# BILINGUAL DATABASE SOFTWARE FRAMEWORK FOR THIRUKKURAL

**Ravi Lourdusamy[1] and Merlin Florrence Joseph[2]**
Dept. of Computer Science, Sacred Heart College, Tirupattur, Tamilnadu, India
[1]raviatschc@yahoo.com [2] merlinflorrence@gmail.com

**Abstract:**

Tamil literature has a rich and long literary tradition spanning more than two thousand years. The oldest extant works show the signs of maturity indicating an even longer period of evolution. This article presents a software framework for developing bilingual applications for such Tamil literary. A desktop application is designed to validate the software framework using the Tamil literature Thirukkural. Further the application for thirukkural is developed to support Tamil and English languages. The application is designed with effective search methods of Kural with a user friendly Graphical User Interface.

**Index Terms:** Bilingual, Framework, Unicode

## 1. INTRODUCTION

In Tamil, there are plenty of literature which consists of rich culture and greater values. To expose the values of Tamil literatures there is a need to develop software applications which supports bilingual or multilingual, so that any of the language users can taste the values and the virtues of Tamil culture. At present in the field of Tamil computing very few desktop and web based applications are available on Tamil literature. The proposed application is developed for Thirukkural which belongs to the Sangam age of Tamil literary. This literary has the rich values and it has been translated into many other languages all over the world. It is one of the most famous Tamil literatures and most readable book in the world.

The aim of this article is to give readers a comprehensive overview of bilingual database and to develop a framework for a bilingual database application. An application is developed adapting to the framework and further the application is also evaluated. This article intends to describe the structure of the bilingual database framework. Database systems need to be efficient with respect to multilingual data. Databases can be classified according to the types of it contents: bibliographic, full-text, numeric, and images. In computing, databases are sometimes classified according to their organizational approach. The most prevalent approach is the relational database, a tabular database in which data are defined so that it can be reorganized and accessed in a number of different ways. The database which is implemented to support two languages is called bilingual database and the database that supports more than two languages is known as multilingual database.

The article is further organized as follows. In section 2 related works on multilingual database are narrated in a nutshell. In section 3 the software framework developed for a bilingual database application is presented. In Section 4 the software framework is validated by implementing a desktop application adopting the framework. The conclusions are proposed in section 5.

## 2. RELATED WORKS

A database is a collection of information that is organized so that it can easily be accessed, managed, and updated. But it is quite difficult to store and manipulate more than one

language in database. This section attempts to review several research works related to the bilingual and multilingual information retrieval system.

Saraswathi.S et.al.,[1] have developed a generic platform for bilingual information retrieval which can be extended to any foreign or Indian language. Search for the solution of the query is not performed in a specific predefined set of standard languages but is chosen dynamically on processing the user's query. This system is tested for information retrieval using the keywords alone and with the use of concept words obtained from the ontology. Isao Goto et.al.,[2] have developed the database that is composed of news texts used for international broadcasting services. The multilingual database is compiled from the news texts belonging to twenty two languages. The research work describes the features of the database and the process of compiling the database and also demonstrates a multilingual translation aid system intended to support news translation. Jeffrey Sorensen et.al.,[3] have proposed a new architecture that supports all types of database such as MSSQL, MySQL, Sybase, MSAccess etc., This architecture also supports Unicode and full-text indexing. Two techniques are used to retrieve the information namely string comparison and N-gram indexing. This system uses Unstructured Information Management Architecture (UIMA) and many of the components to provide access to a multilingual database documents.

Shafi S.M et.al.,[4] have presented the design of an interface for developing a database of 'medieval manuscripts' for accommodating data elements in multilingual and multiscript medium using Unicode character sets. The design of the multilingual interface for medieval manuscripts exhibits the exploitation of the opportunities offered by the information technology without any constraints for the users. Victor Zue et.at.,[5] have developed multilingual conversational systems that support human-computer interactions. It is based on the premise that a common semantic representation can be extracted from the input for all languages, at least within the context of restricted domains. In design of such systems, language dependent information is separated from the system kernel as much as possible, and encoded in external data structures. Helmat Berger et.al.,[6] have proposed an approach for a multilingual natural language query interface that allows the formulation of queries on tourism information. The interface is based on n-gram language identification and a keyword-based query interpretation where natural language analysis is performed for identification of the requested objects. The web-based interface is designed to allow for easy addition of alternative languages and for easy adaptation to other application domains.

Martin Andreson et.al.,[7] have proposed a framework for building a FEderated MUltilingual database System(FEMUS). The term federated meant the global system that provides the functionalities to include as components, different heterogeneous database systems cooperating together. This system is responsible for transforming the global queries and updates into statements for the component schema. In a multidatabase system, local components are kept separate, such that no global federated schema exists. Petraki E et.al.,[8] have proposed the utilization of a priori conceptual relations between terms that exist independently of any documents through a controlled vocabulary known as thesaurus, which incorporates both terms and the conceptual relations among them. The research work also discusses the integration of multilingual thesauri in the set-theoretic FDB (Frame DataBase) data model, which offers a universal schema for all applications. European Statisticians Sharing Advisory Board and Uniter Nation Economic commission for Europe(UNECE) Secretariat[9], with input and peer review from participants in the 2011 joint UNECE/Eurostat/ OECD meeting on management of Statistical Information Systems (MSIS) declared some principles to develop multilingual applications.

Swpna.N et.al.,[10] described some of the most important areas of information retrieval, in particular, Cross-lingua Information Retrieval (CLIR) and Multilingual Information Retrieval (MLIR) where CLIR deals with asking questions in one language and retrieving documents in different language and MLIR deals with asking questions in one or more languages and retrieving documents in one or more different languages. It also explains a description on cross-lingual and multilingual information retrieval, its challenges and current methods, techniques and evaluation tracks to overcome problems for efficient and resourceful searching. Thenmozhi D et.al.,[11] have developed a Cross Lingual Information Retrieval (CLIR) system that helps the users to pose the query in one language and retrieve the documents in another language. It addresses the issue of translating the given query in Tamil to English using Machine Translation approach. It also uses a Morphological Analyzer to obtain the root terms of source query. The system exhibits a dynamic learning approach wherein any new word that is encountered in the translation process could be updated to the bilingual dictionary. Jialun Qin et.al.,[12] have developed and evaluated a multilingual English-Chinese Web portal in the business domain. A dictionary-based approach has been adopted that combines phrasal translation, co-occurrence analysis, and pre- and post-translation query expansion. The proposed approach can be applied in other web based applications or digital libraries.

Abu Sayed Md. Et.al.,[13] have proposed a translator-based approach for handling multilingual data that stores data in theoretical information way with minimum redundancy. In this research work, the algorithms are developed for inserting the multilingual data into a single non-redundant database and querying and updating the database. The algorithms have been evaluated by syntactic data sets generated by a data generation program. The system is implemented for two languages: English and Bangla. In this system, data has been stored in a central server and clients can perform different operations in the database dynamically in a distributed environment. A Thirukkural application called DailyKural is developed by Guru Kathiresan[14]. It was developed as a desktop application to support Tamil and English languages. The application is implemented with few search methods. The application allows the user to view one Kural at a time. There is no feature to view the whole Kural under the category of Chapter, Chapter Group and Section.

## 3. BILINGUAL SOFTWARE FRAMEWORK

The main aim of this section is to present the overview of bilingual software framework. The proposed framework is developed for the Tamil literary, Thirukkural. The application is developed with few reusable components such as: Database structure, Graphical User Interface design and programming code. The objectives of the proposed system are: to present a comprehensive overview of bilingual database applications, develop a software framework for Tamil literary, and develop a desktop application for Thirukkural based on framework and to validate the framework using bilingual database.

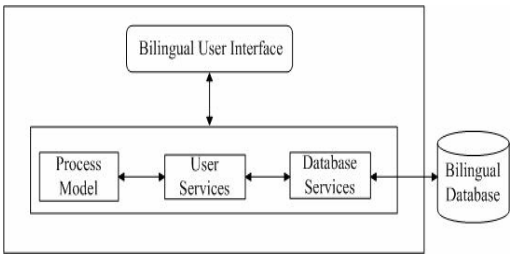


Figure 3.1 Bilingual Software Framework

The software framework is depicted in figure 3.1. In the figure, bilingual interface represents the graphical user interface of the proposed software. Process model includes the process involved in retrieving the data from database based on the user's query. The framework provides user services which means user can install the application in their system wherever they want to use. It allows the user to maintain the framework in the end system. The framework utilizes the bilingual database to retrieve data that is developed with this application.

## 4. IMPLEMENTATION AND EVALUATION

The aim of this section is to describe the development phases involved in implementing the software framework for Thirukkural. The system is implemented as a desktop application using C#.Net Windows Forms as front end and MSSQL database as backend. To support Tamil fonts in MSSQL, Unicode UTF-8 is used. The proposed framework is developed with four modules. They are: 1) Selection of language, 2) Searching methods of Kural, 3) Exporting to MSWord and MSExcel and 4) Facts about Thirukkural and Thiruvalluvar. Selection of language is the first module that is used to select the language either Tamil or English. The list of Kural is categorized under chapter and the list of chapters is structured in hierarchy form using treeview control. By selecting the specific chapter from the treeview, user can view the Kural.

The second module of this application is searching methods of Kural, which provides seven different methods to search Kural efficiently. They are: search kural by kural's number, chapter's number, name of the chapter, and name of the section, name of the chapter group, starting and ending word of the Kural. The third module is exporting Kural from application to MSWord and MSExcel. This module allows the user to export the Kural that they have selected into word format or excel sheet. The system also contains the interesting facts about Thirukkural and Thiruvalluvar. The source of Thirukkural in English is taken from G.U.Popes Thirukkural Translation and commentary[15] and the source of Thirukkural in Tamil is taken from M.Karunanithi's commentary[16].

The system is evaluated using two assessment method. They are: Criteria-based assessment and Tutorial-based assessment. In criteria-based assessment method, the usability and maintainability of the application is evaluated. The installation process of the application is clearly defined and the end user's expectations are evaluated in this assessment. In tutorial-based assessment method, the software is evaluated by the experience of the user with the application. The application is also tested with the Visual Studio Test Professional which is used to identify the bugs and assures the quality of the application. The application is evaluated by the Ph.D Scholars of Department of Tamil, Sacred Heart College. The feedback given by the scholars are analyzed and the application is revised.

## 5. CONCLUSION

A software framework for a bilingual database application is presented in detail. The framework is evaluated by developing a bilingual desktop application for Thirukkural using C#.Net Windows Forms as front end and MSSQL database as backend. The application is portable and the user can install and access the application in any system. The Graphical User Interface designed in the application is simple and user friendly. Effective search methods are incorporated in the application to search the bilingual database based on a criteria.
The application has several reusable components so that it can be adopted for any other Tamil literary by replacing Thirukkural. The bilingual database used by the application can be

redesigned as a multilingual database by adding other languages like Malayalam, Hindi etc., To perform the semantic-based search, the framework can be further implemented by using ontology as backend and C#.Net or Java as front end.

**REFERENCES**


[1] Saraswathi.S, Asma. Siddhiqaa.M. Kalaimagal.K and Kalaiyarasi.M, "Bilingual Information Retrieval System for English and Tamil", Journal of Computing, Vol 2 issue 4, April 2010 ISSN 2151-9617.

[2] Isao Goto, Naoto Kato and Terumasa Ehara, "A Multilingual news database and its application to a translation memory system", In proceeding of: 6th Natural Language Processing Pacific Rim Symposium Post-Conference Workshop on Language Resources in Asia, 11/2001

[3] Jeffrey Sorensen and Salim Roukos, "Rethinking Full-Text Search for Multilingual Databases", IEEE, Computer Society Technical Committee on Data Engineering, 2007.

[4] Shafi S.M and Nadim Akhtar Khan, "Designing an Interface for Multilingual and Multiscript Database of Medieval Manuscripts", TRIM V2(2) July-Dec, 2006.

[5] Victor Zue, Stephanie Seneff, Joseph Polifroni, Helen Meng, and James Glass, "Multilingual human-computer interactions: from information access to language learning", DARPA, N66001-94-C-6040, 1996.

[6] Helmat Berger, Michael Dittenbach, Dieter Merkl and Werner Winiwarter, "Providing multilingual natural language access to tourism information", 2011.

[7] Martin Andreson, Yann Dupont, Markus Tresch and Haiyan Ye, " FEMUS: a FEderated MUltilingual Database System", 1991

[8] Petraki E, Kapetis C, and Yennakoudakis E.J, "Conceptual database retrieval through multilingual thesauri", Computer Science and Information Technology 1(1):1932, 2013, DOI:10.13189/ csit.2013.010103.

[9] The principles and guidelines on Building Multilingual Application for official statistics, United Nations Economic Commision for Europe.

[10] Swpna.N, Padmaja Rani, Kiran Kumar, "A survey on the cross and multilingual information retrieval", National Conference on Research Trends in Computer Science and Technology-2012, Vol 3, Issue (1) NCRICST, ISSN 224 -071X.

[11] Thenmozhi D. and Aravindan C, "Tamil-English Cross Lingual Information Retrieval System", 2009.

[12] Jialun Qin, Yilu Zhou, Michael Chau and Hsinchun Chen, "Supporting Multilingual Information Retrieval in Web Applications: An English-Chinese Web Portal Experiment," IIS-9817473, 1999/4 – 2002/3.

[13] Abu Sayed Md. Latiful Hoque and Mohammad Shamsul Arefin," Multilingual Data Management In Database Environment", Malaysian Journal of Computer Science, Vol. 22(1), 2009.

[14] DailyKural Software developed by Guru Kathiresan, Murugappan P, [Online]. (URL:http://www.dailykural.com/). (Accessed 12 May 2010).

[15] Dr. K. Kalyanasundaram, "மதுரைத் தமிழ் இலக்கிய மின்தொகுப்புத் திட்டம், Project Madurai, 2002.

[16] Gokulnath Murugesan, [Online]. (URL: www.gokulnath.com/thirukkurals/). (Accessed 21, November 2005).