ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2017.0227

LOSSLESS TEXT COMPRESSION FOR UNICODE TAMIL DOCUMENTS

B. Vijayalakshmi and N. Sasirekha
Department of Computer Science, Vidyasagar College of Arts and Science, India

Abstract

Data compressions for different world languages including Indian languages are in high need and demand. Tamil language is one of the longest-surviving classical languages in the world. Usage of Tamil language for communication and storage was increased due to the digitization of government documents and orders. Lossless text compression process for Tamil language document involves substituting an ASCII character in place of Unicode Tamil characters, since the size of an ASCII character is one byte where as a Unicode character size range between 1 byte to 4 bytes depends on the encoding file storage type. The decompression process involves the reverse of compression technique (i.e) replacing ASCII characters with Unicode characters. This paper describes about the architecture of compression and decompression process for Tamil text documents.

Keywords:

Compression, Decompression, Unicode, ASCII and Substitution

1. INTRODUCTION

Data Compression is the process of converting an input data stream to another data stream that has a smaller size. Data compression will encode or replace the original information or representation by a fewer bit characters which is reduced in size. Compression is very useful because it reduces the usage of resources required to store and transmit. This compressed file can be reversed to obtain the original file. It is called as decompression. The two types of compression technique available are lossy compression and lossless compression. The lossy compression results in some loss of data from the original while performing the decompression process. The lossless compression on the other hand will retain its original file exactly without any loss of data. There are many compression types available for text, image, audio and video etc.

This paper deals with lossless text compression for Tamil documents. Text compression is a decrease in the quantity of bits required to signify the data. Compressed data can accumulate less storage capacity, enlarge the velocity of communication and diminish the cost for storage hardware and network bandwidth [9] [17] [21].

Text Compression can be as simple as removing all unnecessary characters, inserting a particular recurring character by a string characters and substituting a minor bit string for a repeatedly taking place bit string. Lossless text compression enables the re-establishment of a file to its original position without the loss of a single bit of data, when the file is uncompressed [11] [18] [26].

There are many compression techniques available for different world and Indian languages. Languages like English, German, France, Chinese etc. has developed many compression techniques. Lots of research is going on in the development of compression process for Indian languages. Many developments in compression technique were going on in other languages, such as Japanese and Chinese [7]. There is a high demand to do compression for different Indian languages. One of the popular Indian languages is Tamil. There was no exclusive compression technique available for Tamil language.

Tamil is an abugida language. An abugida is a kind of syllabify in which the vowel is altered by modifying the foundation consonant symbol, so that all the forms that correspond to a given consonant plus each vowel be similar to one another. Amharic, Hindi and Burmese are also abugida languages. The Tamil script has 12 vowels, 18 consonants and 1 aytam character (neither vowel nor consonant). Apart from that a set of 216 combining letters formed by adding vowel marker to the consonant. Totally there are 247 characters available in Basic Tamil script [1]. Some vowels require the basic shape of the consonant to be altered in a way that is specific to that vowel. Others are written by adding vowel specific suffix to consonant, specific prefix to consonant and both suffix and prefix to a consonant.

The most widely used characters for Tamil document is Unicode characters. It is the widely suitable character set collections for almost all languages and its characters. This Unicode characters are used for data storage and transmission for Tamil documents also. There is no compression technique available for Tamil language. This paper fills the gap by doing compression for Tamil documents. A Tamil Unicode characters need two bytes whereas an ASCII character occupies one byte for a character [10] [16]. The technique can do compression by replacing a single character in place of repeated words for Unicode characters (16 to 32 bits). The compression architecture provided here will take a Tamil Unicode type text file as input and compress the file itself to nearly 50%.

2. RELATED WORKS

Many compression techniques are available for English language and European languages. But for Indian languages a special kind of compression technique should be designed and developed.

Linkon et al. proposed a Modified LZW dictionary based index compression technique for Bangle language [4]. Gleave et al. [8] represent a modified technique of byte-based compressors to operate directly on Unicode characters. Kulkarni et al. [15] explains the method of compression and decompression enhancement with the usage of character and cluster table. In [19] Ramachandran et al. describes the importance of Tamil as a language becomes a Computational Intelligence for Societal Development language in Developing Countries. In [20], the Malayalam text compression by variable length encoding is explained. The Unicode character is represented by less number of bits. Seethalakshmi [23] presented the importance of Unicode encoding technique which was followed for Tamil documentation and software. Jiří [24] represents lossless data compression for Czech language. The grammatical rule and properties for natural languages are different from English language, so compression should be specially designed. In [25], a font is created by mapping ASCII character with Unicode characters. For Indian languages the combination of characters can be replaced by ASCII characters. In [27], the character set of Tamil language is represented. The Unicode characters are universally accepted encoding technique for representing text as well as transmission.

3. EXISTING COMPRESSION TECHNIQUES FOR TAMIL DOCUMENT

The existing compression techniques were mainly deals with European languages like English, French and Germany [3], etc. The Indian languages make use of these compression techniques to reduce the storage space. The major problem is that will performing decompression process many characters are wrongly decode which gives a meaningless output. Many improvisation and development were carried out for the compression process of Indian languages. Text Compression for Guajarati and Malayalam was carried out with the dictionary approach and bit replacement reduction technique respectively [22], [13].

The existing lossless compression techniques like WinZip, a popular Windows program that compresses files then it packages them in an archive. Archive file formats that support compression include ZIP and RAR. The bzip2 and gzip formats are widely used for compressing individual files for English documents. But for natural languages like Tamil a special compression technique should be needed to design.

4. PROPOSED COMPRESSION TECHNIQUE FOR TAMIL DOCUMENT

In computing, a character encoding technique is used to represent a collection of characters used both for transmission and storage in memory. Depending on the abstraction level, context, the corresponding code points and the resulting code space may be represented as bit patterns, octets, natural numbers and electrical pulses, etc. A character encoding is used in computation, data storage, and transmission of textual data. Many encoding types like ANSI, UTF-8, UTF-16, UTF-13 etc. are available.

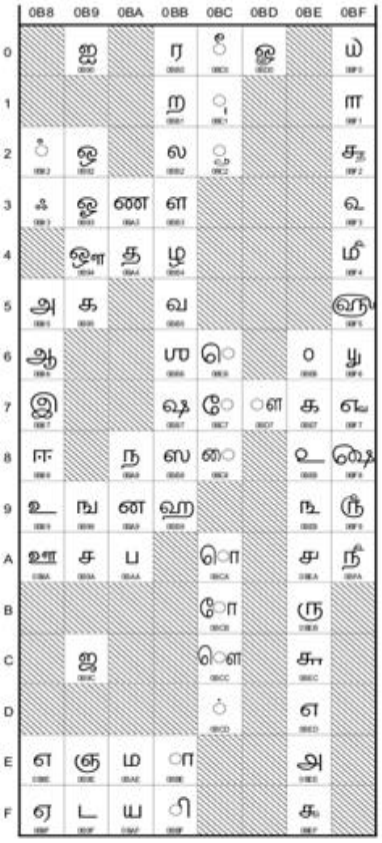
4.1 TAMIL DOCUMENTS IN DIGITAL FORM

Over 65 million Tamils in India and 80 million worldwide, Millions of petitions, commercial transaction registrations, birth/death records are generated in Tamil language every year. The Tamil Nadu government continuously involved in the process of digitizing its billions of records. The government of Tamil Nadu issued an order to use Unicode as current standard for Tamil encoding [28]. The encoding techniques available for Tamil characters are ISCII (7 bits), TSCII/TAB (7bits), TAM (8 bits), 7 bit – Unicode and Proprietary encodings (7/8 bits) [29].

The limitations in the above encoding techniques are it is insufficient to represent all Tamil characters and it is inefficient to store, transmit and retrieve the documents. These problems can be solved by the Unicode Tamil characters set.

4.2 UNICODE TAMIL CHARACTERS

The Unicode is the most acceptable industrial standards for storing, transmitting and documentation [1][14]. It was developed in conjunction with the Universal Coded Character Set (UCS) standard and published as the Unicode Standard. The latest version of Unicode contains a repertoire of more than 128,000 characters covering 135 modern and historic scripts, as well as multiple symbol sets.



	0B8	0B9	0BA	0BB	0BC	0BD	0BE	0BF
0		ஐ		ர	◌ீ	ௐ		௰
1				ற	◌ு			௱
2	◌ஂ	ஒ		ல	◌ூ			௲
3	ஃ	ஓ	ண	ள				௳
4		ஔ	த	ழ				௴
5	அ	க		வ				௵
6	ஆ			ஶ	◌ெ		௦	௶
7	இ			ஷ	◌ே	◌ௗ	௧	௷
8	ஈ		ந	ஸ	◌ை		௨	௸
9	உ	ங	ன	ஹ			௩	௹
A	ஊ	ச	ப		◌ொ		௪	௺
B					◌ோ		௫	
C		ஜ			◌ௌ		௬	
D					◌்		௭	
E	எ	ஞ	ம	◌ா			௮	
F	ஏ	ட	ய	◌ி			௯	

Fig.1. Unicode Version 10.0 for Tamil Characters

Unicode is designed to represent almost all characters in every language in the world [25]. All the characters of Tamil language are now encoded as per the Universal Principle of Unicode. The Tamil characters are range from U+0B80 to U+0BFF in Unicode character set [27].

It is large enough to encompass all characters that are likely to be used in general text interchange, including those in major international, national, and industry character sets. Unicode occupy more space in memory during storage [23]. The Fig.1 shows the recent Unicode version for Tamil Unicode character set [29].

The proposed system involves substituting an ASCII character in place of a Unicode Tamil character, since the size of an ASCII character is one byte (8 bits) where as a Unicode character size range between 1 byte (8 bits) to 4 bytes (32 bits) depends on the encoding technique. To store the file with .txt extension the following encoding types are available, they are ANSI encoding, UTF encoding, Unicode and Unicode big endian encoding. The bits required to store each characters for the above encoding file type are 8 bits, 8 to 32 bits, 16 bits and 16 to 32 bits respectively.

4.3 ARCHITECTURE OF COMPRESSION

To perform lossless data compression for natural languages, a special techniques is needed to design exclusively for particular language (Tamil) is needed which is different from English or other languages [24]. The proposed substitution method performs lossless text compression for Tamil language documents in an effective way. The lossless text decompression will reconstruct the document which was exactly same in the original document [5]. This method reduces almost 50% of storage space so that it is suitable for transformation and also saves the hard disk memory. Substituting a predefined available content to the natural.

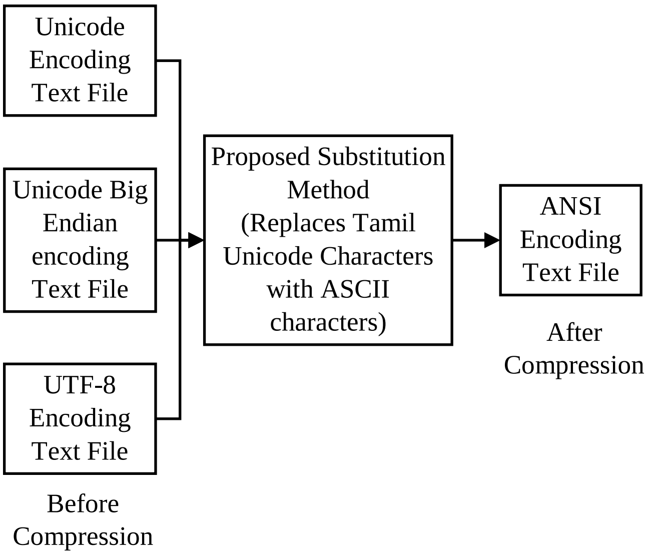

Unicode Encoding Text File
Unicode Big Endian encoding Text File
UTF-8 Encoding Text File
Before Compression
Proposed Substitution Method (Replaces Tamil Unicode Characters with ASCII characters)
ANSI Encoding Text File
After Compression

Fig.2. Lossless Compression Technique

Architecture language is getting increase due to the usage of internet [6]. The Fig.2 shows lossless technique before and after proposed substitution method. The text file which contains the Tamil documents with any one of the encoding technique like Unicode, Unicode big endian or UTF-8 will be given as input to the proposed method. A dictionary which contains collection of Unicode characters indexed with ASCII characters are used for compression and decompression process [16]. The Unicode Tamil characters (16 bits to 32 bits) will be replaced with ASCII characters (8 bits) using proposed substitution method. The compressed file will contain only the ASCII characters. This ensures 50% of compression. The compressed file will be stored in ANSI encoding type.

The compressed file will be reduced to 50% from its original size. This file can be further compressed by any one of the lossless compression technique like Run-length, Huffman or Lempel-ziv etc. that results in again 20% to 40% reduction of storage space [12].

The decompression involves in doing the same method in reverse process by giving an ANSI file as input [22]. This ANSI file contains compressed data with collection of unreadable ASCII characters. The reverse process of substitution method will be performed (i.e.) replacing Tamil Unicode characters in the place of ASCII characters. The resultant decompressed file may be in any one of the encoding type given above in Fig.3.

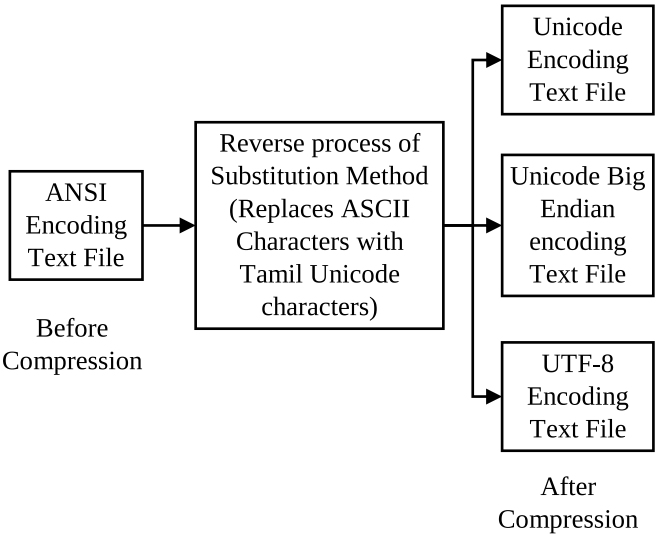

ANSI Encoding Text File
Before Compression
Reverse process of Substitution Method (Replaces ASCII Characters with Tamil Unicode characters)
Unicode Encoding Text File
Unicode Big Endian encoding Text File
UTF-8 Encoding Text File
After Compression

Fig.3. Architecture of Decompression Technique

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

The compression and decompression process was developed as a web application using ASP.NET. In future it is easy for the users to do compression and decompression in online itself. ASP.NET is an open-source server-side web application framework designed for web development to produce dynamic web pages. It was developed by Microsoft to allow programmers to build dynamic web sites, web applications and web services [2]. This ASP.NET web application is developed from Microsoft Visual Studio.

The following steps show how the compression process takes place for a Tamil word.

Table.1. Combination of Unicode characters for a single Tamil character

Tamil Character in text file	Combination of Unicode characters	Unicode 16 bit
தி	த	0BA4
	ி	0BC0
ரு	ர	0BB0
	ு	0BC1
க்	க	0B95
	்	0BCD
கு	க	0B95
	ு	0BC1
ற	ற	0BB1
ள்	ள	0BB3
	்	0BCD

i. The word திருக்குறள் seems to have 6 characters, but actually it is the combination of 11 Unicode characters listed in the Table.5.1.

ii. The Unicode characters combination of the word [23] திருக்குறள் is given below த ◌ீ ர ◌ு க ◌் க ◌ு ற ள ◌்

iii. The proposed substitution method will replace the existing Unicode Tamil characters with ASCII characters for the above example. The Table.1 shows the Unicode for the Tamil characters for the word திருக்குறள். The replacement of a Unicode character by an ASCII character is shown in the Table.2.

iv. Now the word திருக்குறள் will be replaces as öÿÜÑ£á after substitution method. The actual size of திருக்குறள் is 11 bytes in a text file with Unicode encoding type. After compression öÿÜÑ£á is 6 bytes when it is stored as ANSI encoding type text file.

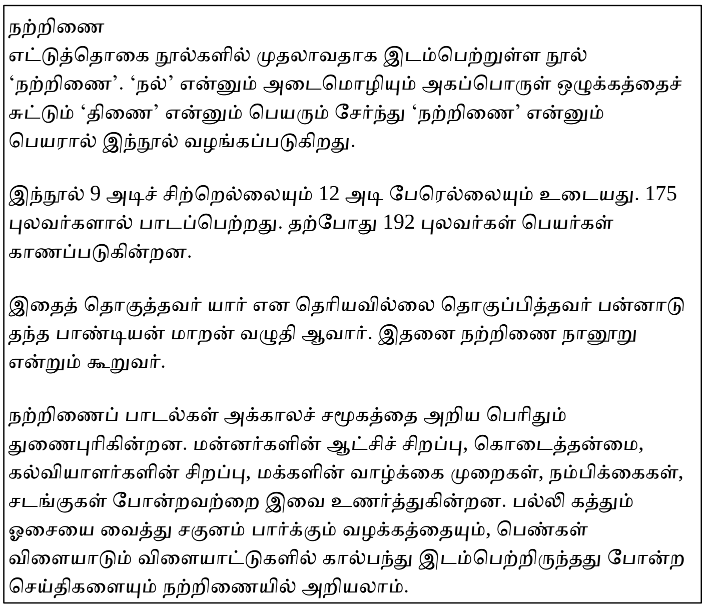
Table.2. Replacement of Unicode and ASCII character of the word திருக்குறள்.

Unicode (16 bit)	Combination of Unicode characters	Tamil Character in text file	ASCII Character	ASCII Code (8 Bits)
0BA4	த	தி	Ö	148
0BC0	◌ீ			
0BB0	ர	ரு	Ÿ	152
0BC1	◌ு			
0B95	க	க்	Ü	154
0BCD	◌்			
0B95	க	கு	Ñ	165
0BC1	◌ு			
0BB1	ற	ற	£	156
0BB3	ள	ள்	Á	160
0BCD	◌்			

The above is the example that shows how the size of a Tamil Unicode character reduced from 11 bytes to 6 bytes after compression. The actual compression process was carried out to the file named natrinai.txt is of size 3082 bytes is given as input to the application shown in Fig.4. This is a file that contains Unicode Tamil characters and stores as a Unicode encoding file type.

After compression the output is given as bharathiar_cprs.txt file which is of size 1540 bytes, shown in Fig.5. This compressed file contains ASCII characters and stored as ANSI encoding file type automatically by the application. The reverse process will do the decompression effectively.

The percentage of compression is calculated by comparing the size of file before and after compression [20]. The Table.3 shows the percentage of compression of Tamil files with the corresponding size in bytes before and after compression. All the files compression percentage is almost 50%. The decompression process is also successful.



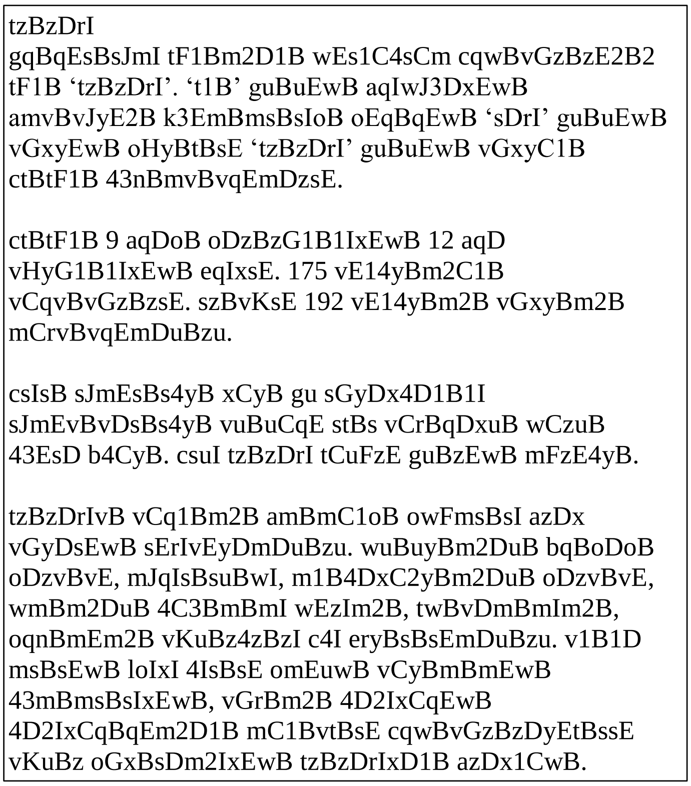
நற்றிணை
எட்டுத்தொகை நூல்களில் முதலாவதாக இடம்பெற்றுள்ள நூல் ‘நற்றிணை’. ‘நல்’ என்னும் அடைமொழியும் அகப்பொருள் ஒழுக்கத்தைச் சுட்டும் ‘திணை’ என்னும் பெயரும் சேர்ந்து ‘நற்றிணை’ என்னும் பெயரால் இந்நூல் வழங்கப்படுகிறது.

இந்நூல் 9 அடிச் சிற்றெல்லையும் 12 அடி பேரெல்லையும் உடையது. 175 புலவர்களால் பாடப்பெற்றது. தற்போது 192 புலவர்கள் பெயர்கள் காணப்படுகின்றன.

இதைத் தொகுத்தவர் யார் என தெரியவில்லை தொகுப்பித்தவர் பன்னாடு தந்த பாண்டியன் மாறன் வழுதி ஆவார். இதனை நற்றிணை நானூறு என்றும் கூறுவர்.

நற்றிணைப் பாடல்கள் அக்காலச் சமூகத்தை அறிய பெரிதும் துணைபுரிகின்றன. மன்னர்களின் ஆட்சிச் சிறப்பு, கொடைத்தன்மை, கல்வியாளர்களின் சிறப்பு, மக்களின் வாழ்க்கை முறைகள், நம்பிக்கைகள், சடங்குகள் போன்றவற்றை இவை உணர்த்துகின்றன. பல்லி கத்தும் ஓசையை வைத்து சகுனம் பார்க்கும் வழக்கத்தையும், பெண்கள் விளையாடும் விளையாட்டுகளில் கால்பந்து இடம்பெற்றிருந்தது போன்ற செய்திகளையும் நற்றிணையில் அறியலாம்.

Fig.4. File natrinai.txt before compression



tzBzDrI
gqBqEsBsJmI tF1Bm2D1B wEs1C4sCm cqwBvGzBzE2B2 tF1B ‘tzBzDrI’. ‘t1B’ guBuEwB aqIwJ3DxEwB amvBvJyE2B k3EmBmsBsIoB oEqBqEwB ‘sDrI’ guBuEwB vGxyEwB oHyBtBsE ‘tzBzDrI’ guBuEwB vGxyC1B ctBtF1B 43nBmvBvqEmDzsE.

ctBtF1B 9 aqDoB oDzBzG1B1IxEwB 12 aqD vHyG1B1IxEwB eqIxsE. 175 vE14yBm2C1B vCqvBvGzBzsE. szBvKsE 192 vE14yBm2B vGxyBm2B mCrvBvqEmDuBzu.

csIsB sJmEsBs4yB xCyB gu sGyDx4D1B1I sJmEvBvDsBs4yB vuBuCqE stBs vCrBqDxuB wCzuB 43EsD b4CyB. csuI tzBzDrI tCuFzE guBzEwB mFzE4yB.

tzBzDrIvB vCq1Bm2B amBmC1oB owFmsBsI azDx vGyDsEwB sErIvEyDmDuBzu. wuBuyBm2DuB bqBoDoB oDzvBvE, mJqIsBsuBwI, m1B4DxC2yBm2DuB oDzvBvE, wmBm2DuB 4C3BmBmI wEzIm2B, twBvDmBmIm2B, oqnBmEm2B vKuBz4zBzI c4I eryBsBsEmDuBzu. v1B1D msBsEwB loIxI 4IsBsE omEuwB vCyBmBmEwB 43mBmsBsIxEwB, vGrBm2B 4D2IxCqEwB 4D2IxCqBqEm2D1B mC1BvtBsE cqwBvGzBzDyEtBssE vKuBz oGxBsDm2IxEwB tzBzDrIxD1B azDx1CwB.

Fig.5. File natrinai_cprs.txt after compression

Table.3. File size before and after compression

File Name	Original File Size (bytes)	Compressed File Size (bytes)
Bharathi.txt	14762	7406
Chennaithagaval.txt	4266	2150
Ettuthokkai.txt	5634	2832
Natrrinai.txt	3082	1540

Pathittrupathu.txt	13074	6536
Vairamuthuvaralaru.txt	26222	13226
Kadavulvazthu.txt	3098	1548
Bala_kandam.txt	664380	334191
Arranya_kandam.txt	112890	56872
Ayodhya_kandam.txt	544088	275295
Kitkindha_kandam.txt	497856	252033
Sundara_kandam.txt	898288	453401
Yutha_kandam.txt	624112	313973

This compression technique almost reduces nearly 50% from the original files. The compressed file can be retained to original file by decompressing it.

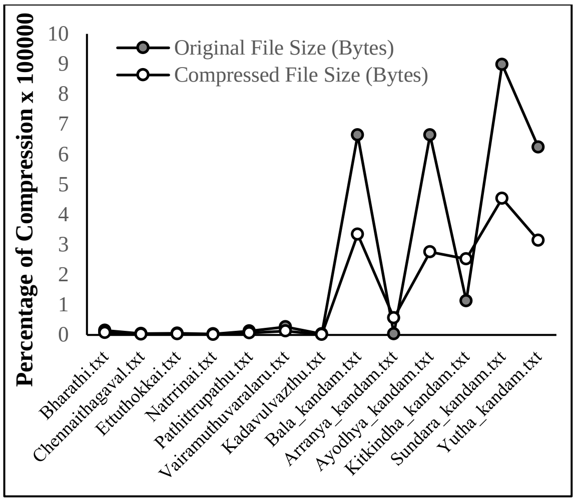


Fig.6. File size before and after compression with percentage of compression

The Fig.6 shows the variation of storage capacity of above mentioned files in Table.3. The deviation clearly shows that the compressed file saves the storage capacity than the original files.

Table.4. Average percentage of file compression.

File Name	Percentage of Compression
Bharathi.txt	49.83%
Chennaithagaval.txt	49.6%
Ettuthokkai.txt	49.73%
Natrrinai.txt	50.03%
Pathittrupathu.txt	50.01%
Vairamuthuvaralaru.txt	49.56%
Kadavulvazthu.txt	50.03%
Bala_kandam.txt	49.70%
Arranya_kandam.txt	49.62%
Ayodhya_kandam.txt	49.4%
Kitkindha_kandam.txt	49.38%
Sundara_kandam.txt	49.69%
Yutha_kandam.txt	49.53%
Average Percentage	49.7%

The Table.4 shows the list of compression percentage of files in Table.3 and the average percentage of file compression. The average percentage is 49.7%; this is due to the replacement of ASCII in the place of Unicode characters.

6. CONCLUSION AND FUTURE WORK

Tamil is a Dravidian language spoken by millions of people in India and all over the world. It is the first Indian state language of Tamil Nadu. There is a high need of storing the Tamil documents in digital form. Many applications are developed for both computers and mobile phones. New technologies are needed to preserve literature, artistic and scientific work of mankind digitally. This lossless compression technique surly paves a way to store the Tamil documents in minimum storage. Almost the compressed document will be reduced to 50%. This technique can be applied to other abugida languages too.

The compression technique works perfectly if the original document contains only Tamil characters. This is due to while performing decompression there may be a chance to substitute Unicode Tamil character wrongly to an ASCII character. The perfection can be further enhanced by placing a special character before the ASCII character in the original document before compression. Further the compression can be enriched by finding the frequency of occurrence of every Tamil character in Tamil documents, so that it can be applied effectively to the proposed system.

REFERENCES

[1] Ajantha Devi and S.Santhosh Baboo, “Embedded Optical Character Recognition on Tamil Text Image using Raspberry Pi”, *International Journal of Computer Science Trends and Technology*, Vol. 2, No. 4, pp. 11-15, 2014.

[2] https://en.wikipedia.org/wiki/ASP.NET, Accessed on 2017.

[3] Arafat Awajan and Enas Abu Jrai, “Hybrid Techniques for Arabic Text Compression”, *Global Journal of Computer Science and Technology*, Vol. 15, No. 1, pp. 23-27, 2015.

[4] Linkon Barua et al., “Bangla Text Compression based on Modified Lempel-Ziv-Welch Algorithm”, *Proceedings of IEEE International Conference on Electrical, Computer and Communication Engineering*, pp. 113-118, 2017.

[5] Guy E. Blelloch, “Introduction to Data Compression”, PhD Dissertation, Computer Science Department, CarNegie Mellon University, 2001.

[6] Eibe Frank, Chang Chui and Ian H. Witten, “Text Categorization using Compression Models”, Available at: https://www.cs.waikato.ac.nz/~eibe/pubs/Frank_categorization.full.pdf.

[7] S. Hewavitharana and H.C. Fernando, “A Two Stage Classification Approach to Tamil Handwriting Recognition”, *Proceedings of Tamil Internet*, pp. 118-124, 2002.

[8] Adam Gleave and Christian Steinruecken, “Making Compression Algorithms for Unicode Text”, *Proceedings of Data Compression Conference*, pp. 22-25, 2017.

[9] Goetz Graefe and Leonard D. Shapiro, “Data Compression and Database Performance”, *Proceedings of Symposium on Applied Computing*, pp. 11-15, 1991.

[10] Svend Juul and Morten Frydenberg. "UNICODE2ASCII: Stata modules to translate between Unicode and ASCII", Available at: https://ideas.repec.org/c/boc/bocode/s458080.html, Accessed on 2016.

[11] Harsimran Kaur and Balkrishan Jindal, "Lossless Text Data Compression using Modified Huffman Coding-A Review", *Proceedings of International Conference on Technologies for Sustainability-Engineering, Information Technology, Management and the Environment*, pp. 1017-1025, 2015.

[12] S.R. Kodituwakku and U.S. Amarasinghe, "Comparison of Lossless Data Compression Algorithms for Text Data", *Indian Journal of Computer Science and Engineering*, Vol. 1, No. 4, pp. 416-425, 2010.

[13] Anish Kumar, Anish, Sk Sakir Ali and Debashis Chakraborty, "Text Database Compression using Replacement and Bit Reduction", *Proceedings of International Conference on Computer Science and Information Technology*, pp. 409-416, 2012.

[14] Shihjong Kuo, "Processors, Methods, Systems, and Instructions to Transcode Variable Length Code Points of Unicode Characters", U.S. Patent, 2017.

[15] Mahesh Dattatray Kulkarni et al., "System and Method for Compression and Decompression of Text Data", U.S. Patent, 2017.

[16] Lishamol Philip and K.M. Abubeker, "LiBek II: A Novel Compression Architecture using Adaptive Dictionary", *Proceedings of International Conference on IEEE Emerging Technological Trends*, pp. 212-218, 2016.

[17] Radu Radescu, "Transform Methods used in Lossless Compression of Text Files", *Romanian Journal of Information Science and Technology*, Vol. 12, No. 1, pp. 101-115, 2009.

[18] J. Nelson Raja, P. Jaganathan and S. Domnic. "A New Variable-Length Integer Code for Integer Representation and its Application to Text Compression", *Indian Journal of Science and Technology*, Vol. 8, No. 24, pp. 11-14, 2015.

[19] R. Ramachandran and Ashik Ali, "Social Challenges faced by Technology in Developing Countries: Focus on Tamil Nadu State", Available at: http://shura.shu.ac.uk/15773/1/Ashik%20Ali%20CISDIDC2017.pdf

[20] S. Divakaran, C.L. Biji, C.Anjali and Achuth Sankar S. Nair, "Malayalam Text Compression", *International Journal of Information Systems and Engineering*, Vol. 1, No. 1, pp. 7-11, 2013.

[21] David Salomon, "*Data Compression: The Complete Reference*", 4th Edition, Springer, 2007.

[22] Sandip V Maniya and M.J. Sheth, "Compression Technique based on Dictionary Approach for Gujarati Text", *International Journal of Engineering Research and Development*, Vol. 4, No. 8, pp. 101-108, 2012.

[23] R. Seethalakshmi et.al., "Optical Character Recognition for Printed Tamil Text using Unicode", *Journal of Zhejiang University Science*, Vol. 6, No. 11, pp. 1297-1305, 2005.

[24] J. Sevcik and J. Dvorsky, "Techniques of Czech Language Lossless Text Compression", *Proceedings of International Conference on Computer Information Systems and Industrial Management*, pp. 813-816, 2016.

[25] Siva Jyothi Chandra, Ashlesha Pandhare and Mamatha Vani, "Multilingual Font Creation by Mapping Unicode to ASCII", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 5, No. 9, pp. 12-18, 2015.

[26] James A. Storer, "*Image and Text Compression*", Springer, 2012.

[27] J. Venkatesh and C. Suresh Kumar, "Tamil Handwritten Character Recognition using Kohonon's Self Organizing Map", *International Journal of Computer Science and Network Security*, Vol. 9, No. 12, pp. 156-161, 2009.

[28] www.tamilvu.org/doc_file/it_e_5_2013.pdf, Accessed on 2017.

[29] www.unicode.org/charts/PDF/U0B80.pdf, Accessed on 2017.