IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

ELECTRONIC DICTIONARY FOR SCHOOL STUDENTS REFERENCE WITH TAMIL LANGUAGE POETRY

Dr. K.Umaraj, Department of Linguistics, Madurai Kamaraj University

V.Selvakumar, Department of Linguistics, Madurai Kamaraj University

Abstract

This paper focuses on preparation of electronic dictionary for school students studying Tamil language poetry. The e- dictionaries incorporate an annotated corpus, providing a wealth of context, including historical insights and historical artifacts. Poetry is a verbal art form. It requires reading, writing and understanding. The study also explores the potential benefits and limitations of integrating technology into language learning, specifically Tamil poetry. The information technology revolution has changed the attitude towards learning and internet has become a desired delivery tool and web based language learning has become more user friendly to day. Learning of poetry is part of activity in language learning situation. The present paper proposes to develop e-dictionary for learning and enhancing the vocabularies with the annotated collected corpus from the Tamil poetry.

Keywords: Electronic Dictionary, database. Tamil poetry, Language skills, Language learning, Essential tools, Information technology.

Introduction

In today's digital age, Electronic dictionaries have become an indispensable tool for students. Whether for in class or at home use, electronic dictionaries offer several advantages over traditional print dictionaries such as faster search portability, and enhanced functionally. The electronic dictionary can become an essential tool not only for students but also for language enthusiasts and experts. With instant access to thousands of words and meanings, an electronic dictionary will enable students and even grow-ups to enhance their language skills and elevate their understanding of Tamil language poetry.

Tamil language poetry is an important part of the Tamil language and culture. Tamil literature is rich in history and tradition, with a vast collection of poetry, prose, and drama. The inclusion of Tamil language poetry in an electronic dictionary can provide students with a deeper understanding of the language, culture, and literature. The proposed electronic dictionary aims to bridge the gap between traditional and modern learning methods, combining the benefits of technology with the richness of Tamil literature.

Corpus

In linguistics, a corpus refers to a large and diverse collection of texts or speech data systematically are analyzed and annotated according to particular features of the language. Corpora can be used as a source of linguistic analysis for language learning, statistical analysis and machine learning applications among other research areas.

A Corpus can be either written or spoken text which could be abstracted from different language genres such as new papers, books, speeches and social media posts. Linguists use corpora to analyze the use and structure variation across different levels of speech and writing, including vocabulary, syntax and discourse. The analysis of corpora has significant implications for fields such as language teaching, Natural language processing and the development of language technology.

Annotated Corpus

An Annotated Corpus in linguistics is a collection of texts that have been marked up or annotated with linguistic information, such as part-of-speech tags (POS tagging), syntactic structures, or semantic roles. There are several benefits to using an annotated corpus in linguistic research. First, it provides a large and standardized data set that can be used to train and test language models, as well as to verify the accuracy of these models. Second, it allows researchers to study how language is used in various contexts and to analyze patterns and variations in language use. Finally, it provides a valuable resource for creating tools that can help people learn and use language more effectively. Creating an annotated corpus can be time-consuming and expensive process, particularly when it involves manually annotating large amounts of text. Additionally, the quality and consistency of annotations can vary between annotators or annotation schemes, which can affect the reliability and validity of research findings. Despite these challenges, annotated corpora continue to be an essential resource for linguistic research and language technology development. They enable researchers to better understand how language works, how it is used in different contexts, and how it can be effectively analyzed and processed by machines.

Previous works

There are number of electronic dictionaries for Tamil have been web published by different research institute and commercial organizations as follows:

i) Ex-PAC organization published an English – English – Tamil electronic dictionary. It has 22000 head words and 35000 sub words,

ii) Tamil lexicon and Muthu Shanmugam Pillai's – Tamil – Tamil edition was available in the Tamil virtual university website,

iii) Winslow and Muthu Shanmugam Pillai's online dictionaries,

iv) Crea's Tamil – Tamil – English dictionary. It was web enabled recently. This dictionary has 21000 head words, 38000 citation, 1700 Srilankan Tamil words, 342 pictures and 1892 pages,

v) Moli trust published 'tarkalat Tamil maraputttotar akarathi' in electronic form on September 2009.

vi) An electronic technical term in Tamil has also been developed by Chellapan, Radha. It has different kinds of retrieval and browsing facility,

vii) Corpuraiyal is on online dictionary containing 20000 root words. Each entry in the dictionary includes Tamil root words with English equivalent, different meaning of the worked and the associated syntactic category, and

viii) Visayacard Paul (2004) developed on electronic dictionary for Tamil named Multidimensional smart dictionary, a project work of central institute of Indian language.

Electronic Dictionary and Annotated Corpus

Electronic dictionaries and annotated corpus in linguistics are related concepts, as they both involve the study of language and language processing.

1. **Corpus based dictionary creation**

   Annotated corpus can be used to create electronic dictionaries that are more comprehensive and accurate than traditional dictionaries. These dictionaries can be generated by compiling large amount of text data and analyzing it to identify common phrases, collocations and usages.

2. **Lexical resource development:**

   An annotated corpus can also be used to develop lexical resources, such as lexicons or thesaurus, which can be used to improve the accuracy of machine translation, text-to-speech synthesis and other natural language processing tasks. By analyzing a corpus, it is possible to identify synonyms, antonyms, and other relationships between words, which can inform the creation of these resources.

3. **Language modeling:**

   Annotated corpus is often used for language modeling, which involves using statistics to identify patterns in language use. These models can be used to inform machine learning algorithms, which can be used to generate text, classify text, or perform other automated language processing tasks.

Limitation

Corpus are collected from the Tamil Literature, Patiṉeṇkīḻkkaṇakku nūlkaḷ. The poetry 'Iṉṉā nāṟpatu', 'iṉiyavai nāṟpatu', 'kār nāṟpatu'and 'kaḷavaḻi nāṟpatu' and 35057 corpus have been collected.

Methodology

Texts in terms of data are collected from the poetry which the students are learning in the school situation for preparing E-dictionary. After collecting the texts, grammatical information and meaning with citations are given. Thereafter the content will be put into database and an advanced search engine will be developed. The data will be encoded in Unicode notation after i) collecting and selecting the poetries, ii) annotating corpus, iii) translating texts, iv) building a data base, v) entry creation and vi) designing the user interface.

In the creation of database, the data analysis is based on the analysis of data in the format of data, size of the data and required database management system. Data architecture is to construct the data redundancy in different normal forms to make a database and avoid the duplicate values and null values of any database. The database design is used to construct the database where in columns are needed, which size of the data and which format are to be followed to prepare database.

The electronic dictionary for school student's dictionary database contains different fields which are serial number, word, grammatical category in Tamil, Tamil and English language meaning.

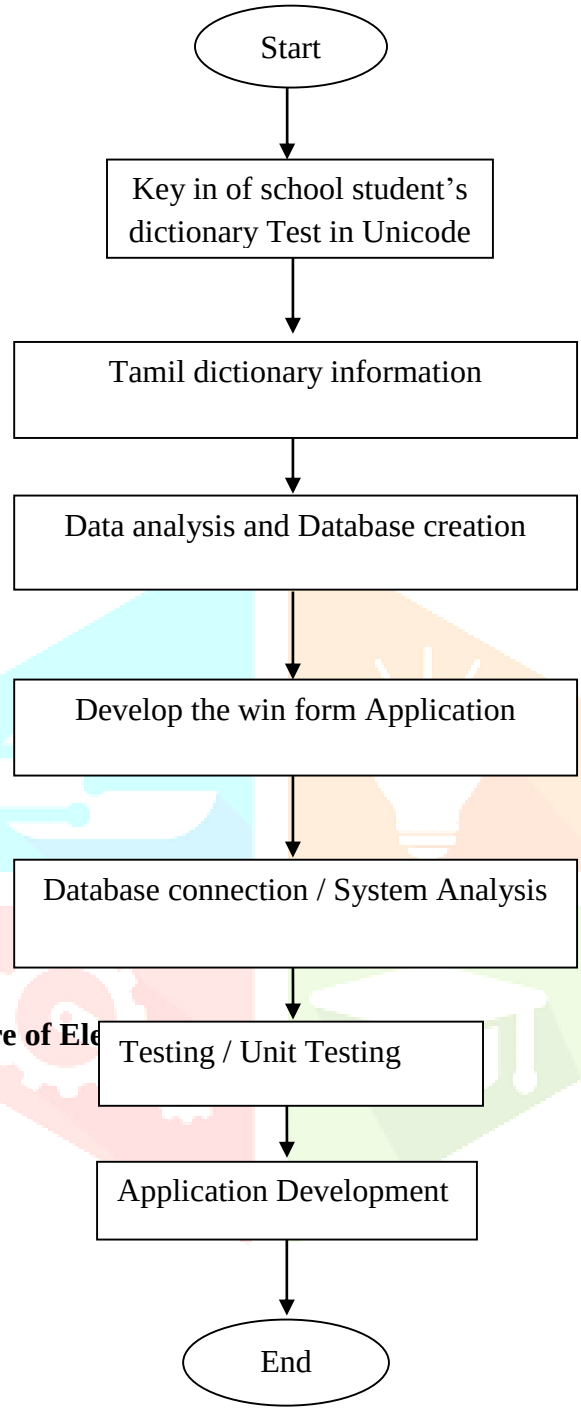
The database is developed using MS Access 2016, the key used in the school students dictionary is composed by UNICODE format; the Unicode format is used to access on web in all platforms. The following five fields such as word, grammatical category in Tamil, meaning in English, concordance, book are entered in UNICODE format.

Unit Testing - In the testing level, UNIT testing is applied.

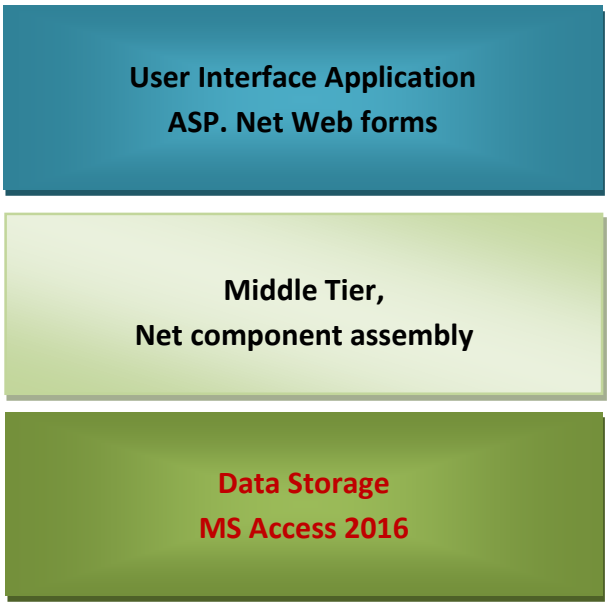
Application development

The successful completion of software developed and unit testing, the program is to be developed as an application software with pre-request tools like Net frame work 4.0 and windows installer. The Electronic dictionary for school student's dictionary software is to develop using data flow diagram.

Data flow diagram



Architecture of Ele



Approaches for identifying the Head words in dictionary

Head word means abstraction or representative of the grammatical forms which actually occur in different sentences. Head word is the word under which a set of related dictionary entries appear. Head word begins with a separate entry in a dictionary.

Morphological analysis: Tamil is an agglutinative language, which means that words are formed by adding affixes to the root word. Morphological analysis is therefore an important approach for identifying the head word in a Tamil poetry dictionary. For example, in the Tamil word கண்+கொடு "kan+kodu", the root word is 'கண்', 'kan' and it means 'eye'. So 'கண்', 'kan' would be the head word.

Frequency analysis: Frequency analysis is another important approach for identifying the head word in a Tamil poetry dictionary. However, in Tamil poetry, there may be different forms of the same word used based on the context, poetic style, or the period in which it was written. Therefore, frequency analysis may not always provide a clear indication of the head word.

Etymological analysis: Tamil has a long and rich literary history, and many words have evolved over time. Etymological analysis can be used to trace the origin and history of a word to determine its root form. For example, in the Tamil word 'காது', 'kaadu' which means "ear," the root word 'கா', 'kaa' which means "to hear."

Semantic analysis: In Tamil poetry, words are often used symbolically or metaphorically. Semantic analysis can be used to identify the underlying meaning of a word and its root form. For example, in the Tamil word 'மலர்' 'malar' which means 'flower' the root word is 'மல்', 'mal' which means "to blossom".

Combination of approaches: As with other languages, a combination of these approaches can be used to identify the head word in a Tamil poetry dictionary. Morphological analysis and etymological analysis are particularly important, given the nature of the Tamil language and its literary tradition.

Ex. ஆற்றும்<VB>துணையால்<NN>அறம்<NN>செய்கை<NN>முன்<NN>இனிதே<NN>

பாற்பட்டார்<NN>கூறும்<VB>பயமொழி<NN>மாண்பு<NN>இனிதே<NN>

வாய்ப்பு<NN>உடையாராகி<NN>வலவைகள்<NN>அல்லாரைக்<VB>

காப்பு<NN>அடையக்<VB>கோடல்<NN>இனிது<VB>

ஆற்றும் ஆற்று|+verb|+fin|+sim|+strong|+fut=உம்|+3sgn=∅,

துணையால்- துணை+noun+inst, அறம்+noun+nom, செய்கை+noun+sg+nom,

முன்+noun+nom, இனிதே+noun+sg+nom, பாற்பட்டார்+noun+sg+nom,

கூறும் கூறு|+verb|+fin|+sim|+strong|+fut=உம்|+3sgn=∅, பயமொழி+noun+sg+nom, மாண்பு+noun+nom,

இனிதே+noun+sg+nom, வாய்ப்பு+noun+sg+nom, உடையாராகி+noun+sg+nom, வலவைகள்

வலவை+noun+pl+nom,

அல்லாரைக்-அல்லாரை+verb+nonfin+sim+inf+sandhik, காப்பு+noun+sg+nom,

அடையக்- அடை|+verb|+nonfin|+sim|+inf=அ|+sandhi-k=க், கோடல்+noun+nom ,

இனிது- இனி+verb+nonfin+sim+vpart.

Collection of Corpus in Poetries - Tamil, Tamil and Meaning in English

	Tamil	Tamil	English
1.	இன்னா	துன்பம்	suffering
2.	துறவோர்	முனிவர்	Saint
3.	ஒழுகு	நடத்தல்	practical way of Life
4.	பந்தம்	சுற்றம்	relatives
5.	வனப்பு	அழகு	beauty
6.	இல்	மனையில்	house, plot
7.	புகல்	நுழைதல்	to enter
8.	செறல்	கோபம்,கொல்லுகை	anger

9. ஓம்பல்	பாதுகாத்தல்	protection
10. மறம்	வீரம்	bravery
11. சுரம்	பாலைவனம்	desert
12. குஞ்சரம்	யானை	elephant

Conclusion

The proposed electronic dictionary for school students with Tamil language poetry and annotated corpus has the potential to be a valuable resource for language learning. The dictionary can help students to improve their vocabulary, understanding of Tamil language and literature, and overall language proficiency. The representation of the meaning of the word is an important characteristic of the E-dictionary; the developer can ensure that the words exhibit their meanings in such a way that they can be easily conceived by the users on the effectiveness of the online dictionaries over the paper dictionaries. But it must be noted that different types of dictionaries are considered suitable for different types of users according to their needs and expectations. There cannot be just a single dictionary that can satisfy all the different needs of its users. Therefore the developers of the e-dictionaries should pay extra attention so that they do not give away any incorrect translations and the language used so that the meanings of the words can be easily understood. It can be said that the work to be done in improving the dictionaries and making them more user friendly is never ending. Some of these areas can be the illustrative example, use of parts of the speech, pictorial representation of words and providing audio facilities to make the users aware of the cultural items that they have never come across, etc. The studies have revealed undoubtedly, there is a great need for improvement in e-dictionaries regarding entries and their presentation; nevertheless It is to be noted that the establishment of an Internet based dictionary certainly is a lengthy and challenging process, which takes years and might probably never be finished due to the continuous development of language. It is without controversy that online dictionaries will be improved and extended in the course of time web-based resources, in particular, internet dictionaries will become more popular and the number of demands will be increased.

Acknowledgement

We the authors of this article acknowledge the Madurai Kamaraj University - RUSA for providing Computer and Printer so as to execute research in the project area.

References

Ramakrishnan.S (Ed), *Kriyavin tarkalat Tamil akarathi (Dictionary of Contemporary Tamil)* Chennai Mozhi – Cre – A, 2008.

Fell baum Wordnet: *An electronic lexical database, Cambridge (Massachuseuts):* MIT Press. 1998.

K.Umaraj, *An electronic dictionary for pathinenkeelkannaku literatur*e, Madurai Kamaraj University, Madurai, 2013.

Cruse, D.A., *Lexical Semantics*, Cambridge University Press, Cambridge, 1986.

Chidambaranatha Chettiyar (Edr.), *English Tamil Dictionary*, University of Madras, Reprint, 1998.

Misra, B.C., *Lexicography in India*, Central Institute of Indian Languages Publication, Manasangangotri, Mysore, 1980.

Vaiyapuri Pillai,S., *History of Tamil Lexicography*, Tamil Lexicon- Vol. XXXVII, University of Madras, Madras, 1936.