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

K.S. Raghavan and A. Neelameghan

1. Background

The Government of India established the Centre of Excellence for Classical Tamil (CECT) about two years ago "with a view to promoting the cause of Classical Tamil". Currently CECT is hosted by the Central Institute of Indian Languages, Mysore, India (CIIL), where there have already been active programmes and projects for fostering Tamil. A project has been initiated to create a digital library of Tamil classics of the Sangam period[•] (Sharada and Manju Alka, 2007). Thesauri and similar controlled vocabularies have been part of standard cataloguing practice in libraries and bibliographic databases. There are efforts to apply these tools for indexing digital hypertext resources via metadata resource descriptors. Metadata sets such as Dublin Core include elements for representing the subject of a resource in addition to other data elements. Such standards often recommend that, where possible, the subject element be taken from a relevant thesaurus. In a comparative study of multilingual thesauri - InfoDEFT and Esser's EXPO 2000 thesauri - Jorna and Davies (2001) remarked that: " …. multilingual tools are getting importance as increasingly diverse groups from different cultural and linguistic backgrounds seek access to equally diverse pieces of information". In the expanding globalization scenario "Conflict arises in the minds of men" through misinterpretation and misunderstanding of messages from different cultures, classes of people and linguistic groups (UNESCO) and it is therefore important to device means, methods and tools for improving inter-cultural and inter-faith exchange of ideas.

[•] Many historians refer to the Tamil literature from c. 300 BC to 300 A.D. as Sangam literature; Sangams were Tamil academies, which according to Tamil legends, enabled poets and authors to gather periodically to publish their work (Source: *Wikipedia*)

Jorna and Davies also mention the problems of developing a multilingual thesaurus for different user groups. Considering all these, it was felt that a bilingual thesaurus should be developed as part of the digital library project to support information indexing and retrieval. A project was therefore initiated to design and develop a Tamil-English bilingual thesaurus covering the domain of classical / ancient Tamil Studies. The project is a part of the Government of India Plan Scheme for Classical Tamil being implemented by the CIIL. It is expected that this thesaurus will play a key role not only in indexing but also in facilitating end users to navigate and retrieve meaningful results from the massive amounts of information in the digital library. This ongoing project started in April 2006 and in the course of the project several issues and findings have emerged. This paper seeks to discuss some of these issues, which have much wider implications.

2. The Present Study

Tamil has a glorious hoary past. Studies and researches in Tamil (including classical Tamil language and literature) are not confined to India. There have been and are several such efforts in academic and research centres in many parts of the world for many decades now. The output of knowledge resources from these activities has been substantial. The CIIL project to develop a bilingual thesaurus for classical Tamil is perhaps the first attempt of its kind.

Typically a thesaurus is a type of controlled indexing vocabulary, in which index terms are restricted to a controlled set of terms. Many thesauri exist, covering a variety of subject domains; e.g. the MeSH, the Art and Architecture Thesaurus, INSPEC Thesaurus, etc. These tools and techniques developed primarily for handling library materials in *print-on-paper* format have been fairly successful. However, questions have been raised about the utility and value of these traditional tools for knowledge organization in the digital environment.

There have been deliberations on the future of traditional tools in the context of networked resources and the nature of information retrieval on the Web; concurrently within the Web community there has been a growing interest in vocabulary-based techniques. Harpring (1999) gives an overview of the Getty's vocabularies with examples of their use in Web retrieval interfaces and collection management systems.

Design and development of bilingual / multilingual thesauri in culture-specific domains (especially in the humanities and social sciences fields) present many problems arising from the very nature of these disciplines. Initial examination of the domain of the bilingual thesaurus being developed indicated that over 1,25,000 (one hundred and twenty-five thousand) Tamil terms may constitute the corpus. Creating records by entering terms in 'A to Z' sequence was abandoned in favour of adopting a subject-wise approach using a source such as the *Tamil Lexicon.* Based on an examination of the corpus of terms, the major disciplines / basic classes have been identified and defined according to the schedule of Basic Classes of *Colon Classification.* This paper discusses the issues specific to the construction of such thesauri and their use in information retrieval:

a) Thesauri and other tools for knowledge organization are based on explicit recognition of the implicit conceptual relationships between concepts represented by terms in a vocabulary. 'Relationship' is an abstraction belonging to or characteristic of two entities or parts together (WordNet 2.0). Determination and categorization of semantic association between a pair of concepts is at the core of this process. What are the issues that culture-specific domains present in recognizing and categorizing conceptual relations? What are the problems in identifying and representing conceptual relations in culture-specific domains; The schema of relations widely employed in thesauri (Equivalence, Hierarchical and Associative relations) have largely emerged

based on the experience of building thesauri in science and technology; These are domains in which semantic closeness between concepts has been fairly clearly defined and by and large identification of conceptual relationships is fairly straightforward. Often the same schema is applied to culture-specific domains also; the adequacy of the existing schema of conceptual relationships in defining and representing the range of relationships encountered in culture-specific domains is an important issue;

b) The abstract nature of the concepts encountered in and associated with Humanities in general and culture-specific domains; rarely can we relate these concepts to concrete referents. What are the linguistic issues relevant to a bilingual / multi-lingual thesaurus in culture-specific domains? A large number of concepts encountered in culture-specific domains are those that have some meaning in the life of the members of the community belonging to the culture. A language is a product of, and reflects the culture of the particular community (ies). In other words, it is the culture and lifestyle prevalent among the members of a particular community that necessitates and results in the formation of lexemes / expressions (words / terms) for concepts associated with that culture and lifestyle. There is therefore the major problem of identifying corresponding and equivalent concepts (terms) in the other languages of the thesauri for concepts that are easily expressed in the source language.

c) Quite early in the course of developing the thesaurus, the near impossibility of comprehensively representing the entire range of related terms (hierarchically and associatively related terms) to any given concept was realized. Two major problems related to:

    - The large number of synonyms and near synonyms for many of the descriptors; and
    - The large number of homographs present in Tamil.

As a result most of the thesaurus records tended to get very large with many synonyms and related terms. This necessitated a re-examination of the need to make every thesaurus record complete. It was thought necessary to keep the size of each thesaurus record within reasonable limits without compromising on its effectiveness. The problem also raised the question of exploring the feasibility of arranging the large number of related terms, within a record, in a logical sequence rather than in alphabetical sequence. Certain devices were experimented with. These related largely to recognizing the complementary role of different K.O. tools and exploiting the availability of many lexical tools on the Web. The value addition that could be made to the thesaurus records by making use of available lexical tools is a characteristic and novel feature of the thesaurus being developed.

d) It has been planned to employ Tamil script as well as Roman transliteration of Tamil terms as descriptors in the thesaurus with view to facilitate its wider use. Issues of transliteration have been addressed keeping in mind the need to facilitate automatic conversion of descriptors in Tamil to Tamil script.

e) The process of developing the thesaurus has been an interactive one and at regular intervals discussions with domain scholars in CIIL is being held to obtain feedback and inputs. During such interactions, the need for inclusion of titles of Tamil classics, the commentaries and respective authors as descriptors was recognized. This has been accommodated.

References

[1] Harpring, P. (1999) "How forcible are the right words: overview of applications and interfaces incorporating the Getty vocabularies". Proceedings of Museums and the Web 1999, Archives and Museum Informatics. http://www.archimuse.com/mw99/papers/harpring/harpring.html

[2] Jorna, Kerstin and Davies, Sylvie (2001). Multilingual thesauri for the modern world – no ideal solution? Journal of Documentation, 57(2); 284-295

[3] Sharada, B.A. and Manju Alka (2007). Development of Classical Tamil Digital Library: CIIL experience (*Abstract in* Prasad, ARD and Madalli, Devika. *Ed*. ICSD 2007. – Bangalore: DRTC, 2007. – p. 351