Kanimozhi - a computer language in Tamil

N. Senthilraja, B. Amutha M. Ponnavaikko

SRM University, Chennai

Abstract

Kanimozhi is a Programming language in Tamil designed for writing Programs in Tamil language. The keywords for kanimozhi is observed from tamil language for writing programs. The existing Tamil Compilers available are using Unicode for Tamil language processing. Unicode Tamil has 31 code positions only out of 247 Tamil Characters. These 31 characters include 12 vowels, 18 Agra-uyirmey and one atom. Five Grantha Agra-uyirmey are also provided code space in Unicode Tamil. The other Tamil Characters have to be rendered using a separate software. Only 10% of the Tamil Characters provide code space in the Present Unicode Tamil. 90% of the Tamil Characters that are used in general text interchange donot provide the necessary code space. This problem was solved in TACE16 encoding. And TACE16 is efficient in terms of data storage application, sorting index structures and processing speed. Kanimozhi uses TACE16 unicode standards.

1. Introduction

TACE16 is a 16 bit character encoding technique where all Tamil characters can be represented through a single character. There is no specific Tamil Compiler available to fulfill all these necessities as of today. We have been enthroned to design a Tamil Compiler with unique codes for all the Tamil characters and to design a compiler for Tamil which will be used for executing Tamil programs. There are 5 phases in the Tamil Kanimozhi design. They are Lexical Analyzer, Syntax Analyzer, Semantic Analyzer, Code Generator and Optimizer. As a part of the Tamil compiler design, we have completed the design of the Lexical analyzer and achieved good results.

2. Existing attempts at Programming languages in Tamil

Many solutions were suggested earlier for Programming languages in Tamil. We will review two of the proposed solutions **Swaram** and **Ezhil**.

Swaram a static typed Tamil Programming language was introduced first for Tamil language in 2003. It has feature set related to C Programming language. Swaram allows mixed English and Tamil identifiers so that it can access external libraries in English. Keywords will be in tamil. Swaram is not publicly available, which severly limits language development, system use, community support and improvement.

Ezhil also a static typed Tamil Programming language was introduced in 2008. Ezhil incorporates most of the concepts from Swaram.

Both these Programmin Languages uses Unicode Tamil. Unicode Tamil is a 8 bit encoding which requires multiple code points for the most used characters. Some of the limitations of Unicode Tamil are storage size doubles, Security vulnerabilities, Ambiguous combinations(requires normalization) and Simple counting, sorting, searching inefficient.

3. TACE16 Encoding

Tamil All Character Encoding(TACE16) is a 16-bit unicode based character encoding scheme for Tamil language. TACE16 character encoding scheme not only overcomes all the issues with the present Unicode encoding standard for Tamil language which are mentioned above, but also provides additional advantage over major performance improvements in both processing time and processing space which are the major factors in affecting the efficient and speedy execution of any computer based program. It uses Tamil99 and Tamil Typewriter keyboard layouts, which are approved by Tamil Nadu Government, and maps the input keystrokes to its corresponding characters of TACE16 scheme.

The TACE16 has the code positions as given in table below:

	xx0	xx1	xx2	xx3	xx4	xx5	xx6	xx7	xx8	xx9	xxA	xxB	xxC	xxD	xxE	xxF	xy0	xy1	xy2	xy3	xy4	xy5	xy6	xy7	xy8	xy9	xyA	xyB
0		க்	ங்	ச்	ஞ்	ட்	ண்	த்	ந்	ப்	ம்	ய்	ர்	ல்	வ்	ழ்	ள்	ற்	ன்	ஜ்	ஶ்	ஷ்	ஸ்	ஹ்	க்ஷ்		◌	◌
1	அ	க	ங	ச	ஞ	ட	ண	த	ந	ப	ம	ய	ர	ல	வ	ழ	ள	ற	ன	ஜ	ஶ	ஷ	ஸ	ஹ	க்ஷ		க	மீ
2	ஆ	கா	ஙா	சா	ஞா	டா	ணா	தா	நா	பா	மா	யா	ரா	லா	வா	ழா	ளா	றா	னா	ஜா	ஶா	ஷா	ஸா	ஹா	க்ஷா		கா	ஹ்ம
3	இ	கி	ஙி	சி	ஞி	டி	ணி	தி	நி	பி	மி	யி	ரி	லி	வி	ழி	ளி	றி	னி	ஜி	ஶி	ஷி	ஸி	ஹி	க்ஷி		கி	யு
4	ஈ	கீ	ஙீ	சீ	ஞீ	டீ	ணீ	தீ	நீ	பீ	மீ	யீ	ரீ	லீ	வீ	ழீ	ளீ	றீ	னீ	ஜீ	ஶீ	ஷீ	ஸீ	ஹீ	க்ஷீ		கீ	ஊ
5	உ	கு	ஙு	சு	ஞு	டு	ணு	து	நு	பு	மு	யு	ரு	லு	வு	ழு	ளு	று	னு	ஜு	ஶு	ஷு	ஸு	ஹு	க்ஷு		கு	ெ
6	ஊ	கூ	ஙூ	சூ	ஞூ	டூ	ணூ	தூ	நூ	பூ	மூ	யூ	ரூ	லூ	வூ	ழூ	ளூ	றூ	னூ	ஜூ	ஶூ	ஷூ	ஸூ	ஹூ	க்ஷூ		கூ	ே
7	எ	கெ	ஙெ	செ	ஞெ	டெ	ணெ	தெ	நெ	பெ	மெ	யெ	ரெ	லெ	வெ	ழெ	ளெ	றெ	னெ	ஜெ	ஶெ	ஷெ	ஸெ	ஹெ	க்ஷெ		கெ	ை
8	ஏ	கே	ஙே	சே	ஞே	டே	ணே	தே	நே	பே	மே	யே	ரே	லே	வே	ழே	ளே	றே	னே	ஜே	ஶே	ஷே	ஸே	ஹே	க்ஷே		கே	
9	ஐ	கை	ஙை	சை	ஞை	டை	ணை	தை	நை	பை	மை	யை	ரை	லை	வை	ழை	ளை	றை	னை	ஜை	ஶை	ஷை	ஸை	ஹை	க்ஷை		கை	
A	ஒ	கொ	ஙொ	சொ	ஞொ	டொ	ணொ	தொ	நொ	பொ	மொ	யொ	ரொ	லொ	வொ	ழொ	ளொ	றொ	னொ	ஜொ	ஶொ	ஷொ	ஸொ	ஹொ	க்ஷொ		கொ	
B	ஓ	கோ	ஙோ	சோ	ஞோ	டோ	ணோ	தோ	நோ	போ	மோ	யோ	ரோ	லோ	வோ	ழோ	ளோ	றோ	னோ	ஜோ	ஶோ	ஷோ	ஸோ	ஹோ	க்ஷோ		கோ	
C	ஔ	கௌ	ஙௌ	சௌ	ஞௌ	டௌ	ணௌ	தௌ	நௌ	பௌ	மௌ	யௌ	ரௌ	லௌ	வௌ	ழௌ	ளௌ	றௌ	னௌ	ஜௌ	ஶௌ	ஷௌ	ஸௌ	ஹௌ	க்ஷௌ		கௌ	
D	ஃ																								ஸ்ரீ			
E																												
F																												

The encoding is Universal since it encompasses all characters that are found in general Tamil text interchange. The encoding is very efficient to parse.

The characters can be processed by simple arithmetic operations.

It is very efficient to divide a vowel-consonant(UyirMei) character into its corresponding vowel and consonant. This is very efficient in terms of performance over large data.

Also it is very efficient to find whether a character is vowel or consonant or vowel-consonant(UyirMei) or numbers.

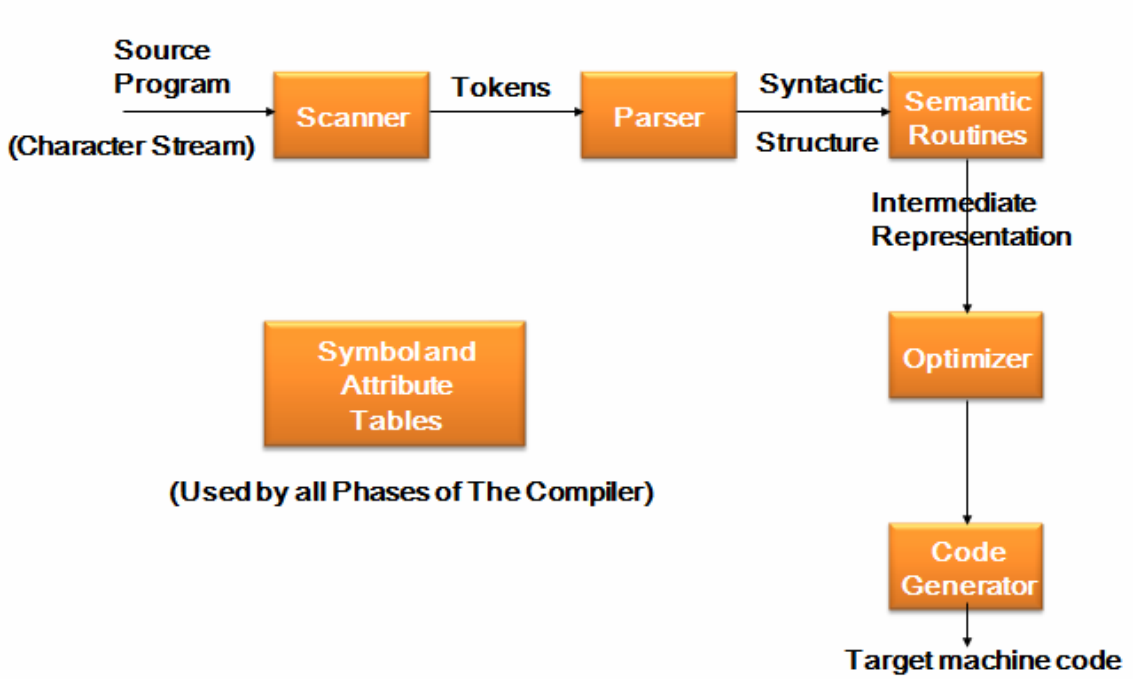
4. Kanimozhi – a Tamil Programming language

Kanimozhi a Tamil Programming language uses TACE16 16 bit encoding.

Kanimozhi program is completely written in Tamiil language. All the keywords are represented in Tamil language. It uses all the existing operators in current programming languages. Some of the keywords and the equivalent tamil version are shown in the table below.

Keywords	**Keywords in Kanimozhi**
If statement	எனில் கூற்று
If then else statement	எனில்….இல்லெனில் கூற்று
Goto statement	செல் கூற்று
For statement	ஆக கூற்று
While statement	உண்மையெனில் கூற்று

The phase diagram of Kanimozhi is as follows:



i) Scanner

The Lexical Analyzer is the first phase of Kanimozhi. It begins the analysis of the source program by reading the input, character by character, and grouping characters into individual words and symbols (tokens). We have completed the design of Lexical Analyzer.

For example consider the following statement.

```
உண்மையெனில் கூற்று(ப>=ம)
{
   ப = ப - 3;
}
```

When the above statetment is passed to Lexical Analyzer, it splits the statement into tokens.

LEXEME	TOKEN
KEYWORD	உண்மையெனில் கூற்று
LPAREN	(
ID	ப
COMPARISON	>=
ID	ம

RPAREN	)
LCURLY	{
ID	␣
ASSIGNMENT	=
ID	␣
ARTIHMETIC	-
INTEGER	3
SEMICOLON	;
RCURLY	}

ii) Parser

iii) Semantic analyzer

iv) Code Optimizer

v) Code Generator

Conclusion

Kanimozhi will be useful for most of people who are expert in Tamil and for rural tamil people but could not use Computers to the core because of the language gap. They need a Programming language to use the Computer in an easy and efficient way. The Research is to design and develop Kanimozhi as a full fledged Tamil Compiler for Tamil people intending to use Compiler efficiently. Our goal is to spread the computer functionalities to all the people who are not bound to English language which in turn will help to develop Tamil language.

References

[1] S.G. Ganesh, G.R. Prakash and K.K. Ravi Kumar, An overview of 'Swaram' : A programming language in Tamil, Proceedings of Tamil Internet Conference, 2003

[2]Mutthaiya Annamalai, (2008) Ezhil Project:
http://students.uta.edu/mx/mxa6471/download.html.

[3] Laurel Peterson, (1999) Principles of Compiler design, Pearson Publication.

[4] Dr. M. Ponnavaikko, Mani M. Manivannan and Manoj Annadurai, Review of Tamil Unicode, http://www.unicode.org/L2/L2007/07193-tamil-pres-2.pdf