Design and Development of a Bilingual Thesaurus for Classical Tamil Studies
Experiences and Issues

Abstract
Based on experiences with the design and development of a Tamil-English bilingual thesaurus this paper discusses with examples: (1) some issues related to vocabulary management in multilingual thesauri in culture-specific domains; (2) special aspects of the Tamil language in this regard; (3) alternative ways of linking certain descriptors to lengthy lists of NTs and RTs; (4) advantages of integrated use of two or more knowledge organization tools; and (5) use of the bilingual thesaurus for certain types of research in Tamil. Issues related to equivalence, non-hierarchic associative relationships, homographs, NTs are discussed. The paper suggests that integrated use of two or more knowledge organization tools adds value.

Introduction
Bilingual and multilingual thesauri are useful tools in cross language information retrieval that require translation of queries and document representations from input language to other languages. Jorna and Davies (2001) have remarked that: “... multilingual tools are gaining importance as increasingly diverse groups from different cultural and linguistic backgrounds seek access to equally diverse pieces of information”. Even within the Web community, there is growing interest in vocabulary-based techniques. Harpring (1999) gives an overview of the Getty’s vocabularies with examples of their use in Web search interfaces.

Recently, the Government of India established a Centre of Excellence for Classical Tamil (CECT) to promote ‘the cause of Classical Tamil’. The Centre has initiated a project to build a digital library of Tamil classics of the Sangam period1 (Sharada and Manju Naika, 2007) and, as part of this, a Tamil-English bilingual thesaurus covering the domain of ‘Tamil Studies’. Some issues relevant to multilingual thesauri in culture-specific domains have been reported in another paper. (Neelameghan and Raghavan, 2005) This paper discusses the following aspects of multilingual thesaurus construction based on the experience in working on this ongoing project:

• Vocabulary management in bilingual and multilingual thesauri in culture-specific domains;
• Special aspects of the Tamil language in this regard;
• Alternative ways of linking descriptors to lengthy lists of NTs and RTs;
• Advantages of integrated use of two or more knowledge organization tools; and
• Use of the bilingual thesaurus for certain types of research in Tamil.

The Approach
Tamil, a Dravidian language with its own script, is independent of both Sanskrit and Latin and has a literary tradition going back to around 300 B.C. The domain of Tamil Studies covers a wide range of subjects including practically all areas of humanities, many social

1. Many historians refer to the Tamil literature from ca. 300 B.C. to 300 A.D. as Sangam literature; Sangams were Tamil academies, which enabled poets and authors to gather periodically to publish their work (Wikipedia).
sciences and a few sciences. As there is no information retrieval thesaurus covering the
domain, the thesaurus is being built using a ‘bottom up’ approach starting with a corpus
of 125,000 Tamil terms identified using available lexical tools. The principal source has
been the *Tamil Lexicon* of the University of Madras (1924–1936), web editions of which
are available (University of Chicago, Dictionaries of South-Asian Languages (DSAL);
University of Cologne, Tamil Lexicon (COTL)). In building the thesaurus a modular ap-
proach based on the schedule of main classes of Colon classification was employed. This
helped to have adequate control, even though integrating records for terms in different
disciplines could raise problems of handling duplicate entries at a later stage. The features
of the thesaurus are described in another paper (Neelameghan and Raghavan, 2007). The
thesaurus is being maintained as a WINISIS database with the following fields:

<table>
<thead>
<tr>
<th>Tag</th>
<th>Name</th>
<th>Repeatability</th>
<th>Sub-fields</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Descriptor</td>
<td></td>
<td>abcde</td>
</tr>
<tr>
<td>11</td>
<td>Descriptor1</td>
<td></td>
<td>a</td>
</tr>
<tr>
<td>2</td>
<td>SN</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>US</td>
<td>a</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>UF</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>BT</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>NT</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>RT</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>900</td>
<td>Links</td>
<td></td>
<td>abcdedf</td>
</tr>
</tbody>
</table>

The Field 11, *Descriptor1* is being used for inputting the descriptor in a format required
for hyper-linking with the DSAL database. Field 1 and all the words (excluding apparatus
words) in field 2 are indexed. Terms in all the fields except fields 2 and 4 are hyper-linked
either to the corresponding thesaurus record or to DSAL’s *Tamil Lexicon*. Some of the
issues relevant to the bilingual thesaurus are discussed below.

**Script and Transliteration**

Tamil terms were transliterated and entered in Roman script. The major concern was to
adopt a transliteration standard that supported automatic and consistent conversion to
Tamil script. The COTL transliteration scheme was adopted as it met these requirements.
In other words terms in the Tamil script could be added as a separate field to the database
at a later date without having to key-in the terms.

**Equivalence Relationship**

A language is a product of, and reflects the culture of a community. It is the culture and
lifestyle of the members of a community that necessitate the formation of lexemes/expres-
sions (words/terms). Concepts in culture-specific domains have meaning in the life of the
members of the community belonging to the culture. Unless communities speaking two
different languages share the same culture, it is likely that certain concepts in culture-
specific domains may have unique verbal expressions only in a particular language. For
example:

- *cAttunARRu* = Young plants planted in place of the dead ones
- *aSTAgkaputti* = Eight Kinds of Knowledge
- *cARvAkam* = cAruvAka’s materialistic philosophy which says perception is the only
  source of knowledge
In this thesaurus two issues relating to equivalence relations had to be handled:

- The presence of a large number of synonyms in Tamil for many of the concepts
- The difficulty in finding equivalent concepts (and terms) in the English language

The notion of ‘equivalence relationship’ had to be extended to include both synonyms in Tamil and terms in English denoting near equivalent/semantically close concepts. In practice different strategies, depending on the situation, were employed:

- To use the nearest semantically equivalent English term;
- To use an appropriate phrase/compound term in English that represented the concept;
- To use the original term itself in case of concepts almost accepted as proper nouns.

The second issue related to ‘equivalence relationship’ was the occurrence of several terms in Tamil with closely similar connotations. Searches for two synonymous Tamil words, ‘tAmarai’ and ‘kamalam’ and their English equivalent, ‘lotus’ carried out in the online Tamil Lexicon (http://dsal.uchicago.edu/dictionaries/tamil-lex/) retrieved records in which the input string occurred either as an independent word or embedded in another word. More than 200 of the 327 terms were related to ‘tAmarai’ and 82 of these were synonyms!

<table>
<thead>
<tr>
<th>Search Term</th>
<th>No. of Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>tAmarai</td>
<td>327 Entries with tAmarai as entry word or in the explanation</td>
</tr>
<tr>
<td>kamalam</td>
<td>36 entries with kamalam as entry word or in the explanation</td>
</tr>
<tr>
<td>Lotus</td>
<td>309 entries with Lotus as entry word or in the explanation</td>
</tr>
</tbody>
</table>

### Homographs

Many of the Tamil terms were found to be homographs. For example, ‘tAmarai’ has been used to indicate:

- Lotus plant;
- Lotus flower;
- Lotus as a shape (entities in the shape of a lotus);
- Lotus-like properties (e.g., soft like lotus petals)

Some other examples of homographs in Tamil are:

- appu = Thigh; Father; Loan; Debt; Domestic male servant; Water; Trumpet tree; Sixth division of day
- iTimpam = Baby; Misery; Spleen; Egg of birds; Castor plant.

The evolution in the meaning and connotation of terms in Tamil probably explains the presence of a number of homographs. The Tamil words kurinchi, mullai, marutam, neitl, and palai, for example, denoted the plants, creepers, trees (all flora) found in abundance in the region concerned. Over a period of time the regions where these plants were found to grow came to be referred to by these same terms. Subsequently even the lifestyle characterizing the region came to be referred to by the term. This probably explains the use of the names of the regions for the codes of life prevalent in various regions. (Neelameghan, 2008). The real meaning is to be understood in the context. In building the thesaurus it became necessary to extensively use role operators to indicate the different connotations and contexts of the Descriptor. For example:

iTimpam (baby); iTimpam (castor); iTimpam (egg); iTimpam (misery); iTimpam (spleen)
Hierarchical Relationship

The general guideline adopted for hierarchical relationships between concepts was to treat types or varieties as also parts of an entity as narrower terms. In reality a part may be part of any level (or order). The example below represents hierarchical relationship:

- **Tâmarai** (lotus)
  - **MirunALam** (Stalk of the Lotus)
    - **TâmaraimuL** (thorny portion of the stalk of the lotus)

Parts of lower levels are labelled as NT1, NT2, NT3, etc. The major issue in handling hierarchical relationship arose from the occurrence of a large number of NTs to many descriptors; e.g., *utti* (literary device) has over 32 NTs; *cangka-nUl* (Tamil classics) has 41 NTs, *AmaNakku* (castor plant) has 55 NTs, *nAyanAr* (Saiva saints) has 63 NTs, *ceyyuLLami* (figure of speech) has over 180 NTs. The usual practice is to list all NTs, and to treat NTs to the same descriptor as RTs to each other. For example:

- **AmaNakku** (castor plant)
  - SN Castor plant; Ricinus communis
  - BT *tAparavastu* (plant)
  - NT *acAram* (castor plant)
  - NT *amanTalam* (castor plant)
  - NT *amanTam* (castor plant)

    ... ... ...

Clicking on a NT, say *acAram* (castor plant) will link to the record:

- **acAram** (castor plant)
  - SN Castor plant variety
  - BT *AmaNakku* (castor plant)
  - BT2 *tApara-vastu* (plant)
  - RT *amanTalam* (castor plant)
  - RT *amanTam* (castor plant)

    ... ... ...

This approach resulted in long schedules and a large number of linked entries. To reduce the size of the thesaurus without compromising on the links (e.g. to the related varieties of castor plants, in the above example), in all cases involving five or more NTs, the full list of NTs is given in one schedule under the descriptor with hyper-links to this record from each of the NTs. For example, the record for *acAram* (castor plant) as descriptor will be:

- **acAram** (castor plant)
  - SN Castor plant variety
  - BT *AmaNakku* (castor plant)
  - BT2 *tApara-vastu* (plant)
    - RT *amanTalam* (castor plant)
    - RT *amanTam* (castor plant)

    ... ... ...

Lateral Relationships (Non-Hierarchical Associative Relationships)

Conceptually lateral relationships (LRs) have remained a grey area. Theoretically a concept may have conceptual relationship with any other concept. However, from a pragmatic point of view it is both necessary and helpful to identify conceptual relationships that have the potential to contribute to enhancing information processes. While it may be difficult to measure the impact of providing for lateral relationships on retrieval effectiveness, it is generally conceded that associative relationships are important in a thesaurus to support navigation and for enhancing search. In reality many terms are associated with
any given term. Which of the LRs should be handled by the vocabulary has been the subject of discussion. A general rule could be that all near-permanent relationships between concepts should be handled by the vocabulary. A few efforts at defining and categorizing LRs have been reported (Neelameghan & Ravichandra Rao, 1975; Neelameghan & Maitra, 1978; Neelameghan, 2001). However, the schema was applied in the design of a substantive thesaurus only recently (Moholt, 2001). A comprehensive attempt to examine the semantics of relationships in knowledge organization has also been reported. (Bean, 1996; Bean & Green, 2001) The work related to the present bilingual thesaurus was used to identify LRs in culture-specific domains. A taxonomy of LRs between concepts has been developed. (Neelameghan & Raghavan, 2005; Neelameghan & Raghavan, 2006) An earlier version of this taxonomy has been mapped to their nearest equivalents in FrameNet highlighting the characteristics and also the varying degrees of compatibility between the two schemes. (Green & Bean, 2006) Working on the present thesaurus has also helped in continuously evaluating the schema. For example, one kind of LRs that had to be handled was:

- **Transformation/Metamorphosis**: Relationship arising from a change either in the external form or internal nature of an entity: Transformation of an individual following a mystical experience; Transformation of an individual after adapting another way of life (e.g., **grihasthaAsrama** to **sanyAsAsrama**); Telekinesis; Resurrection, etc.

Theoretically, a case for explicitly marking different kinds of LRs in a thesaurus could be made. However, at this stage, it is not clear as to how these could be used for enhancing navigation and retrieval; small scale experiments are on to obtain a feedback from the users on the utility of specifying the nature of LRs between a pair of concepts (terms) instead of merely labelling all as RTs. Such an approach could lead to:

- Improved precision in IR;
- Better mapping of a semantic domain; and
- Knowledge discovery.

### Web of Relationships and Value Addition

The building of the thesaurus led to the recognition of a complex network of relationships between concepts. Given this, the near impossibility of providing, for any given concept, the complete network of relationship as also a full description was realized. For handling the web of relationships two other options have been explored:

- **Maintaining especially long lists of NTs and RTs as separate files and hyper-linking the corresponding thesaurus record(s) to these files**;
- **Linking thesaurus records to appropriate lexical tools on the Web.** Every record in the Tamil-English thesaurus has been linked to the page containing the corresponding lexical record in the online **Tamil Lexicon (DSAL)** (Figure 2).

Discussions with subject specialists identified some other special requirements; e.g. the need for listing, as descriptors, titles of all the 41 Tamil classics of the period, commentaries on these, etc. In handling this it was realized that listing of a large number of titles as NTs in an alphabetical sequence is not helpful. Instead grouping the NTs in a logical sequence as in the schedule of classics in Colon Classification was found to be more helpful to users.

Neelameghan & Parthasarathy (2008) based on their study of literary devices (**utti**) in Tamil scholarly communication show that knowledge organization tools (KOTs), such as thesauri linked to lexicon, can also aid knowledge discovery and researches in the subject domain covered by the KOT.
Conclusions

There is increasing awareness of the dynamic and context-dependent character of languages especially in the Humanities. This suggests the importance of tools providing conceptual information that explain the use and meaning of terms and also indicate their relationships to other concepts. Relationships are important for disambiguation in terms of, showing the different meanings that terms can assume in a particular knowledge domain. Such tools are also of value in supporting formulation of well-structured queries that are known to yield better results than poorly structured queries. Bilingual and Multilingual thesauri especially in culture-specific domains raise several issues that need to be adequately addressed:
• It is helpful to use appropriate mechanisms and web-based lexical tools for adequate and comprehensive handling of relationships between concepts especially in culture-specific domains.
• It is difficult to find equivalent terms with one-to-one correspondence in a bilingual thesaurus especially in the Humanities.
• It is indeed difficult to implement parallel and identical hierarchies of related terms (BTs, NTs and RTs) for many concepts in two or more languages.

References