Building and Analysis of Tamil Lyric Corpus with Semantic Representation

Karthika Ranganathan karthika.cyr@gmail.com
Geetha T V tv_g@hotmail.com
Department of Computer Science and Engineering, Anna University, Chennai, 600025, India

Abstract

In the new era of modern technology, the cloud has become the library for many things including entertainment, i.e, the availability of lyrics. In order to create awareness about the language and to increase the interest in Tamil film lyrics, a computerized electronic format of Tamil lyrics corpus is necessary for mining the lyric documents. In this paper, the Tamil lyric corpus was collected from various books and lyric websites. Here, we also address the challenges faced while building this corpus. A corpus was created with 15286 documents and stored all the lyric information obtained in the XML format. In this paper, we also explained the Universal Networking Language (UNL) semantic representation that helps to represent the document in a language and domain independent ways. We evaluated this corpus by performing simple statistical analysis for characters, words and a few rhetorical effect analysis. We also evaluated our semantic representation with the existing work and the results are very encouraging.

1 Introduction

Tamil (pronounced as Tamizh) is one of the ancient and classical Indian languages spoken in Tamil Nadu. At the beginning of the 21st century, about 66 million people were speaking Tamil language, one of the oldest languages surviving from 300 BCE (Annamalai and Steever, 1998). Some people in Sri Lanka, Malaysia, Singapore, Mauritius, Fiji Islands, Canada and South Africa are also currently speaking Tamil language [1]. According to the history of Tamil grammar and lexical changes, the period can be categorized into three, old Tamil (300 BCE-700 CE, Sangam period), Middle Tamil (700 – 1600) and modern from 1600 (Thomas, 1998). Very few literature exists from the old Tamil containing 2381 poems by 473 poets and about 102 of which remained anonymous (Shinu, 2003).

Scholars from the Sangam period dealt with emotional and material topics such as war, governance and trade. These literatures are difficult to understand and require literates to explain. In the late 19th century, Tamil literature has been simplified so that everyone can understand, which has created interests amongst Tamil people. Poets like Subramaniya Bharathi utilized the power of Tamil language to create awareness of freedom for both women and British India. In the last century, songs in Tamil films have also played a vital role in interacting with people on social and political issues.

In the modern form, Tamil cinema can be classified as a combination of drama and songs, where the latter is composed with music according to a different genre of the lyrics. In the recent past, the encroachment of foreign languages, Tamil film lyrics are mixed with other language words which was a challenging task to build the lyric corpus.

[1] http://salc.uchicago.edu/tamil-at-chicago

Mining lyric documents facilitates the users to get the lyric characteristics, such as emotion, genre and a lyricist and also search the lyrics on the web. The linguistic features are however inadequate for identifying lyric characteristics, especially for morphologically rich languages like Tamil. For extracting higher level features, we chose semantic representation based on Universal Networking Language (UNL), which defines the conceptual structure in the form of a semantic graph [2]. To the best of our knowledge, no work has been done previously to build a Tamil lyric corpus with semantic graph representation.

The rest of the paper is organized as follows. Section 2 describes the related work. In section 3, we discuss the linguistic issues faced while building the Tamil lyric corpus. Section 4 explains about building the Tamil lyric corpus and illustrates some statistics about it. The evaluation of the corpus is given in Section 5. Finally, Section 6 concludes the paper with future work.

2 Related Work

In this section, we discussed related work carried out for the creation of corpus of Tamil and other languages. Shikhar et al. (2012) created a raw corpus for Assamese language. They build the corpus with a total of 1.5 Million words from their main categories such as Media, learned material and literature. In another work, Sarkar et al. (2007) explained the procedure and issues of automatic corpus creation for Bangla language. Vandana and Dash (2018) developed the Newspaper text corpus for Hindi language. They considered 4 online websites of Newspapers with a fixed time span of 10 years (2007-2016) and built a million-word corpus of the Hindi newspaper texts.

The first corpus for Tamil language was created by Technological Development for Indian Languages (TDIL), the Department of Electronics, Govt. of India. This consists of more than 3 million words from various domains (Francis et al., 1995). Later on, various research organizations are working for the improvement and annotated of Central Institute of Indian Languages (CIIL) Tamil corpus (Rajendran, 2006). Rachakonda et al. (2011) developed an annotation corpus of discourse connectives and their arguments by using 2,00,000 sentences of Tamil encyclopedia articles. Tamil Emotional speech corpus was built by Vijesh (2013) and identified five emotions of Happy, Sad, Anger, Fear, and Neural using GMM classifier. Analysis of lyrics based on the usage of words, concepts co-occuring and rhymes has been done by (Ranganathan et al., 2011). The same author has developed the lyric visualization tool to visualize each lyric characteristic such as emotion, genre, rhyming features, pleasantness and rhetorical style based on statistical modelling (Ranganathan et al., 2013). Chinnappa and Dhandapani (2021) built a new Tamil lyric corpus with a dataset of 5449 songs and identified the lyricist of the song using the BERT model. In our corpus, the number of lyrics is high and also the semantic representation is available for the document. This semantic information is useful to translate the Tamil lyric into other natural languages and also to identify the semantic textual similarity (Singh and Bhattacharyya, 2012), relation extraction (Nguyen and Ishizuka, 2006), search engine (Saviya shree et al., 2013) and sentiment analysis (Rani, 2014).

A language independent semantic representation of UNL has been used to convey the word and content based information of the document (Uchida et al., 1999). Parteek Kumar (2012) proposed a UNL based MT system for Punjabi language by developing Punjabi EnConverter, Punjabi DeConverter and a web interface for online EnConversion and DeConversion process. The author addressed the limitations of the Punjabi Machine Translation. To the best of our knowledge, there is no semantic graph representation for Tamil lyric documents. For Tamil language, Sridhar et al. (2016), have proposed the English-Tamil Machine translation system using UNL. They also developed the sentence formation algorithm to rearrange the Tamil words

[2] http://www.undl.org/unlsys/unl/unl2005

into correct sentences. In another work, UNL semantic representation has been developed by (Jagan et al., 2011) for the tourism domain using rule based approach. This work is not suitable for lyric semantic representation, since the relations and the contextual information in the lyric documents are different. Although, several corpora have been developed for Tamil language and non-Tamil languages for different domains, there is not enough corpus size available to perform applications, such as automatic generation of lyrics, lyric similarity, emotion, genre identification or other Natural Language Processing (NLP) related tasks.

3 Linguistic aspects of Traditional and Lyric Documents

From the viewpoint of linguistic criteria, lyric documents have more interesting issues when compared to traditional documents. The processing of lyrics varies from normal text processing. Usually, the document conveys information about a particular topic, whilst lyrics meant to convey emotion and feelings. In the traditional documents, the sentences sometimes follow the standard subject–verb–object (SVO) structure where the subject comes first, the verb second, and the object third, whereas in the lyric documents, either the verb will be followed by a consecutive noun or the verb may not present in the sentence at all. In addition, some lyrics may contain compound words, colloquial words and in some cases numeral words.

Also, the lyric document contains specific properties such as rhyme, meter, freshness and pleasantness to convey emotions and amenable to music. Special attention was given to defining and extracting these features. Rhyme scheme is used as strings for letters, where each line corresponds to the repetition of same syllables. In Tamil lyrics, the internal and external rhyme schemes are presented with three variations, alliteration, rhyme and end-rhyme, known in Tamil literature as (mōṉai) (first two letters are identical for two words in a rhyme), (etukai) (second two letters are identical for two words in a rhyme) and (iyaipu) (last two letters of the two words in a rhyme are identical) (Rajam, 1992). The freshness of the word based on various timelines facilitated to identify the characteristics of the lyrics. Generation of rhythmic pattern by grouping the strong and weak beats together is measured by using the meter score. Moreover, semantic relations between words of the lyric may be similar to those of documents, however, additional context and semantic attributes need to be extracted to convey the lyric characteristics. The need of tackling the above unique features of lyrics require construction of a large corpus and careful categorization is important for computerization.

4 Building Lyric Corpus for Tamil Language

In the recent days, most of the researchers focus on the automatic extracting and processing of lyrics (Rafael et al., 2014). Tamil films consist of several thousand lyrics; however, a huge amount of these lyrics is not yet computerized. Hence, it is very important to build a Tamil film lyric corpus. In this paper, a Tamil lyric corpus has been built by tagging the un-annotated corpus and used to determine the lyric characteristics.

Figure 1 shows the diagram for building a Tamil lyric corpus. For building the lyric corpus, we follow four steps, i) Data collection, ii) Data refinement, iii) Data Tagging and iv) Data validation.

4.1 Data collection

Lyric data are collected from books and websites. A large number of lyrics are downloaded from the Tamil website, such as Thenkinnam [3]. Crawler was used to retrieve the lyrics from the websites. Most of the lyric websites contain the information about the lyrics such as movie name, year and singers and in some cases the lyricist name, composer name. In addition, we have also

[3] http://thenkinnam.blogspot.in/

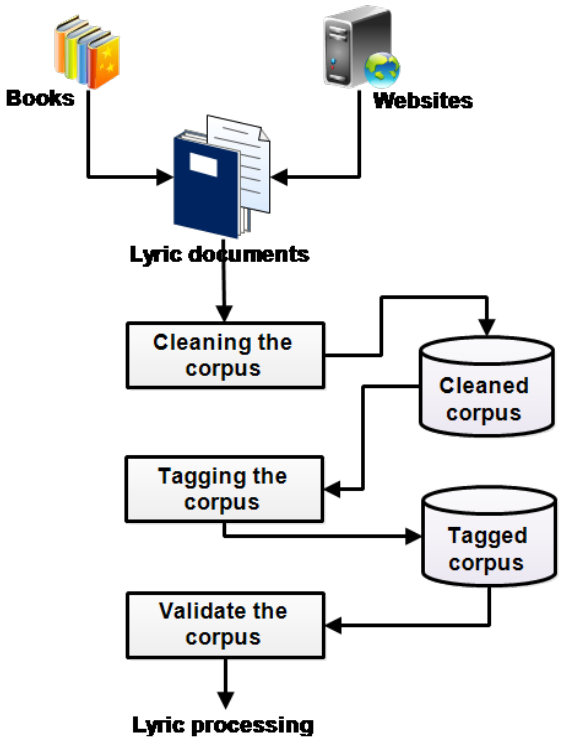

Figure 1: Building a Tamil lyric corpus

collected the Transliterated English documents of Tamil film lyrics. By applying heuristic rules, these documents have been converted into Tamil lyric documents. Only the legendary authors printed books were available to extract the data for Tamil film lyrics. For our corpus creation, we used the books from different lyricists Kannadasan, Vaali and Vairamuthu.

The collections of data were saved as a file with an index number and the title of the song. This helped in avoiding duplication of the songs whilst collecting the lyrics. However, at some instances, there are two different songs with a slight variation in the words of the lyrics. For example, in the Tamil film "7g rainbow colony", a song "நினைத்து நினைத்து பார்த்தேன் – niṉaittu niṉaittu pārttēṉ (I remembered and remembered)" appeared twice with the same title, written by the same author and the music were composed by the same musician. For these types, the entire lyrics of the songs were analysed and the additional information which made it distinct was gathered and stored in the corpus with the emotional tag, happy or sad for identifying the sentiment polarity of the words. This helped in avoiding the elimination of songs with only slight variation in lyric content.

4.2 Data refinement

After collecting freeform data, a substantial cleaning process is required to resolve the linguistic problems such as spelling variations, spelling errors, dialectual variations, foreign langauges encroachment and joined words/lines problems by applying heuristic rules. In addition, single phase lyrics and undefined Universal Character Set Transformation Format (UTF)) characters were filtered out from the data collection.

4.3 Data tagging

The cleaned lyric data are stored in the database. In our work, we used a XML format to store the information obtained from the collected data. This format is mainly used to retrieve the data and also to add or remove the data easily. Normally, in Indian film lyrics including Tamil lyrics, the structure of a lyric is composed of three parts. The first part of the song is called "Pallavi", which represents the theme of the song. "Anupallavi" is the second part, which comes after the "pallavi" and it is optional in most of the cases in lyrics. The final part of the lyric is the "charanam", which has been used to convey the detailed information of the song. Note that in few lyrics more than one charanam is also present. Hence, for each lyric, along with the information obtained, the pallavi, anupallavi and charanam parts are also tagged and represented in the XML format, an exemplar has been shown in Table 1 for a Tamil lyric example:

4.4 Data validation

Validation has been carried out manually by 15 Tamil linguistic experts in the Tamil Computing Lab (TaCoLa lab - Anna University, India) and the statistics of the data cleaned is shown in Table 2 and lyric corpus set is shown in Table 3.

This corpus has been used in many of lyric processing techniques. The main focus is on Tamil film lyric mining since the existence of numerous Tamil songs and the availability of Tamil lyrics available on the web makes this mining an interesting issue. Understanding lyrics and identification of lyric characteristics, from the lyric data set are challenging issues.

4.5 Semantic representation - UNL

The semantic processing of any natural language is represented using Universal Networking Language (UNL), which helps to construct the semantic graph for each sentence as a hypergraph, in which the nodes represent the universal words (concept) and the link represent the relations (exist between concepts). UNL consists of three components, namely, Universal Words (UWs) – representing the meaning of a word or a sentence, UNL relations – representing the relationship between two different concepts in a sentence, and UNL attributes – representing the mood, tense, aspect, etc. (Uchida et al., 1999).

A UW is made up of a character string as head word (an English-language word) followed by a list of constraints. The headword is an English word that is interpreted as a label for a set of concepts. However, the constraints list restricts the interpretation of a UW to a specific concept included within the basics of UW and a set of 58 semantic relations that connects two different UWs within a sentence. An exemplar of a lyric sentence and its semantic representation is shown in Figure 2.

Example: நான் மலரோடு தனியாக ஏன் இங்கு நின்றேன் - Nāṉ malarōṭu taṉiyāka ēṉ iṅku niṉṟēṉ (Why am I standing here alone with the flower)

Figure 2 shows the UNL semantic constraints, UNL relations and UNL attributes used for the example lyric line. Here, icl>person, agt>thing, pof>plant, aoj>thing and icl>place indicate the semantic constraints, which helps to identify whether the concept is a person, or place, or object, etc. @verb represents the UNL attributes and man, obj, agt and plc represents semantic relations between the two concepts.

5 Evaluation

5.1 Analysis of corpus

We present a detailed analysis of our Tamil lyric corpus. This analysis has been used in the lyric search engine and lyric processing tasks, namely, emotion, genre, lyricist, and similarity of lyrics. For lyric analysis, we used frequency occurrence of character and word usage. Also, the usage of rhetorical effects such as metaphor and simile in the corpus has been identified. In

No	Description	Example
1	Title of the lyric	<தலைப்பு> என்னுயிரே என்னுயிரே </தலைப்பு> <Title> oh my heart </Title> <Talaippu> eṉṉuyirē eṉṉuyirē </talaippu>
2	Name of the movie	<படம்> உயிரே</படம்> <Paṭam> uyirē </paṭam> <Movie name> Soul </Movie name>
3	Composer name for the lyric	<இசை>A.R. ரஹ்மான் </இசை> <Icai>A.R. Rahmāṉ </icai> <Composer> A.R.Rahman </Composer>
4	Lyricist name for the lyric	<வரிகள்> வைரமுத்து </வரிகள்> <Varikaḷ> vairamuttu </varikaḷ> <Lyricist> Vairamuthu </Lyricist>
5	Singer name for the lyric	<பாடியவர்> - </பாடியவர்> <Pāṭiyavar> - </pāṭiyavar> <Snger> - </Singer>
6	Tune name of the lyric	<ராகம்> - </ராகம்> <Rākam> - </rākam> <Tune> - </Tune>
7	Rhythm name of the lyric	<தாளம்> - </தாளம்> <Tāḷam> - </tāḷam> <Rhythm> - </Rhythm>
8	Year of the lyric	<வருடம்>1998</வருடம்> <Varuṭam>1998</varuṭam> <Year> 1998 </Year>
9	Opening or first unit of the lyric	<பல்லவி> என்னுயிரே என்னுயிரே etc. </பல்லவி> <Pallavi> eṉṉuyirē eṉṉuyirē etc </Pallavi> <First unit> Oh my heart ! etc. </First unit>
10	Second unit of the lyric	<அனுபல்லவி> நம் காதலிலே வரும் etc.</அனுபல்லவி> <Aṉupallavi> nam kātalilē varum etc.</Aṉupallavi> <Second unit> We enlight the world etc. </Second unit>
11	Third unit of the lyric	<சரணம்-1> கைகள் நான்கும் etc. </சரணம் - 1> <Caraṇam-1> kaikaḷ nāṉkum etc. </Caraṇam - 1> <Third unit – 1> before intimate etc. </Third unit -1> <சரணம்-2> என்னுயிரே என்னுயிரே etc. </சரணம் - 2> <Caraṇam-2> eṉṉuyirē eṉṉuyirē etc. </Caraṇam - 2> <Third unit – 2> Oh My heart! etc. </Third unit -2>

Table 1: Lyric Tagging

Lyric Corpus	Statistics
Total songs before clean-up	15394
Duplicate (deleted)	64
Single – phrase lyric (deleted)	10
Undefined UTF characters (deleted)	34
Total songs after clean-up	15286

Table 2: Data Clean up Statistics

Lyric Corpus	Statistics
Documents	15286
Lines	212004
Words	917185

Table 3: Lyric corpus set statistics

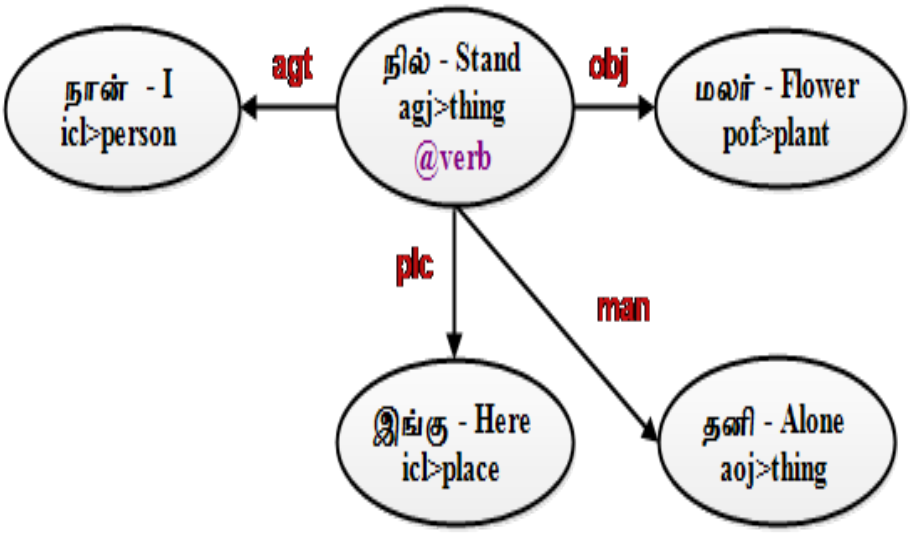

Figure 2: Semantic representation of Tamil lyric sentence

the figurative language, metaphor and simile reveal the feelings of the people strongly. These effects conveyed the meaning of the lyric with a minimal set of words.

5.1.1 Statistical analysis of corpus

Table 4 shows the Top 5 list of frequency distribution of character for Tamil corpus. In Tamil lyric, the least frequently occurred letter was vowel consonant னை and the most frequently occurred letter was a consonant க்.

Table 5 shows the Top 5 list of frequency words from the corpus without function words. Here, we are considering the concept of each word as additional count. For example, the word மலர் – flower has concepts புஷ்பம், மாமலர், பூ and அலரி.

In addition, a large number of lyric documents usually contain many compound and colloquial words (Lestari, 2019). By using (Umamaheswari et al., 2011), we have identified 2528 compound words and 317 colloquial words from the lyric corpus.

Character	Percentage (%)
க் - k	3.28
ப் - p	2.95
த் - t	2.56
ல் - l	2.43
ட் - ṭ	2.16

Table 4: Top 5 frequency characters

Words	Percentage (%)
காதல் - kātal (Love)	2.84
கண் - kaṇ (Eye)	2.19
பூ - pū (Flower)	1.93
நிலவு - Nilavu (Moon)	1.84
மனம் - Maṉam (Mind)	1.68

Table 5: Top 5 frequency words

No	Metaphor	Simile
	மலர்ப்பாதம்	மீன் போன்ற கண்கள்
1	Malarppātam	mīṉ pōṉṟa kaṇkaḷ
	Flower foot	Eyes like fish
	ரோஜாப்பூ கன்னம்	குயில் போல பாடு
2	rōjāppū kaṉṉam	kuyil pōla pāṭu
	Rosy cheeks	Sing like Quail
	மான்விழி	வில் போன்ற புருவம்
3	māṉviḻi	vil pōṉṟa puruvam
	Deer eye	Bow like eyebrow
	பூவிதழ்	நிலவை போன்ற முகம்
4	pūviṯaḻ	nilavai pōṉṟa mukam
	Flower lips	Moon like face
	அன்னநடை	பூ போன்ற முகம்
5	aṉṉanaṭai	pū pōṉṟa mukam
	Swan walk	Flower like face

Table 6: Top 5 list of Metaphor and Simile

5.1.2 Rhetorical effect analysis of corpus

Table 6 shows the Top 5 list of metaphor and simile used in the lyric corpus. Here, the metaphors present in the lyric documents are generally in the form of noun-noun and verb-noun compounds. Most lyricists used cue words such as போன்ற - pōṉṟa (like) and போல - pōla (like) for comparisons.

5.2 Comparison of the lyric semantic graph with existing work

From Table 7, our approach results in a high F-measure compared to the rule-based approach (Jagan et al., 2011).

In some cases, lyrics do not always have a verb and therefore the existing work does not form complete semantic graphs. Here, the relations and the contextual information of the documents are different.

Approach	F-measure
Our system	0.61
Existing system	0.52

Table 7: Comparison of our system with existing system

6 Conclusion and Future Work

In this paper, we described the Tamil lyric corpus and the linguistic issues faced during the process. This corpus has been stored in the XML format to add or remove the data easily. We believe that this is the first attempt in creating the Tamil lyric corpus. This paper also discussed the semantic representation for extracting the context level information from the documents. We carried out the evaluation by performing statistical and rhetorical analysis of lyric corpus which resulted in promising results.

In future, we have planned to increase the corpus size of the lyric documents by crawling in the web and annotating various lyric characteristics of emotion, genre, lyricist inferred based on the semantic features. We have also planned to build the semantic representation using a semi-supervised approach. We also planned to extend our work to other morphologically rich languages.

References

Annamalai and Steever, S. B. (1998). *Modern Tamil:The Dravidian Languages*. Taylor and Francis.

Chinnappa, D. and Dhandapani, P. (2021). Tamil Lyrics Corpus: Analysis and Experiments. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 1–9, Kyiv.

Francis, E., D, J. B., and Ganesan (1995). *Final Report Development of Corpora of Texts of Indian Languages in Machine Readable Form, Part II (Tamil, Telugu, Kannada, Malayalam)*. CIIL.

Jagan, B., Geetha, T. V., Parthasarathi, R., and Karky, M. (2011). Morpho-Semantic Features for Rule- based Tamil Enconversion. *International Journal of Computer Applications (IJCA)*, 26(6):11–18.

Kumar, P. (2012). *UNL-based machine translation system for Punjabi language*. Ph.D thesis, Thapar Institute of Engineering and Technology.

Lestari, F. D. (2019). *An Analysis of Compound Word in the Selected Song Album of Taylor Swift*. Ph.D thesis, STKIP PGRI SIDOARJO.

Nguyen, P. and Ishizuka, M. (2006). A statistical approach for Universal Networking Language-based relation extraction. In *Proceedings of the International Conference on Research, Innovation and Vision for the Future*, pages 153–160, Ho Chi Minh City, Vietnam.

Rachakonda, Teja, R., and Sharma, D. M. (2011). Creating an annotated tamil corpus as a discourse resource. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 119–123, Portland, Oregon, USA.

Rafael, R., Almeida, M. A., and Carlos, N. S. J. (2014). The ethnic lyrics fetcher tool. *EURASIP Journal on Audio, Speech, and Music Processing*, 27(1):1–10.

Rajam, V. (1992). *A Reference Grammar of Classical Tamil Poetry*. American philosophical society.

Rajendran, S. (2006). *A Survey of the state of the art in Tamil language technology*. Language in India.

Ranganathan, K., Barani, B., and Geetha, T. V. (2013). A Tamil lyrics search and visualization system. *Lecture Notes in Computer Science*, 8281(1):513–527.

Ranganathan, K., Geetha, T. V., Parthasarathi, R., and Karky, M. (2011). Lyric mining Word, rhyme and concept co-occurrence analysis. In *Proceedings of the Tamil Internet Conference*, pages 276–281, Philadelphia, USA.

Rani, S. (2014). *Rule Based Sentiment Analysis System*. Ph.D thesis, THAPAR UNIVERSITY.

Sarkar, A. I., Shahriar, D., Pavel, H., and Khan, M. (2007). *Automatic Bangla corpus creation*. PAN Localization Working Papers.

Saviya shree, K. V., Umamaheswari, E., Jagan, B., Geetha, T. V., and Parthasarathi, R. (2013). Concept Based Search Engine (CBSE) System for Tamil and English. In *Proceedings of the International Tamil Internet Conference*, pages 105–111, University Of Malaya, Kuala Lumpur, Malaysia.

Shikhar, S., Bharali, H., Deka, A. G. R., and Barman, A. (2012). A structured approach for building assamese corpus: insights, applications and challenges. In *Proceedings of the 10th Workshop on Asian Language Resources*, pages 21–28, Mumbai, India.

Shinu, A. (2003). Chera, chola, pandya: Using archaeological evidence to identify the tamil kingdoms of early historic south india. *Asian Perspectives*, 42:207 – 223.

Singh, J.and Bhattacharya, A. and Bhattacharyya, P. (2012). janardhan: Semantic textual similarity using universal networking language graph matching. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (SEM)*, pages 662–666, Montreal, Canada.

Sridhar, R., Sethuraman, and Krishnakumar (2016). English to Tamil machine translation system using universal networking language. *Sādhanā*, 41(1):607–620.

Thomas, L. (1998). *Old Tamil:The Dravidian Languages*. Taylor and Francis.

Uchida, H., Zhu, M., and Senta, T. (1999). *A gift for a millennium*. The United Nations University.

Umamaheswari, E., Ranganathan, K., T V Geetha, Parthasarathi, R., and Karky, M. (2011). Enhancement of Morphological analyzer with compound, numeral and colloquial word handler. In *Proceedings of the 9th International Conference on Natural Language Processing*, pages 177–186, Anna University, Chennai, India.

Vandana and Dash, N. S. (2018). *Creation and Compilation of Hindi Newspaper Text Corpus*. Language in India.

Vijesh, J. C. (2013). Building and evaluation of tamil emotional speech corpus. In *Proceedings of the 5th National Conference on Signal Processing Communication and VLSI Design*, pages 389–392, Coimbatore, Tamil Nadu, India.