JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
http://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Ramalingam, A., & Navaneethakrishnan, S. C. (2021). A discourse-based information retrieval for Tamil literary texts. *Journal of Information and Communication Technology, 20*(3), 353-389. https://doi.org/10.32890/jict2021.20.3.4

A Discourse-Based Information Retrieval for Tamil Literary Texts

[1] Anita Ramalingam & [2]Subalalitha Chinnaudayar Navaneethakrishnan
[1&2]Department 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 (Natpu – 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: நட்பு சிதையாமல் இருக்க என்ன செய்ய வேண்டும்?

English Transliteration: Naṭpu citaiyāmal irukka eṉṉa ceyya vēṇṭum?

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

Example 2: Retrieved Thirukkural Couplet

Thirukkural Couplet:
உடம்பை தனித்தழியப் புள்பறந் தற்றே
உடம்பொடு உயிரிடை நட்பு.

Tamil Explanation: உடலுக்கும் உயிருக்கும் உள்ள உறவு மட்டைக்கும் பறவைக் கூஞ்சுக்கும் உண்டான உறவு போன்றதுதான்.

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.

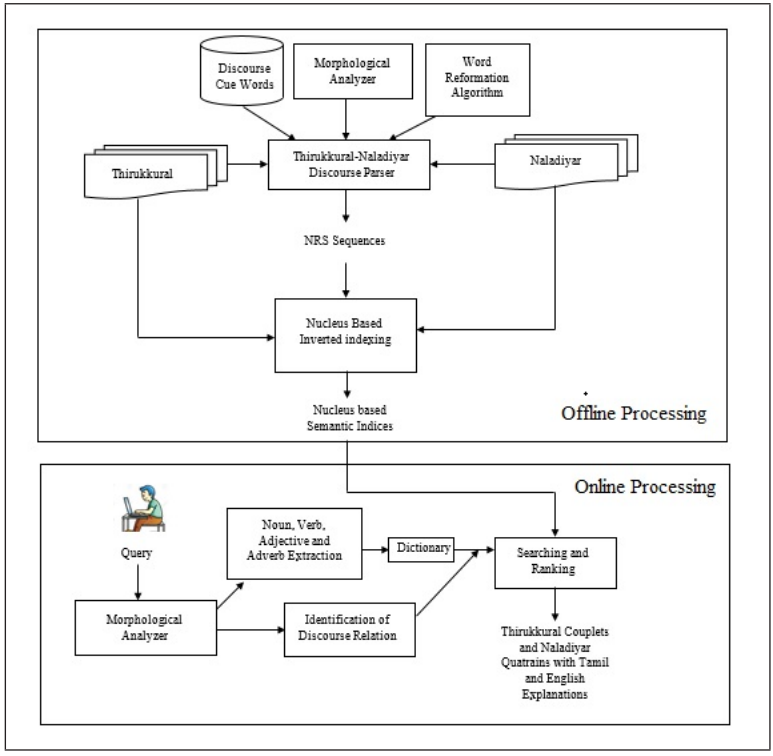
This subsection discusses the computational work on Tamil. It is observed that much of the existing research carried out have been restricted to the Tamil expository (essay-type) texts, with very few studies undertaken on Tamil literary texts. The IR system, implemented thus far for Tamil, has limited 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 (Tin̲rarr̲ē – As if eaten)" needed to be morphologically analyzed to separate the cue word, "அற்ற (Ar̲ru

– 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 "சய்வதால் (Ceyvatāl – By doing)" is given in Example 3.

Example 3:

Word: "சய்வதால் (Ceyvatāl – By doing)"

Morphological Analyzer Output:
<சய்வதால்>:சய் < Verb > வ் < Future Tense Marker > அதால் <Conditional Suffix>

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 (Ārṟiṉ – If you do)", indicated that the *Thirukkural* couplet contained a *Condition* discourse relation.

Example 4:

Thirukkural Couplet:
வீழ்நாள் படாஅமை நன்றாற்றின்
அஃதொருவன்
வாழ்நாள் வழியடைக்கும் கல்.

English Transliteration:
Vīḻnāḷ paṭā'amai naṉṟāṟṟiṉ aṅtoruvaṉ
vāḻnāḷ vaḻiyaṭaikkum kal.

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>:உண் < Verb > ட் < Past Tense Marker > உ < Verbal Participle Suffix > ஆல் < Instrumental Case >

ii. Verb + "இன்"
Example: கூறின் (Kūṟiṉ – Says the)
<கூறின்>:கூற < Verb > இன் < Genitive Case >

iii. Verb + "இல்"
Example: அடக்கில் (Aṭakkil – In Suppressive)
<அடக்கில்>: அடக்க < 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

Discourse Relation	Hand-Coded Rules	Sample Cue Words
Antithesis	Verb + இன்+ உம் Verb + உம் Conjunctions like ஆயினும்	ஊட்டினும் (Ūṭṭiṉum – Feed) சென்றோரும் (Ceṉṟōrum – Those) எனினும் (Eṉiṉum – However)
Condition	Verb + ஆல் Verb + இன் Verb + இல் Conjunctions like ஆயின்	உண்டால் (Uṇṭāl – Eats) உரைப்பின் (Uraippiṉ – Text of) பழிக்கில் (Paḻikkil – If blame) எனின் (Eṉiṉ – If the)
Elaboration	Conjunctions indicating example	போல் (Pōl – Like) அற்ற (Aṟṟu – Like)
Otherwise	Anti-conditional	இல்லாயின் (Illāyiṉ – Nullity of the) பறொவிடின் (Peṟāviṭiṉ – Unless the)
Explanation	Conjunctions indicating reason	அதனால் (Aṭaṉāl – So) பயத்தலால் (Payattalāl – Because it produces)
Joint	Conjunctions indicating “and”	மற்றும் (Maṟṟum – And) இரண்டும் (Iraṇṭum – Both)
List	Words indicating numbers of more than one	இவை (Ivai – These) ஐந்த (Aintu – Five)
Contrast	Conjunctions indicating “but”	ஆனால் (Āṉāl – But) அன்றால் (Aṉṟāl – But then)
Background	Pronoun + இன் + உம் Noun + இன் + உம் Noun + இன்	அதனினும் (Aṭaṉiṉum – Even more) நீரினும் (Nīriṉum – Strait) வறுமையின் (Vaṟumaiyiṉ – Of poverty)
Solutionhood	Conjunctions indicating answers for the question	என்ற (Eṉṟu – That) என்பத (Eṉpatu – The)

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 "சீர் (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

அசையும் சீரும் இசையொடு சேர்த்தி
வகுத்தனர் உணர்த்தல் வல்லோர் ஆறே.

English Transliteration:
Acaiyum cīrum icaiyoṭu cērtti
vakuttaṉar uṇarttal vallōr āṟē.

Ilampooranar Urai:
அசையையுஞ் சீரையும் ஓசையோடு
சேர்த்திப் பாகுபாடுணர்த்தல்
வல்லோர்கள் நெறி.

Meaning in English:
It is the nature of the poets to unite the tokens (*cirs*) and the building blocks of *cirs* (*asai*) to form a prosody.

For instance, the phrase "நுனியின் கரும்புதின் றற்றே (Nuṉiyiṉ karumputiṉ ṟaṟṟē – Like the cane of the tip is eaten)" in *Naladiyar* had only two words, "நுனியின் (Nuṉiyiṉ – Of the tip)" and "கரும்புதின்றற்றே (karumputiṉṟaṟṟē – Like eating sugarcane)" It was noted that the second word was further split into two words, "கரும்புதின் (karumputiṉ – Of sugarcane)" and "றற்றே (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:

<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, "அற்ற (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 "அற்ற (Arṟu – 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 "சென்றேமாயும் (Ceṉṟōrum – Even gone)". The discourse parser identified the NRS sequence, which is given in Figure 3.

Example 5:

Naladiyar Quatrain:
யானை எரத்தம் பொலியக் கடநிழற்கீழ்ச்
சனேதைத் தலவைராய்ச் சன்றோரும் -
ஏனை
வினைஉவப்ப வறோகி வீழ்வர்தாம்
கொண்ட
மனையாளை மாற்றார் கொள.

English Transliteration:
Yāṉai eruttam poliyak kuṭainiḻaṟkīḻc
cēṉait talaivarāyc ceṉṟōrum - 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: சன்றோரும்
Nucleus: யானை எரத்தம் பொலியக் கடநிழற்கீழ்ச் சனேதைத் தலவைராய்ச் சன்றோரும்
Satellite: ஏனை வினைஉவப்ப வறோகி வீழ்வர் தாம் கொண்ட மனையாளை மாற்றார் கொள.
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:

Naladiyar Quatrain:
தகள்நீர் பெருஞ்செல்வம் தோன்றியக்கால்
தொட்டுப்
பகடு நடந்தகூழ் பல்லாரோ டுண்க;
அகடுற யார்மாட்டும் நில்லாத செல்வம்
சகடக்கால் போல வரும்.

English Transliteration:
Tukaḷnīr peruñcelvam tōn̲riyakkāl toṭṭup
pakaṭu naṭantakūl̲ pallārō ṭuṇka;
akaṭuṟa yārmāṭṭum nillātu celvam
cakaṭakkāl pōla varum.

Meaning in English:
When by blameless means thou hast acquired great wealth, then eat with others rice imported by oxen, for wealth never remaineth in the centre with anyone, but changes its position like a cart-wheel.

Figure 4

NRS Sequences for Example 6

1. தகள்நீர் பரெஞ்சல்வம்
தோன்றியக்கால் தொட்டப் பகட
நடந்ததூழ் பல்லாரோ டண்க

Connective: தோன்றியக்கால்
Nucleus: தகள்நீர் பரெஞ்சல்வம் தோன்றியக்கால் தொட்டப்
Satellite: பகட நடந்ததூழ் பல்லாரோ டண்க
Discourse relation: Condition

2. அகடற யார்மாட்டும் நில்லாத
சல்வம் சகடக்கால் போல வரம்.

Connective: போல
Nucleus: சல்வம் சகடக்கால் போல வரம்.
Satellite: அகடற யார்மாட்டும் நில்லாத
Discourse relation: *Elaboration*

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 "நட்பு (Natpu – Friendship)".

Figure 5

Semantic Indexing

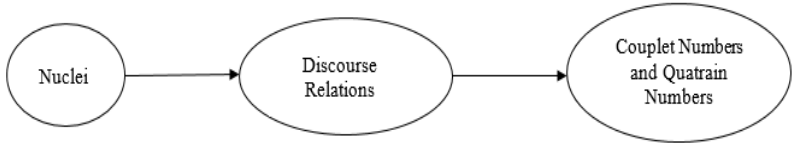


Table 2

Indexing for the Sample Word "நட்பு (Natpu – Friendship)"

Discourse Relations	*Thirukkural* Couplets along with their Number	*Naladiyar* Quatrains along with their Number
Antithesis	790, 794	223
Condition	821, 830	232
Elaboration	338, 781, 788, 800, 813, 821, 1122	12, 216, 232, 237
Otherwise	795	219
Explanation	790, 821	12
Joint	802, 1122	219, 370, 371
Contrast	786, 788, 817, 830, 874, 1165	77, 219, 223, 237, 370, 371
Background	781, 816, 817	12, 219
Solutionhood	789, 790, 794, 795, 801	77

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: "சயெ்வதால் (Ceyvatāl – By doing)"
   <சயெ்வதால்> :சயெ் < Verb > வ் < Future Tense Marker > அதால் <Conditional Suffix>
2. Verb + "ஆல்"
   Example: "மடிந்தால் (Muṭintāl – If possible)"
   <மடிந்தால்>: மடி < Verb > ந்த் < Past Tense Marker > உ <Verbal Participle Suffix > ஆல் < Instrumental Case >
3. Verb + "இன்"
   Example: "சயெ்யின் (Ceyyin – If do)"
   <சயெ்யின்>: சயெ் < 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

Question Words	Discourse Relation
ஏன் (Ēṉ – Why) எதற்கு (etaṟku – Why) எதற்காக (etaṟkāka – Why)	*Explanation*
எப்பொழுது (eppoḻutu – When) எவ்வாறு (evvāṟu – How) எவை (evai – What)	*Condition, Antithesis, Otherwise*
எதனை (Etaṉai – What) எதை (Etai – What)	*Background, Elaboration*
எவை (Evai – which) எத்தனை (Ettaṉai – How many)	*Joint, List*
என்ன (Eṉṉa – What) யார் (Yār – Who)	*Solutionhood, Contrast*

Similarly, other relations could be identified from user queries as well. Algorithm 1 was used for identifying the discourse relations of *Thirukkural c*ouplets 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") or (Verb + "இன்")) 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 + போல "(Pōla – Like)"

   then the discourse relation is Elaboration

6. If the word in query has
   i. Solutionhood question word + Solutionhood cue word

   then the discourse relation is Solutionhood

7. If the word in query has
   i. Solutionhood question word + Solutionhood cue word + Negative Feature

   then the discourse relation is Contrast

8. If the word in query has
   i. Explanation question word

   then the discourse relation is Explanation

9. If the word in query has
   i. List question word

   then the discourse relation is List

10. 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 (Iṉṟi – 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:

<வாழ்வதால்>: வாழ்வு < Noun > அதால் < Conditional Suffix >

As the word "வாழ்வதால் (Vāḻvatāl – Because of living)" had a conditional suffix, the discourse relation used for retrieving results was *Condition.*

Example 7:

User query: உதவி சயெ்த வாழ்வதால் உண்டாகும் நன்மை என்ன?
English Transliteration: Utavi ceytu vāḻvatāl uṇṭākum Naṉmai eṉṉa?
Meaning in English: What are the benefits of helping and living?

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.

<உதவி>: உதவி < Noun >
<சயெ்த>: சயெ் < Verb >
<வாழ்வதால்>: வாழ்வு < Noun >
<உண்டாகும்>: உண்டாக < Verb >
<நன்மை>: நன்மை < Adjective >

The synonyms of the above words were also fetched and used for searching. For instance, for the query word, "உதவி (Utavi – Help)", synonyms namely, "தணை (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

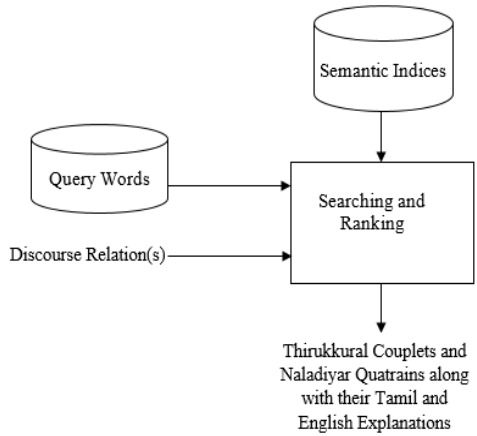
"உதவி (Utavi – Help)" and its synonyms. Other query words like "செய் (Cey – Do)", "வாழ்வு (Vāḻvu – Life)", "உண்டாக (Uṇṭāku – become)", and "நன்மை (Naṉmai – Benefit)", along with their synonyms, were grouped under the "*non-head*" set.

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($H_i \in$ SemanticIndices)
7. Add the couplet number or quatrain number of the semantic index to the set res_1

(continued)

```
8    .      end
9.   end
10.  for each N_i in Non-head = {N_1, N_2,… N_m }
11.       if(N_i ∈ SemanticIndices)
12.            Add the couplet number or quatrain number of the semantic
     index to the set res_2
13.     end
14.  end
15.  Case (1)                                          // Ranking
     Strategy level-1
16.  for p=1 to r, where r is the size of the set res_1
17.       for q=1 to s, where s is the size of the set res_2
18.            if (res_1[p] == res_2[q])
19.                 Display Thirukkural couplet or Naladiyar quatrain, along
     with its Tamil and
                    English Explanations
20.            end
21.      end
22.  end
23.  Case (2)                                          // Ranking
     Strategy level-2
24.  for p=1 to r, where r is the size of the set res_1
25.       if res_1[p] is not displayed in Case (1)
26.            Display Thirukkural couplet or Naladiyar quatrain, along with
     its Tamil and
                  English Explanations
27.       end
28.  end
29.  for q=1 to s, where s is the size of the set res_2
30.       if 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
     Strategy level-3
35.  Let DisRelOther be the set of discourse relations not identified from
     the query
36.  for each discourse relation in DisRelOther
37.       Let SemanticIndices={ S_1, S_2,…S_p } be the semantic indices
     corresponding to
       DisRelOther
38.       for each H_i in Head = {H_1, H_2,… H_n }
39.            if(H_i ∈ SemanticIndices)
40.                 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).
41.            end
42.       end
43.       for each N_i in Non-head = {N_1, N_2,… N_m }
```

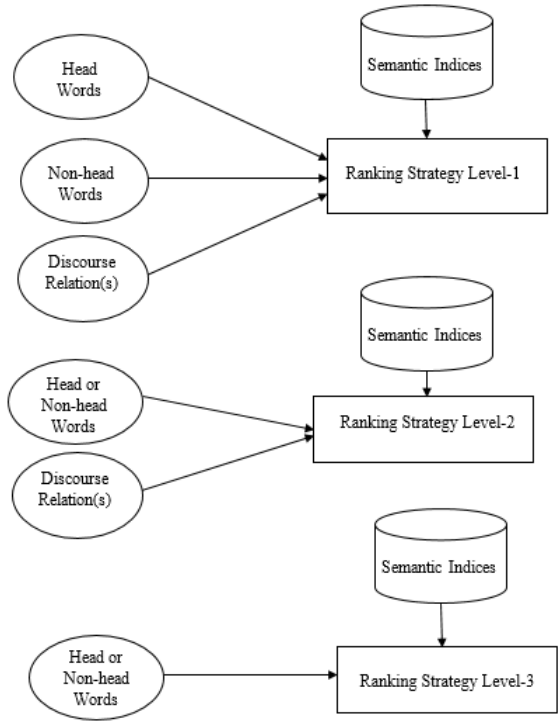
(continued)

```
44.            if(N_i ∈ 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 (Nanmai – Benefit)" in the *head* set and *non-head* set, respectively, along with the *Condition* discourse relation.

Example 8:

Thirukkural:

பயன்தூக்கார் செய்த உதவி நயன்தூக்கின்
நன்மை கடலின் பெரிது.

Tamil Explanation: என்ன பயன் கிடைக்கும் என்ற எண்ணிப் பார்க்காமலே, அன்பின் காரணமாக ஒருவர் செய்த உதவியின் சிறப்பு கடலை விடப் பெரிது.

English Explanation: If we weigh the excellence of a benefit which is conferred without weighing the return, it is larger than the sea.

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 அல்லால் துணைவேண்டா எஞ்சாமை
எண்ணி இடத்தால் சயின்.

Tamil Explanation: ஒரு சயெலுக்கரிய வழி மறகைளகைக் கறயைின்றிச் சிந்தித்துச் சயெ்யுமிடத்த, அஞ்சாமை ஒன்றதைத் தவிர, வறே தணை தவேயைில்லை.

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 உதவி இவன்தந்தை
என்நோற்றான் கொல்எனும் சொல்.

Tamil Explanation: ஆகா! இவனைப் பிள்ளையாகப் பற்றத் இவன் தந்தை பற்ற பரெம்பறே, என்ற ஒரு மகன் புகழப்படுவதான், அவன் தன்னுடயை தந்தைக்குச் சயெ்யக்கூடிய கமைம்மாற எனப்படும்.

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

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

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:

User query: சினத்தின் விளைவு
English Transliteration: Ciṉattiṉ viḷaivu
Meaning in English: The effect of anger

Figure 8

Retrieved Thirukkural for Example 11

Thirukkural:
தன்னைத்தான் காக்கின் சினங்காக்க காவாக்கால்
தன்னையே கொல்லுஞ் சினம்.

Tamil Explanation: ஒருவன் தன்னைத்தானே காத்துக் கொள்ள வேண்டுமானால், சினத்தைக்கைவிடவேண்டும்.இல்லையேல் சினம், அவனை அழித்தவிடும்.

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.

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

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:

$$\text{AP@10} = \frac{1}{\text{TP}} \sum_{k=1}^{10} \text{P@k} * \text{R@k} \quad (2)$$

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:

$$\text{MAP} = \frac{1}{\text{N}} \sum_{i=1}^{\text{N}} \text{N AP}_\text{i} \quad (3)$$

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

Average Precision for Google, Keyword-based, and Discourse-based Searches

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

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 "தீ நட்பு (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

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

Abraham, S. A. (2003). Chera, Chola, Pandya: Using archaeological evidence to identify the Tamil kingdoms of early historic South India. *Asian Perspectives*, *42*(2), 207–223. https://doi.org/10.1353/asi.2003.0031

Adigalasiriyar (1985). *Tolkāppiyam: Poruḷatikāram - ceyyuḷiyal* [Tolkappiyam: Book of semantics – Chapter of poetry]. Tamil University Thanjavur.

Agosti, M., Marchesin, S., & Silvello, G. (2020). Learning unsupervised knowledge-enhanced representations to reduce the semantic gap in information retrieval. *ACM Transactions on Information Systems (TOIS)*, *38*(4), 1–48. https://doi.org/10.1145/3417996

Anandan, P., Saravanan, K., Parthasarathi, R., & Geetha, T. V. (2002, December). Morphological analyzer for Tamil. *In International Conference on Natural Language Processing*.

Anita, R., & Subalalitha, C. N. (2019a, July). An approach to cluster Tamil literatures using discourse connectives. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)* (pp. 1-4). IEEE. https://doi.org/10.1109/ICESIP46348.2019.8938315

Anita, R., & Subalalitha, C. N. (2019b, December). Building discourse parser for Thirukkural. In *Proceedings of the 16th International Conference on Natural Language Processing (ICON-2019) IIIT Hyderabad, India: NLP Association of India* (pp. 18–25).

Fauzi, M. A., Arifin, A. Z., & Yuniarti, A. (2017). Arabic book retrieval using class and book index based term weighting. *International Journal of Electrical & Computer Engineering*, *7*(6), 2088–8708. http://doi.org/10.11591/ijece.v7i6.pp3705-3710

Giridharan, R., Vellingiriraj, E. K., & Balasubramanie, P. (2016, April). Identification of Tamil ancient characters and information retrieval from temple epigraphy using image zoning. In *2016 International Conference on Recent Trends in Information Technology* (*ICRTIT*) (pp. 1–7). IEEE. https://doi.org/10.1109/ICRTIT.2016.7569600

ilearnTamil live online Tamil tuition: Tamil to Tamil Dictionary. (2018, June 15). Retrieved from https://ilearntamil.com/tamil-to-tamil-dictionary/

Liu, L., Liu, L., Fu, X., Huang, Q., Zhang, X., & Zhang, Y. (2018). A cloud-based framework for large-scale traditional Chinese medical record retrieval. *Journal of Biomedical Informatics*, *77*, 21–33. https://doi.org/10.1016/j.jbi.2017.11.013

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, *8*(3), 243–281. https://doi.org/10.1515/text.1.1988.8.3.243

Meng, L., Tan, A. H., & Wunsch II, D. C. (2019). Online multimodal co-indexing and retrieval of social media data. In *Adaptive Resonance Theory in Social Media Data Clustering*, (pp. 155–174). Springer, Cham. https://doi.org/10.1007/978-3-030-02985-2_7

Saravanan, M. S. (2020). Semantic document clustering based indexing for Tamil language information retrieval system. *Journal of Critical Reviews*, *7*(14), 2999–3007. https://dx.doi.org/10.31838/jcr.07.14.563

Prasath, R., Sarkar, S., & O'Reilly, P. (2015, April). Improving cross language information retrieval using corpus based query suggestion approach. In *International Conference on Intelligent Text Processing and Computational Linguistics*, (pp. 448 457). Springer, Cham. https://doi.org/10.1007/978-3-319-18117-2_33

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

Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2020.06.009

Sankaralingam, C., Rajendran, S., Kavirajan, B., Kumar, M. A., & Soman, K. P. (2017, September). Onto-thesaurus for Tamil language: Ontology based intelligent system for information retrieval. In *2017 International Conference on Advances in Computing, Communications and Informatics* (*ICACCI)* (pp. 2396–2396). IEEE. https://doi.org/10.1109/ICACCI.2017.8126206

Subalalitha, C. N. (2019). Information extraction framework for Kurunthogai. *Sādhanā*, *44*(7), 1–6. https://doi.org/10.1007/s12046-019-1140-y

Subalalitha, C. N., & Anita, R. (2016). An approach to page ranking based on discourse structures. *Journal of Communications Software and Systems*, *12*(4), 195–200. http://dx.doi.org/10.24138/jcomss.v12i4.78

Tekli, J., Chbeir, R., Traina, A. J., & Traina Jr, C. (2019). SemIndex+: A semantic indexing scheme for structured, unstructured, and partly structured data. *Knowledge-Based Systems*, *164*, 378–403. https://doi.org/10.1016/j.knosys.2018.11.010

Thenmozhi, D., & Aravindan, C. (2018). Ontology-based Tamil–English cross-lingual information retrieval system. *Sādhanā*, *43*(10), 1–14. https://doi.org/10.1007/s12046-018-0942-7

Zamani, H., Dehghani, M., Croft, W. B., Learned-Miller, E., & Kamps, J. (2018, October). From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (pp. 497–506). https://doi.org/10.1145/3269206.3271800