A Discourse-Based Information Retrieval for Tamil Literary Texts

1Anita Ramalingam & 2Subalalitha Chinnaudayar
Navaneethakrishnan

1&2Department of Computer Science and Engineering
SRM Institute of Science and Technology, India

anitaramalingam17@gmail.com
subalalitha@gmail.com

Received: 30/10/2020 Revised: 11/1/2021 Accepted: 12/1/2021 Published: 11/6/2021

ABSTRACT

Tamil literature has many valuable thoughts that can help the human community to lead a successful and happy life. Tamil literary works are abundantly available and searchable on the Internet. However, the existing search systems follow a keyword-based match strategy that fails to satisfy user needs. This necessitates the demand for a focused information retrieval system that semantically analyzes the Tamil literary text, which will eventually improve the search system performance. This paper proposes a novel information retrieval framework that uses discourse processing techniques in aiding semantic analysis and representation of Tamil literary texts. The proposed framework was tested using two ancient literary works, the Thirukkural and Naladiyar, which were written in 300 Before the Common Era (BCE). The Thirukkural comprises 1,330 couplets, each 7 words long, while the Naladiyar consists of 400 quatrains,
each 15 words long. The proposed system, tested with all the 1,330 Thirukkural couplets and 400 Naladiyar quatrains, achieved a mean average precision (MAP) score of 89 percent. The performance of the proposed framework was compared with Google Tamil search and a keyword-based search, which was a substandard version of the proposed framework. Google Tamil search achieved a MAP score of 56 percent while keyword-based method achieved a MAP score of 62 percent. It showed that the discourse processing techniques could improve the search performance of an information retrieval system.

**Keywords:** Discourse parser, Morphological Analyzer, Inverted indexing, Ranking, Tamil information retrieval.

**INTRODUCTION**

Tamil language and literature have a long and glorious tradition, with the written form of the Tamil language dating back to 600 Before the Common Era (BCE). Classics of Tamil literature, such as the Tholkappiyam, date back to 300 BCE (Abraham, 2003). Tamil classical literature breathes ethical, moral, and philosophical values that are to be explored for the good of society. A 2017 Google survey determined that Tamil is more widely used than any other Indian language on the web. Tamil language has significantly transformed over these years in terms of script and style. More language tools exist to handle the current Tamil language; however, the literary types of text are yet to be handled. Though many literary types of text are available online, they are not completely accessed by the users due to the lack of language tools. This paper addresses such issues and has attempted to enhance the current language tools such as Morphological Analyzer. Most of the existing search systems use the keyword-based match strategy to retrieve all types of Tamil text. This paper puts forth a semantic analysis that can go well with both current and literary types of Tamil text in order to increase the performance.

Tamil literary works such as the Thirukkural and Naladiyar are didactic in nature, conveying information using words of a fixed length. Both Thirukkural and Naladiyar have three distinct sections on virtue, wealth, and love. The Thirukkural has 133 chapters and 10 two-line couplets in each, while the Naladiyar has 40 chapters and 10
four-line quatrains in each. In both couplets and quatrains, all lines have 4 cirs, except the last with 3. A cir can be a single word or a combination of more. It can also contain part of a word, while the rest are appended to the next to sustain style and prosody (Adigalasiriyar, 1985). These words are not handled by the Tamil Morphological Analyzer (Anandan et al., 2002). To the best of the authors’ knowledge, no Tamil Morphological Analyzer is able handle this type of literary text. This is because Tamil literary texts are distinctive, with no linguistic resources to process them computationally. The proposed approach attempts to deal with this problem by using a word reformation algorithm that identifies and separates words from cirs for further processing.

In the proposed work, the discourse structure for the Thirukkural and Naladiyar is built using the rhetorical structure theory introduced by Mann, Thompson, and Matthiessen at the University of Southern California (Mann & Thompson, 1988). The theory captures the coherence between text fragments using discourse relations and forms a discourse structure called a rhetorical structure. The building blocks of discourse structure are a nucleus, satellite, and discourse relations. The nucleus carries the requisite information in the text, the satellite transports all the additional information supporting the nucleus, with the discourse relation connecting the two.

The proposed work comprises two components: offline and online processing. Offline processing involves constructing a discourse parser that captures the semantic relations in the couplets and quatrains of the Thirukkural and Naladiyar, resulting in a discourse structure. Inverted indexing is then performed, based on the discourse structure. In online processing, a user query is processed to make it compatible with the index structure. It is then matched with the indices to retrieve the couplets and quatrains requested, with relevant explanations. The retrieved results are ranked according to their semantic relevance to the query, using discourse relations.

In the traditional keyword-based search, Thirukkural couplets containing query words are retrieved. Since no importance is ascribed to the semantic interpretations of the query, irrelevant results are obtained. For instance, the keyword-based search retrieves the Thirukkural couplet in Example 2 in the top-ranking position, as it
contains the keyword “நட்பு (Naṭpu – friendship)”, even though it is irrelevant to the query in Example 1. The proposed discourse-based search system also retrieves this *Thirukkural* couplet. However, it pushes the couplet to the 34th position in the ranking by giving importance to the discourse relations present in the *Thirukkural*. Thus, the incorporation of discourse relations is one of the most substantial reasons for retrieving relevant results and filtering irrelevant ones.

**Example 1:**

**User query:** நாட்பு சிதையாமல் இருக்க என்ன செய்ய வேண்டும்?

**English Transliteration:** Naṭpu citaiyāmal irukka eṇṇa ceyya vēṇṭum?

**Meaning in English:** What can be done to keep friendships intact?

**Example 2: Retrieved Thirukkural Couplet**

**Thirukkural Couplet:** குடம்பைதனித்துஒழியப் புள்பறந் தற்றே உடம்பொடு உயிரிடை நட்பு.

**Tamil Explanation:** உடலுக்கும் உயிருக்கும் உள்ள உறவும் முட்டைக்கும் பறவைக் குஞ்சுக்கும் உண்டான உறவுபோன்றதான்.

**English Explanation:** The love of the soul to the body is like (the love of) a bird to its egg which it flies away from and leaves empty.

The contribution of the proposed work is four-fold:

1. It identifies semantic discourse relations using hand-coded rules that are exclusively designed to process Tamil literary texts.
2. It implements a Word Reformation Algorithm to handle literary texts.

3. It proposes a discourse-based indexing technique to semantically retrieve relevant Tamil text and literature information from the web.

4. It proposes a discourse-based search and rank algorithm.

The rest of the paper is organized as follows. The next section describes the related works, followed by the proposed work. The last section presents the results and discussion, followed by a section that concludes the paper and offers directions for future work.

RELATED WORKS

This section surveys state-of-the-art studies from two perspectives. The first subsection describes work on information retrieval systems developed in three languages: English, Chinese, and Arabic. Meanwhile, the second subsection examines computational work carried out on Tamil language and literature.

Works on Information Retrieval Systems

Fauzi et al. (2017) proposed an information retrieval (IR) system to retrieve information from Arabic *fiqh* texts. In all, 13 Arabic *fiqh* e-books were used as a dataset. The pages of the 13 books were treated as documents, and an inverse book frequency for ranking was developed. They tested their work with precision, recall, and F-measure evaluation metrics and achieved 76 percent precision, 74 percent recall, and 75 percent F-measure. Zamani et al. (2018) proposed a standalone neural ranking model for document retrieval, using an inverted indexing technique to index the documents collected. Their work was tested using newswire and web collections. Precision, mean average precision (MAP), normalized discounted cumulative gain, and recall were used as evaluation metrics. Liu et al. (2018) proposed a parallel indexing method to index traditional Chinese medical reports. A multifactor ranking model was applied for ranking, and the medical reports were displayed using a template-based visualization method.
Meng et al. (2019) proposed a scheme for indexing and retrieving social media data. Experiments were undertaken on two image datasets and an e-commerce dataset. Tekli et al. (2019) proposed an approach for semantic indexing to retrieve information from structured, unstructured, and semi-structured data. An algorithm developed to facilitate searching was tested on the Internet Movie Database (IMDb) movie dataset. Their work was evaluated using precision, recall, F-score, and MAP metrics, and produced 0.3636 precision, 0.2085 recall, 0.2815 F-score, and 0.1393 MAP. Samia and Khaled (2020) introduced an Arabic plagiarism detection system. Semantic indexing was carried out using Arabic ontology and part-of-speech (POS) tagging. The AraPlagDet corpus was used to evaluate their work using precision, recall, and F-score metrics, and achieved 82.4 percent precision, 93.2 percent recall and 87.5 percent F-score. Agosti et al. (2020) proposed an unsupervised neural framework for information retrieval. Semantic indexing, synonymy, and polysemy were used to eliminate semantic gaps, indicating a mismatch between document terms and queries. The TREC CDS and OHSUMED collections were used to test their work, with precision and recall as the evaluation metrics.

The abovementioned works have targeted building an IR system for English, Chinese, and Arabic languages. These IR systems are capable of semantically interpreting the query, whereas Indian language-based IR systems are still at the keyword-based search level. This is due to the fact that language tools like ontology, part-of-speech tagging, and Morphological Analyzers are yet to be comprehensively developed for most Indian languages. Although the proposed semantic IR framework has been tested with Tamil literary text, the framework can also be used for expository Tamil text. Furthermore, the techniques adapted by the proposed IR framework can be extended to any language.

**Computational Works on Tamil**

Prasath et al. (2015) proposed a cross-language IR approach for a given user query in another language. A corpus-driven query suggestion approach for re-ranking was used on Tamil and English news collections of the FIRE corpora, with precision as the evaluation metric. Subalalitha and Anita (2016) introduced a page ranking algorithm based on discourse relations to retrieve web pages. The
rhetorical structure theory was applied to ascertain semantic relations between web pages, as well as hyperlinks in the web pages. In all, 500 Tamil and 50 English tourism web pages were tested, using precision as the evaluation metric. Giridharan et al. (2016) proposed a scheme to retrieve information from ancient Tamil texts inscribed in temples, and transform the epigraphy into current Tamil digital texts, along with their meaning. The Brahmi database was used as a dataset, and an accuracy of 84.57 percent was obtained.

Sankaralingam et al. (2017) put forward a methodology for information retrieval for Tamil texts, using ontology to convert ontological structures into visual representations that aid retrieval. Lexical and semantic relations such as homonymy, synonymy, antonymy, and meronymy were used on their 50,000-word general domain dataset. Thenmozhi and Aravindan (2018) proposed a cross-lingual IR system using ontology. Tamil queries were translated into English and the relevant documents were retrieved in English. Ambiguity was eliminated from Tamil and English queries with word-sense disambiguation and ontology, respectively. A Tamil-English bilingual dictionary, a Tamil Morphological Analyzer, and a named entity database were used to translate Tamil queries into English. Their methodology was evaluated for the agricultural domain, with precision as the evaluation metric. Anita and Subalalitha (2019a) developed an approach to cluster Thirukkural couplets using discourse connectives as features. The K-means clustering machine learning algorithm was used. Cluster purity, the Rand index, precision, recall, and F-score were employed as evaluation metrics to obtain 79 percent purity, 92 percent overall Rand index, 79 percent precision, 80 percent recall, and 79 percent F-score.

Subalalitha (2019) proposed an information extraction scheme for the Tamil literary work, Kurunthogai. Details pertaining to food, flora, fauna, vessels, waterbodies, noun unigrams, verb unigrams, adjective-noun bigrams, and adverb-verb bigrams were extracted. A Tamil Morphological Analyzer tool was used to extract N-grams. Precision was used as the evaluation metric and 88.8 percent was obtained. Saravanan (2020) introduced a cluster-based Tamil document retrieval system using semantic indexing. The K-means algorithm was applied on a dataset taken from the Tamil Language Consortium Repository. F-score, used as an evaluation metric, achieved 60 percent.
This subsection discusses the computational work on Tamil. It is observed that much of the existing research carried out have been restricted to the Tamil expository (essay-type) texts, with very few studies undertaken on Tamil literary texts. The IR system, implemented thus far for Tamil, has limited itself to Tamil expository texts. This research focuses on information retrieval on two classic Tamil literary texts, the *Thirukkural* and the *Naladiyar*. Moreover, while English has an SVO (Subject-Verb-Object) sentence structure, Tamil uses either the SVO or SOV (Subject-Object-Verb) structure. Tamil literature, on the other hand, follows neither pattern, which explains why existing Tamil search systems are still keyword-based. Tamil literary texts currently need non-existent advanced natural language processing (NLP) tools for text processing, making it more difficult to process them than standard essay-type ones. The proposed system tries to modify existing Tamil grammar tools to adapt them to the literary text for which the IR system is constructed. The word reformation algorithm described in the proposed work section is designed to ease the construction of the discourse structure. Furthermore, the search and ranking strategy is meticulously designed to tap the semantics incorporated into the discourse structure.
THE PROPOSED WORK

The architecture for the proposed work is shown in Figure 1.

Figure 1

*Architecture for the Proposed Work*

*Thirukkural* couplets and *Naladiyar* quatrains were given as input to the *Thirukkural-Naladiyar* discourse parser, with the entire process separated into offline and online processing. Discourse parsing and indexing were executed in the former, while query processing, ranking, and searching are undertaken in the latter. Cue words or signal words were used to identify the nucleus, satellite, and discourse relations (Anita & Subalalitha, 2019b). Occasionally, however, cue words need a morphological analysis in order to assist in discourse parsing. For example, the word “தின்றற்றே (Tiṉṟaṟṟē – As if eaten)” needed to be morphologically analyzed to separate the cue word, “அற்று (Aṟṟu...
Like)” from the rest in order to identify the nucleus, satellite, and discourse relations. The proposed work used the Tamil Morphological Analyzer proposed in Anandan et al. (2002). The Morphological Analyzer failed to identify some discourse cues as the Tamil literary text had a different style. Therefore, the word reformation algorithm was proposed in this paper to find the discourse cues. The proposed discourse parser focused on identifying the ten discourse relations that were predominantly observed in Thirukkural couplets and Naladiyar quatrains, namely Antithesis, Condition, Otherwise, Explanation, Contrast, Elaboration, Background, Joint, List, and Solutionhood. In this paper, the nucleus, discourse relation, and satellite were altogether referred to as Nucleus Relation Satellite (NRS) sequences.

All the nuclei were separated from the NRS sequences. It was observed that the length of the nucleus varied from three to eight words. The words were tokenized, and their root forms (lemmas/lemmata) were indexed with the discourse relations pointing to their corresponding Thirukkural couplets and Naladiyar quatrains.

On the other side, in online processing, the query was obtained from the user, and Morphological Analysis was performed after tokenizing the query. The Morphological Analyzer divided the words into morphemes and also gave the POS information of each morpheme. From this morphological information, discourse relations and grammatical components such as nouns, verbs, adjectives, and adverbs were identified. For instance, the Tamil Morphological Analyzer output for the word “செய்வதால் (Ceyvatāl – By doing)” is given in Example 3.

**Example 3:**

<table>
<thead>
<tr>
<th>Word: “செய்வதால் (Ceyvatāl – By doing)”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morphological Analyzer Output:</td>
</tr>
<tr>
<td>செய்வதால்:செய்&lt; Verb &gt;வ்&lt; Future Tense Marker &gt;அதால்&lt; Conditional Suffix&gt;</td>
</tr>
</tbody>
</table>

From the output shown in Example 3, the grammatical component verb and the discourse relation Condition were identified. Question
words presented in the query were also analyzed to identify discourse relations. In all, 60 question words were taken for analysis, and the sample question words are shown as in Table 3. The dictionary was used for retrieving the synonyms of the words indicating the grammatical components. These words and discourse relations identified from the query were matched with the semantic indices, which consisted of the nuclei words and discourse relations produced in offline processing. The matched Thirukkural couplets and Naladiyar quatrains were retrieved and ranked according to their query-index similarity.

**Discourse Parser Construction**

The Thirukkural discourse parser (Anita & Subalalitha, 2019b) was enhanced to explore more discourse relations that were present in both the Thirukkural and the Naladiyar. The enhanced component of the Thirukkural-Naladiyar discourse parser is described in this section.

**Cue Words**

Cue words, or signal words, are collected by manually analyzing about 600 Thirukkural couplets and 100 Naladiyar quatrains. These cue words offer clues for identifying discourse relations. In Example 4, the cue word, “ஆற்றின் (Āṟṟiṉ – If you do)”, indicated that the Thirukkural couplet contained a Condition discourse relation.

**Example 4:**

<table>
<thead>
<tr>
<th>Thirukkural Couplet:</th>
<th>English Transliteration:</th>
</tr>
</thead>
<tbody>
<tr>
<td>வீழ்நாள் படாஅமை நன்றாற்றின் அஃதொருவன் வாழ்நாள் வழியடைக்கும் கல்.</td>
<td>Vīḻnāḷ paṭā’amai naṉṟāṟṟiṉ aḥtoruvaṉ vāḻnāḷ vaḻiyaṭaikkum kal.</td>
</tr>
</tbody>
</table>

**Meaning in English**
If one allows no day to pass without some good being done, his conduct will be a stone to block up the passage to other births.
Cue words for all the ten discourse relations were found based on hand-coded rules. The rule for identifying the discourse relation can be any one of the following forms:

1. POS tag + one or more case suffixes
2. Conjunction cue words

Examples for POS tag + one or more case suffixes for the *Condition* discourse relation are:

i. Verb + “ஆல்”
Example: உண்டால் (Uṇṭāl – Eats)
<உண்டால்> : உண்டால் <Verb> ட் <Past Tense Marker>
<அள்> <Verbal Participle Suffix> ஆல் <Instrumental Case>

ii. Verb + “இன்”
Example: கூறின் (Kūṟiṉ – Says the)
<கூறின்> : கூறு <Verb> இன் <Genitive Case>

iii. Verb + “இல்”
Example: அடக்கில் (Aṭakkil – In Suppressive)
<அடக்கில்> : அடக்கு<Verb>இல்<Instrumental Case>

Examples for conjunction cue words for the *Condition* relation are:

i. எனின் (Eṉiṉ – If the)

ii. பெறின் (Peṟiṉ – If Received)

iii. ஆற்றின் (Āṟṟiṉ – If you do)

The entire rule set for identifying ten discourse relations is given in Table 1.
Table 1

Rules and Examples for Cue Words

<table>
<thead>
<tr>
<th>Discourse Relation</th>
<th>Hand-Coded Rules</th>
<th>Sample Cue Words</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Antithesis</strong></td>
<td>Verb + இன் + உம்</td>
<td>இன் + உம் (டினும – Feed)</td>
</tr>
<tr>
<td></td>
<td>Verb + உம்</td>
<td>உம் (சென்றோரும் – Those)</td>
</tr>
<tr>
<td></td>
<td>Conjunctions like</td>
<td>ஊட்டினும் (டினும – Feed)</td>
</tr>
<tr>
<td></td>
<td>ஆயினும்</td>
<td>ஆயினும் (டினும – Feed)</td>
</tr>
<tr>
<td></td>
<td>எனினும</td>
<td>எனினநும் (டினும – Feed)</td>
</tr>
<tr>
<td><strong>Condition</strong></td>
<td>Verb + இன்</td>
<td>இன் (வெளியில் – Eats)</td>
</tr>
<tr>
<td></td>
<td>Verb + உம்</td>
<td>உம் (வெளியில் – Eats)</td>
</tr>
<tr>
<td></td>
<td>Verb + இல்</td>
<td>இல் (வெளியில் – Eats)</td>
</tr>
<tr>
<td></td>
<td>Conjunctions like</td>
<td>ஊட்டினும் (டினும – Feed)</td>
</tr>
<tr>
<td></td>
<td>ஆயினும்</td>
<td>ஆயினும் (டினும – Feed)</td>
</tr>
<tr>
<td></td>
<td>எனினும</td>
<td>எனினநும் (டினும – Feed)</td>
</tr>
<tr>
<td><strong>Elaboration</strong></td>
<td>Conjunctions indicating example</td>
<td>போல (போல – Like)</td>
</tr>
<tr>
<td></td>
<td>அற்று</td>
<td>அற்று (அற்று – Like)</td>
</tr>
<tr>
<td><strong>Otherwise</strong></td>
<td>Anti-conditional</td>
<td>இல்லாயினும (இல்லாயினும – Nullity of the)</td>
</tr>
<tr>
<td></td>
<td>பெறாவிடினும (பெறாவிடினும – Unless the)</td>
<td></td>
</tr>
<tr>
<td><strong>Explanation</strong></td>
<td>Conjunctions indicating reason</td>
<td>அதனால் (அதனால் – So)</td>
</tr>
<tr>
<td></td>
<td>பயத்தலால் (பயத்தலால் – Because it produces)</td>
<td></td>
</tr>
<tr>
<td><strong>List</strong></td>
<td>Conjunctions indicating “and”</td>
<td>மற்றும் (மற்றும் – And)</td>
</tr>
<tr>
<td></td>
<td>இரண்டும்</td>
<td>இரண்டும் (இரண்டும் – Both)</td>
</tr>
<tr>
<td></td>
<td>Words indicating numbers of more than one</td>
<td>ஐந்து (ஐந்து – Five)</td>
</tr>
<tr>
<td><strong>Contrast</strong></td>
<td>Conjunctions indicating “but”</td>
<td>அன்றால் (அன்றால் – But)</td>
</tr>
<tr>
<td></td>
<td>அன்றால்</td>
<td>அன்றால் (அன்றால் – But)</td>
</tr>
<tr>
<td><strong>Background</strong></td>
<td>Pronoun + இன் + உம்</td>
<td>இன் + உம் (அதனினும – Even more)</td>
</tr>
<tr>
<td></td>
<td>Noun + இன் + உம்</td>
<td>இன் + உம் (நீரினும – Strait)</td>
</tr>
<tr>
<td></td>
<td>Noun + இல்</td>
<td>இல் (வறுமையின் – Of poverty)</td>
</tr>
<tr>
<td><strong>Solutionhood</strong></td>
<td>Conjunctions indicating answers for the question</td>
<td>என்று (என்று – That)</td>
</tr>
<tr>
<td></td>
<td>என்பது</td>
<td>என்பது (என்பது – The)</td>
</tr>
</tbody>
</table>

Each cue word signals a discourse relation and hence, after identifying the cue words, they are mapped to their corresponding discourse relations.
Word Reformation Algorithm

In Tamil literature, letters are grouped to form a “சிர் (Cīr – Tidy)”. Even though a cir looks like a word, it is not always equivalent to one, at times containing more than a word or part of it. A cir is used in place of words to retain intonation in poetic form, as explained in song number 1,268 of the Tamil grammar treatise, Tholkappiyam (Adigalasiriyar, 1985), which is given in Figure 2.

Figure 2

Tholkappiyam Song Number 1,268

For instance, the phrase “நுனியின் கரும்புதின் டட்டே (Nuṉiyiṉ karumputiṉ ṟaṟṟē – Like the cane of the tip is eaten)” in Naladiyar had only two words, “நுனியின் (Nuṉiyiṉ – Of the tip)” and “கரும்புதின் டட்டே (karumputiṉṟaṟṟē – Like eating sugarcane)” It was noted that the second word was further split into two words, “கரும்பு (karumputiṉ – Of sugarcane)” and “டட்டே (ṟaṟṟē)”, to preserve the prosody of the poem. These literary works were written on palm leaves, and readers had to memorize them, owing to the lack of resources, and thus prosody helped them in the memorization process. The Tamil Morphological Analyzer
failed to identify morphemes from this poetic line since it operated on word tokens separated by white spaces. The analyzer tried to identify morphemes from the three tokens: “நுனியின் (Nuṉiyiṉ – Of the tip), “கரும்புதின் (karumputiṉ – Of sugarcane)”, and “றற்றே (ṟaṟṟē)”. Since “றற்றே” was not a meaningful word, the output of the morphological analysis threw up an error, as follows:

```xml
<றற்றே>:<Error>
```

Once the words were meaningfully reformed, the analyzer divided the words into morphemes. This was done by merging the word that threw up the error in the output with the previous word, to ascertain if a meaningful word could be formed. Thus, the word reformation yields two meaningful words, “நுனியின் (Nuṉiyiṉ – Of the tip)” and “கரும்புதின்றற்றே (karumputiṉṟaṟṟē – Like eating sugarcane)”, by merging the third token with the second.

Next, the analyzer split the words in question into morphemes. The cue word, “அற்று (Aṟṟu – Like)”, from “கரும்புதின்றற்றே (karumputiṉṟaṟṟē – Like eating sugarcane)”, was identified and retrieved by the Morphological Analyzer, signaling the presence of the Elaboration relation. The word reformation algorithm was tested with both Thirukkural couplets and Naladiyar quatrains. Of the 7,263 errors thrown up by the analyzer, 4,181 were rectified by the algorithm, and the rest were caused by unidentified and agglutinative Tamil literary words. The word reformation algorithm helped procure additional cue words for the cue word identification task. The 2,041 cue words found without word reformation increased to 3,049, following its application.

Word reformation occurs in three stages: cue word identification of discourse parsing, indexing, and searching modules in the proposed work. Following the identification of cue words using the word reformation algorithm, NRS sequences were identified. The next subsection describes the process of nucleus identification.

**Nucleus Identification**

Thirukkural couplets had a single discourse relation, whereas Naladiyar quatrains had one or two. If the quatrain had delimiters
such as a hyphen (-), comma (,), or no symbol in the middle, the Thirukkural-Naladiyar discourse parser would generate a single NRS sequence. If there were delimiters such as a semicolon (;) or an exclamation mark (!) in the middle, the delimiter would split the quatrain into two text spans, and the Thirukkural-Naladiyar discourse parser would generate two NRS sequences.

The discourse parser analyzed the Thirukkural couplets and Naladiyar quatrains one after the other to check for cue words. The cue word was explicit in certain couplets and quatrains but merged with others in specific cases. With explicit cue words, the discourse relation was directly identified. Where they were not explicit, the Tamil Morphological Analyzer (Anandan et al., 2002) identified and extracted them. For example, the cue word “அற்று (Aṟṟu – Like)”, indicative of the Elaboration discourse relation, was recognized from the word “தின்றற்றே (tiṉṟaṟṟē – Like eating)”.

Following the identification of the discourse relation, the Thirukkural-Naladiyar discourse parser determined the nucleus from the couplets and quatrains. The nucleus might appear before or after the cue word in the text (Anita & Subalalitha, 2019b). Cue words were categorized into two sets, cue1 and cue2, based on their position in the respective couplets and quatrains. In the cue1 set, the nucleus preceded the cue word but succeeded it in the cue2 set. It was observed that the length of the nucleus was half that of the couplet and quatrain text spans, while the remainder formed the satellite. Once the nucleus was determined, NRS sequences were identified for every couplet and quatrain.

Since the Naladiyar quatrain in Example 5 had a hyphen in line 2, it had a single discourse relation, identified by the discourse parser as the Antithesis, using the cue word “சென்றோரும் (Ceṉṟōrum – Even gone)”. The discourse parser identified the NRS sequence, which is given in Figure 3.
Example 5:

Naladiyar Quatrain:

Yāṉai eruttam poliyak kuṭainiḻaṟkīḻc
cēṉait talaivarāyc ceṅṟōrum - ēṉai
viṉaiulappa vēṟāki vīḻvartām koṇṭa
maṉaiyāḷai māṟṟār koḷa.

Meaning in English:

Even those who have marched as generals, mounted on the back of an elephant and shaded by the umbrella, when the effect of evil deeds works their ruin, will suffer a change of state, and while their wives are enjoyed by their foes, will fall forever.

Figure 3

NRS Sequence for Example 5

Connective: கொண்டான்
Nucleus: யானை எருத்தம் பொலியக்
kuṭainiḻaṟkīḻc
cēṉait talaivarāyc ceṅṟōrum -
viṉaiulappa vēṟāki vīḻvartām koṇṭa
maṉaiyāḷai māṟṟār koḷa.
Satellite: ஏனை வினைஉவப்ப வேறாகி
vīḻvartām koṇṭa
maṉaiyāḷai māṟṟār koḷa.
Discourse Relation: Antithesis

Given that the Naladiyar quatrain in Example 6 had a semicolon in line 2, it had two discourse relations. The discourse parser split the quatrain into two. It identified the Condition discourse relation using the cue word “தோன்றியக்கால் (Tōṉṟiyakkāl – If
appearing)” in the first part, and the *Elaboration* discourse relation using the cue word “போல (Pōla – Like)” in the second. As a result, it produced the NRS sequences as given in Figure 4.

Example 6:

\[
\text{Naladiyar Quatrain:} \\
\text{Tukaḷnīr peruñcelvam tōṉṟiyakkāl toṭṭup} \\
\text{pakaṭu naṭantakūḷ pallārō ṭuṇka;} \\
\text{akaṭuṟa yārmāṭṭum nillātu celvam} \\
\text{cakaṭakkāl pōla varum.}
\]

\[
\text{English Transliteration:} \\
\text{Tukaḷnīr peruñcelvam tōṉṟiyakkāl toṭṭup} \\
\text{pakaṭu naṭantakūḷ pallārō ṭuṇka;} \\
\text{akaṭuṟa yārmāṭṭum nillātu celvam} \\
\text{cakaṭakkāl pōla varum.}
\]

\[
\text{Meaning in English:} \\
\text{When by blameless means thou hast acquired great wealth,} \\
\text{then eat with others rice imported by oxen, for wealth never} \\
\text{remaineth in the centre with anyone, but changes its position like} \\
\text{a cart-wheel.}
\]
Similarly, NRS sequences are identified for all the *Thirukkural* couplets and *Naladiyar* quatrains.

**Discourse-based Inverted Indexing**

Indexing is a key process in an IR system. Efficient indexing increases the retrieval speed and accuracy of the results retrieved. Most indexing techniques use keywords or synonyms to index a text document. Much more attention to semantics can be engendered with an indexed NRS sequence representing a tiny semantic graph. The proposed system used this tiny semantic graph to index *Thirukkural* couplets and *Naladiyar* quatrains to retrieve semantically closer results for user queries. As nuclei carried necessary information in the text, they were used as pointers to the NRS sequences in the proposed indexing technique. Grammatical categories that included nouns,
verbs, adjectives, and adverbs were accorded importance in the nuclei and used for indexing. The Tamil Morphological Analyzer identified these grammatical categories.

Figure 5 shows a representation of the semantic index. Inverted indexing was undertaken by checking all the nuclei with the couplets and quatrains of every discourse relation. The discourse relations, as well as related couplets and quatrains, comprised the indices. The number of a particular couplet and quatrain indicated the poem/song number in the *Thirukkural* and *Naladiyar*. Table 2 shows the indexing for the word “இந்து (Naṭpu – Friendship)”.

**Figure 5**

*Semantic Indexing*

![Semantic Indexing Diagram](image)

**Table 2**

*Indexing for the Sample Word “இந்து (Naṭpu – Friendship)”*

<table>
<thead>
<tr>
<th>Discourse Relations</th>
<th>Thirukkural Couplets along with their Number</th>
<th>Naladiyar Quatrains along with their Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antithesis</td>
<td>790, 794</td>
<td>223</td>
</tr>
<tr>
<td>Condition</td>
<td>821, 830</td>
<td>232</td>
</tr>
<tr>
<td>Elaboration</td>
<td>338, 781, 788, 800, 813, 821, 1122</td>
<td>12, 216, 232, 237</td>
</tr>
<tr>
<td>Otherwise</td>
<td>795</td>
<td>219</td>
</tr>
<tr>
<td>Explanation</td>
<td>790, 821</td>
<td>12</td>
</tr>
<tr>
<td>Joint</td>
<td>802, 1122</td>
<td>219, 370, 371</td>
</tr>
<tr>
<td>Contrast</td>
<td>786, 788, 817, 830, 874, 1165</td>
<td>77, 219, 223, 237, 370, 371</td>
</tr>
<tr>
<td>Background</td>
<td>781, 816, 817</td>
<td>12, 219</td>
</tr>
<tr>
<td>Solutionhood</td>
<td>789, 790, 794, 795, 801</td>
<td>77</td>
</tr>
</tbody>
</table>
Query Processing, Searching, and Ranking

Once a query was received from a user, the possible discourse relations were figured out and the grammatical components (nouns, verbs, adjectives and adverbs) were identified by the Morphological Analyzer.

Query Processing – Identification of Discourse Relations

Discourse relations from a query were worked out in two ways, using both keywords and question words. The words in the query were analyzed to understand the discourse relation so as to retrieve search results. For instance, a word in the query might have one of the following cue words that signal a Condition discourse relation:

1. Conditional suffix: “அதால்” (atāl – If the)
   Example: “செய்வதால்” (Ceyvatāl – By doing)
   <செய்வதால்> : செய் < Verb > வ் < Future Tense Marker > அதால் < Conditional Suffix >

2. Verb + “ஆல்”
   Example: “முடிந்தால்” (Muṭintāl – If possible)
   <முடிந்தால்> : முடி < Verb > உ < Past Tense Marker > ஆல் < Verbal Participle Suffix > உ ஆல் < Instrumental Case >

3. Verb + “இன்”
   Example: “செய்யின்” (Ceyyiṉ – If do)
   <செய்யின்> : செய் < Verb > ய் < Sandhi > இன் < Genitive Case >

Thirukkural couplets and Naladiyar quatrains connected by the condition discourse relation were retrieved, along with their explanations in Tamil and English. Question words also played an important role in identifying discourse relations. For example, Tamil question words referring to the English question word “why” normally indicated an explanation discourse relation. Table 3 shows a few Tamil question words and their discourse relations.
Table 3

Some of the Tamil Question Words and Discourse Relations

<table>
<thead>
<tr>
<th>Question Words</th>
<th>Discourse Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ஏன் (Ēṉ – Why)</td>
<td>Explanation</td>
</tr>
<tr>
<td>எதற்கு (etaṟku – Why)</td>
<td></td>
</tr>
<tr>
<td>எதற்காக (etaṟkāka – Why)</td>
<td></td>
</tr>
<tr>
<td>எப்பொழுது (eppoḻutu – When)</td>
<td>Condition, Antithesis, Otherwise</td>
</tr>
<tr>
<td>எவ்வாறு (evvāṟu – How)</td>
<td></td>
</tr>
<tr>
<td>எவை (evai – What)</td>
<td></td>
</tr>
<tr>
<td>எதனை (Etaṉai – What)</td>
<td></td>
</tr>
<tr>
<td>எதை (Etai – What)</td>
<td>Background, Elaboration</td>
</tr>
<tr>
<td>எவை (Evai – which)</td>
<td></td>
</tr>
<tr>
<td>எத்தனை (Ettaṉai – How many)</td>
<td>Joint, List</td>
</tr>
<tr>
<td>என்ன (Eṉṉa – What)</td>
<td></td>
</tr>
<tr>
<td>யார் (Yār – Who)</td>
<td></td>
</tr>
</tbody>
</table>

Similarly, other relations could be identified from user queries as well. Algorithm 1 was used for identifying the discourse relations of *Thirukkural* couplets and *Naladiyar* quatrains from user queries.

Algorithm 1: Discourse Relation Identification from Search Queries

1. If the word in query has any of the following
   i. Conditional suffix or (Verb + “ஆல்”) or (Verb + “இன்”)
   ii. Condition question word
   then the discourse relation is Condition

2. If the word in query has any of the following
   i. (Conditional suffix or (Verb + “ஆல்”) or (Verb + “இன்”))
      and Negative feature
   ii. Antithesis question word and Negative feature
   then the discourse relation is Antithesis

(continued)
3. If the word in query has any of the following

   Negative Conditional suffix
      i.   Cue word for Otherwise discourse relation
   iii. Otherwise question word and (Conditional suffix or (Verb +
         “ஆல்”)) and Negative feature
      then the discourse relation is Otherwise

4. If the word in query has any of the following
   i.   Background question word + “விட (Viṭa – Than)”
   ii.  Background question word + “காட்டிலும் (kāṭṭilum –
         Than)”

   then the discourse relation is Background

5. If the word in query has
   i.   Elaboration question word + Solutionhood cue word
   then the discourse relation is Elaboration

6. If the word in query has
   i.   Solutionhood question word + Solutionhood cue word +
       Negative Feature
       then the discourse relation is Contrast

7. If the word in query has
   i.   Explanation question word
       then the discourse relation is Explanation

8. If the word in query has
   i.   List question word
       then the discourse relation is List

9. If the word in query has
   i.   Joint question word
       then the discourse relation is Joint

In Algorithm 1 above, negative features indicated negativity. Adjectives like “கடும் (Kaṭum – Severe)”, adverbs like “இன்றி (Iṉṛi – Without)”, and negative finite verbs like “அல்ல (Alla –
         Not)” fell into this category.

In Example 7, the Tamil Morphological Analyzer output for the word “வாழ்வதால் (vāḻvatāl – Because of living)” is given below:
As the word “வாழ்வதால் (Vāḻvatāl – Because of living)” had a conditional suffix, the discourse relation used for retrieving results was *Condition*.

**Example 7:**

<table>
<thead>
<tr>
<th>Tamil Query</th>
<th>English Transliteration</th>
<th>Meaning in English</th>
</tr>
</thead>
<tbody>
<tr>
<td>உதவி செய்து வாழ்வதால் உண்டாகும் நன்மை என்ன?</td>
<td>Utavi ceytu vāḻvatāl uṇṭākum Naṉmai eṉṉa?</td>
<td>What are the benefits of helping and living?</td>
</tr>
</tbody>
</table>

**Query Processing – Interpreting Query Words**

The grammatical components namely noun, verb, adjective, and adverb from the query words were to be identified for further processing. The Tamil Morphological Analyzer identified these grammatical components from the user query, apart from which its semantics were interpreted using the Tamil dictionary (ilearnTamil live online Tamil tuition: Tamil to Tamil Dictionary, 2018). Grammatical components identified from the query in Example 7 are given below.

<table>
<thead>
<tr>
<th>Tamil Word</th>
<th>English Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>உதவி</td>
<td>Help</td>
</tr>
<tr>
<td>செய்து</td>
<td>Do</td>
</tr>
<tr>
<td>வாழ்வதால்</td>
<td>Because of living</td>
</tr>
<tr>
<td>உண்டாகும்</td>
<td>Occur</td>
</tr>
<tr>
<td>நன்மை</td>
<td>Benefit</td>
</tr>
</tbody>
</table>

The synonyms of the above words were also fetched and used for searching. For instance, for the query word, “உதவி (Utavi – Help)”, synonyms namely, “துணை (Tuṇai – Sub)”, “கொடை (Koṭai – Donation)”, and “நன்றி (Naṉṟi – Thanks)” were identified. This improved the quality of the search results. The head noun, which referred to the noun occurring first in the query, was identified from all the nouns in the query. The head noun and its synonyms were grouped together as a single set, titled “head”, as in Example 7, with the word
“உதவி (Utavi – Help)” and its synonyms. Other query words like “செய் (Cey – Do)”, “வாழ்வு (Vāḻvu – Life)”, “உண்டாகு (Uṇṭāku – become)”, and “நன்மை (Naṉmai – Benefit)”, along with their synonyms, were grouped under the “non-head” set.

**Discourse-based Searching and Ranking**

Figure 6 depicts the searching and ranking process based on discourse techniques. Discourse relations identified from the query, along with the preprocessed query words, were used for the search. These were matched with the semantic indices, and the results were retrieved using Algorithm 2.

**Figure 6**

*Searching and Ranking Process*

---

**Algorithm 2: Discourse-based Search and Rank**

1. Let DisRel be the discourse relations identified from the query
2. Let SemanticIndices = \{S_1, S_2, \ldots S_p\} be the set of indices that matches DisRel.
3. Let Head = \{H_1, H_2, \ldots H_n\} denote the set of Head words and its synonyms.
4. Let Non-head = \{N_1, N_2, \ldots N_m\} denote the set of Non-head words and its synonyms.
5. for each H_i in Head = \{H_1, H_2, \ldots H_n\}
6.        if(H_i ∈ SemanticIndices)
7.               Add the couplet number or quatrain number of the semantic index to the set res,
8.       end
9.   end
10.  for each \( N_i \) in Non-head = \{N_1, N_2, \ldots N_m \}
11.    if(\( N_i \in \text{SemanticIndices} \))
12.       Add the couplet number or quatrain number of the semantic index to the set \( \text{res}_2 \)
13.    end
14.  end
15.  Case (1) // Ranking
16.  Strategy level-1
17.  for p=1 to \( r \), where \( r \) is the size of the set \( \text{res}_1 \)
18.    for q=1 to \( s \), where \( s \) is the size of the set \( \text{res}_2 \)
19.      if (\( \text{res}_1[p] == \text{res}_2[q] \))
20.         Display Thirukkural couplet or Naladiyar quatrain, along with its Tamil and English Explanations
21.    end
22.  end
23.  Case (2) // Ranking
25.  for p=1 to \( r \), where \( r \) is the size of the set \( \text{res}_1 \)
26.    if \( \text{res}_1[p] \) is not displayed in Case (1)
27.      Display Thirukkural couplet or Naladiyar quatrain, along with its Tamil and English Explanations
28.    end
29.  for q=1 to \( s \), where \( s \) is the size of the set \( \text{res}_2 \)
30.    if \( \text{res}_2[q] \) is not displayed in Case (1)
31.      Display Thirukkural couplet or Naladiyar quatrain, along with its Tamil and English Explanations
32.  end
33.  end
34.  Case (3) // Ranking
35.  Strategy level-3
36.  Let DisRelOther be the set of discourse relations not identified from the query
37.  for each discourse relation in DisRelOther
38.    Let \( \text{SemanticIndices} = \{ S_1, S_2, \ldots S_p \} \) be the semantic indices corresponding to \( \text{DisRelOther} \)
39.    for each \( H_i \) in Head = \{H_1, H_2, \ldots H_n \}
40.      if(\( H_i \in \text{SemanticIndices} \))
41.         Display the Thirukkural couplet or Naladiyar quatrain, along with its Tamil and English Explanations, if it is not already displayed in Case (1) or Case (2).
42.    end
43.  end
44.  for each \( N_i \) in Non-head = \{N_1, N_2, \ldots N_m \}
44. if \( N_i \in \text{SemanticIndices} \)
45.     Display the Thirukkural couplet or Naladiyar quatrain, along with its Tamil and English Explanations, if it is not already displayed in Case (1) or Case (2).
46.     end
47. end
48. end

Figure 7 shows the ranking strategy, levels 1 – 3, of the proposed system.

Figure 7

*Ranking Strategy Levels*

The proposed system used a three-level ranking strategy as shown in Algorithm 2 and Figure 7. In Ranking Strategy Level 1, query words in the *head* and *non-head* sets, as well as their discourse relation(s), were matched with the semantic indices. Couplets and quatrains...
retrieved using Ranking Strategy Level 1 were given the highest priority, as the combination of (head, non-head, discourse relation) present in the couplets and quatrains was semantically most relevant to the query. Explanations, in Tamil and English, of the couplets and quatrains were also retrieved alongside. For instance, the Thirukkural couplet in Example 8 was retrieved using Ranking Strategy Level 1 for the query in Example 7. It was observed that this couplet was most semantically similar to the query, because it matched with the words “உதவி (Utavi – Help)” and “நன்மை (Naṉmai – Benefit)” in the head set and non-head set, respectively, along with the Condition discourse relation.

**Example 8:**

<table>
<thead>
<tr>
<th>Thirukkural:</th>
<th>பயன்தூக்கார் செய்த உதவி நன்மை கடலின் பெரிதும்.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tamil Explanation:</td>
<td>என்னப் பயன் கிடைக்கும் என்று எண்ணிப் பார்க்காமலே, அன்பின் காரணமாக ஒருவர் செய்த உதவின் சிறப்பு கடலை விடப் பெரிதும்.</td>
</tr>
<tr>
<td>English Explanation:</td>
<td>If we weigh the excellence of a benefit which is conferred without weighing the return, it is larger than the sea.</td>
</tr>
</tbody>
</table>

The *Thirukkural* in Example 9, retrieved through Ranking Strategy Level 2, was semantically similar to the query, although not as similar to that retrieved by Ranking Strategy Level 1. Ranking Strategy Level 2 attempted to partially tie query words with the couplets and quatrains. This was done by matching them either with a head set word or a non-head set word of the query, along with the discourse relation.
Example 9:

Thirukkural:

அஞ்சாமை அவ்வளவு சமானமில்லா நீங்கும் தேவை தவிர வேண்டாது.

Tamil Explanation: ஒரு செயலுக்குரிய வழி முறைகளைக் குறையின்றிச் சிந்தித்துச் செய்யுமிடத்து, அஞ்சாமை ஒன்றைத் தவிர, வேறுதேவை தவிர வேண்டாது.

English Explanation: You will need no other aid than fearlessness, if you thoroughly reflect (on what you are to do), and select (a suitable) place for your operations.

The Thirukkural couplet in Example 10 was retrieved through Ranking Strategy Level 3, which tried to match the query words and their synonyms with the semantic indices containing the words, plus any discourse relation. This ensured that the semantics in the discourse structure was utilized by the ranking scheme as well.

Example 10:

Thirukkural:

மகன்தந்தைக்கு ஆற்றும் உதவியும் நேர்த்தரங்கம் தொடர்புப் பொருளைச் செய்யும்.

Tamil Explanation: ஆகா! இவனைப் பிள்ளைப் பெற்றது இவன் தந்தை பெற்றபெரும்பேறு, என்று ஒரு மகன் தந்தைக்குச் செய்யப்பட்டும். அன்று கத்தையா தொக்கியதற்கும் வானையாளும் காதலைம் செய்யப்பட்டது.

English Explanation: (So to act) that it may be said “by what great penance did his father beget him,” is the benefit which a son should render to his father.
Nevertheless, discourse relations could not be identified from all queries. In such cases, the words in the query and their synonyms were matched with the semantic indices and the *Thirukkural* couplets and *Naladiyar* quatrains were retrieved and ranked based on the popularity of the discourse relations fetched from the indices. The popularity of the discourse relations was calculated based on their frequency in both the *Thirukkural* and *Naladiyar* as shown in Table 4. Thus, the proposed approach obtained results using discourse-based relations, even if the query was incomplete.

### Table 4

**Number of Couplets and Quatrains in Discourse Relations**

<table>
<thead>
<tr>
<th>Discourse Relation</th>
<th>Number of <em>Thirukkural</em> Couplets</th>
<th>Number of <em>Naladiyar</em> Quatrains</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Condition</em></td>
<td>379</td>
<td>157</td>
<td>536</td>
</tr>
<tr>
<td><em>Contrast</em></td>
<td>209</td>
<td>205</td>
<td>414</td>
</tr>
<tr>
<td><em>Antithesis</em></td>
<td>141</td>
<td>182</td>
<td>323</td>
</tr>
<tr>
<td><em>Elaboration</em></td>
<td>152</td>
<td>150</td>
<td>302</td>
</tr>
<tr>
<td><em>Explanation</em></td>
<td>97</td>
<td>90</td>
<td>187</td>
</tr>
<tr>
<td><em>Otherwise</em></td>
<td>93</td>
<td>87</td>
<td>180</td>
</tr>
<tr>
<td><em>Solutionhood</em></td>
<td>53</td>
<td>92</td>
<td>145</td>
</tr>
<tr>
<td><em>Joint</em></td>
<td>83</td>
<td>59</td>
<td>142</td>
</tr>
<tr>
<td><em>Background</em></td>
<td>64</td>
<td>39</td>
<td>103</td>
</tr>
<tr>
<td><em>List</em></td>
<td>59</td>
<td>10</td>
<td>69</td>
</tr>
</tbody>
</table>

Given that the *Condition* discourse relation had the highest number of couplets and quatrains, the results were retrieved from it for the matching indices. After that, they were retrieved from the *Contrast* discourse relation, and so on.
The user query in Example 11 was syntactically incomplete, containing no clue to help identify the discourse relation. The Condition discourse relation was given first preference, and the results were retrieved using it, given that it connected the highest number of couplets and quatrains. The Thirukkural couplet retrieved is as given in Figure 8.

Example 11:

<table>
<thead>
<tr>
<th>User query: சினத்தின் விளைவு</th>
</tr>
</thead>
<tbody>
<tr>
<td>English Transliteration: Ciṉattiṉ viḷaivu</td>
</tr>
<tr>
<td>Meaning in English: The effect of anger</td>
</tr>
</tbody>
</table>

Retrieved Thirukkural for Example 11

Thirukkural: 
தன்னைத்தான் காக்கின் சினங்காக்க் காவாக்கால்
தன்னையே கொல்லுஞ் சினம்.

Tamil Explanation: ஒருவன் தன்னைத்தானே காத்துக் கொள்ளவேண்டுமானால்
சினைக் கைவிடவேண்டும். இல்லையே இல்லையே
அவனை அழித்துவிடும்.

English Explanation: If a man would guard himself, let him guard against anger; if he do not guard it, anger will kill him.

The next section describes the evaluation of the proposed work.

RESULT AND DISCUSSION

The proposed method was tested on all 1,330 Thirukkural couplets and 400 Naladiyar quatrains. An enhanced version of the existing Thirukkural discourse parser (Anita & Subalalitha, 2019b) was
employed to construct the proposed *Thirukkural-Naladiyar* discourse parser. The parser identified NRS sequences in *Thirukkural* couplets and *Naladiyar* quatrains, and constructs a discourse structure for all 1330 *Thirukkural* couplets and 400 *Naladiyar* quatrains.

The Tamil dictionary used to obtain query word synonyms had 95,291 words. On average, four synonyms were identified for a word and 40 indices for a query, pointing to many NRS sequences.

The discourse-based indexing and search system was tested using 15 queries, and the proposed discourse-based IR system was compared with the Google Tamil search engine. In order to justify the advantages of incorporating semantics into the proposed system, a keyword-based search was also performed on the proposed system, and the results were compared. The proposed method was evaluated using precision (P), average precision (AP), and MAP metrics computed using Equations 1, 2, and 3.

Precision (P) = \[
\frac{\text{Number of Thirukkural couplets and Naladiyar quatrains correctly retrieved}}{\text{Total number of Thirukkural couplets and Naladiyar quatrains retrieved}}
\] (1)

Precision for the *Thirukkural* couplets and *Naladiyar* quatrains was computed using Equation 1. AP@10 is the average precision for the top 10 retrieved results, which is computed using Equation 2:

\[
\text{AP}@10 = \frac{1}{\text{TP}} \sum_{k=1}^{10} P@k \times R@k
\] (2)

where TP indicates the number of relevant *Thirukkural* and *Naladiyar* quatrains retrieved; k@10 is the top 10 results retrieved from the search system; P@k is the precision at the kth retrieved result, and R@k is the relevance at k. Relevance is equal to 1 if the result is relevant, and 0 if it is not. The MAP, which is the mean of all queries, is computed using Equation 3:

\[
\text{MAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i
\] (3)

where N is the number of queries, which is 15. The MAP score for the Google Tamil search was 0.56 and 0.62 for the keyword-based method, while the MAP score for the proposed discourse-based search system was 0.89.
Table 5 shows the average precision values for the Google Tamil search, keyword-based search, and the proposed discourse-based search, which indicated a performance comparison of the proposed method with the Google Tamil search and keyword-based method, where Qi indicates query i. The precision and MAP scores achieved by the proposed work were higher than those of the Google Tamil search and keyword-based methods, owing to the use of discourse relations, semantic indices, and the discourse-based search and rank algorithm.

### Table 5

<table>
<thead>
<tr>
<th>Queries</th>
<th>Google Search</th>
<th>Keyword-based Search</th>
<th>Discourse-based Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>0.5</td>
<td>0.61</td>
<td>0.82</td>
</tr>
<tr>
<td>Q2</td>
<td>0.49</td>
<td>0.49</td>
<td>0.91</td>
</tr>
<tr>
<td>Q3</td>
<td>0.55</td>
<td>0.66</td>
<td>0.91</td>
</tr>
<tr>
<td>Q4</td>
<td>0.54</td>
<td>0.59</td>
<td>0.86</td>
</tr>
<tr>
<td>Q5</td>
<td>0.67</td>
<td>0.71</td>
<td>0.81</td>
</tr>
<tr>
<td>Q6</td>
<td>0.52</td>
<td>0.66</td>
<td>0.84</td>
</tr>
<tr>
<td>Q7</td>
<td>0.57</td>
<td>0.65</td>
<td>0.89</td>
</tr>
<tr>
<td>Q8</td>
<td>0.7</td>
<td>0.74</td>
<td>0.94</td>
</tr>
<tr>
<td>Q9</td>
<td>0.43</td>
<td>0.47</td>
<td>0.96</td>
</tr>
<tr>
<td>Q10</td>
<td>0.55</td>
<td>0.65</td>
<td>0.95</td>
</tr>
<tr>
<td>Q11</td>
<td>0.57</td>
<td>0.64</td>
<td>0.93</td>
</tr>
<tr>
<td>Q12</td>
<td>0.41</td>
<td>0.47</td>
<td>0.98</td>
</tr>
<tr>
<td>Q13</td>
<td>0.67</td>
<td>0.71</td>
<td>0.85</td>
</tr>
<tr>
<td>Q14</td>
<td>0.52</td>
<td>0.54</td>
<td>0.91</td>
</tr>
<tr>
<td>Q15</td>
<td>0.66</td>
<td>0.67</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Both the Google search and keyword-based search were built around keyword matching. In Google, mostly the entire chapter of *Thirukkural*...
matching the keywords present in the query were retrieved by skipping the semantically relevant ones. For example, if a query carried the word “நட்பு (Naṭpu – Friendship)”, the Google search retrieved a chapter on “நட்பு (Naṭpu – Friendship)”, in addition to a chapter on “தீநட்பு (Tī naṭpu – Evil Friendship)” from the *Thirukkural*, which explained the reason for the low MAP score.

In order to justify the power of the semantics underlying the discourse relations, the proposed approach was further tested by skipping discourse relations and indexing only keywords and synonyms. It was observed that even the top results retrieved by the modified search system were found to be irrelevant. This is because the search was based solely on words and their meanings, bypassing the semantic connections between them, culminating in a low MAP score. The low scores achieved by the keyword-based search revealed that discourse relations played a vital role in picking up semantically closer *Thirukkural* couplets and *Naladiyar* quatrains.

**CONCLUSION**

Tamil literary classics, dating prior to 300 BCE, are a treasure trove of invaluable information. Computationally analyzing them to make them accessible to today’s generation is inevitable. In this paper, a discourse relation-based semantic indexing, searching, and ranking mechanism for the Tamil literary texts, *Thirukkural* and *Naladiyar*, has been attempted. The discourse structures of both have been constructed by enhancing the existing *Thirukkural* discourse parser (Anita & Subalalitha, 2019b).

Unlike existing Tamil search systems, the proposed search system retrieved *Thirukkural* couplets and *Naladiyar* quatrains that were semantically closer to a user query, along with their explanations in Tamil and English. This is largely owing to the fact that the proposed system is greatly tailored that the semantic interpretations of the *Thirukkural* and *Naladiyar* are captured through discourse relations. Furthermore, the proposed ranking strategy ranked the retrieved couplets and quatrains by matching the semantic components of both the query and the discourse-based indices. The proposed strategy was evaluated using the MAP score and compared with the traditional methods.
keyword-based search and Google search. The results were better than those produced by state-of-the-art approaches and were largely driven by the discourse relations used in indexing and ranking.

The proposed system can be further extended to retrieve other Tamil classics, with a few modifications in their discourse structure. The proposed system lays a strong foundation for the computational analysis of Tamil literary works to build various NLP applications, such as Question Answering Systems and Summary Generation Systems, to name a few.

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES


Samia, Z., & Khaled, R. (2020). Multi-agents indexing system (MAIS) for plagiarism detection. Journal of King Saud University-


