transliteration scheme find it convenient to use this lexicon while an average user often finds it difficult to learn the new transliteration scheme. Thus the average user is at a loss for an online dictionary despite the existence of such a good resource on the internet. Kiriyaavin Tarkalat Tamil Akarati at University of Chicago is available only to

its university community and not for the public [6]. The Alternative Dictionary project, being a resource for slang words, lists only one person on its Tamil team and its current status of the collection is an unknown. Its short focus (of slang words) and the one-person team are difficult to convince that it will become a complete Tamil dictionary. Among the Tamil dictionaries available online, only one dictionary by Umar, accepts search input in Tamil characters. But this, being a voluntary effort has only a limited number of Tamil words in its collection and so is not incomplete.

It is disconcerting to see that Tamil dictionaries can exist, which cannot accept or display Tamil characters and yet be called 'Tamil dictionary'. This is true for other forms of resources in Tamil as well – like the large volumes of Tamil literature available in transliterated format, compiled during the early days of computers in universities and research institutes. These collections are based on some old transliteration scheme and hence cannot be displayed in Tamil characters. The problem here is how to use these existing resources and enable their display in Tamil characters.

This paper is organized as follows. The next section presents a solution to this problem in case of the Lexicon at Cologne followed by a discussion on the implemented design. Then the advantages in this approach are presented. In the final section, some possible future work is suggested.

## 2. Problem and its Solution

The OTL interface at Cologne is designed to accept transliterated Tamil words (or English) and search for the matching (or equivalent) Tamil words in the collection. Anyone on the internet can search for Tamil words in this dictionary using a web browser and do not require other special software for this purpose. This has been implemented using the most widely used Common Gateway Interface (CGI) standard [7]. In this method, a request from an internet user is passed to the web server at Cologne, with the search word and other necessary parameters for the search.

To enable Tamil input and output in the Tamil Lexicon, one may propose the idea of converting the entire database into Tamil coding and developing a new Tamil interface. But this argument fails to remember the fact that the present user community who are comfortable with existing transliteration scheme also has to move and so may not feel comfortable with this transition. Therefore this is not a favorable solution to implement.

A possible solution is to sandwich the OTL interface between two character conversion layers, and simulate the use of Tamil characters for input and output. These new layers are non-intrusive, non-destructive and remain outside (i.e., wraps) the original interface. Therefore, this kind of interface is called a *wrapper*. Here, it is a *Tamil wrapper* to the Online Tamil Lexicon. It is important for a wrapper to be non-intrusive and non-destructive, since its sole function (simply put) is only that of a translator and not an 'intelligent agent' – meaning, it should not modify the query or the result by a large extent.

With this solution, the new Tamil interface to OTL forms a layer outside the original interface and will accept Tamil words to search in the dictionary. The search request from the user in Tamil characters is converted to OTL scheme and is passed to the web server at Cologne. The corresponding result is modified very minimally, for the purpose of displaying Tamil characters on a web browser and is returned to the user.

## 3. Implementation

The solution proposed above has been implemented successfully and is found to work as expected. Figure 1 shows a block diagram representation of this implementation.

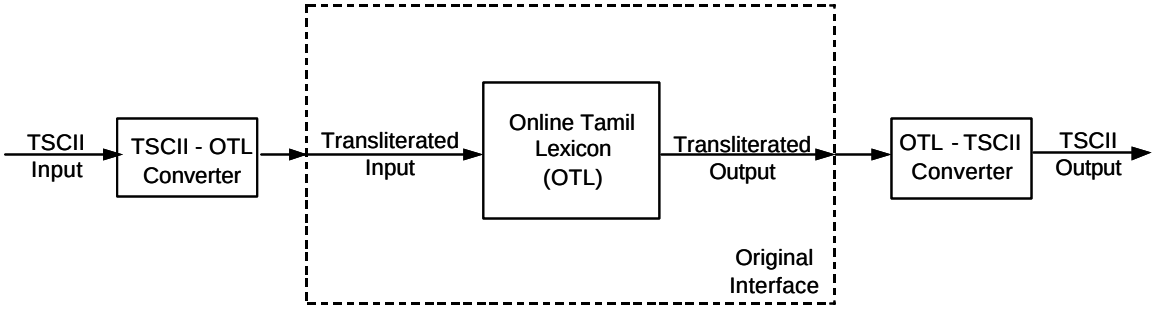



**Fig. 1.** Implementation of the Tamil interface to Online Tamil Lexicon.

The present implementation of the wrapper uses the Tamil Standard Code for Information Interchange (TSCII) scheme and so can convert characters between TSCII and OTL coding schemes. When the input word to search the dictionary is in Tamil, the wrapper converts the TSCII word to a representation in OTL scheme, using a lookup table method [8]. This converted word is passed on to the original OTL interface and the search results (transliterated words only) are converted back to TSCII coding for display in the web browser.

This wrapper has been implemented in Perl and is fully functional. It is available for use from,

`http://www.ee.vt.edu/~anbumani/tamildict/`

Figure 2 shows the interface to the wrapper, which can accept Tamil characters to search the lexicon. Figure 3(a) shows the results in Tamil for an example search, after minimal modification of the actual result. Figure 3(b) shows the dictionary entry for one of the results in Tamil – after conversion by the wrapper.

## 3. Advantages

The proposed wrapper technique has inherent advantages. The wrapper was developed independently and does not interfere with the already existing system. This is true for the wrapper technique in general, since the goal is to merely translate the input and output, between a scheme native to the existing system and a scheme, easy for the common user.

In most cases, once the character conversion procedure – the backbone of the interface, is well implemented, the wrapper need not be rewritten when small changes occur in the original system. The existing user base can still use the original system and be comfortable, while average users can also utilize the valuable resource, using a system that is widely popular and convenient.

The character conversion between two coding schemes is very fast and takes negligible amount of time. But the one disadvantage in case of independent wrapper implementation on the internet (as in the OTL case) is the slow network response. The wrapper interface located on a web server has to depend on the original interface available on a remote server. So there is some amount of network traffic involved in its use. This leads to slow responses at the wrapper

end. But this problem can be circumvented if the wrapper is also hosted locally on the server where the resource is available. This completely eliminates unnecessary delay in the network traffic and so will benefit the end user.

## 4. Further Work

This concept of wrappers can be extended in many ways to utilize already existing resources for Tamil. The wrapper is best suited for implementation, where a resource is only available in some proprietary format or in an archaic coding scheme.

One such example is the large number of Tamil newspapers and newsmagazines available on the internet. Since these websites do not yet follow one of the two recommended standards (TSCII/TAB) for Tamil coding, they require the user to install a new font on the computer, which is not always feasible in restricted working environments. The easy solution is to develop a simple wrapper for these websites. Such a wrapper can convert the proprietary format coding to TSCII/TAB coding on the fly. Thus the user is relieved of the dependency on a proprietary encoding using the wrapper technique.

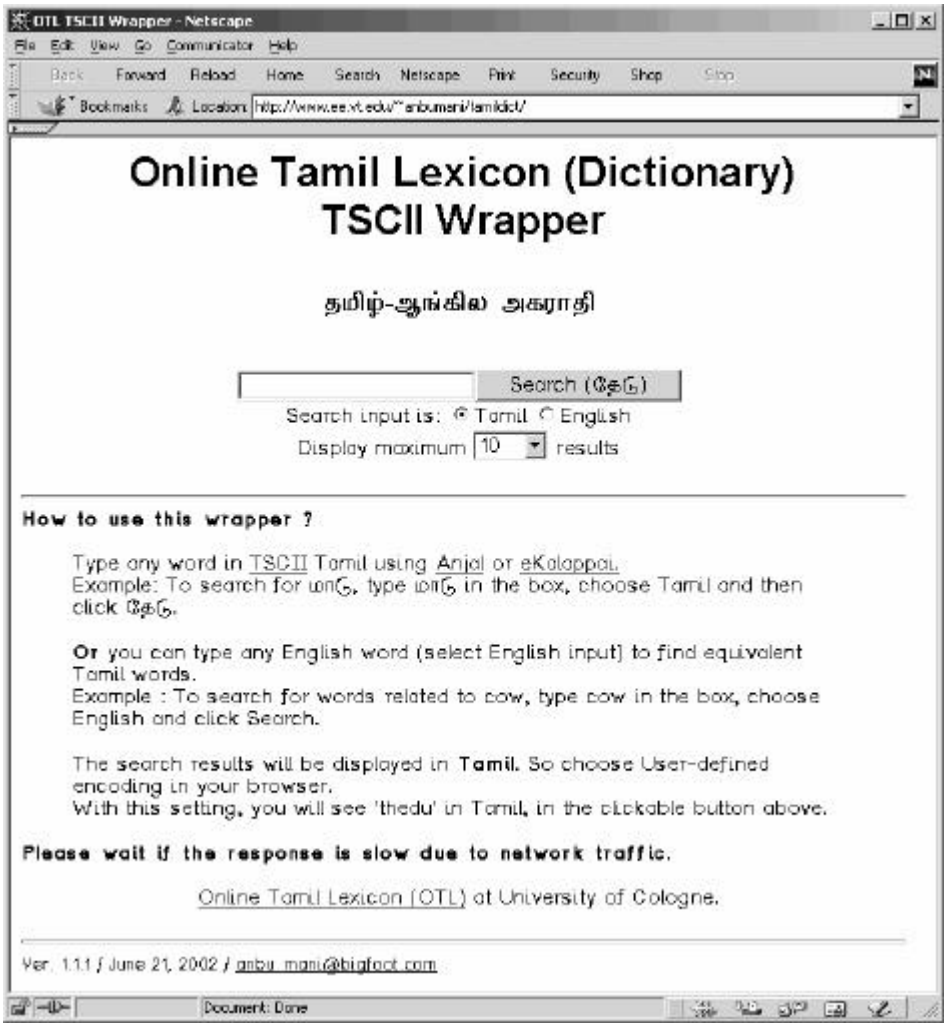
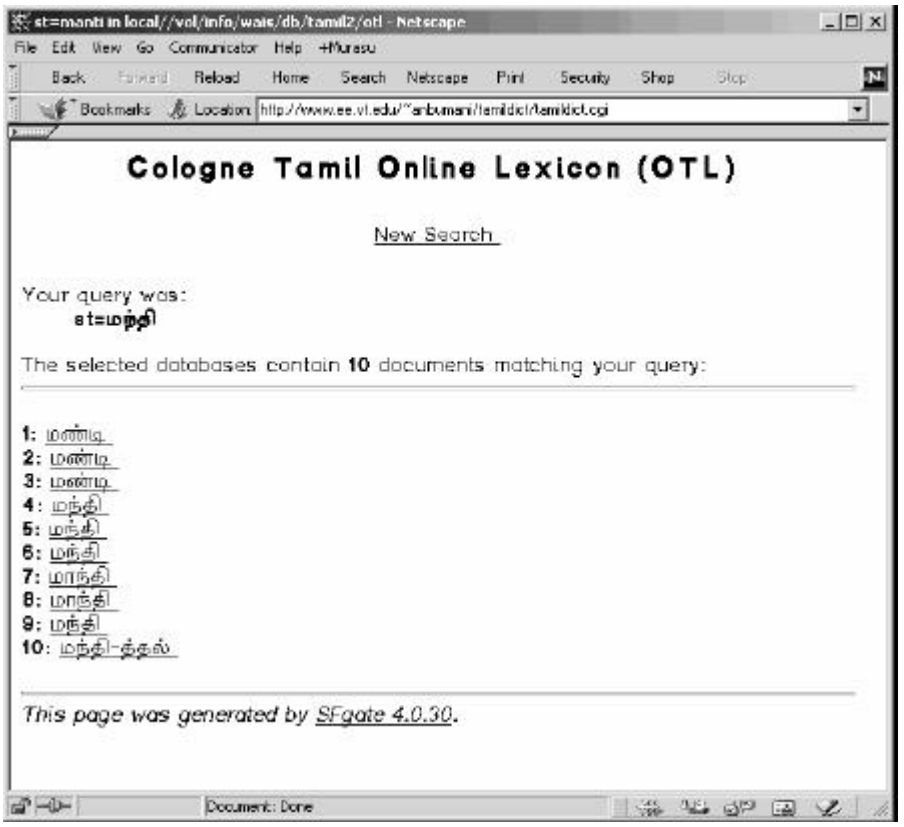
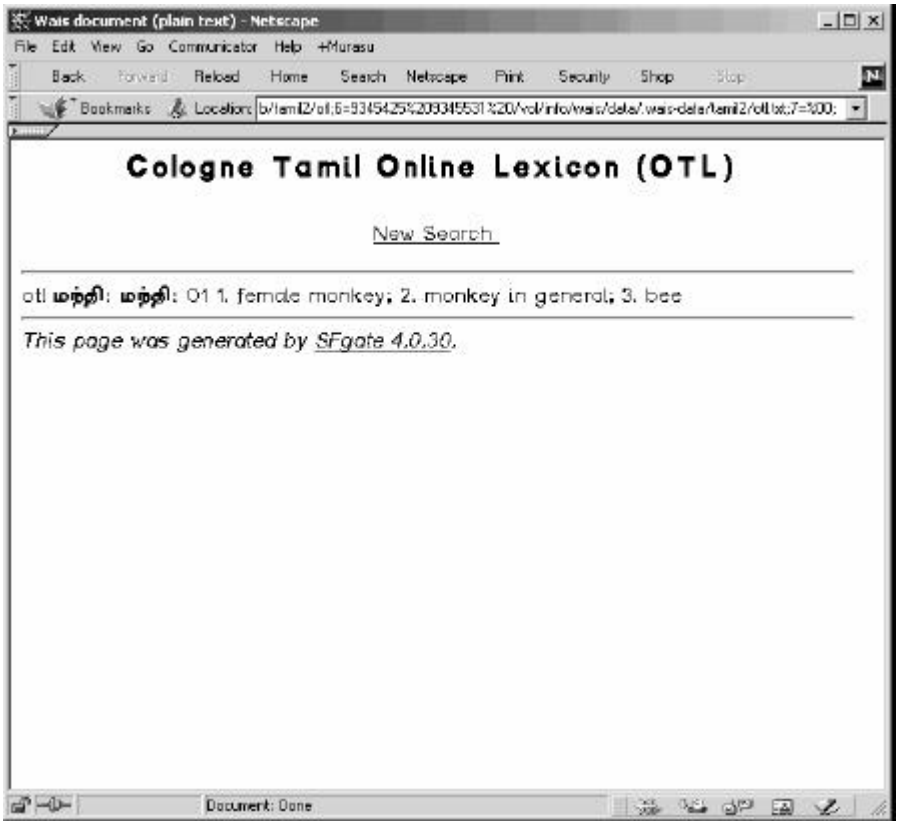



**Fig. 2.** Tamil (TSCII) interface to the OTL.




(a)




(b)

**Fig. 3.** OTL output modified by the Tamil wrapper: (a) results from a search; (b) dictionary entry of a word.

Similarly, wrappers can be implemented for other systems (like databases), where the underlying system is in a format other than TSCII/TAB coding. One could argue that a simple conversion of the database itself to the recommended coding is a better choice. This is simple to perform has to be agreed. But in some cases, this decision to convert character coding is not made by one person but is vested with a hierarchy of individuals. It is also possible that there are many related applications, which need to be rewritten if the conversion were to happen and so the mere thought of coding conversion is dreadful. Therefore in these circumstances, a simple wrapper implementation will be the most simple, logical, and inexpensive solution.

## 5. Conclusion

It can be concluded that the wrapper solution is best suited for implementation in the problems described earlier. Through this paper, a wrapper solution to the Online Tamil Lexicon was presented and its implementation was discussed in detail. The developed interface enables a user to utilize the comprehensive Tamil dictionary at University of Cologne, through a Tamil (TSCII) interface. The advantages of this design far outweigh the other possible solutions. The comparative merits of designing wrappers to other Tamil resources and directions for future work were also presented

An interested reader can get the source code of the implemented wrapper by sending an email to the author.

## References


[1] Online Tamil Lexicon, University of Cologne,
`http://www.uni-koeln.de/phil-fak/indologie/tamil/otl_search.html`

[2] Kriyavin Tarkalat Tamil Akarati, University of Chicago
`http://ets.lib.uchicago.edu/Databases/Tamil/`

[3] Alternative Tamil Dictionary (TAD),
`http://www.notam02.no/~hcholm/altlang/ht/Tamil.html`

[4] Umar, *Tamil Dictionary*,
`http://www24.brinkster.com/umarthambi/tamil/ETamil_search.asp`

[5] Tamil Computing Words, `http://www.tcwords.com`

[6] Digital Activities List, University of Chicago,
`http://www.lib.uchicago.edu/e/dl/diglist.html`

[7] Specifications and Documentation to CGI, World Wide Web Consortium (W3C),
`http://www.w3.org/CGI/`

[8] D. Sivaraj, *Anjal to TSCII Converter*,
`http://www.tamil.net/people/sivaraj/convert.html`