மின்மஞ்சரி

அக்டோபர் 2010

உலகத் தமிழ்த் தகவல் தொழில்நுட்ப மன்றம் (உத்தமம்)

http://www.infitt.org

ஆசிரியர்

கவிஅரசன் வா.மு.சே.,

தலைவர் (பொறுப்பு) உத்தமம்

இணை ஆசிரியர்

முனைவர் G.S. மகாலட்சுமி

கணினி அறிவியல் மற்றும் பொறியியல் துறை,

அண்ணா பல்கலைக்கழகம், சென்னை—25.

மின்மஞ்சரி ஆசிரியர் குழு

ஆசிரியர்

கவிஅரசன் வா.மு.சே., **தலைவர் (பொறுப்பு) உத்தமம்**

இணை ஆசிரியர்கள்

முனைவர் கு. சூ. மகாலட்சுமி, அண்ணா பல்கலைக் கழகம்,சென்னை

முனைவர் வா.மு.சே. ஆண்டவர், பச்சையப்பர் கல்லூரி, சென்னை

முனைவர் இரா. செல்வராஜ், தலைவர் தமிழ்மணம், அமெரிக்கா

முனைவர் மாலா நேரு, அண்ணா பல்கலைக் கழகம், சென்னை

முனைவர் அ.முத்துக்குமார், குமரகுரு பொறியியல் கல்லூரி, கோவை

பொறிஞர். திருமதி அனு கணேஷ், தமிழ் ஆசிரியை, அமெரிக்கா

முனைவர் செ. செந்தில்குமார், அண்ணா பல்கலைக் கழகம், சென்னை

முனைவர் ஞானபாரதி, மத்தியத் தொழில் ஆய்வு நிறுவனம், சென்னை

பொறிஞர். பாலசுந்தரராமன், கணினி அறிஞர், பெங்களூரு

பொறிஞர். வெற்றிப்பாண்டியன், இணைச் செயலர், வாசிங்டன் தமிழ்ச்சங்கம்

ஆசிரியர் பக்கம்

உத்தமம் தொடர்ந்து வெளியிட்டுவரும் மின்மஞ்சரியின் வடிவமைப்பு இம்முறை தமிழரின் பண்பாட்டுப் பழமையை ஒத்து வடிவமைக்கப் பட்டுள்ளது.

கணினி அறிஞர்கள் நவீன முறைகளைத் தமிழில் பயன்படுத்தும் தன்மையும், தனது பழமையைப் போற்றி மகிழ்வர் என உணர்த்தும் வண்ணமும் அமைத்து வடிவமைக்க உதவிய இணை ஆசிரியர் மகாலட்சுமிக்கு ஆசிரியர் குழுவின் சார்பில் எனது நன்றி உரித்தாகுக.

மூன்றே மாதங்களில் இவ்வருடத்தின் இரண்டாம் மலர் வெளிவருவது பெரும் மகிழ்வையளிக்கின்றது.

தமிழ்க் கணினி அறிஞர், ஆசிரியர், சான்றோர், மாணவர், பயனர், நுகர்வோர் எனப் பல பிரிவினர் விரும்பிப் படிக்கும் வண்ணம் நமது இதழின் தரத்தை மேலும் உயர்த்த முயல்வோம்.

மின்மஞ்சரியின் அடுத்த இதழ், பொங்கல் மலர் 2011 ஆக, உத்தமத்தின் பத்தாண்டு வரலாற்றைப் பதிப்பிக்கும் வண்ணம் மிகச் சிறப்பாகப் பதிப்பிக்கும் எண்ணமும் உள்ளது.

தங்களின் ஆதரவை என்றும் போல் தந்துதவ வேண்டுகிறோம்.
நட்புடன்
வா.மு.சே. கவிஅரசன்
ஆசிரியர், மின்மஞ்சரி.

உள்ளடக்கம்

INTERNATIONAL FORUM FOR INFORMATION TECHNOLOGY IN TAMIL (INFITT)

Tamil is an ancient language but it will remain ancient unless it takes advantage of Information Technology. Information, as technology, got its start in the United States in the early 1950's. The language of choice was English. Handling of Tamil text materials in computers, in Tamil script form, became possible in the mid-eighties through the advent of a handful of Tamil fonts and text editors. Free fonts, for use in all three of the commonly used computer platforms (Windows, Mac and Unix), and the emergence of the World Wide Web, as a powerful medium for information exchange, started from the mid-nineties. It enabled the widespread exchange of Tamil digital materials worldwide through Internet-based communication tools (Email, Web). Tamil Information Technology efforts expanded phenomenally and became firmly established amongst Tamils worldwide thereby introducing a phase transition in the Tamil, and through it, Tamil culture.

The first International Conference devoted to Tamil Computing-TamilNet 97- organized at the National University of Singapore, in June 97, laid the foundations for bringing together, annually, key players in the field of Tamil Information Technology. The primary issues for attention were delineated and the urgent need for standardization of basic elements, such as font encoding and associated keyboards, was recognized. The second International Conference-Tamilnet 99- was organized in Chennai, Tamil Nadu, India in February 1999. It brought together representatives of Tamil speaking peoples and took a landmark step in recommending official glyph encoding and keyboard layout standards. The third International Conference -Tamil Inaiyam 2000- was held in Singapore during 22-24 July 2000.

As a follow up to Tamilnet 99, National Steering Committees have been formed in Tamil Nadu, Sri Lanka and Singapore and several regional committees being established, to promote Tamil Information Technology in the respective countries and regions. Substantial amounts of monetary and human resources have been made available to develop standards and software in key areas of Tamil Information Technology.

There is an urgent need to coordinate the efforts of various national and regional initiatives, through an International Forum for Information Technology in Tamil (hereinafter referred to as INFITT), consisting of regional and national representations, concerned institutions and individuals.

Many Asian languages already have such International Steering Committees. An example

is the Asia-Pacific Networking Group (AP-NG? devoted to Chinese. In addition to China, there are regional representatives from North America, Europe, country and individual representatives from Singapore, Hong Kong, Taiwan etc. The chairmanship is rotated and all decisions are collectively made. Another type of association is the Unicode Consortium, a voluntary grouping of representatives of major hardware and software industry that coordinates and leads world-wide efforts on the emerging Unicode standard. INFITT is an international body created to meet the felt needs of the Tamil Information Technology community.

Article 1

Introduction

The Tamil speaking people around the world, presently of about 80 million, strongly feel that the preservation and development of their culture and language critically depends on their ability to come to terms and taking advantage of evolving Information Technology. The International Forum for Information Technology in Tamil (INFITT) is founded for the purpose of coordinating Tamil Information Technology efforts worldwide, furthering the growth of on-line Tamil content globally, and facilitating the development of new Information Technologies for the advancement of Tamil culture, language, education and skill development, especially through global information infrastructure.

The INFITT will provide a forum and a mechanism for coordinating such activities through conferences, workshops, symposia, meetings, working groups, commissioned research and studies and other means, legal and appropriate.

Article 2

Status

1. The INFITT is established with the support of governmental and non-governmental organizations, agencies and institutions, the private sector, foundations, research and education institutions, industry and concerned individuals all focused on the promotion and development of Tamil culture and language through Information Technology.

2. The INFITT shall operate as a non-profit non-governmental autonomous organization, international in status, and non-political in management, staffing and operations. The INFITT shall be organized for coordination, research, development and education.

3. INFITT does not have a regulatory or enforcement role. Its primary functions are advisorial, promotional, facilitative and liaison with bodies and individuals concerned with the promotion of Tamil Information Technology

4. The collective decisions of the INFITT shall be in the form of recommendations to multinational agencies, nations, states, economies, organizations, institutions and individuals.

Article 3

Aims

1. The mandate of the INFITT is to promote Tamil culture and language through Information Technology,

computing, Tamil Internet, Tamil electronic content by the coordination of their regional, national, international and individual efforts and resources worldwide.

2. The aims and key objectives of INFITT are,

1. To organize and coordinate the efforts of its own policy and technical groups, various regional and national steering groups, independent groups (including Internet-based 3 organizations) industry and individuals and, within a global framework, facilitate, dialogue and promote cooperation and collaboration among various groups and individuals;

2. To identify key application areas for development of Tamil Information Technology, to define broad guidelines for their implementation and to provide technical assistance wherever possible;

3. To develop norms and standards for Tamil computing, including text and data handling, across different platforms, and the development of "open source-application programming interfaces (OS-API);

4. To promote education and the dissemination of Tamil Information Technology knowledge regarding Tamil computing;

5. To organize "Tamil Internet" ("Tamil Inaiyam") conferences regularly (preferably annually) in different parts of the world in cooperation with concerned national organizations.

6. To provide a mechanism for the Tamil Information Technology community to be represented collectively at international, regional and national Information Technology and networking organizations and their conferences or meetings, and to act as a liaison body and a voice for Tamil Information Technology in these bodies.

Article 4

Guiding Principles

1. The INFITT shall serve as an international catalyst, forum and resource devoted to developing competence and expertise in Tamil Information Technology.
2. The INFITT will complement its activities with those of other international and national societies, associations and institutions, industry and individuals, that have similar aims. Its activities will, wherever appropriate, be planned and implemented in collaboration with such societies, associations, institutions and individuals.
3. The INFITT will promote the standing of Tamil Information Technology in the global arena through liaison and cooperation with other international bodies.

Article 5

Activities

1. In fulfilling the aforementioned aims and functions, in the spirit of the guiding principles, the lNFITT shall engage in a range of activities including:

• holding meetings and arranging lectures, training courses, workshops seminars, symposia and conferences;

• commissioning the publishing and dissemination of books, periodicals, reports and research and working papers through print and electronic on-line and other media;

• establishing and maintaining contact with individuals and institutions with expertise in relevant fields through collaborative research, seminars, exchange visits, sabbatical attachments and likewise;

• commissioning studies and other projects on behalf of or in collaboration with other organizations and institutions;

• maintaining offices, information resources (including websites, databases, archives) and other facilities as may be necessary for its proper functioning;

• assist in development of international standards and norms in formats recognized by the lnformation Technology industry, such as Internet Drafts (ID) and Requests for Comments (RFC), in all key areas of Tamil Information Technology;

• and taking such other actions as may INFITT further and fulfill the aims and objectives

2. The INFITTs activities, programs and plans shall be reviewed periodically, taking into account the changing needs of developing and developed economies and the INFITT's capacities in meeting these needs.

Article 6

INFITT and its Organs

1. The INFITT is a corporate body, consisting of a General Body of members, having a common scale.
2. It can sue and be sued
3. The organs of INFITT are

i) The General Body,

ii) the General Council, and

iii) An Executive Committee including a Secretariat

Article 7

Members of General Body

1. Members of General Body may consist of

1. A state;

2. A regional body;

3. A national body;

4. Non-governmental organizations, agencies, institutions, the private sector, research and education institutions or industry;

5. Individuals.

2. Anyone who subscribes to the Aims of INFITT and abides by its rules and regulations can become a member of General Body.

3. General Council of INFITT shall periodically fix the membership fee of members in each of these categories. Continuation of membership would require prompt payment of membership fee.

4. No partisan group including a state, economy, country, non-governmental organization, agency, institution, company, industry, or organization shall have majority representation in the General Body.

5. Conduct of members. Members shall perform all duties for the INFITT as well as conduct their own professional activities in an ethical and professional manner. The INFITT may recommend disciplinary action for conduct of any member inconsistent with the purposes of the INFITT.

6. Resignation. Any member may resign at any time. Such resignation shall be made in writing and shall take effect at the time specified therein, or, if no time is specified, at the time of receipt by the Chairman or Secretary.

7. Removal. Any member can be removed by the Executive Committee subject to approval by the General Council.

8. Each member shall have voting rights as per one.

Article 8

INFITT General Council

1.1 The membership of the General Council shall consist of 51 members constituted by geographical, institutional and individual representation.

1.2 The geographical representation will be as follows:
India- 16
Sri Lanka – 06

North America – 08
Malaysia – 04
Singapore – 03
Europe – 05
Australasia – 01
Middle East & Africa – 02

1.3 The number of Institutional Members and Members-at-large in the General Council will be 6 and they will be elected by the General Body.

1.4 Wherever there is a national or a regional steering committee, its nomination to the General Council shall be limited to 50% of that geographical unit.

1.5 The rest of the geographical representatives will be elected by their respective geographical members of the General Body.

1.6 The duration of membership to the General Council shall be limited to two years

1.7 The General Council will elect the executive committee from its membership

1.8 The General Council will review and accord its approval wherever necessary the decisions of the executive committee.

2. The General Council may appoint sub or ad-hoc committees or working groups composed of appropriate technical experts to address specific topics of Tamil Information Technology as it may deem necessary for the performance of its functions. These committees/working groups will function under the direction of the Executive Committee.

3. The General Council will work largely as an Internet working group, except when it meets in annual "Tamil Internet" conferences, or through a mailing list (a closed list open only to members).

4. The Members of the General Council shall meet at least once a year in person, preferably during the "Tamil Internet" Conference. At this Meeting, amongst other things, the General Council shall elect or renew the office-bearers of the Executive Committee in accordance with the rules laid down in article 9.

Article 9

Executive Committee

1.1 The Executive Committee shall consist of 9 members, the Chair, the Vice-Chair, the Secretary, the immediate Past Chairman and three other Member all being members of the General Council.

1.2 The Executive Director of the Secretariat shall serve as the Secretary of the Executive Committee.

1.3 Except for the Past Chairman and the Secretary all other offices and members of the Executive Committee will be elected by the General Council from amongst its members.

2.1 The members of the Executive Committee will hold office for two years.

2.2 The Chair and Vice-Chair shall hold office for one year. In general, the Vice-Chair is Chair-elect and succeeds to the Chair the following year.

2.3 The election of the Chair and Vice-Chair may be by direct ballot or through rotation.
2.4 Tamil Nadu shall have a permanent position in the Executive Committee.

3. Executive Committee Members, upon assumption of their post, shall serve in their personal capacity and are not considered, nor do they act, as official representatives of their source states, countries, economies, institution, community or organization.

4. To this end, the Executive Committee shall:

1. define objectives and approve plans to meet INFITT aims and monitor the achievement of these aims;

2. formulate policies to be followed by the Secretary in pursuing the specified objectives;

3. ensure INFITT's cost-effectiveness, financial integrity, and accountability;

4. approve INFITT's program and budget;

5. appoint an external auditor and approve an annual audit plan;

6. approve INFITT's broad organizational framework and that of the Secretariat;

7. approve the INFITTs fund raising and resource mobilization strategies, policies and programs, and promote such fund raising and resource mobilization activities;

8. perform all acts which may be considered necessary, suitable and proper for the attainment of any or all of the aims of the INFITT as set forth in the articles herein. In special cases or routine matters, Executive Committee may meet informally via teleconferencing or on-line asynchronous means, and take decisions.

9. The Chairman of the Executive Committee shall preside over a meeting of the General Council.

Article 10

INFITT Secretariat

1. The Executive Director shall head the Secretariat and shall report to the Chairman of the Executive Committee and through him to the Executive Committee

2. The Executive Director shall be responsible to the Executive Committee for coordinating the operation and management of the INFITT and for ensuring that its programs and objectives are properly developed and carried out. The Secretary shall work closely with

the financial committees of the INFITT to coordinate fund raising and resource mobilization activities.

3. The Executive Director shall implement the policies determined by the Executive Committee.

4. The Executive Director shall be the legal representative of the INFITT.

Article 11

Capacity

The INFITT shall have the following capacity:
1. to receive, acquire or otherwise lawfully obtain from any governmental authority or from any corporation, company, association, person, firm, foundation, other entity or individual, whether international, regional or national, such charters, licenses, rights, concessions or similar assistance - financial or otherwise - as are conducive to and necessary for the attainment of the aims of the INFITT without compromising its neutrality and non-partisan, non-governmental and non-profit role in Tamil Information Technology worldwide.
2. to receive, acquire or otherwise lawfully obtain from any governmental authority or from any corporation, company, association, person, firm, foundation or other entity, whether international, regional or national, by donation, grant, exchange, device, bequest, purchase or lease, either absolutely or in trust, contributions consisting of such properties, real, personal, or mixed including funds and valuable effects or items as may be useful or necessary to pursue the aims and activities of the INFITT and to hold, operate, administer, use, invest, sell, convoy or dispose of the said properties in accordance to the principles laid down elsewhere in this document;
3. to enter into agreements and contracts;
4. to employ persons according to its own regulations;
5. to institute, and defend in, local proceedings; and
6. to perform all acts and functions as may be found necessary, expedient, suitable or proper for the furtherance, accomplishment or attainment of any and/or all of the purposes and activities herein stated, or which shall appear, at any time, as conducive to or necessary and useful for the aims and activities of the INFITT.

Article 12

Financing

1. The budget of the INFITT shall be funded by signatories of the Establishment Agreement for INFITT, international, regional and national organizations, public and private institutions and individuals which wish to make financial and other voluntary contributions to it. The INFITT may receive contributions from other sources. It may also receive contributions and gifts toward the establishment of an endowment program.

2. The financial operation of the INFITT shall be governed by financial regulations to be developed by the Secretariat and approved by the Executive Committee.
2.1 The budget for the INFITT will be approved, annually, by the Executive Committee.

2.2 The budget shall consist of two components: a "core" budget covering activities and resource requirements which are central and critical to the operational effectiveness and sustainability of INFITT, and a "special " budget which is fully supported by voluntary funds and contributions from donors who support specific initiatives and activities of the INFITT.

2.3 Special initiatives may include named programs, activities or grants provided by donors.

3. The "core" budget shall be supported primarily by fees collected from individuals, institutions and from obligatory contributions provided by signatory economies which shall be calculated in direct proportion to total GDP of each economy/state/country, except that such contribution from any one economy shall be capped at no more than 25% of the core budget.

4. An annual audit of the operations of the INFITT shall be conducted by an independent international accounting firm or by the appointment of Honorary Auditors appointed by the INFITT on the recommendations of the Secretary. The results of such audits shall be made available by the Secretary to the Executive Committee. The audit report with the comments of the Executive Committee shall be circulated to all members of the INFITT.

Article 13

Transparency

1. The deliberations of INFITT shall be conducted in a totally transparent manner.
2.1. The INFITT and its subordinate entities shall operate to the maximum extent feasible in an open and transparent manner consistent with procedures designed to ensure fairness.
2.2. The INFITT shall maintain one or more World Wide Web sites.
3. INFITT shall constantly search for additional transparency policies and transparency procedures designed to provide information about, and enhance the ability of interested persons to provide inputs to the INFITT and subordinate entities. Any such additional transparency policies and procedures shall be widely publicized by the INFITT in draft form, both within the INFITT membership and on a publicly-accessible Internet World Wide Web site maintained by the INFITT. Any such additional transparency policies and procedures may be adopted only after a process for receiving and evaluating comments and suggestions has been established by the INFITT Executive Committee, and after due consideration of any comments or suggestions received by the INFITT.

Article 14

Relationships with Other Organizations

1. In order to achieve its objectives in the most efficient way, the INFITT may enter into agreements for cooperation with relevant national, regional or international organizations, foundations and agencies, both public and private, and with individuals.

Article 15

Amendments

I. This Constitution may be amended by a two-thirds majority of all voting Members of the General Body, provided notice of such a proposed amendment together with its full text shall have been mailed to all members at least eight weeks in advance of the ballot.
2. The Quorum required for constitutional amendments shall be 50% of the General Body membership.
3. Voting may be carried out by electronic means or as designated according to procedures instituted by the Executive Committee.

Article 16

Dissolution

1. The INFITT may be dissolved by a three-fourths majority of all voting Members, if it is determined that the purposes of the INFITT have been achieved to a satisfactory degree or if it is determined that the INFITT can or will no longer be able to function effectively.

2. The Quorum required for dissolution of INFITT shall be 50% of the General Body membership.

3. In the case of dissolution, any land, physical plant and other assets situated in participating economies, and made available to the INFITT, and permanent fixed capital improvements thereon shall revert to their legal owner. The other assets of the INFITT shall be transferred for use for similar purposes or distributed to institutions having purposes similar to those of the INFITT in the participating economies.
3.1. The decision of the Executive Committee, in all these matters, will be final.

Article 17

Transitional Arrangements

1 The participants at the "Tamil Internet conference, convened to approve the Establishment Agreement, shall elect a pro-tem chairman .

2 The INFITT shall come into being with the approval of the constitution by the participants at a special meeting held during the Tamil Internet2000 conference in Singapore.

3. The first General Council shall be constituted in consultation with the participants of that meeting.

4. The first Executive Committee will be established by the first General Council.

5. The first Executive Committee will decide on the location of the Secretariat and the appointment of Executive Director.

6. Within one year, the first General Council will take necessary steps to establish the General Body

Developing Concept Maps to Aid Pre-collegiate Learning and Assessment in Tamil

Mahalakshmi G.S.[1] and Sendhilkumar S [2]
[1]Department of Computer Science and Engineering,
[2]Department of Information Science and Technology,
Anna University, Chennai 600025, India
mahalakshmi@cs.annauniv.edu, thamaraikumar@cs.annauniv.edu

Abstract: With the rapid development of computer information technology and network technology, the current teaching models are being dramatically changed. Concept map and collaborative learning in teaching are paid more and more attention by teachers and students. Concept Map is the visual expression of the concepts and the relationships between in a certain domain. It can not only integrate tacit knowledge, but also cluster explicit knowledge. Therefore, it is important to research how to construct concept maps. However, all these progress are internationally acclaimed since they were developed to suit for languages which drew wide attention, say English. And none of them were brought forward to Indian curriculum of education, even a thought process at the research level. These motivations led us to look for development of concept maps for Tamil medium students which can be prestigiously stated as a pioneering research in the discipline of school education in India.

1. Introduction

With the wide spread applications of information technology to education at all levels, we need to focus on learner-centric knowledge management in order to complement the conventional learning system. To enhance self-learning and assessment especially in learning subjects that require high knowledge retention, application of concept maps to promote education has been a proven success worldwide.

1.1 Academic Significance

The first difficulty someone who attempts to comprehend a text faces is to understand what it is all about. That is, to grasp the global sense of the communication, understand its elements and the relationships among them. In this context, the student may understand some of the concepts involved in the definition. These concepts are linked by words forming whole sentences that seem to make sense. However, trying to understand the overall conceptual structure is more difficult. It is

probably easier for many students to grasp a whole sense of the concept frame of reference when faced with a graph like the structure. This is due to the powerful visual effect that a graph has in order to facilitate understanding of a concept or a conceptual structure.

The interaction with the school teachers revealed the decay of conceptual methodology of teaching in the present educational curriculum, which was slowly taken over by the 'memorizing' mentality of the students due to societal and household anxiety. One can remember that the ancient methodology of education was not this; rather education should harness the young minds and inculcate researching mentality for anything they get introduced to, in their schools, with the aim of knowing further, and not just memorize and sit for exams to score high marks. Therefore, new instructional methods and techniques to increase retention are the need of the hour in pre-collegiate education in India. Methods improving students' perception of the field of study do have an impact in the assessment of their understanding of respective subjects.

1.2 Societal and Educational Factors

The use of concept maps in education provides an opportunity to move teaching and learning from memorization and repetition to reflection and critical thinking. However, Indian curriculum for higher secondary and high school education though have included more revolutionary methodologies for teaching, lack in efficient methods for self learning and assessment. In addition, societal awareness of students who learn the subjects through regional languages, especially in rural regions, is definitely few steps behind those who learn the same subject via English.

✷Recently activity based learning has emerged as an innovative discipline for primary education and the same is successfully implemented in primary schools under Government of Tamilnadu.

✷In addition, the chief minister of Tamilnadu has announced that steps will be taken for converting engineering education in Tamil from the forthcoming academic year.

1.3 Contribution to Knowledge

1.Concept Maps **help to improve understanding of a given subjec**t and facilitate building student's own knowledge, as long as the student has the opportunity to use, criticize, analyze, question or improve expert's maps or Concept Maps generated by his own peers.
2.The implementation of concept maps in the classroom allows both the teacher and the student **discovering and describing meaningful relations among the concepts** object matter of the study, making it possible to **create connections** between them and the context in which activities are developed.
3.The concept map helps the learners to have a **better overview of the course** and what aspect he/she should pay attention.
4.Concept maps constructed are very useful for teachers as an **evidence** of the way as

each one of the parties involved in the process assumes his/her own learning. From their **follow-up and analysis**, experiences can be designed to help their students overcome weaknesses or to reinforce strengths acquired in learning process.

This motivated us to apply technological advancement and research in contributing better methodologies for education with a special focus on students whose medium of instruction is any regional language (for our study: Tamil) other than English. Although concept maps have been proven to be a successful resource for grade improvement in abroad, their use is little explored in Indian education. The detailed literature analysis conducted for the same revealed the fact that almost no work reported the development and use of concept maps to promote education in regional language – Tamil.

In this context this paper concentrates on development of concept maps that eliminates the need for memorization and helps the students with active participation, to learn the subjects in their respective regional languages.

2. Objective

With the objective of applying ICT in Regional Language Education, in this paper we express the methodology of developing concept maps for various subjects at the higher secondary/pre-collegiate level. The objectives of the proposed work for developing concept maps is

1.to enhance pre-collegiate teaching and learning with an eye on self-learning and understanding.

2.to uplift the quality of rural education, both teaching and learning, by applying technological innovation

3.to aid in conceptual understanding of descriptive theory subjects

The idea is as follows: Natural Language Processing (NLP) techniques have been successfully used to automatically extract concept words from text through a detailed analysis of their content. Later, sentence level analysis is done to construct concept maps at the primary level. The primary concept maps are later merged to form the matured level of concept maps. Concept maps thus constructed shall be visualized if needed, via visualization tools.

3. Related Work

In the last decade, the electronic learning became a very useful tool in the students' education from different activity domains. The accomplished studies indicate that the students substantially appreciate the e-learning method, due to personalized instruction, informational content standardization, real time access to qualitative information resources and friendly interfaces. They don't consider it as a replacement of the traditional learning.

As it is known, an essential aspect in the learning process (either electronic or traditional) is the possibilities to evaluate the students. It is very important both for professor and student to test the understanding degree of the course. One of the best possibilities is to ask questions from the studied course. It is tested this way the degree of understanding of each studied material and the integration of new knowledge with the previous ones (that should already be known). These facts will have as a result an in-depth understanding of the learning materials.

Here we discuss the study and research done in connection with the proposed topic, by various experts outside India.

3.1 Question Generation for Learning Evaluation

Taking into consideration the high number of learning material existing in electronic format, the importance of the testing and evaluation systems has increased. Most of these systems use tests that were generated by teachers that permit a good evaluation and pursuance of the student evolution. In the last years, new preoccupations appear for automatic question generation. It's a subclass of Natural Language Generation (NLG) that is very important in a series of areas as: learning environment, data mining or information extraction. For example in [Andrenucci & Sneiders, 2005] it is introduced a template based approach to generate questions on four types of entities. It is considered that his approach failed in producing questions that can enhance the students' knowledge level.

The authors [McGough et. Al., 2008] present an interesting solution to the problem of presenting students with dynamically generated browser-based exams with significant engineering mathematics content. They introduce WTML (Web Testing Markup Language), which is an extension of HTML. A very interesting approach is found in [Wang et. Al., 2008]. Here, the main idea is to generate the questions automatically based on question templates which are created by training on many medical articles. This idea has advantages (easiness in building medical learning system, no additional work to build the question database or grading), but also disadvantages: the generated questions are factual and maybe less meaningful than the manual questions, time consuming to parse the articles and obtain the semantic interpretation, missing some important information.

Taking into account the advantages and disadvantages of the presented solutions, [Liana Stanescu et. Al., 2008] tried to design and implement a software instrument (Test Creator) that permits generation of questions based on electronic materials that students have. The solution implies teachers to have a series of tags and templates that they have to manage. These tags can be used to generate questions automatically.

3.2 Concept Maps applied for Question Generation

Concept maps are a result of Novak and Gowin's (1984) research into human learning and knowledge construction. Novak (1977) proposed that the primary elements of knowledge are concepts and relationships between concepts are propositions. Novak (1998)

defined concepts as "perceived regularities in events or objects, or records of events or objects, designated by a label." Propositions consist of two or more concept labels connected by a linking relationship that forms a semantic unit.

Concept maps are a graphical two-dimensional display of concepts (usually represented within boxes or circles), connected by directed arcs encoding brief relationships (linking phrases) between pairs of concepts forming propositions. The simplest concept map consists of two nodes connected by an arc representing a simple sentence such as 'flower is red,' but they can also become quite intricate.

One of the powerful uses of concept maps is not only as a learning tool but also as an evaluation tool, thus encouraging students to use meaningful-mode learning patterns. Concept mapping may be used as a tool for understanding, collaborating, validating, and integrating curriculum content that is designed to develop specific competencies. Concept mapping, a tool originally developed to facilitate student learning by organizing key and supporting concepts into visual frameworks, can also facilitate communication among faculty and administrators about curricular structures, complex cognitive frameworks, and competency-based learning outcomes. To validate the relationships among the competencies articulated by specialized accrediting agencies, certification boards, and professional associations, faculty may find the concept mapping tool beneficial in illustrating relationships among, approaches to, and compliance with competencies.

However, the only issue is that the learner must choose to learn meaningfully. The one condition over which the teacher or mentor has only indirect control is the motivation of students to choose to learn by attempting to incorporate new meanings into their prior knowledge, rather than simply memorizing concept definitions or propositional statements or computational procedures. The indirect control over this choice is primarily in instructional strategies used and the evaluation strategies used. Instructional strategies that emphasize relating new knowledge to the learner's existing knowledge foster meaningful learning. Evaluation strategies that encourage learners to relate ideas they possess with new ideas also encourage meaningful learning.

3.3 Concept Maps in E-learning

Recent researches have demonstrated the importance of concept map and its versatile applications especially in e-Learning. For example, while designing adaptive learning materials, designers need to refer to the concept map of a subject domain. Moreover, concept maps can show the whole picture and core knowledge about a subject domain. Research from literature also suggests that graphical representation of domain knowledge can reduce the problems of information overload and learning disorientation for learners. However, construction of concept maps typically relied upon domain experts in the past; it is a time consuming and high cost task.

Concept maps creation for emerging new domains such as e-Learning is even more challenging due to its ongoing development nature. The aim of Chen et. Al. [2006] is to construct e-Learning domain concept maps from academic articles. They adopt some relevant journal articles and conference papers in e-Learning domain as data sources, and

apply text-mining techniques to automatically construct concept maps for e-Learning domain. The constructed concept maps can provide a useful reference for researchers, who are new to the e-Leaning field, to study related issues, for teachers to design adaptive learning materials, and for learners to understand the whole picture of e-Learning domain knowledge.

3.4 Concept Map Mining

There is yet another approach [Villalon and Calvo, 2009] for automatic concept extraction, using grammatical parsers and Latent Semantic Analysis. Essays, as any other text, represent both the knowledge and the writing skills of its author, hence, an Automatic Concept Map from Essay (ACME) should reflect both. Therefore, the words for the concepts and relations must be extracted literally from the document, and the hierarchy of concepts must reflect the importance of the concepts relative to what was written in the particular document. However, the performance is related to the way concepts are chosen by humans. We believe that understanding this phenomenon and using it for the automatic selection of concepts could lead to big improvements.

3.5 Patents

✷US Patent 5506937 - Concept map based multimedia computer system for facilitating user understanding of a domain of knowledge [12]
✷A computer system having an explanation facility for facilitating user understanding of concepts underlying a domain of knowledge which enables a user to interact with and explore the domain of knowledge. The explanation facility utilizes a concept-map based representation of a domain of knowledge and several icons to control the mode of output of information from the computer system. Each concept map has concept nodes which represent concepts in the domain of knowledge, links between the concept nodes, and icons. The icons are positioned at the concept nodes and represent alternative modes of output of information from the computer system. A user desiring more information about a concept node can select one of the icons corresponding to the mode of output of information desired. Modes of output of information include audio, video (images and movies), text, concept maps, and combinations of the foregoing. Through the use of concept maps and icons that control modes of output of information, a user may navigate the domain of knowledge and retrieve information specific to the user's particular needs.

4. Methodology for Automatic Concept Map construction

Natural Language Processing (NLP) techniques shall be successfully used to automatically extract concept words from ‘text’ through a detailed analysis of their content. Later, sentence level analysis is done to construct concept maps at the primary level. The primary concept maps are later merged to form the matured level of concept maps. Concept maps thus constructed shall be visualized if needed, via visualization tools.

Initially the descriptive passages in Tamil are parsed using Vaanavil Tamil Parser [Saravanan et al, 2004], a tool for parsing Tamil sentences. Later the parsed output is analysed with Atcharam [Anandan et al, 2001] a tool for morphological analysis for identification of nouns and root words of verbs. In addition, the adverbs and adjectives are also extracted. The final set of concept words emerging out are made to undergo a filtration process where unwanted words are filtered out heuristically.

The outcome of the analysis is embedded as a dependency graph with concept words as entries across rows and columns, and look-up with relations between the concepts listed. The generated concept maps (Fig 2) shall be visualized at the output.

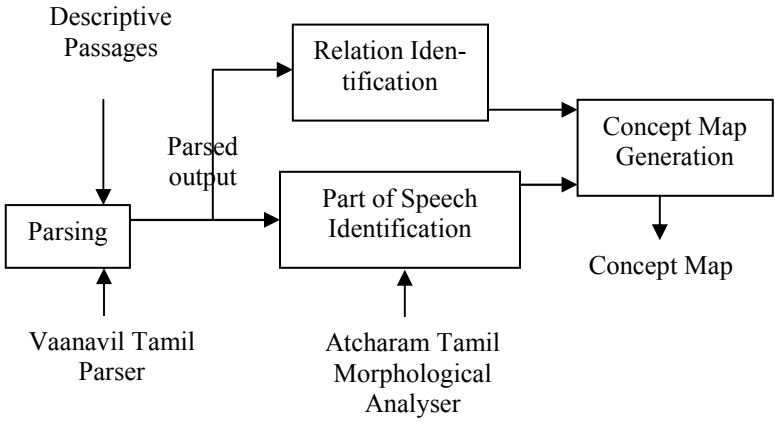


Figure 1. Block Diagram of Automated Concept Map Development

Input:

நான் ஒரு அழகான மயிலை பார்த்தேன். அது தன் தோகையை விரித்து ஆடியது. அதை நான் என் சிறிய கைப்பேசியில் புகைப்படம் எடுத்தேன். அதைப் பார்த்து என் நண்பர்கள் ஆச்சரியப்பட்டனர்.

Output:

Concept-words:

concept-word 0 நான் -N concept-word 1 அழகான -rp
concept-word 2 மயிலை -N
concept-word 3 பார்த்தேன் -V concept-word 4 தோகையை -N
concept-word 5 விரித்து -K
concept-word 6 ஆடியது -V concept-word 7 நான் -N
concept-word 8 சிறிய -V
concept-word 9 கைப்பேசியில் -N concept-word 10 புகைப்படம் -N
concept-word 11 எடுத்தேன் -V concept-word 12 அதைப் -V
concept-word 13 நண்பர்கள் -Nconcept-word 14 பார்த்து -Ad
concept-word 15 ஆச்சரியப்பட்டனர் -V

Figure 2. Screenshot – Concept Words at the final output

5. Results and Discussions

The concept words extraction across the passages has been performed across three different approaches. In Approach1 the nouns and verbs alone are extracted from the passage. In Approach2 the other key grammar forms like adverbs, adjective, relative pronouns are included. In Approach3 the extracted concept words from appraoch2 are being filtered heuristically for optimum results. For testing the effectiveness of these three different approaches we have proposed, we use the Concept Word Factor (CWF) for each passage. The original set of words along with the extracted set of words is presented to experts and their opinions are noted. We have proposed a method to calculate CWF using the equation,

Concept-word factor,

CWF = ((Original words – concept words)/Original words) * expert opinion

original words = no. of words in the passage

concept-words = no. of words generated from answer evaluation system

expert opinion = effectiveness of the concept-words in a scale of 1 – 5

We took a sample set of twenty questions from the Tamilnadu Government 9th std state board tamil book and obtained the extracted concept-words and the expert opinion for each of them. From the obtained results, we plotted the following graph.

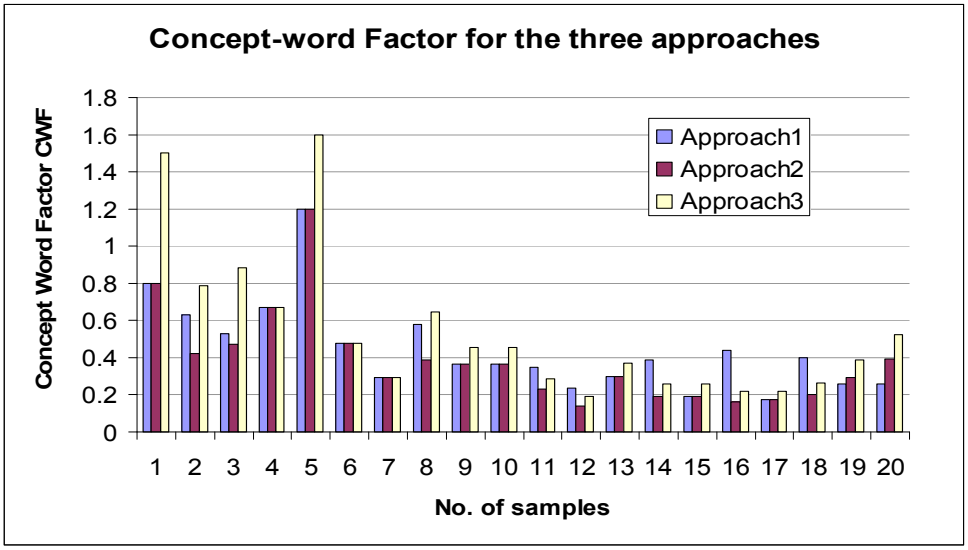


Figure 3.Concept-Word Factor across the three approaches

From Fig.3 we can infer that by large, the final approach3 (filtered approach) is more effective in capturing the required concept words in a passage. But since we use heuristics based approach for filtering, for certain cases the final approach3 is less effective than the basic approach1.

6. Conclusion

If teaching-learning educational process is considered as a goal through which students can get a meaningful learning of stated concepts, which extend and articulate their network of relations and can apply them in different contexts, it is necessary that teachers include tools to speed up act performance of agents involved in the construction of the new knowledge. In our case, applying a concept map tool in the classroom will allow students being themselves more motivated to carry out proposed activities and to participate in the construction of their own knowledge.

The methodology of developing concept maps discussed in the paper shall be (i) *Extended to impart concept map based* learning in other regional languages; (ii) Applied to generate associated concept animations for enriching automated content development; (iii) Applied to automated answer evaluation thereby taking part in self-assessment activity of pre-collegiate examinations; (iv) Used to dynamically generate questions and further continue the answer evaluation process in an e-learning setting; and (v) Applied for automatic document summarization

References

1. Anandan, P., Ranjani Parthasarathy & Geetha, T.V., 2001. "Morphological Analyser for Tamil", *ICON 2002*, RCILTS-Tamil, Anna University, India.

2. Andrenucci A., .Sneiders, E., "Automated Question Answering: Review of the Main Approaches", in *Proceedings of the 3rd International Conference on Information Technology and Applications (ICITA'05)*, July 4-7, Sydney, Australia, IEEE, Vol. 1, 2005, pp.514-519.

3. Jorge Villalon, and Rafael A. Calvo, "Concept Extraction from student essays, towards Concept Map Mining", Ninth IEEE International Conference on Advanced Learning Technologies, 2009. pp.221-225

4. Liana Stanescu, Cosmin Stoica Spahiu, Anca Ion, Andrei Spahiu, "Question generation for learning evaluation", Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 509 – 513, 2008 IEEE

5. McGough J., Mortensen J., Johnson J., Fadali S., "A web-based testing system with dynamic question generation". *LNCS 1611-3349*, 2008, pp. 242-251.

6. Nian-Shing Chen a, Kinshuk b, Chun-Wang Wei a, and Hong-Jhe Chen, "Mining e-Learning domain concept map from academic articles", Computers & Education, Vol. 50 (2008) 1009–1021, Elsevier

7. Novak, JD. A theory of education. Ithaca, NY: Cornell University Press, 1977.

8. Novak J., Learning, creating and using knowledge. Concept Maps™ as facilitative tools in schools and in corporations. London: Lawrence Erlbaum, 1998.

9. Novak J. and Gowin, Learning how to learn. New York and Cambridge, UK: Cambridge University Press, 1984.

10. R. Saravanan, Ranjani Parthasarathi and Geetha T.V., 'Vaanavil – Parser for Tamil', Resource Center for Indian Languages – Tamil, Dept. of Computer Science and Engineering, Anna University Chennai, India, 2004.

11. Wang W., Tianyong H., Wenyin L., "Automatic Question Generation for Learning Evaluation in Medicin", in *LNCS Volume 4823*, 2008, pp. 242-251.

Tamil Transliteration Standard for Pan-Indic Scripts

Vinodh Rajan S
vinodh@virtualvinodh.com

1.1 Introduction

Transliteration is a method of representing the text of a writing system in another writing system. It does not reflect how the alphabets are pronounced in the Source Script, but only replicates the way the text is written in the source script. Ideally in Transliteration, the source spelling is preserved in the target script, by replacing the characters in a one-to-one manner. However, in Transcription, the text is rendered in the target script by reproducing the text based on how it is pronounced not how it is written.

Rendering *செய்தாய்* as *ceytāy* is Transliteration, where as *seydāy* is Transcription.

1.2 Indic & Tamil Scripts

All Modern Indic Scripts are ultimately derived from *Brahmi*, the script of the Asokan inscriptions. Brahmi was used all over the Indian sub continent to write the regional languages. At the Initial stages virtually, all languages were written employing this single script. However, as time passed regional variations were developed and at one point of Time, each variant of the Brahmi evolved into different regional scripts giving birth to the modern Indic scripts.

Unlike other Indian languages, Tamil adapted the Brahmi Script albeit in a modified form, to suit its grammar. Brahmi was adapted as a phonemic script. Each letter denoted a phoneme. A pho-

neme can have different allophones based on the consonants that precede or follow it. So, a letter can be pronounced in several ways based on its position in a word.

(Tamil à IPA Transcription available at http://www.virtualvinodh.com/tamil-ipa)

Here a single phoneme */க/ - /k/* denotes several allophones *[k] [g] [x]* depending upon the context. Similarly other consonants also have several allophones.
Therefore, Tamil script does not have separate letters for voiced consonants such as *ga, ba, da etc*. Tamil also lacks aspirated consonants such as *kha, gha, dha etc*, and Vocalic vowels ऋ ɾ ॠ ɾː ऌ ɭ/ ॡ ɭː since they do not appear in the native language.

Tamil Word	Transliteration	IPA Transcription
கால்	*k*āl	*k*ɑ:l
வங்கம்	vaṅ*k*am	ʋʌŋ*g*ʌm
பகல்	Pa*k*al	pʌ*x*ʌl

1.3 Transliteration Standards

Many Transliteration standards such as IAST, ISO 5919 exist to standardize the way in which Indic Scripts are represented using Latin characters. IAST primary deals with Sanskrit, and ISO with all the Indic Languages. ISCII prescribes a standard that enables Devanagari script to represent the letters of all the Indian languages including Tamil. In most of the cases, special diacritic marks are introduced to extend the native character and enable it to represent a foreign character set.

Tamil Script: பன்னூறாயிரம் விதத்தில் பொலியும் அவலோகிதன் மெய்த்தமிழ்

ISO 5919: paṉṉūṟāyiram vitattil poliyum avalōkitaṉ meyttamiḻ

ISCII Devanagari: पन्नूराियरम् िवतत्तिल् पॊलियुम् अवलोकितन् मॆय्त्तमिऴ्

But such a standard is absent for Tamil. At present the Tamil Script is inept of expressing letters existing in other Indic Languages. But, at times, mostly in scholarly works, it is necessary to express other Indic Languages in Tamil script, with the originality of the source preserved. In English, IAST/ISO enables such lossless transliteration into Latin Script. However, such a situation is not possible for Tamil scholars, who often resort to a lossy approximate transliteration.

Hence it is highly essential to develop a Transliteration standard for the pan Indic orthographies. Standards such as these are necessary to perform a mutual lossless transliteration between Tamil and other pan-Indic scripts. This would greatly enable us to transliterate

Indic language documents, webpages, etc. into Tamil Script & vice versa. Scholars can represent other Indic languages in Tamil script without distorting the source. It also enables people to read, learn and represent other Indic languages completely in Tamil script itself.

1.4 Existing Conventions

Saurashtra and some Sanskrit publications have long since adopted the method of using superscript & subscript numerals to transliterate their text into Tamil script. In this method, numerals are used to denote the non-existent *Varga* consonants such *kha, ga, gha, etc.*

कामदेवाय विद्महे पुष्पबाणाय धीमहि

तन्ोऽङ्गः प्रचोदयात्

kāmadēvāya vidmahē puṣpabāṇāya dhīmahi

tannō'naṅgaḥ pracōdayāt

காமதே³வாய வித்³மஹே புஷ்பபா³ணாய தீ⁴மஹி

தந்நோ'நங்க³: ப்ரசோத³யாதே³த் (Superscript)

காமதே₃வாய வித்₃மஹே புஷ்பபா₃ணாய தீ₄மஹி

தந்நோ'நங்க₃: ப்ரசோத₃யாதே₃த் (Subscript)

This has a respectable level of acceptance among the readers. However there are many variations in representing the letter श ś*a*, *Anusvara* and *Vocalic letters*. In some cases, these letters are not uniquely represented at all. Hence, leading to a lossy transliteration. They also don't conform to the usage whether the superscript numerals or the subscript numerals are to be employed consistently. Each publication may follow its own convention of subscript/superscript numerals

1.5 Tamil Transliteration Standard

The present method of is mainly concerned with Sanskrit alone. However, the Transliteration standard for Tamil must enable to support all the nuisances of Indic Languages such as, Malayalam *cillākṣarās,* Sinhala *saññaka (Pre-Nasalized)* consonants. These characters must be enabled to be represented in the Tamil Script. Diacritic signs for many other letters such as vocalic vowels, *anusvāra, anunāsika,* must be standardized.

The proposed transliteration standard is given below:

Vowels											Dra.
ISO	a	ā	i	ī	u	ū	ṛ	ṝ	ḷ	ḹ	e
Devanagari*	अ	आ	इ	ई	उ	ऊ	ऋ	ॠ	ऌ	ॡ	ऐ
Tamil	அ	ஆ	இ	ஈ	உ	ஊ	ரு«	ரூ«	லு«	லூ«	எ
ISO	ē	ai	o	ō	au	aṁ	aṁ	aḥ			
Devanagari*	ए	ऐ	ओ	ओ	औ	अं	अँ	अः			
Tamil	ஏ	ஐ	ஒ	ஓ	ஔ	அம்°	அம்¿	அ:			

Vowels	Neo & Si.	Sinh.	Neo.
ISO	æ	ǣ	ô
Devanagari/ Sinhala	ऍ/ෑ	ෑ	ऑ
Tamil	எ¬	ஏ¬	ஆ¡

Consonants										
ISO	ka	kha	ga	gha	ṅa	ca	cha	ja	jha	ña
Devanagari*	क	ख	ग	घ	ङ	च	छ	ज	झ	ञ
Tamil	க	க^2	க^3	க^4	ங	ச	ச^2	ஜ	ஜ^2	ஞ
ISO	ṭa	ṭha	ḍa	ḍha	ṇa	ta	tha	da	dha	na
Devanagari*	ट	ठ	ड	ढ	ण	त	थ	द	ध	न
Tamil	ட	ட^2	ட^3	ட^4	ண	த	த^2	த^3	த^4	ந
ISO	pa	pha	ba	bha	ma	ya	ra	la	va	
Devanagari*	प	फ	ब	भ	म	य	र	ल	व	
Tamil	ப	ப^2	ப^3	ப^4	ம	ய	ர	ல	வ	
Devanagari*	ष	श	स	ह	ळ	ऱ	न	क़	य़	फ़
Tamil	ஷ	ஸ^2	ஸ	ஹ	ள	ற	ன	ழ	ய·	ஃப
ISO	ṅga	ñja	ṇḍa	ṅda	ṁba					
Sinhala	හ	ඡ	ඬ	ඳ	ඹ					
Tamil	ங்°க^3	ஞ்°ஜ	ண்°ட^3	ந்°த^3	ம்°ப^3					
ISO	qa	k̲ha	ġa	za	ṛa	ṛha				
Devanagari	क़	ख़	ग़	ज़	ड़	ढ़				
Tamil	ஃக	ஃக^2	ஃக^3	ஃஜ	ஃட^3	ஃட^4				

In addition to the above letters, Chillu letters in Malayalam are differentiated form their Virama forms using ¦. Superscript numerals have been employed consistently.

* Indicate equivalent characters in other Indic languages also.

1.6 Sample Transliterated Texts in Tamil

Devanagari

ॐ □मो भगवते अपरिमितायुर्ज्ञा□सुवि□श्चिततेजोराजाय तथागतायार्हते सम्यक्संबुद्धाय, तद्यथा, ॐ सर्वसंस्कार परिशुद्ध धर्मते गगणसमुद्गते स्वभावपरिशुद्धे महा□यपरिवारे स्वाहा

ஓம் ° நமோ ப⁴க³வதே அபரிமிதாயுர்ஜ்ஞாநஸு²விநிஶ்சிததேஜோராஜாய ததா²க³தாயார்ஹதே ஸம்யக்ஸம் °பு³த்³தா⁴ய, தத்³யதா², ஓம் ° ஸர்வஸம் °ஸ்கார பரிஶு²த்³த⁴ த⁴ர்மதே க³க³ணஸமுத்³க³தே ஸ்வபா⁴வபரிஶு²த்³தே⁴ மஹாநயபரிவாரே ஸ்வாஹா

Kannada

ಕಾವೇರಿ ಕರ್ನಾಟಕದ ಜೀವನದಿ. ಕೊಡಗು ಜಿಲ್ಲೆಯ ಪಶ್ಚಿಮ ಘಟ್ಟದಲ್ಲಿ ತಲಕಾವೇರಿಯೆಂಬ ಸ್ಥಳದಲ್ಲಿ ಉಗಮಿಸುವ ಈ ನದಿ, ಮೈಸೂರು ಜಿಲ್ಲೆಯ ಮೂಲಕ ತಮಿಳುನಾಡಿಗೆ ಹರಿದು ಬಂಗಾಳಕೊಲ್ಲಿಯನ್ನು ಸೇರುತ್ತದೆ.

காவேரி கர்நாடகத³ ஜீவநதி³. கொட³கு³ ஜில்லெய பஸ்²சிம க⁴ட்டத³ல்லி தலகாவேரியெம் °ப³ ஸ்த²ளத³ல்லி உக³மிஸுவ ஈ நதி³, மைஸூரு ஜில்லெய மூலக தமிளுநாடி³கெ³ ஹரிது³ ப³ம் °கா³ளகொல்லியந்து³ ஸேருத்ததெ³.

Malayalam

എല്ലാ ബുദ്ധന്മാരുടെയും കാരുണ്യത്തിന്റെ മൂ□ത്തിമദ്ഭാവമായ ബോധിസത്വമാണ് അവലോകിതേശ്വര□. ലോകേശ്വര□, ലോകനാ□ ഈശ്വര□ എന്നൊക്കെ അറിയപ്പെടുന്ന അഥവാ ബുദ്ധ□ ആണ് അവലോകിതേശ്വര□.

എല്ലാ பு³த்³த⁴ந்மாருடெயும் ° காருண்யத்திந்றெ மூர்¦த்திமத்³பா⁴வமய போ³தி⁴ஸத்வமாண் அவலோகிதேஸ்²வரந்¦. லோகேஸ்²வரந்¦, லோகநார்¦ ஈஸ்²வரந்¦ எந்நொக்கெ அறியப்பெடுந்ந அத²வா பு³த்³த⁴ந்¦ ஆண் அவலோகிதேஸ்²வரந்¦.

Sinhala

□□□ □□□ □□□□□ □□ □□□□ □□ □□□□ □□□□□□ □□□□□ □□□□□□ □□□ □□□□□□□ □□□□□ □□□□□□ □□□ □□□□□ □□□ □□ □□□□□□□□ □□□□□□□□ □□□□□□□□□□□ □□□□□□□ □□□□□ □□□□□□□□.

லொவ அந் ரடசல் ஹா ஸஸந் °த³ந கல ஶ்ரீ லம் °காவே விவித⁴ பெதெ³ஸ் துள தி³வியந் விஸா²ல வஸ²யெத் அந் ஸதுந் ஸமக³ தம ஸ்வபா⁴விக வாஸஸ்தா²ந கெ²தா³க³நிமிந் ஜீவத்வந ஆகாரய தெ¬³கியஹெ¬க.

1.6 Aksharamukha – Script Converter

Aksharamukha (http://www.virtualvinodh.com/aksharamukha), a PHP based web transliteration application that works based on the proposed standard has been developed. This application converts all the Indic orthographies and also several Roman Transliteration standards such as ISO 5919, IAST, Harvard-Kyoto, Velthuis etc. into lossless Tamil script & vice versa.

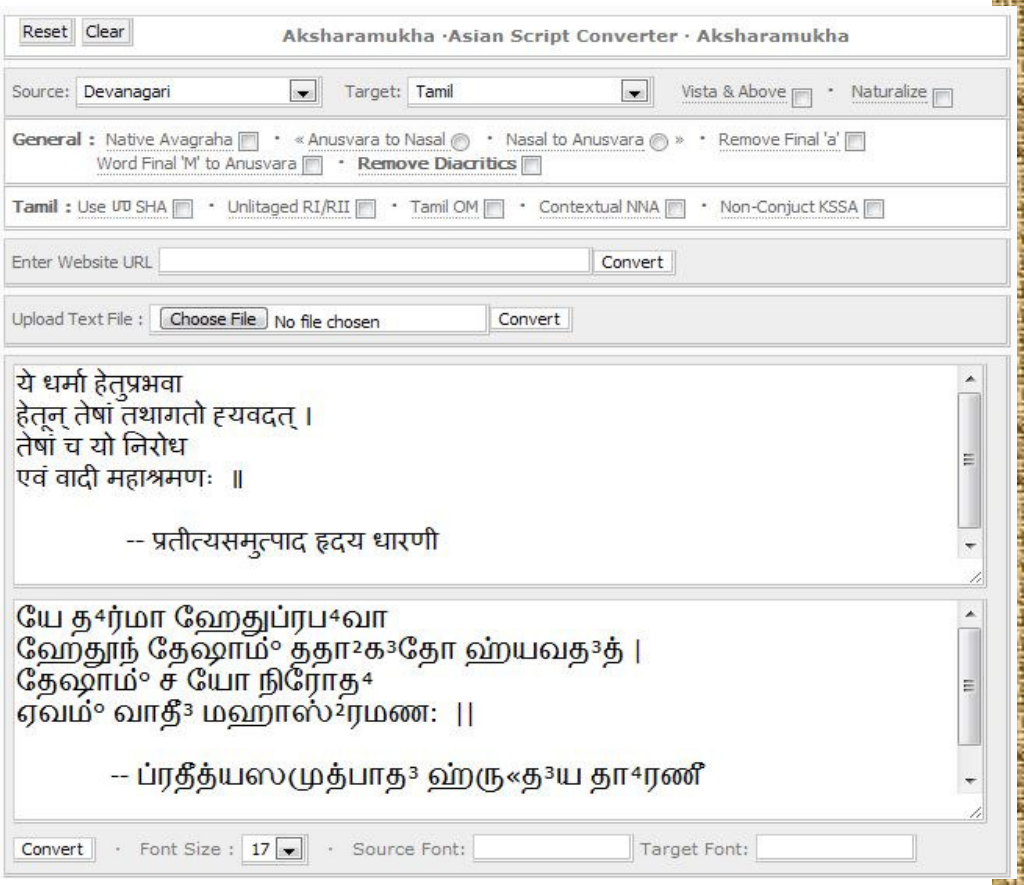
Other than transliteration, it supports *transcription* of Tamil into other scripts, by introducing voiced and unvoiced consonants, and other nuisances in the Target Script based on Tamil grammatical conventions.

It also enables *anusvāra* - *mellinam* replacement, and other additional features during transliteration/transcription such as *Tamil OM, Contextual usage of Tamil Letter NNNA, Unligated Tamil Consonants RI/RII etc.* Details instructions can be found at http://www.virtualvinodh.com/tamil . These options can be used to naturalize the source text in the Target script as far as possible.

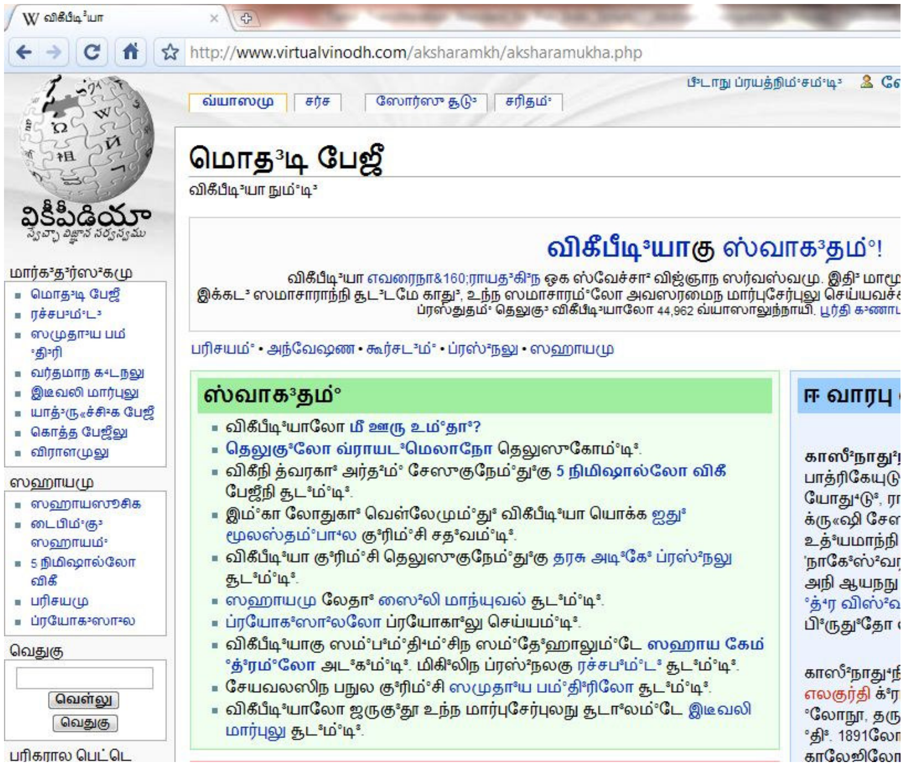
The converter also supports website Transliteration. It can be used to convert Website in other Indic languages into Tamil Script and also vice versa. It is highly useful for reading website, when the visitor has the knowledge of the Language but not the script.

The converter is open source project, whose details can be found at https://launchpad.net/aksharamukha

Screenshots of the Tool can be seen below.



Script Converter Page



Telugu Wikipedia in Tamil Transliteration (Website Transliteration)

Automated comic generation for tamil rhymes

Sendhilkumar S., Divya R., Ilakiya R. and Prashanth P.

Department of Information Science and Technology,
Anna University, Chennai 600025, India
thamaraikumar@cs.annauniv.edu

Abstract: Human communication is based on verbal and non-verbal behavior. Facial expressions convey more information or provide additional expressiveness for the dialogue delivered. Rendering of rhymes is no exception. Expressiveness takes a major role in making things interesting and fresh in the minds of young children. With the advent of information technology, many content development audio-visual diskettes are in the market which makes the rendering of tamil rhymes much easier. However, things might get bored if a child happens to see the same 'cake' for 'pat-a-cake' rendering.

We have attempted at developing a framework for automatic creation of cartoon/comic layouts for Tamil rhymes. The idea is as follows: An image corpus with annotated image / effects / descriptions is constructed. The annotations on images are not only textual but also pictorial. Using these annotations, one might understand the parts of the image which may undergo effects morphing at the comical layout generation.

Initially, the rhymes at the input are analysed by language processing tools for Tamil, and the keywords are inferred. The grouping of rhyme lines is also done at this stage by semantically analyzing the content. The extracted keywords are fed into the image search engine and the corresponding set of images is identified. The dependency between the resulted images is also identified with the help of parsed rhymes. The dependency graph thus constructed is utilized to generate comical layouts for dialogues, foreground, background and visual effects.

Initially foreground is decided by inferences from the input rhyme. Next to this, the effects layout is to be decided. i.e. 'Jo Jo Naaikutty, Thulli vaa Naaikutty' leads to identifying the image of 'Naaikutty' from the image corpus along with the zones of effects that can be superimposed over the image of 'Naaikutty', as indicated in the annotations. When foreground information is to be decided, the layout generator, marks the 'Naaikutty' image as foreground, and then superimposes the effects over the image so as to appear as if the little dog is jumping. This shall be achieved by making feather marks above and beneath the image. Later, the background layout is decided by associating the grouped lines of rhymes with the previous or following concepts in the rhymes.

This is an iterative process and once when it is done, the dialogues like the dog wooing or barking, shall be added at the dialogue layer.

References

[1] Léon J. M. Rothkrantz and Ania Wojdel, A Text Based Talking Face, a chapter in: Text, Speech and Dialogue, LNCS, Springer Berlin/Heidelberg, Vol. 1902/2000, pp. 213-240, 2000.

[2] Preethi Jayaram, Savitha A. and G.S. Mahalakshmi, Multimedia Rhymes, Indian Conference on Intelligent Systems, ICIS 2007, DMI college of Engg., Allied Publishers, Chennai, 2007.

Analysis of Touching characters in the Keywords of Tamil Document Images

Abirami.S and Vasantha Kumar. P
Dept of Information Science and Technology,
Anna University, Chennai -600 025.
abirami@annauniv.edu, abirami_mr@yahoo.com

Abstract: This paper attempts to identify, analyze and resolve the touching characters present in the keywords of Tamil imaged documents. Resolving of touching characters in the Tamil word images can improve the recall rate of retrieval while enabling search from Tamil document images.

1. Introduction

Document images are gaining popularity and importance in the recent decades, since a lot of efforts has been made to build digital libraries, which digitize high-volume archives of paper documents (patents, historical and business documents) to provide the public with a free and easy online access. As a matter of fact, many organizations currently practice business on document image archives. However, such archives are often poorly indexed which makes them unfit for IR purposes. In addition, text retrieval techniques cannot be applied directly over imaged documents. As a result, the user experiences a lot of difficulty in retrieving relevant information from imaged data and there arose a need for efficient document image retrieval tools.

Document Image Understanding (or interpretation) is the formal representation of the abstract relationships indicated by the two dimensional arrangement of symbols [9]. Researches have addressed two different techniques [4] to identify or understand the text from document images, namely, Optical Character Recognition (OCR) and Keyword Spotting technique. Optical Character Recognition (OCR) lies at the core of the discipline of pattern recognition. The objective is to develop computer algorithms to identify the characters of the alphabet. Optical Character Recognition deals with the machine recognition of characters present in an input image obtained using scanning operation [4]. It refers to the process by which scanned images are electronically processed and converted into an editable text (ASCII representation). Standard text retrieval techniques could be applied over the recognized text to retrieve information from the document images.

In contrast, Keyword Spotting approaches understand the text at the word level of document images. These approaches understands the image properties of text at word level and converts it into an intermediary representation instead of converting the entire document image into ASCII representations character by character. Since, information retrieval is concerned with the keyword, (i.e.) obtaining a query word from the user and retrieving the documents relevant to the user query, the Keyword Spotting technique has gained its popularity. By applying Keyword Spotting technique, relevant document images could be fetched by matching the word image representations of documents directly with the user query word.

Numerous keyword spotting techniques and IR systems have been reported to retrieve the information from Roman and Chinese document images [7] [8] and some of the Indian languages such as Hindi [3] [5] and Telugu [6][2] . However, feature extraction techniques discussed in the literature are specific and language dependent and cannot be applied to the Tamil since the shapes of the Tamil characters are complicated and varied. In addition, Tamil text recognition systems could not be utilized for information retrieval since it suffers inherent weaknesses, namely (1) Failure in the discrimination of a set of characters that closely resemble with others in the character set, (2) Restrictions in font faces and sizes [10] and (3) Post processing to correct the errors occurred during the recognition. Information retrieval would lead to a poor performance in these systems since spell check is essential after recognition.

These outstanding problems in Tamil text recognition systems motivated researches to the idea of developing LR-TB-FS technique [1], to retrieve information from Tamil document images. Here, the idea is based on the assumption that the technique devised would extract features of word images across various font faces and font sizes instead of training the shapes of the characters. Basic features are extracted by traversing through the vertical centroid area and horizontal zones of characters in word images to record their black and white disposition rates and similarity has been computed between the query word and the word images to retrieve documents.

LR-TB-FS technique [1] generates feature string for the keywords by making use of the following processes:

Feature String generation

Input	:	Word image object.
Output	:	Feature String.
Process Logic	:	

Primitive P with seven attributes is calculated.

- No of vertical lines are counted using vertical line identification.
- No of horizontal lines is calculated using horizontal line identification.
- Vertical transition rate is calculated.
- Horizontal transition rate is calculated in the Ascender, Middle and Descender Zones.
- Lower Outline of the primitive has been calculated.

2. Touching Character Analysis

In order to generate a feature string and to identify the keywords at the word level, characters which touch with one another should also be dealt in the LR-TB-FS technique and this requires Touching character analysis.

Apart from the presence of regular characters, there is a possibility of touching characters (characters touching each other and appear to be a single character in character segmentation) in the document images. Feature String generation for touching characters using the LR-TB-FS algorithm [1] would reduce the recall rate considerably in the retrieval process, since the primitives for the touching characters are different from those of the regular characters.

Therefore, in order to handle touching character, an additional heuristic has been proposed here during the identification of horizontal lines. Initially, pixel width of the vertical lines of characters has been determined based on the font size of the characters. While identifying the horizontal line, the Black Run Length of the entity in a horizontal orientation has been determined (Here, the Black run length corresponds to the total number of black pixels in a row arranged horizontally).

If the horizontal black run length of the entity exceeds the pixel width pertaining to the vertical line boundary (identified based on the font size) and arises with a vertical black run length (black pixel arrangement in vertical direction) in a perpendicular direction, a touching character exists in that position and the whole entity can be divided into isolated entities with respect to that position. Primitives (seven attributes) are identified for isolated entities using the LR-TB-FS algorithm.

3. Conclusion

LR-TB-FS technique which was proposed earlier to extract features of the Tamil word image generates Feature String without any explicit character conversion to improve the IR. Neither a post-processing algorithm after feature generation nor a spell check procedure has been involved in this technique and this achieves a significant performance in IR. In addition to this, touching character analysis in the word images through their horizontal black run length and the pixel width of the attribute boundaries can further resolve the touching character issues. The touching character analysis could improve the recall rate of the Tamil information retrieval process considerably.

References:

1.Abirami .S, Manjula.D (2009), “Feature string-based intelligent information retrieval from Tamil document images”, Int. J. Computer Applications in Technology, Vol. 35, Nos. 2/3/4, 2009, pp150-165.

2.Balasubramanian A. and Jawahar C.V. (2006), ‘Textual search in graphics stream of PDF’, International Conference on Digital Libraries. Pp.1-10.

3.Chaudhury S., Sethi G., Vyas A. and Harit G. (2003), 'Devising Interactive Access Techniques for Indian Language Document Images', Proceedings of the Seventh International Conference on Document Analysis and Recognition, pp.885-889.

4.Doermann D. (1998), 'Indexing and Retrieval of Document Images: A Survey', Journal of Computer Vision and Image Understanding, Vol. 70, No.3, pp.287-298.

5.Harit G., Jain R. and Chaudhury S. (2005a), 'Improved Geometric Feature Graph: A Script Independent Representation of Word Images for Compression and Retrieval', Proceedings of the Eighth International Conference on Document Analysis and Recognition, pp. 421-425.

6.Jawahar C.V., Meshesha M. and Balasubramanian A. (2004a), 'Searching in Document Images', Proceedings of the International conference on Visualization, Graphics and Image Processing, pp. 622-627.

7.Lu Y. and Tan C.L. (2002b), 'Word spotting in Chinese document images without layout analysis', Proceedings of the International Conference on Pattern Recognition, pp. 57-60.

8.Lu Y. and Tan C.L. (2004a), 'Information Retrieval in Document Image Databases', IEEE Transactions on Knowledge and Data Engineering, Vol.16, No.11, pp. 1398-1410.

9.Nagy G. and Seth S. (1984), 'Hierarchical representation of optically scanned documents', Proceedings of the International Conference on Pattern Recognition, pp. 347-349.

10. Seethalakshmi R., SreeRanjani T.R., Balachandar T., Abnikant Singh., Markandey S., Ritwaj R. and Sarvesh K. (2005), 'Optical Character Recognition for printed Tamil text using Unicode', Journal of Zhejiang University Science, Vol. 6A, No.11, pp. 1297-1305.

Adapting Nannool To Identify Additional Defeat Strategies And Refutation Techniques For Argument Gaming

Mahalakshmi G.S. and Geetha T.V.

Department of Computer Science and Engineering, Anna University, Chennai - 25
mahalakshmi@cs.annauniv.edu, tvgeedir@cs.annauniv.edu

Abstract

Argument gaming is a process where arguments and counter-arguments are exchanged as moves and counter-moves during rational discussions which follow the 'tarka' style of argumentation. Moves shall be the proposed arguments and the counter-moves shall be the refutations put forth for the proposed arguments. Refutations are generated after thorough analysis of arguments in search of any reason fallacies. Tarka Sastra recommends various refutation techniques with which every argument can be refuted. These techniques shall be classified under three strategies of defeat, namely, attack, expand and change, depending on the policy of refutation. However, to maintain the nativity of argumentation for knowledge sharing, we propose various other refutation techniques which can be adapted from Nannool, a major treatise on Tamil literature and evolve additional defeat strategies. The logic about 'how to state one's ideas while authoring a book' is seen from the argumentation perspective, which produces interesting refutations additional to those defined traditionally in Tarka Sastra.

Keywords: *Refutation, argumentation, Nannool, Fallacy*

1. Introduction

Richard Nordquist [Richard, 2008], in his Grammar & Composition defines refutation as follows: Refutation can be defined as the part of an argument wherein a speaker or writer anticipates and counters opposing points of view. Refutation is the process of attacking, weakening, tearing down or destroying the argument of an opponent [David Porter, 1954]. The prerequisite for refutation are, Argument Analysis, Defect Analysis, and Defeat Analysis [Mahalakshmi et. al., 2008]. Argument analysis procedures simply interpret the arguments by mapping them to Nyaya logics, the recommended standard to interpret arguments traditionally [Mahalakshmi et. al., 2006a]. As a result of argument analysis, every argument is split into its constituent elements of arguments. The defect analysis procedures apply defect exploration algorithms over the argument to explore argument defects [Mahalakshmi et. al., 2007]. Argument defects are generally fallacies present in the

reason component of the proposed argument. After the defects are identified, the defects are evaluated and populated into a defect set [Mahalakshmi et. al., 2006b]. The defeat analyser takes the arguments as input and tries to identify the best possible defeat strategy [Mahalakshmi et. al., 2008] applicable over the elements of arguments. The defeat strategy and respective refutation technique is identified and the counter-arguments are constructed with the opposing view. This process of repeated moves of arguments and counter-arguments make argument gaming. The objective of argument gaming is knowledge sharing.

During vedic times, 'tarka' style of argumentation was conducted only to improve one's scholarly knowledge. In this paper, by comparing the generation of arguments with the process of authoring thesis, we arrive at the interesting recommendations of various disciplines and schools of Indian philosophy, whose thoughts on narrating a thesis, shall be applied to modern argumentation scenario.

Dignaga, the reputed Buddhist logician, was the first to give much thought on fallacies of thesis [Kandasamy, 2000]. Dharmakirti, the author of Nyayabindu, has emphasized [Kandasamy, 2000] that a valid thesis should not be:

- A fact already proved
- A fact, although not yet proved, but adduced as a reason, not as a consequence
- A fact, which the disputant himself does not intend to prove on that occasion
- It must not necessarily be a fact explicitly stated and
- It must not be a fact impossibly by self-contradiction

Other fallacies are enumerated in the section of Buddhist Tamil epic Manimekalai that deals with the principles of Buddhist logic [Kandasamy, 2000]. According to Dignaga [Kandasamy, 2000], a thesis should be a valid proposition which the disputant himself believes, which the bonafide really intends to prove. From the aforesaid definition, it becomes clear that if there is any deviation, it will result in a fallacy of thesis.

Tarka Sastra [Virupakshananda, 1994] depicts argumentation as a means to preach knowledge in vedic times. Tarka sastra lists various refutation techniques which are to be followed for argumentation. In this paper, we have made a first attempt to classify the traditional refutation techniques of Tarka Sastra under various strategies of defeat namely, attack, expand and change. We have also utilized ideas of fallacies of thesis from Tamil literature, Nannool [U.Ve.Saaminaatha Iyer, 1995], and adapted to argument gaming. In other words, we have identified additional refutation techniques and evolved new defeat strategies to be applied to argument gaming. The following section discusses in detail about traditional refutation techniques and their classification into various defeat strategies.

2 Refutation techniques based on Tarka Sastra

According to Vidyabhusana [S.C.Vidyabhusana, 1988], refutation, or a point of defeat, an

occasion for rebuke or a place of humiliation, arises generally from misemployment of the proposition or any other part of an argument. There are various points of defeat, recommended by Nyaya Sastra [S.C.Vidyabhusana, 1988]. The traditional definition of every one of these refutation techniques are given below.

✷Hurting the proposition – This occurs when one admits in one's own example, the character of a counter-example.
✷Shifting the proposition – This arises when a proposition being opposed one defends it, by importing a new character to his example and counter-example.
✷Opposing the proposition – This occurs when the proposition and its reason are opposed to each other.
✷Renouncing the proposition – If one disclaims a proposition when it is opposed, it will be called "renouncing the proposition".
✷Shifting the reason – This occurs when the reason of a general character being opposed, one attaches a special character to it.
✷Shifting the topic – This is an argument which sets aside the real topic of discussion and introduces one which is totally irrelevant.
✷The meaningless – This is an argument which is based on a non-sensical combination of letters in a series
✷The unintelligible – This is an argument, which although repeated three times, is understood neither by the audience nor by the opponent.
✷The incoherent – The incoherent is an argument which conveys no connected meaning on account of the words being strung together without any syntactical order.
✷ The inopportune – This is an argument, the parts of which are mentioned without any order of precedence. Since the meaning of an argument is affected by the order in which its parts are arranged, the person who overlooks the order cannot establish his conclusion and is therefore rebuked.
✷Saying too little – If an argument lacks even one of its parts, it is called "Saying too little". As all the five parts or members are essential, a person who omits even one of them should be scolded as "Saying too little"
✷ Saying too much – This is an argument which consists of more than one reason or example.
✷Repetition - This is an argument in which the word or the meaning is said over again.
✷Silence – This is an occasion for rebuke which arises when the opponent makes no reply to a proposition although it has been repeated three times by the disputant within the knowledge of the audience
✷Ignorance – Ignorance is a non-understanding of a proposition
✷Non-ingenuity – This consists in one's inability to hit upon a reply
✷Evasion – Evasion arises if one stops an argument in the pretext of going away to attend another business
✷Admission of an opinion – This consists in charging the opposite side with a defect by admitting that the same defect exists on one's own side
✷Overlooking the censurable – This consists in not rebuking a person who deserves rebuke
✷Censuring the non-censurable – This consists in rebuking a person who does not deserve rebuke.
✷Deviating from a tenet – A person who after accepting a tenet departs from it in the course of his disputation, is guilty of "deviating from a tenet".
✷Semblance of a reason – It is the manner of submitting a reason over the discussion, as a matter of fact of support for a previously stated exceptional subject or exceptional reason.

3 Refutation techniques based on Nannool

Additional refutation strategies partially identical to Nyaya are described in Nannool. Nannool is a composition of collection of aphorisms in Tamil, with every segment discussing about different objectives and issues in connection with fallacies of thesis, organizing and interpreting the contents, intelligent ways of re-phrasing contents at place of duplications, teaching and learning methodology, qualities of teacher and learner and the like [S. Ilavarasu, 1999; U.Ve. Saaminatha Iyer, 1995; Vellai Vaarananaar, 1974]. It is this property of knowledge sharing from Nannool, which is mapped to procedural argumentation adapted from Nyaya.

Apart from various categories of style and nature of arguments as specified by Nyaya's primitive and advanced reasoning, Nannool also states further parameterization and classification of nature of arguments which when combined with Nyaya school of argument structure, frames a common grammar of handling arguments involved in knowledge sharing. The perspective of Nannool is about offline arguments which an author follows while authoring a book. However, we have modeled the perspective to suit argumentative discussion for knowledge sharing. (Note: Throughout this section, we have used (in order), the section of tamil verse, the name we coined for that technique, and the explanation from the argumentation perspective).

Nannool states that any thesis is expected to contain several special qualities for its existence. Seven types of policies with which the ideas of the thesis are coherently organized, ten defects which have to be avoided while authoring the thesis, ten techniques which enhance the presentation quality of the thesis which eventually makes the thesis appealing to the reader, thirty two techniques for describing and discussing about any thing within the thesis, etc. are to name a few. Out of the above-mentioned qualities, we have considered seven policies, ten defects and thirty two techniques of authoring a thesis and modeled those to suit argumentation.

There are seven policies of defeat as per Nannool. Each of these policies are defined below, as applicable to argumentation.
utanpatal - Admission of other's opinion
✷*maRuthal* - Opposing the proposition, or opposing other's opinion
✷*piRar tham matham mEr kondu kaLaivu* - Accept other's opinion, quote the same, and refuse at a later point by supplying proper reason
✷*thAn nAtti thanAdu niRuppu* - Introduce a subject and support it with reasons
✷*iruvar mARukOL oru thalai thuNivu* - Justify and conclude a subject that has equal falsivity
✷*piRar nUl kutram kAttal* - Highlighting the defects of other's arguments, and
✷*piRithodu padAn than madham koLaL* - Evasion.

Nannool states that there are ten fallacies which reduces the marvel of the thesis. We adapt these ten fallacies as the occasions for rebuke in argumentation. They are:
✷*kunrakUral* - saying too little – statement with selection or restriction
✷*migai padakkUral* - saying too much – over-statement
✷*kUriyadu kUral* - Repetition – repetition of the same statement
✷*mArukoLakkUral* - Contradiction – statement of contrary
✷*vazhU Col puNarthal* – Meaninglessness – usage of meaningless words in the argument
✷*mayanga vaithal* – Doubtful – presenting an argument which is not easily and clearly

understandable for the opponent
✷*vetRena thoduthal* – Unintelligible – presenting different perspectives in the argument which is of no use at that instant
✷*matRondru virithal* - Shifting the topic – elaborating about a different topic in the argument while there is a necessity of explaining an important perspective or subject
✷*sendRu theindu iRuthal* - Renouncing the proposition – presenting an argument which, though presents an additional perspective, reduces the beauty of the previously proposed argument, for which this is the support
✷*nindru payan inmai* – Untimely – presenting a good argument with not very suitable ideas to the current situation.

Nannool also lists 32 techniques for authoring a good thesis which is appealing to the reader. We have modeled those techniques to suit the construction of arguments by a proponent or opponent. The definitions as per our interpretation, are as follows:

✷*nudali pugudal* - Causation - This occurs when an argument carries only the cause for a previously stated effect
✷*othu muRai vaippu* – Modularisation – arranging the sections of argument(s) to follow a semantic thread with respect to the previous arguments
✷*thoguthu ccuttal* – Summarisation – presenting the summary of pre-stated ideas in the argument
✷*vaGuthu kAttal* - Classification - This occurs when a pre-stated argument summary is explained in detail by mentioning the sub-classifications
✷*muDithu kkAttal* – one-time argument - presenting an argument which has the start to end of a particular idea
✷*muDivu idam kUral* - Referential Annotation – presenting an argument which has the incomplete idea, and which either explicitly or implicitly indicates an instant in the future, where the idea will be explained in detail.
✷*thAn yeduthu mozhidal* - Reference to conclusion – presenting an argument which explicitly refers to previous conclusions of the self or the opponent
✷*piRankOl kUral* - Submission of other's opinion – presenting an argument which portrays the ideas of others' arguments or conclusions
✷*soRporul virithal* - Expansion of a subject - This occurs when a hard-to-interpret concept stated previously is detailed for its meaning in the submitted argument
✷*thodar ccol puNarthal* – Merging – merging diverse ideas in a single argument as a means to support the pre-stated proposition
✷*iRattuRa mozhidal* – Dualism – presenting an argument which possesses a dual meaning
✷*yEduvin muditha*l - Semblance of a reason – presenting an argument which also brings in a suitable reason to substantiate the claim
✷*oppin muditha*l - Conclusion by cases – presenting an idea in an argument which also presents other case-based analogical ideas in the same argument
✷*mAtteRindu ozhugal* - Mapping to a local implication – presenting an argument which shall be interpreted in connection with a pre-stated argument
✷*iRandadu vilakkal* - Rejecting the defeated – presenting an argument which exclude the ideas which were already defeated in the argumentative discussion
✷*yeDiradu pOtRal* - Admission of post-argument – presenting an argument which has to be interpreted in connection with one or more arguments to be stated in the near future
✷*munmozhindu kOdal* - Submission of an opinion – presenting an argument which supports the ideas of a pre-stated argument
✷*Pinnadu niRuthal* - Postponing a reason - This occurs when the mention of a reason

which is very much expected at an argument is postponed to a later stage of discussion

✷*Vigarpathin mudithal* - Difference in conclusion – presenting a conclusive argument with words which are semantically orthogonal

✷*Mudinthathu mudithal* - Summarisation of conclusions – summarizing more than one suitable conclusion with a single argument

✷*Uraitum enral* - Restricting an expansion – presenting an argument which restricts the expansion of an idea until a later point of argumentation

✷*Uraithaam endral* - Avoiding an expansion - This occurs when the arguer avoids expanding a concept in the proposed argument, when it is actually required, for interpretation of the submitted argument, stating that, the expansion has been already included in a pre-stated proposition

✷*Oru thalai thuNidal* - conclusion by bias – adopting one of the pre-stated conclusive arguments with a bias

✷*Eduthu kAttal* - Argument by example – presenting an argument which discusses an example as a support or explanation for a pre-stated argument

✷*Edutha mozhiyin yeida vaithal* – inference by argument – presenting an argument which explains a related idea and also by inference, elaborates more on other related and comparable ideas

✷*Innadu alladu idu yena mozhidal* – argument with multiple interpretation - presenting an argument which states a prominent interpretation of a topic or subject, and also includes other equivalent interpretations within the same argument

✷*Yenjiya collin yeida kUral* – presenting an argument which fulfills the left-out or unexplained part of a pre-stated idea or argument

✷*piRa nUl mudinthathu than udan paduthal* – Admission of conclusion – admitting the conclusions obtained previously out of present or previous argumentative discussions

✷*than kuRi vazhakkam miga yeduthu uraithal* - Repeating an exception - This occurs when the arguer repeats a previously stated argument to highlight the exception present in that argument

✷*collin mudivin apporul mudithal* - Expansion at instance - This occurs when the expansion of hard-to-interpret concepts are done after a detailed introduction of the related concepts

✷*ondrina mudithal thanninam mudithal* - Conclusion by modal – combining similar and analogical ideas in a single conclusion

✷*uythuNara vaippu* - Demanding inference – presenting an argument which implicitly recommends adoption of a particular interpretation for a pre-stated proposition or argument, when the pre-stated argument has multiple interpretations.

This section has discussed the fallacies of thesis as said in Nannool, with a light on application to argumentative discussions. However, these fallacies are a combination of argument fallacies and reason fallacies. Also, the first hand implementation of Indian logic based argumentation system for knowledge sharing shall be challenging due to the fact that, measuring and evaluation of most of the fallacies mentioned above is not yet formally defined in the literature, from argument gaming perspective. In other words, evaluating a thesis by adopting the above ideas is quite challenging. Out of the above refutation techniques recommended by Nyaya and Nannool, we have considered only a subset of 16 refutation techniques from both the Indian traditions, to adapt to argument gaming scenario.

4 Classification of Refutation techniques for Argument Gaming

The selected set of refutation techniques are classified into two kinds: concept refutation, relation refutation (Fig. 1). This classification is made after analysing the component of argument refuted when a particular refutation technique is used. Construction of defeat strategy determination involves distinguishing the defeat strategies into two classes, concept-originated vs. relation originated (refer Fig. 1). Defeat strategies which can be applied to oppose the concepts of the input argument are termed as concept refutations; defeat strategies that are applied to oppose the relation elements of the input argument are termed as relation refutations; there are hybrid defeat strategies which are common to both kinds. Concepts and relations refers to Nyaya logic definition of enriched concepts and relations, in this work.

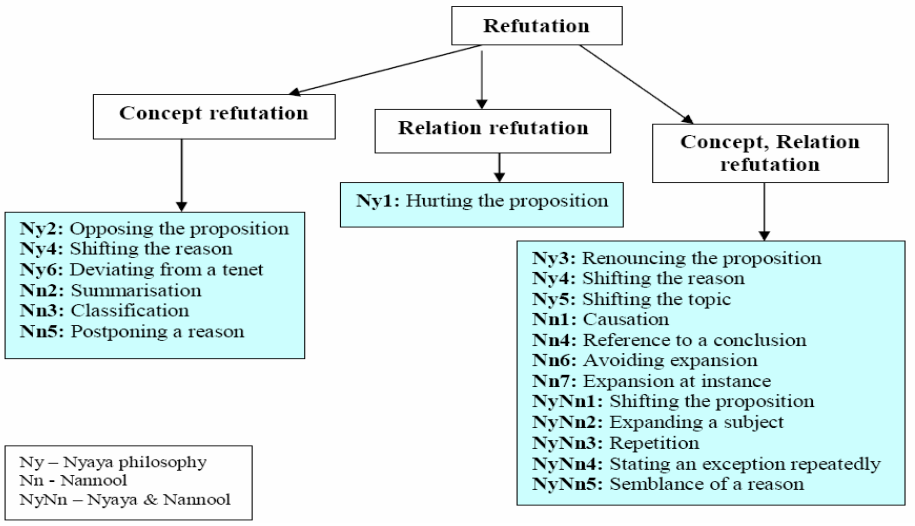


Fig.1 Classification of Refutation

Further, we have made attempts to capture every one of the refutation techniques under five defeat strategies, depending on the motive of refutation. They are namely, 1. Attack, 2. Introduce, 3. Expand, 4. Change, and 5. Repeat.

Though attack, expand and change are the widely accepted defeat strategies [Giacomin et. al., 2001], we have attempted at framing other perspectives of defeat categorisation like, Introduce, Repeat to suit argument gaming for knowledge sharing. The selected refutation techniques are analysed for their motive over any argument and are listed under the five defeat strategies (refer fig. 2). However, a single refutation technique shall be found to possess more than one objective of defeat which is clearly indicated in figure 2. (say, NyNn1 is listed under Introduce and Expand). The formal definitions of defeat strategies according to Nyaya logics is discussed below:

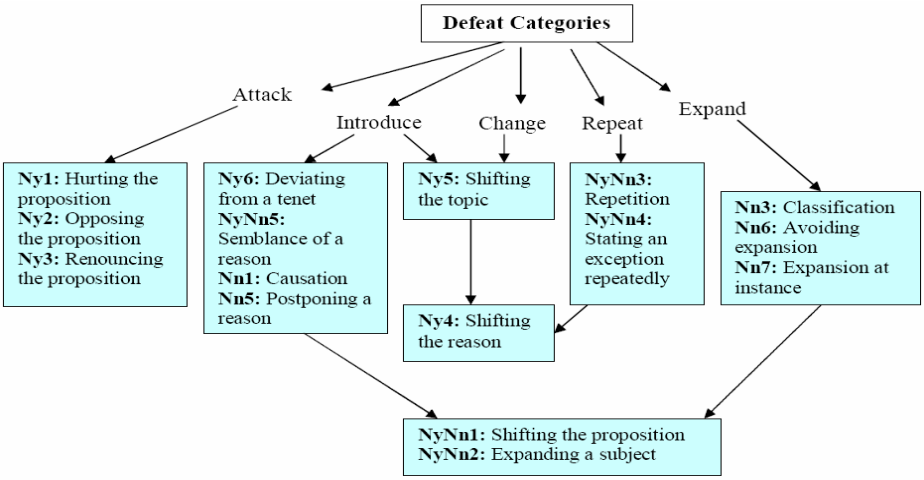


Fig 2. Classification of refutation techniques under various Defeat strategies

Definition 1: (Refutation φ) A refutation φ is said to be a mapping existing between set of counter-arguments to set of arguments. It can be denoted as $\varphi: \{\Gamma\} \rightarrow \{A\}$. The points of refutation (or the portions of argument which is prone to refutation) are denoted by **φ※**

Definition 2: (Attack φ_a) An 'attack' mapping is a function denoted as φ_a.
$\varphi_a = \{ (f: \gamma \rightarrow \alpha)$, such that $\varphi※ = \{C_R \in \alpha , C_S \in \alpha\} \}$.
By this statement, it is implicit that φ※ also includes the relation elements of α.
A refutation is said to attack the proposed argument only when it generates an opposition to Concept: Subject or Concept: Reason

Definition 3: (Introduce φ_i) An 'introduce' mapping is a function $\varphi_i : \gamma \rightarrow \alpha$, such that $\varphi※ = \{C_{OI} \in \alpha , C_{OI\text{-}new} \in \gamma\}$. C_{OI} is the object of inference which is to be proved in the argument α , and $C_{OI\text{-}new}$ is the object of inference that is newly introduced for proof by the counter-argument γ.

Introduce is the idea of including a new topic which is closely related to the topic or subject of discussion, such that, the responses does not deviate very much from the topic of discussion. During discussion, it may so happen that the current object of inference is put aside by the introduction of new object of inference (or) the conclusion of newly introduced object of inference will aid in concluding the previous object of inference and therefore, the discussion then whirls around the proving or disproving of new object of inference.

Definition 4: (Expand φ_e) An 'expand' mapping is a function $\varphi_e : \gamma \rightarrow \alpha$, where $\gamma \triangleq support(\alpha)$. $support(\alpha)$ is the function which supports the concept and relation elements of argument α by strengthening or enriching their definitions.

Expansion is the idea of enriching or strengthening the concept or relation of the input argument.

Definition 5: (Change φ_c) A 'change' mapping is a function $\varphi_c : \gamma \rightarrow \alpha$, where $\gamma \triangleq (\varphi_e(\alpha) \vee \varphi_i(\alpha))$.

Change is the operation of generating a counter-argument which changes one or many of the elements of the input argument of the proponent. Change can happen by Introduce or Expand.

Definition 6: (Repeat φ_r) A 'repeat' mapping is a function $\varphi_r : \gamma \rightarrow \alpha$, where $\gamma \subseteq \alpha$.

Repeat is the idea of counter-argument generation by repeating the part or whole of the input argument

Therefore in argument gaming, the identified refutation techniques, their classification and their categorisation into five defeat strategies are mapped together in terms of the elements of argument that is getting refuted, and is populated into a standard defeat table (Table 1). The refuter consults the defeat table, maps the defective elements of argument with the defeat table, and obtains the appropriate defeat strategy and refutation technique. This refutation technique recommended by the defeat table is later utilised for constructing the effective counter-argument.

Table 1. The structure of Defeat table

Defeat strategies	Refutation techniques			
	Elements of arguments	Elements of arguments	Elements of arguments	Elements of arguments

Analysing the defects in the arguments and exploring upon defeat strategies through which the defects shall be eliminated, form part of the preparatory work for refutation construction. The core process of refutation, is divided into two stages: Mapping and counter-argument generation.

Definition 17: (Mapping) Mapping is defined as the Cartesian product of two sets, defect set Ω and defeat set Δ, denoted as $\Omega \times \Delta$.

$$\Omega \times \Delta = \{(\omega, \delta) \mid \omega \in \Omega \text{ and } \delta \in \Delta\}$$

The best ω-δ pair (hole – defeat pair) is identified after evaluation and the counter-argument is later constructed by the recommended defeat strategy δ which is best suitable

for eliminating the worst hole ω, of the ω-δ pair. There may be more than one counter-argument resulting from a ω-δ pair, and the decision to select the most effective counter-argument is handled by the subsequent higher layers of knowledge sharing architecture. The objective here is not to miss the important and most productive hole, which might play a vital role in highlighting the prominent flaw in the submitted argument. Various constraints are needed to be carefully examined before deciding on the hole set.

In other words, the defect table [Mahalakshmi et. al., 2007] and defeat table allows the identification of best hole-defeat pairs mapped through the elements of argument in the submitted argument. After identifying the best hole-defeat pair, the actual counter-argument needs to be constructed. The best hole-defeat pair may include more than one refutation. But, only one refutation, which is very effective, has to be used for constructing the actual counter-argument. Therefore, selection of optimal refutation is a vital part in counter-argument generation.

Counter-argument Generation :

After determining hole set, Ω, and defeat set, Δ, the gaming agent is now supplied with a pool of refutations out of which the optimal refutation has to be selected. The constraint here is such that, the recommended refutation should cover the best *(h-d)* pair that will help to construct the effective counter-argument. The optimal policy $p_t(b)$ [Mahalakshmi et. al., 2008] takes a belief state b and returns the defeat that maximizes the utility. The gaming agent simultaneously constructs all counter-arguments and identifies one under the optimal refutation policy.

In a scenario, where more than one counter-argument possessing equal strength factor are recommended by an optimal refutation, the counter-arguments are selected in a random manner. The projection of counter-argument receives an immediate reward and an observation. The resulting set of possible counter-arguments, $\{c_1, c_2, \ldots, c_q\}$ are populated into a refutation set, denoted as Γ. The counter-arguments in Γ are arranged in a prioritized manner relating to the variation of concept and relation elements of the constructed arguments.

The selection of one counter-argument from the pool of generated counter-arguments is achieved after the evaluation or reward assignment for every counter-argument. Later, evaluator determines the strength of the constructed counter-argument and also the inference obtained during the hole-finding mechanism of the Defect Analyser. Determination of counter-argument strength depends on various factors: number of equivalent counter-arguments constructed, number and type of hole and defeats, number of best possible *h-d* pairs, amount of information inferred implicitly and the amount of information ignored from the submitted argument etc.

5. Conclusion

This chapter proposed an algorithm of reasoning by predicting the flow of arguments in argument gaming. The entire scenario of argumentation is motivated by Indian logic. The objective is to utilize the presence of reason fallacies in the submitted argument for further generation of counter-arguments. The notion of anticipating the counter-arguments beforehand, transforms the entire argumentation scenario into a pattern of argument gaming. However, calculation of rewards depends on the scenario of implementation setting. More detailed discussion of evaluation methodologies is dealt in the following chapter.

6. References

1.David Porter eds. , Argumentation and Debate, New York, 1954 Massimiliano Giacomin: Self-Stabilizing Distributed Algorithms for Defeat Status Computation in Argumentation. Multi-Agent-Systems and Applications 2001: 137-147

2. S.N. Kandasamy, Indian Epistemology – as expounded in the Tamil Classics , International institute of Tamil studies, Chennai , 2000.

3. G.S. Mahalakshmi and T.V. Geetha (2008b) , Modeling Uncertainty in Refutation Selection – A POMDP based approach, Journal of Uncertain Systems - special issue on "Advances in uncertain theory and its applications", 2008. (in press)

4. G.S. Mahalakshmi and T.V. Geetha: 2006, A Mathematical Model for Argument Procedures based on Indian Philosophy, Proc. of International Conf. Artificial Intelligence and Applications (AIA 2006) as part of the 24th IASTED International Multi-conference Applied Informatics (AI 2006), Austria.

5. G.S.Mahalakshmi and T.V.Geetha: 2007, Navya-Nyaya Approach to Defect Exploration in Argument Gaming for Knowledge Sharing, In proc. of International Conf. on Logic, Navya-Nyaya & Applications - A Homage To Bimal Krishna Matilal (ICLNNA '07), Jadavpur Univ., Calcutta, India.

6. G.S.Mahalakshmi and T.V.Geetha, *Architecture of Indian-logic based Procedural Argumentation System for Knowledge Sharing,* Proceedings of IEEE SMC United Kingdom & Republic of Ireland Chapter Conference on Advances in Cybernetic Systems (AICS 2006), Sheffield Hallam University, UK. September 2006.

7. Richard Norquist, Grammar and Composition, 2008 Prof. Soma Ilavarasu, "Nannool Ezhuthathikaaram", Manivasagar Publishers, Parrys Corner, Chennai, 1999.

8. Dr. U.Ve. SaaminaathaIyer, "Pavanandhi munivar iyatriya Nannool moolamum, Mayilainaathar uraiyum", Dr. U.Ve. SaaminaathaIyer Noolnilaiyam, Besant Nagar, Chennai, 1995.

9. Sathis Chandra Vidyabhusana, A History of Indian Logic – Ancient, Medieaeval and Modern Schools, Motilal Banarsidass Publishers Private Ltd., Delhi, India, ISBN:81-208-0565-8. pp. 84, 1988.

10. K.Vellai vaaraNanaar, "Tolkaapiyam – Nannool Ezhuthathikaaram", Annamalai University, Tamilnadu. Third Edition, 1974.

11. Swami Virupakshananda: 1994, Tarka Samgraha, Sri Ramakrishna Math, Madras.

Identification of Foreign Words in Tamil Scripts

Mohammed Afraz and Sobha Lalitha Devi

AU-KBC Research Center,
MIT, Anna University, Chennai-44
sobha@au-kbc.org,

Abstract:

We present a statistical approach based upon N-grams for context-free identification of transliterated foreign names and borrowed words in Tamil text. The method is purely statistical and does not require the use of any lexicons or linguistic analysis tool for the source languages. It also does not require any manually annotated data for training – we learn from noisy data acquired by over-generation. We report precision/recall results of 80/82 for a corpus of unique words.

1 Introduction

Human language is constantly changing, with new words being created on a daily basis. The native speakers tend to use borrowed foreign terms and foreign names in written texts. Such borrowed words appear as foreign words included in the language and as a transliterated words. The adoption of a foreign name into one language is usually a process of adjusting its original pronunciation to suit the phonological regularities in the target language. This procedure of phonetically "translating" foreign names is called transliteration. One of the main reasons of the importance of transliteration from the point of view of Natural Language Processing (NLP) is that Out Of Vocabulary (OOV) words are quite common since every lexical resource is very limited in practical terms. Such words include named entities, technical terms, rarely used or 'difficult' words and other borrowed words, etc.

The OOV words present a challenge to NLP applications like information Retrieval (IR) systems, Cross Lingual Information Retrieval (CLIR) and Machine Translation (MT). In sample data which was used for testing, we found genres with as many as 5% of the word instances as foreign words. These transliterated words and foreign words require special treatment in NLP and IR systems. For example, in IR, query expansion requires special treatment for foreign words; when tagging text for parts of speech, foreign words appear as unknown words and the capability to identify them is critical for high-precision PoS tagging; in Machine Translation. In Named Entity Rec-

ognition and Information Extraction, the fact that a word is transliterated from a foreign language is an important feature to identify proper names. In this paper we describe a method for identifying foreign words and transliterated words in Tamil, written in the Tamil script. The method is unsupervised, uses easily acquired resources, and is not specific to Tamil. The native Tamil writing system has properties that make it different for the foreign words. The system uses the romanized form (WX notation) for representing Tamil.

In the rest of the paper, we first review previous work related to the task of transliterated word identification, since most of the borrowed word will appear in the transliterated form. We then present our approach: we identify transliteration by comparing the letter structure of words with models trained in a way that captures the sound structure of the language – one in Tamil and one in English, as written in the Tamil writing system. Then we discuss the the results and end the paper with conclusions and future scope.

2 Related Work

There is a growing body of research on automatic extraction of transliterated pairs. Sherif and Kondrak[1] use seed examples and a sentence aligned English/Arabic text to jointly learn a bilingual string distance function and extract transliterated pairs. While this work aims at complete alignment, our task is only the identification of transliterated candidates. The task of identifying transliterated words has been less studied. Stalls and Knight [2] identified the problem – ". . . in Arabic, there are no obvious clues, and it's difficult to determine even whether to attempt a back-transliteration, to say nothing of computing and accurate one" – but don't deal with it directly.

Oh and Choi [3] studied identification of transliterated foreign words in Korean text. They used a corpus of about 1,900 documents in which each syllable was manually tagged as being either Korean or Foreign. However, beside requiring a large amount of human labor, their results are not applicable to Tamil. Nwesri et al. [4] dealt with the identification of transliterated foreign words in Arabic text in the setting of an information retrieval system. They tried several approaches: using an Arabic lexicon (everything which is not in the lexicon is considered foreign), relying on the pattern system of Arabic morphology, and two statistical Ngram models, the better of which was based on Cavnar and Trenkle's rank order statistics [5], traditionally used for language identification. For the statistical methods, training was done on manually constructed lists of few thousands and needed hand written heuristic rules.

Another related field of research is that of language identification, in which documents are classified according to their language. The problem of finding transliterated foreign words can be regarded as performing language identification on individual words instead of documents or sentences. Algorithms that rely on letter-Ngram statistics can be relevant to the foreign words identification task. Two notable works are [5] and [6], both based on letter-level Ngram statistics. Canvar and Trenkle use rank order statistics, and Dunning use Naive-Bayes classification with a trigram language model, and add-one smoothing.

3 Our Approach

We concentrate on identifying borrowed and transliterated words in Tamil text. As most borrowed words come from English sources, we concentrate on finding borrowed words from such origins. We also focus on a context-free approach, which works on words alone, without requiring their context. Our intuition is that Tamil words sound different

than English words, and as letters are closely related to sounds, this should be reflected in their writing. Therefore, we believe that letter level n-gram and syllable level n-gram approaches should be applicable to this problem, given sufficient data and suitable language models, even at the word level.

3.1 About the Tamil language:

Tamil is a member of the Dravidian family of languages and is spoken primarily in the state of Tamil Nadu in India and in Sri Lanka. Tamil script consists of 12 vowels, 18 consonants, and one special character called the aytam in the Tamil script. The vowels and consonants combine to form 216 compound characters, bringing the total number of characters in the script to 247.

Unlike other Indian languages, Tamil does not have signs for voiceless aspirated such as /kh/, voiced /g/, and voiced aspirated stops /gh/, which explains the relatively small number of signs in the Tamil script compared to other Indian languages. To write some of these sounds, some signs have multiple sound values: க stands for both /ka/ and /ga/, ச for both /ca/ and /sa/,த for /ta/, /da/, and ப for /pa/ and /ba/, and so on. Sometimes these phonetic alterations are conditioned by the sound's position in the word such as ப is /pa/ at the beginning of word or after a voiceless consonant, and /ba/ between vowels or after /m/, while other times they are somewhat random such as ச can be both /ca/ and /sa/ at the beginning of a word. It has three /na/ ந ,ன ,ண and two /ra/ர,ற .This confusion is due to phonological changes. The language had some missing sounds to express some foreign names, so Tamil borrowed from Sanskrit letters that added some special letters to Tamil. The below six letters are called Grantha letters and have been used to write loanwords.

ஜ ஸ ஹ ஷ க்ஷ ஸ்ரீ

Our core idea is to identify the foreign words in native language and study the how these phonological changes are reflected in the script. Our model is unsupervised, it does not require any manually tagged data for training. we built two language models, one for the native language and other for the foreign language. For training the language model of Foreign words in Tamil script, we need a corpus of foreign words. Unfortunately, there is no large corpus of transliterated foreign words in Tamil script available, and we manually create and collects the words from web sources like blogs, websites, technical manuals ..etc. we found two typical problems with the foreign words,

a) The phoneme letter equivalence are of many to many relationship i.e., letters can be often be pronounced in different ways and certain sounds can be written with different letters. The following examples in Tamil illustrates this behavior.

இஞ்சினியரிங்

இன்சினியரிங்

இன்ஜினியரிங்

இன்ஜினீயரிங்

b) The phoneme of different languages differs, if a language does not have the sound, it will transliterate this sound by another similar sound and most of the foreign words

which are loanwords from English used the Grantha letters.

ஹேஸ்டிங்

ஹாம்பர்க்

ஹாலிவுட்

ப்ரொவ்சிங்

For training the language model of Foreign words in Tamil script, we created a list of 3046 foreign words. For the native language model we use a collection of Tamil texts from a period of about 100 years ago, which are assumed to have a relatively low frequency of foreign words.

3.2 Unsupervised Approach

With a fair amount of training data for both language models. We build the N-gram model at the letter level and syllable level for both the language models, which results in the language 'sound-patterns'. A word *w* can be represented as,

$w \Rightarrow l_1, l_2, l_3, l_4, \ldots\ldots\ldots\ldots l_i$
or
$w \Rightarrow s_1, s_2, s_3, s_4, \ldots\ldots\ldots\ldots s_i$

where,
l_i = i^{th} letter/character in word *w*
s_i = i^{th} syllable in word w

We use Naive-Bayes classifier method for choosing the best generative language model for a given word: Assuming two generative language models, T_n for generating native Tamil words and T_f for generating foreign words, and an observed word *w*, we would like to find the model T_i that maximizes $P(T_i|w)$. As we don't know $P(T_i|w)$, we use Bayes Formula, and get:

$$P(T_i|w) = \frac{P(w|T_i)P(T_i)}{P(w)}$$

P(w) = 1 (the word is given), and we assume both models are equally probable $P(T_n) = P(T_f)$ so we are left with the problem of evaluating $P(w|T_i)$. By normalizing the results we can get not only a decision, but a probability for each category.

$$P'(T_f|w) = \frac{P(T_f|w)}{P(T_f|w) + P(T_n|w)}$$

$$P'(T_n|w) = \frac{P(T_n|w)}{P(T_f|w) + P(T_n|w)}$$

At the heart of the over-generation mechanism is a procedure in which, given a foreign word in foreign script,outputs many possible transliterations of that word in the native

script. These transliterations are meant for machines to learn from, and need not be 'correct' by human standards. Indeed, most of them will not be accepted as proper transliterations of the word by humans. They do, however, seem to capture the 'essence' of the foreign language in native writing.

All Foreign Words are Not Created Equal Examining foreign words in Tamil text reveals that a very big proportion of them are proper names. Although many English words function also as proper names, the sound patterns of proper names are somewhat different than those of regular English words. As they represent much of the data we want to identify, they deserve a special treatment. For this reason, the resulting model can not reliably decide between 'English' and 'English name', but it does succeed in approximating some of the properties of names which get lost in the mass of all Tamil words.

4 Experimental Settings

For training the Tamil model, we use the prose and mass sections of the ' *Project Mudari* ' (prose, poetry and essays). For the foreign words we use text from Tamil blogs, Technical Tamil documents, Tamil Articles and reviews. A total of 10,355 words, 5055 of them unique. We removed the Morphological inflections from all words. Overall, there are 4810 Tamil words, 628 foreign words of which 271 are foreign proper names, and another 67 words which are ambiguous (these words either have a clear foreign origin but became so common that they sound Tamil, or they have two readings, one Tamil and one borrowed). For the cross validation of experiments described we took for each fold a different 20% of the Tamil words and 20% of the Foreign words for testing, and the rest for training.

As we have build the N-gram model at the letter level and syllable level for both the language models, which results in the probability of sound pattern that can occur with respect to native language. We consider probability of sound patterns at the letter level and the syllable level which clearly differentiates foreign words from the native language.

Table 1: Total number of words used for the experiment

Tamil words	Foreign words	Ambiguous words	Total words
4810	628	67	5055

5 Results

The following tables summarize the results of all the experiments described above. Precision is the ratio between the number of correctly identified foreign words and the total number of words identified as foreign. Recall is the ratio between the number of identified foreign words, and the real number of foreign words. The 68 ambiguous words were excluded from the calculations.

Table 2: Summarizes the precision and recall of the experiment

	Precision	Recall
N-gram letter level	78.44%	80.09%
N-gram syllable level	92.23%	86.22%
Total	**80.67%**	**83.31%**

5 Open Issues and Future Work

Segmentation: Tamil words gets appended with prefixes and suffices. All the results presented here ignored that fact and were on words without prefixes. This is suitable for IR settings, but not good enough for the general NLP application, in which text appear unsegmented. This prefixation can make otherwise 'foreign' words to get tagged as 'Tamil'. Although recent morphological disambiguators for Tamil [7] perform such segmentation with reasonable accuracy (above 98%), they all rely on a Tamil lexicon. By its very nature, such lexicon is bound not to be complete in its coverage of transliterated foreign words, and so the above-mentioned disambiguators unfortunately fall on the wrong side of the pipeline – it is precisely for improving such systems that foreign words identification is needed. Dealing with unsegmented data is an interesting problem, left for future work.

Context: Our method is context free, and assigns probabilities to words. But context does matter in some cases (for example, some words can be either foreign or native, depending on the reading). Context can also help in the segmentation problem. In addition, integrating our method with the output of a POS tagger[8] is also a promising direction.

Different Kinds of Borrowings This work focuses on identification of foreign words, yet no real distinction was made between cognates, borrowings and "real" transliterations. For most parts we can indeed ignore this distinction: "heavily incorporated" cognates and borrowing, which we consider as native words, tend to adopt the Tamil pronunciation (and writing), and are considered Tamil also by our algorithm, while less incorporated cases of borrowing exhibit inconsistent writing patterns, and demand special treatment similar to that of "pure" transliterations[9]. However, finding an algorithmic way of separating these groups is an interesting research question.

6 Conclusion

We presented a method for identification of borrowed words and transliterated names in Tamil text. The method is very loosely supervised, does not require annotated data, and achieves satisfactory results (precision/recall of 80%/82% with a purely statistical

method). We verified that our approach of learning from a large amount of automatically generated data greatly outperforms learning with a small amount of manually annotated data. Giving specific care to identification of proper nouns was shown to improve performance.

References

[1] Sherif, T., Kondrak, G.: Bootstrapping a stochastic transducer for arabic-english transliteration extraction. In: Proc. of ACL. (2007)

[2] B. Stalls and K. Night: Translating Names and Technical Terms in Arabic Text. In: Proc. of the COLING/ACL Workshop on Comp. Approaches to Semitic Languages. (1998)

[3] Oh, J., Choi, K.: A statistical model for automatic extraction of korean transliterated foreign words. Int. J. of Computer Proc. of Oriental Languages 16 (2003)

[4] Nwesri, A.F., Tahaghoghi, S., Scholer, F.: Capturing out-of-vocabulary words in arabic text. In: Proc. of EMNLP2006. (2006)

[5] Cavnar, W.B., Trenkle, J.M.: N-gram-based text categorization. In: Proc. of SDAIR-94. (1994)

[6] Dunning, T.: Statistical identification of language. Technical report, Computing Research Lab, New Mexico State University (1994)

[7]Menaka S, Sobha L (2009), "Optimising the Tamil Morphological Analyser", In 3rd National Conference on Recent Advances and Future Trends in IT - (RAFIT2009), Punjabi University, Patiala, PUNJAB.

[8] Arulmozhi. P, Sobha L, Pattabhi R K Rao. (2006) "A Hybrid POS Tagger for a Relatively Free Word Order Language", In Proceedings of Symposium on Modeling and Shallow Parsing of Indian Languages, Indian Institute of Technology, Mumbai, pp 79-85.

[9] Mohammad Afraz and Sobha L (2008), "English to Dravidian Language Machine Transliteration: A Statistical Approach Based on N-grams", In the Proceedings of International Seminar on Malayalam and Globalization, Trivandrum, Kerala.

[10]Yoav Goldberg and Michael Elhadad "Identification of Transliterated Foreign Words in Hebrew Script" 2001.

DETECTION OF METAPHORS IN TAMIL DESCRIPTIVE PASSAGES

S.Vidhya and G.S. Mahalakshmi

Dept. Of Computer Science and Engineering
Anna University, Chennai–25
mahalakshmi@cs.annauniv.edu

ABSTRACT

In this modern trend, there has been a great impact for poetry reading and writing in both fixed and free order languages. Generally, writers make use of metaphors, similes, parables, and inflections in their works. In specific, Metaphors play an important role in helping the writers express his or her ideas in exaggerated form and in turn, adds beauty to his or her work. Metaphors make the literature works meaningful and lead the readers to a colorful world. Most people who read those poems do not understand the work because the style and the word usage differs from one writer to the another. Many researches are being done in the area of metaphors for fixed and free order languages. In this paper, a system has been proposed which helps the researchers and Tamil workers to understand the style and metaphor type usage in the Tamil context. For this work, computational linguistics is used which includes Natural Language Processing and Understanding, knowledge representation, reasoning of fixed and free order language.

1. INTRODUCTION:

Natural Language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. NLP encompass both text and speech. Work on speech processing has evolved into a separate field. Natural Language generation systems convert information from computer databases into readable human language. Natural Language understanding systems convert samples of human language into more formal representations such as parse trees or first order logic structures that are easier for computer programs to manipulate.

Many problems within NLP apply to both generation and understanding,. Applying NLP techniques to fixed order languages such as English, Arabic is now overtaken with working on those techniques for free order languages such as Tamil, Kanada, Malayalam. Also in fixed and free order languages, area of criticism, metaphors, discourse theory, prose analysis are the major part of study. Metaphors play a major role in language usage both in text and in speech. Works on metaphors is an ongoing research in linguistics. Natural language processing is now applied in all those areas. In specific , In this paper , we concentrate only on the metaphors in Tamil language and in specific the writings of Kannadhasan and Bharathidhasan who follow a similar style of language.

Some of the NLP techniques used in fixed and free order languages are language structure, morphology, syntax, semantic and pragmatic. parser and analyzer analyses the sentence and finds the root words in the sentence. The root words are given to the wordnet and meaning for each is obtained. Then the sense used in the context is analyzed using natural language understanding. Here, section 2 covers all the related works to this proposed work. Section 3 covers the detailed description of the proposed system.

2. Background:

This section covers some of the related works in Natural Language processing and understanding. Some of the NLP techniques discussed below are : Parsing, Morphological analysis, semantic and pragmatics, disambiguation and reasoning.

Parsing :

'Parsing' is the term used to describe the process of automatically building syntactic analysis of a sentence in terms of a given grammar and lexicon. The resulting syntactic analyses may be used as input to a process of semantic interpretation, (or perhaps phonological interpretation, where aspects of this, like prosody, are sensitive to syntactic structure). Parsing in natural language is mostly context dependent. Many parsers have been built for fixed and free order languages [Michael A. Covington]. Tamil language is one of the free word order language due to the tight coupling between the morphological and syntactic levels. parsing of these free order languages is complex. I

n most English sentences, the sentences will be in the order of subject, verb and finally object. But in Tamil sentence, it follows the subject, object and then the verb. However here interchangeable of subject and object is also possible.

1. சிவா சென்னைக்கு வந்தான்.

Tamil Form : siva chennaiku van'taan
English Form : siva Chennai came

2. சென்னைக்கு சிவா வந்தான்.

Tamil form : chennaiku siva va'ntaan
English form : Chennai siva came

In the above examples , the first sentence is in the form of subject , object, verb and the second sentence is in the form of object, subject and verb. Both the sentences makes sense in Tamil form but not in the English form.

Thus free order nature of the Tamil language is made being it a morphologically rich language. In the fixed order languages, the position of the word plays a major role in determining the syntactic function of the word. In particular, in the free order languages like Hindi, the local word groups help in determining the syntactic function. As a result of parsing, the nouns, verbs, adverbs, adjectives are obtained. For some verbs, finding the meaning requires to maintain a large lexical database. Also a parser alone cannot effectively run without morphological information of the words appearing in a sentence. In this sense, we argue that morphological analyzer for a language is a fundamental system towards the construction of machine translation systems.

Morphological analysis :

Morphological analysis is the process of segmenting words into morphemes and analyzing the word formation. Morphological analyzer is used in speech synthesizer, speech recognizer, segmentation, lemmatization, noun decompounding, spell and grammar checker, sentence boundary detection and machine translation. At the out set, a morphological analyzer of a language reveals structures of words in the particular language. Since the use of the correct word form is the basis for grammar of any language, the morphological analysis is of great importance to all the steps in natural language processing.

In other words, morphological analyzers are inevitable in any machine translation system in general and parsers in particular. It should be noted that natural language processing systems without morphological analyzers require a huge lexical database. This reduces the computational efficiency of natural language translation systems. Although, there are so many morphological analyzers available; they are differing from one another due to the differences in structures of words in respective languages.

Morphological analyzers for English language have been developed by many researchers. Koskenniemi's two-level morphology was the first practical and most general model in the history of computational linguistics for the analysis of morphologically complex languages. Koskenniemi' s Pascal implementation of morphological analysis was quickly followed by others. Later on, this is extended to Tamil language. The morphological structure of Tamil is quite complex since it inflect to person, gender, and number markings and also combines with auxiliaries that indicate aspect, mood, causation, attitude etc in verb. Noun inflects with plural, oblique, case, postpositions and clitics suffixes. For the purpose of analysis of such inflectionally rich languages, the root and the morphemes of each word have to be identified. The structure of verbal complex is unique and capturing this complexity in a machine analyzable and gene ratable format is a challenging job.

The formation of the verbal complex involves arrangement of the verbal units and the interpretation of their combinatory meaning. Phonology also plays its part in the formation of verbal complex in terms of morphophonemic or sandi rules which account for the shape changes due to inflection. Classification of Tamil verbs based on tense inflections is evolved. The inflection includes finite, infinite, adjectival, adverbial and conditional forms of verbs (Rajendran.S et al., 2001). Generally SVMTool is developed for POS tagging but now this tool is used in morphological analyzer for classification task using the machine learning approaches.

Support vector approaches have been around since the mid 1990s, initially as a binary classification technique, with later extensions to regression and multi-class classification. Here Morphological problem is converted into classification problem. These classifications can be done through supervised machine learning approach [Kadri Hacioglu and Wayne Ward,2003].

Support Vector Machine is a new approach to supervised pattern classification which has been successfully applied to a wide range of classification problems. SVM is based on strong mathematical foundations and results in simple yet very powerful algorithms. SVMs are learning systems that use a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. SVMTool is an open source generator of sequential taggers based on Support Vector Machine.

For eg:

Input word : படித்தான்

Output : படி <*VERB_ROOT*>

த்த் <*PAST_TENSE*>

ஆன் <*3SM*>

In addition, to improve the quality of the morphological analyzer, especially when semantic analysis of words are taken into account, concept dictionary is used. This dictionary contains further information such as synonyms and antonyms for the words given in the base dictionary. Also the basic requirement for the sense identification is the Natural Language Understanding.

Natural Language Understanding:

Regardless of the approach used for language generation, some common components need to be identified for natural language understanding. The system needs a lexicon of the language and a parser and grammar rules to break sentences into an internal representation. The construction of a rich lexicon with a suitable ontology requires significant effort, *e.g.*, the WorldNet lexicon required many person-years of effort.

The Natural Language Understanding uses the techniques of pragmatics and semantics. Pragmatics studies how the transmission of meaning depends not only on the linguistic knowledge (e.g. grammar, lexicon etc.) of the speaker and listener, but also on the context of the utterance, knowledge about the status of those involved, the inferred intent of the speaker, and so on. In this respect, pragmatics explains how language users are able to overcome apparent ambiguity, since meaning relies on the manner, place, time etc. of an utterance. This level is concerned with the purposeful use of language in situations and utilizes context over and above the contents of the text for understanding

The goal is to explain how extra meaning is read into texts without actually being encoded in them. This requires much world knowledge, including the understanding of intentions, plans, and goals. Some NLP applications may utilize knowledge bases and inferencing modules. For example, the following two sentences require resolution of the anaphoric term 'they', but this resolution requires pragmatic or world knowledge.

e.g.

The city councilors refused the demonstrators a permit because they feared violence.

The city councilors refused the demonstrators a permit because they advocated revolution.

The use of pragmatics in traditional Tamil language is from the ancient Tamil literature. Analyzing the style of those writers is difficult and combining all those past to current usage of language is complex and difficult. So usually semantic theory is considered to guide the comprehension. The interpretation capabilities of a language understanding system depend on the semantic theory it uses. Competing semantic theories of language have specific trade offs in their suitability as the basis of computer automated semantic interpretation. These range from naïve semantics or stochastic semantic analysis to the use of pragmatics to derive meaning from context.

Semantic analysis:

Semantic processing determines the possible meanings of a sentence by focusing on the interactions among word-level meanings in the sentence. This level of processing can include the semantic disambiguation of words with multiple senses; in an analogous way to how syntactic disambiguation of words that can function as multiple parts-of-speech is accomplished at the syntactic level. Semantic disambiguation permits one and only one sense of polysemous words to be selected and included in the semantic representation of the sentence. For example, amongst other meanings, 'file' as a noun can mean either a folder for storing papers, or a tool to shape one's fingernails, or a line of individuals in a queue. If information from the rest of the sentence were required for the disambiguation, the semantic, not the lexical level, would do the disambiguation.

A wide range of methods can be implemented to accomplish the disambiguation, some which require information as to the frequency with which each sense occurs in a particular corpus of interest, or in general usage, some which require consideration of the local context, and others which utilize pragmatic knowledge of the domain of the document. [Xavier Carreras and Llu´ıs M`arquez,2004]

In general, Morphological knowledge provides the tools for building words, while syntactic knowledge combines words to form sentences. Semantic knowledge provides the meaning of a given word, and pragmatic knowledge helps us to interpret the complete sentence in its true context. All of these different linguistic knowledge forms, however, have a common associated problem, their many ambiguities, which is difficult to resolve. One of the main objectives in designing any NLP system, therefore, is the resolution of ambiguity. Furthermore, each type of ambiguity, whether it be structural, lexical, quantifying, contextual or referential, requires its specific resolution procedure.

The resolution of the lexical ambiguity that arises when a given word has several different meanings can be done with the help of sense identification techniques.

Sense identification approaches :

Some of the approaches related with sense identification are deixis, anaphora reference, cataphora reference, semantic interpretation and finally the word sense disambiguation.
In linguistics , **deixis** refers to the phenomenon wherein understanding the meaning of certain words and phrases in an utterance requires contextual information. Words are deictic if their semantic meaning is fixed but their denotational meaning varies depending on time and/or place. Words or phrases that require contextual information to convey any meaning - for example, English pronouns - are said to be deictic. To identify the appropriate nouns when pronouns like his, her, they, all , each used in a context seems to be a difficult task. Many languages share the same issue like that with English.

The universal conventional solution is based on the context, which is always the same—the antecedent is a representative individual of a class, whose gender is unknown or irrelevant. Normally masculine, but sometimes feminine, forms of singular pronouns are supplied, in what is called *generic* usage. The context makes the generic intent of the usage clear in communication.

Example: An ambitious academic will publish as soon as she can.

Unless there is reason to believe the speaker thinks ambitious academics are always female, the use of *she* in this sentence must be interpreted as a generic use. Deixis is closely related to both

indexicality and anaphora, as will be further explained below. Anaphora refers to the way in which a word or phrase relates to other text:

✷An exophoric reference refers to language outside of the text in which the reference is found.
✷A homophoric reference is a generic phrase that obtains a specific meaning through knowledge of its context. For example, the meaning of the phrase *"the Queen"* may be determined by the country in which it is spoken. Because there are many Queens throughout the world, the location of the speaker provides the extra information that allows an individual Queen to be identified.
✷An endophoric reference refers to something inside of the text in which the reference is found.
✷An anaphoric reference, when opposed to cataphora, refers to something within a text that has been previously identified. For example, in *"Susan dropped the plate. It shattered loudly"* the word *"it"* refers to the phrase *"the plate"*.
✷A cataphoric reference refers to something within a text that has not yet been identified. For example, in *"He was very cold. David promptly put on his coat"* the identity of the *"he"* is unknown until the individual is also referred to as *"David"*.

Identifying all these references in a context correctly is a big task. To resolve all these issues word sense disambiguation technique is used which involves assigning a definition to a given word, in either a text or a discourse, that endows it with a meaning that distinguishes it from all of the other possible meanings that the word might have in other contexts.

Knowledge sources for WSD provide data which are essential to associate senses with words. They can vary from corpora of texts, either unlabeled or annotated with word senses, to machine-readable dictionaries, thesauri, deixis, ontologies, etc. Dictionaries gives just all the possible meaning of words. Thesauri is just the contradictory for dictionary. Recent advances in WSD have benefited greatly from the availability of corpora annotated with word senses. Most accurate WSD systems to date exploit supervised method which automatically learn cues useful for disambiguation from hand-labeled data. Although supervised approaches out perform their unsupervised alternatives, they often require large amounts of training data to yield reliable results [T.H. Ng,1997], and their coverage is typically limited to the words for which sense-labeled data exist.
Unfortunately, creating sense-tagged corpora manually is an expensive and labor-intensive endeavor which must be repeated for new domains, languages, and sense inventories.[Roberto Navigli and Mirella Lapata,2010] These resources are alone not enough to find the sense so at the next level Ontologies had been created to specify the relations and reasoning.

Ontology :

The hypernym/hyponym relationships among the noun synsets can be interpreted as specialization relations between conceptual categories. In other words, WordNet can be interpreted and used as a lexical ontology in the computer science sense. However, such an ontology should normally be corrected before being used since it contains hundreds of basic semantic inconsistencies such as (i) the existence of common specializations for exclusive categories and (ii) redundancies in the specialization hierarchy. Furthermore, transforming WordNet into a lexical ontology usable for knowledge representation should normally also involve (i) distinguishing the specialization relations into *subtypeOf* and *instanceOf* relations, and (ii) associating intuitive unique identifiers to each category. WordNet has also been converted to a formal specification, by means of a hybrid bottom-up top-down methodology to automatically extract association relations from WordNet, and interpret these associations in terms of a set of conceptual relations.

There are ontologies available namely NYAYA ontology which specifies the relations and provides a reasoning for those relations. But upto now, these ontologies do not concentrate on context based sentences. Nyaya ontology that is in existence establishes the relations based on three categories namely concept , quality and action. No other attributes are taken into account. This does not help in identifying the context relations.

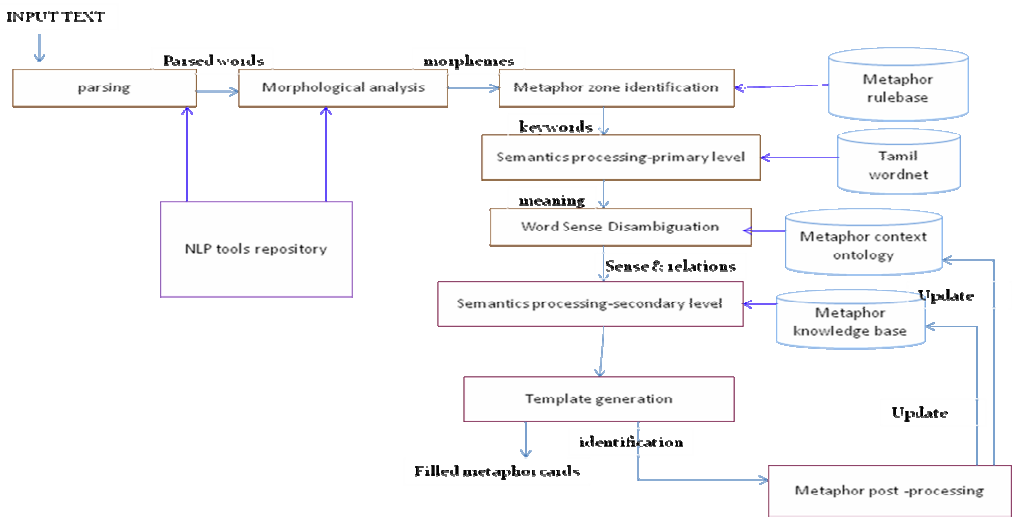
So, in order to specify the context based relations a new ontology is to be developed in the below proposed system. Section 3 covers the detailed description of the system to be developed and the context based ontology creation , metaphor rule base and knowledge base creation for metaphors.

3. DETECTION OF METAPHORS

Metaphors in Tamil language varies largely from the past ancient literature to the present. There is drastic change in the words and their usage from the past to the present. Combining all those features of past and the present is difficult. So, this system is proposed only for the works of kannadhasan and Bharathidhasan of 20th century.

The work on detecting the metaphors in Tamil descriptive passages started with the first step of analysis of the text. This analysis involves parsing and morphological analyzing of the text with the help of Natural Language Processing tools such as Vaanavil and Atcharam. After this process, semantic processing involves identifying the meaning of the root words using Tamil Wordnet and then the sense of the word used in the given context is identified using disambiguation technique and the ontology that has been developed. From the relations and the sense in which the word is used in the context, metaphors types can be identified using the Metaphor knowledge base that has been created. A rule base has been created for metaphor zone identification. For example, a sample of the selectional restriction rules are stored as follows:

Verb=>subject_category;subject_case=>object_category;object_case.



In the next section, we will discuss in detail about the modules in the system.

4. DETAILED DESCRIPTION

The modules used in this system are: Parsing, Morphological Analysis, Handling of Nouns, Verbs, Metaphor zone identification, Semantic processing at primary and advanced level, Template generation and Updation.

Parsing:

Parsing is a complex process due to the ambiguity present at the morphological and syntactic level .Most parsing in natural language is context dependent .The parser is a process, which identifies syntactic constituents of a sentence and represents the same using a parse tree. The parser tool used here is vaanavil.

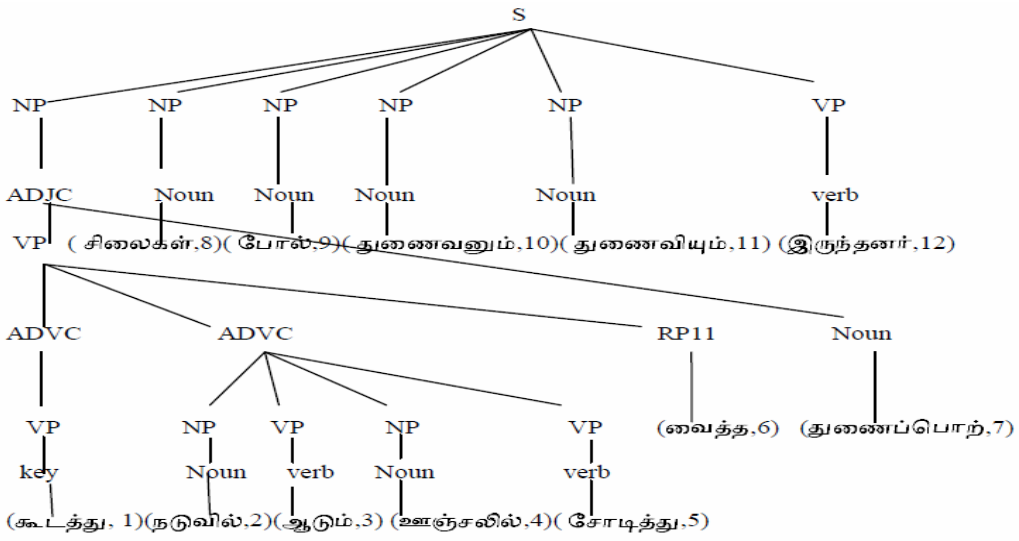
Symbol - Description

S	Sentence
SS	Sub sentence
NP	Noun phrase
VP	Verb phrase
key	Verbal participle
RP	Relative participle
Con	Connective
Cnjn	Conjunction
NC	Noun clause
ADJC	Adjective clause
ADVC	Adverb clause

The words of the input sentence will be shown along with their position in the sentence starting from zero separated by a comma. This gives the words in a parse tree structure.

Sample output :

கூடத்து நடுவில் ஆடும் ஊஞ்சலில் சோடித்து வைத்த துணைப்பொற் சிலைகள் போல் துணைவனும் அன்புகொள் துணைவியும் இருந்தனர்



After the parsing is done for the text , we will get the nouns, verb and adverbs. Those words may or may not be present in the dictionary specified due to their usage. So it is necessary to find the root words for those words. This is done by morphologically analyzing the words.

Morphological Analysis:

For the morphological analysis, it is intended to use the analyser tool which analyses the words and gives the appropriate root words or morphemes. The morphological analyser tool used in this system is ATCHARAM.

Atcharam-morphological analyzer

A Morphological analyzer breaks a word into its root word and associated morphemes. A morpheme is defined as the smallest part of a language that can be regularly assigned a meaning. Tamil is a morphologically rich language in which most of the morphemes coordinate with the root words in the form of suffixes.

Phases of Tamil Morphological Analyzer:
✷Given an input string, the processing starts from right to left to look for suffixes. A list of suffixes is maintained
✷Searches for the longest match in the suffix list
✷Removes the last suffix, determines its tag and adds it with the word's suffix list
✷Checks the remaining part of the word in the dictionary and exits if the entry found
✷According to the identified suffix, generates the next possible suffix list
Repeats steps 2 to 5 with the current suffix list

Before handling the morphology of the words it is necessary to study sandhi rules of Tamil. These come into the picture when a root word and suffix combine, or when one suffix combines with another suffix. These rules help to obtain inflected form from base forms of words.

Handling of Nouns:
Nouns can occur in isolation or can take plural, oblique suffix, case suffixes, postpositions and clitics. In this analyser each suffix is dealt with separately.
Noun Stem + [oblique] + [plural suffix] + [case suffix] + [postpositions] + [clitics]
The system deals the clitics in a single block. It recursively removes clitics one after the other. After removing the clitics suffixes the remaining word enters the postposition block.

One may write the postposition suffix either as a separate word, separated by space or as an additional suffix to the noun. The analyzer should handle both cases. Since the analyzer takes a single word as an input, the dictionary will take care of it. Otherwise the system will remove the postposition suffix and pass the remaining word to the case marker block.

Tamil has eight case suffix entries. But some case suffixes has many variants. For example the dative case 'கு' as other allomorphs like 'க்கு', 'அக்கு', 'உக்கு', 'இற்கு'etc. Finding the dative suffix and adjusting the root word will increase the complexity of the system. To avoid the problem the system treats each variants as a separate case marker entry. The locative case suffix 'இல்' and 'இடம்' almost always occurs with the postposition 'இருந்து'. Hence the analyser deals with 'இலிருந்து'and 'இடமிருந்து'as separate case entries. After removing the case suffix the remaining word will be either a root word or a root word with the oblique suffix or a root word with a plural suffix. The remaining word can be a root word or an inflected from of the root word. If it is inflected form, the root word is obtained by applying the appropriate sandhi rules.
The system considers only two oblique suffixes 'த்த்', 'அற்று' suffix. After removing each

oblique suffix the root has to be adjusted. After removing the oblique suffix 'thth, the root word requires a 'ம்' at the end to get its base form.

e.g. 'மரத்தை': 'மரம் + த்த் + ஐ'

Normally oblique forms are followed by cases.
The suffix 'கள்' is the main plural marker for Tamil. It has some variants like 'ட்கள்' 'ங்கள்' 'ற்கள்'. Removing the morpheme is little bit complex because some root nouns may end with 'கள்', for example 'மக்கள்' 'மகள்'. Because of the possibility all nouns ending in 'கள்' are maintained as a list and checked against before removing 'கள்'. After removing the 'கள்' suffix the system should modify the root word to get its base form.

Handling of verbs:
The simplest form of the verb is obtained by the addition of person, number, gender (PNG) markers along with tense markers. The verbs in Tamil are classified into nine classes depending upon the present, past and future tense markers they take. The markers also depend on the PNG feature that the verb takes. Normally the PNG features of the verb matches with the PNG feature of the subject of the sentence.

In Tamil, auxiliaries indicating aspect, mood etc. are added to the main verb to form a compound form of the verb. The main verb occurs as the first part of the compound form and is in either verbal participle or infinitive form depending upon the auxiliary that occurs immediately after it. The same is true for all auxiliaries except the last one. The last auxiliary takes on the appropriate PNG markers. In the case of some auxiliaries the PNG markers are not added since the auxiliary occurs in a neutral form. In the design of the analyzer nearly thirty auxiliaries are tackled.

Before adding the auxiliary suffix to the verb/auxiliary, the verb/auxiliary have to be changed into either verbal participle form or a infinitive form, which is decided by the next auxiliary. The verbal participle is the tenseless non-finite verb form. It has both a positive and a negative form. The positive verbal participle is formed by the affixation of the verbal participle suffix to the verb stem. The verbal participle suffix is either the combination of the past tense marker of the verb and the enunciative vowel 'u' or the past tense marker alone. The infinitive is formed by the affixation of the infinitive suffix to the verb stem either 'a' in case of weak verb, or 'ka' in case of middle verb, or 'kka' is case of strong verb.

The adjectival participle is the only non-finite verb form that distinguishes tense. Tamil has a past, present and future adjectival participle. The past or present adjectival participle is formed by adding either the past or present tense allomorph to the verb stem and then adding the adjectival suffix 'அ'.

The dealing of auxiliaries is done by grouping them into five categories depending on the suffixes that precede them. The final auxiliary, which takes a PNG and tense marker if any is removed. Then each auxiliary is removed in turn till the main verb is reached.

Sample output of morphological analysis:

நடு
சிலை
போல

துணைவன்
அன்பு
துணைவி

After the main root words of the noun and verbs are found they are taken to the zone identification where it analyses for the presence of any key words that relate to the metaphors

Metaphor zone identification:

The identified keywords are automatically loaded to the template that is generated . Also the noun that is present before the keyword is found by backtracking the sentence and also all the nouns that appear after the keyword are also taken into account .For this a metaphor rule base is created .

Rule base creation:

A rule base used for identifying the metaphor type possible for the key word that occurs in the text.

Sample Contents of rule base:

At present, the verbs list includes both cognitive as well as non-cognitive verbs. Examples of verbs include pAr (to see), kelY (to listen), vA (to come), thEtu (to search), piti (to catch), po (to go), kal (to learn), etc.

The selectional restriction rules are stored as follows:

Verb=>subject_category;subject_case=>object_category;object_case.

When a verb does not take any object, the keyword [no_obj] is used to denote the same. In addition to the subject and object categories, the rule also contains the appropriate case markers to be used for the subject and object nouns. This additional information is stored for use by the Morph Generation component.

Some examples of selectional restriction rules are given below:

pAr=>[+living,+animate,+vertebrate,+mammal,+human];NOM=>[no_obj]
pAr=>[+living,+animate,+vertebrate,+mammal,+human];NOM=>
[+living,+animate,+vertebrate,+mammal,+human];
piti=>[+living,+animate,+vertebrate,+mammal,+human];NOM=>
[living,+concrete,+movable,+artif
act,+solid,+instrument,- vehicle,+implements];NOM
3piti=>[+ living,+animate,+vertebrate,+mammal,+human];NOM=>[no_obj]
similarly metaphors such as

uvamai ani:

[subject=>+living, +animate,mammal,human][pola, pondra ,oppa ,poraya =>[uvamai urubu]
[object=>+living,+animate,mammal,human,concrete,movable,edible]

pinvarunilai ani:

[verb=> +concrete+,movable, +existing]=different meaning => sol pinvarunilai ani
any word[verb]=>[repetition,+same meaning]=>porul pin varu nilai ani

uruvaga ani:
subject=>[+living,+animate,+human,+concrete,+movable][verb]obj=>[+living,+animate,+human,+concrete,+movable]=>word meaning =>same[+,multiple] => uruvaga ani.

Semantic processing:

Next level is the semantic processing of the nouns that relate to the keyword found . Wordnet is used to find the the meaning of the nouns considered. The primary and secondary meaning of the nouns are obtained from the wordnet. There are also possibilities of occurance of more than two meaning and out of which only the most commonly used meaning are chosen and they are updated to the template.

Table 1: semantic processing primary level

Key-word	Obj1	Obj2	Primary meaning	Secondary meaning	what , When , Where How, Why , Color texture , Other features
போல்	சிலை	துணைவன், துணைவி	கல்லில் செதுக்கபட்ட உருவம்	உலோகத்தை உருக்கி செய்தது	
	துணைவன்,துணைவி	சிலை	கணவன், மனைவி	நண்பன், தோழி	

Also its other properties are obtained from the text if any specified such as where it occurs, when , how , and some others physical properties such as color, texture and so on. There may be possibilities of ambiguity between the primary and secondary meaning , the sense in which the words are used need to be found. For this we, do word sense disambiguation**.**

Word sense disambiguation:

The disambiguation algorithm is improved by using better semantic relationships from Word Net. Also it is intended to add more lexical categories such as verb. Also more lexical sources, dictionaries , thesaurus are to be used for disambiguating. An appropriate ontology is developed based on the words and their relationships that exists between them.

Ontology development:

The ontology that is being in use is the Nyaya Ontology. This ontology provides the reasoning and relationships between the words. But , this ontology does not concentrate on the context based reasoning. Therefore , a new ontology has been developed taking into account the some six attributes : substance, quality, activity, generality,particularity and inherence.Also the relations such as isa, has, and so on are established.The sense and the relations in which it is used is identified from the ontology developed

Table 2 : word sense disambiguation template

How, Why,Color, texture, Other features	Where	When	what	Secondary meaning	Primary meaning	Obj2	Obj1	Keyword
	ஊஞ்சலில்			உலோகத்தை உருக்கி செய்தது	கல்லில் செதுக்கப்பட்ட உருவம்	துணைவன், துணைவி	சிலை	போல்
			அன்பு	நண்பன், தோழி	கணவன்,மனைவி	சிலை	துணைவன், துணைவி	
		சிலை	கூடத்து நடுவில்	ஆடும்			ஊஞ்சலில்	

Semantic level processing – advanced level:

In this level, the obtained relations and senses are compared with the rule base that is created . the metaphor rule base is created by using the writings of Kannadhasan and bharathidasan. This contains the related words , the property to which it is related and the metaphors that occurs in it. The sample metaphor knowledge base contains

Table 3 : Metaphor Knowledge Base

ID	word1	word2	property	metaphor
4	சிலை	துணைவன்,துணைவி	உருவம்	உவமை
5	அரசன்	மதி	வழிகாடுதல்	உவமை
6	புதையல்	பாவை	மதிப்பு	உவமை
7	மதி	முகம்	பொலிவு	உவமை
8	சிலை	அம்பு	கூர்மை	உருவகம்

The metaphors are obtained from the knowledge base and then the template is filled with the properties obtained. After the template is filled, the metaphor ontology and the metaphor knowledge base are updated if required .

5. Conclusion:

Thus this paper gives the system that helps the users to identify the metaphors such as parables , similies, polysomy, synonym types. This forms an initiative to the further research in the other fields. Tamil researchers will be benefited more with this initiative work which may even form a base to other researches such as criticism. semantic interpretation and in the field of pragmatics.

References :

1. Roberto Navigli and Mirella Lapata,2010, An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation,IESE 2010, University of Edinburgh,UK.
2. Anandan. P, Ranjani Parthasarathy, Geetha T.V.2002. *Morphological Analyzer for Tamil, ICON 2002,*RCILTS-Tamil, Anna University, India.
3. Shady Shehata 2009, A WordNet-based Semantic Model for Enhancing Text Clustering, IEEE2009, University of Waterloo, Ontario, Canada
4. Banu M, Karthika C, Sudarmani P and Geetha T.V. , 2007, Tamil Document Summarization Using Semantic Graph Method, *ICON2007,Anna University, India*
5. T.H. Ng, "Getting Serious about Word Sense Disambiguation," Proc. ACL SIGLEX Workshop Tagging Text with Lexical Semantics:Why, What, and How?, pp. 1-7, 1997.
6. K. Saravanan, Ranjani Parthasarathi, T. V. Geetha, "syntactic Parser for tamil",International Conference,2003
7. D. Yarowsky and R. Florian, "Evaluating Sense Disambiguation across Diverse Parameter Spaces," Natural Language Eng., vol. 9, no. 4, pp. 293-310, 2002.
8. Michael A. Covington, "A dependency parser for free and fixed order Languages ", University of Georgia Athens. 489
9. Xavier Carreras and Llu´ıs M`arquez, *Introduction to the CoNLL-2005, Shared Task: Semantic Role Labeling*,Proceedings of CoNLL, 2005.
10. Xavier Carreras and Llu´ıs M`arquez, (2004) *Introduction to the CoNLL-2005, Shared Task: Semantic Role Labeling*,Proceedings of CoNLL, 2004.
11. Andrew McCallum, Dayne Freytag, and Fernando Pereira, *Maximum Entropy Markov Models for Information Extractionand Segmentation.* 17th International Conf. on Machine Learning, 2000.
12. Ting Liu, Wanxiang Che, Sheng Li, Yuxuan Hu and Huaijun Liu, *Semantic Role Labeling System using Maximum EntropyClassifier,* CoNLL, 2005.
13. Kadri Hacioglu and Wayne Ward, *Target word detection and semantic role chunking using support vector machines,* 2003.
14. Budditha Hettige,Asoka S. Karunananda, A Morphological Analyzer to Enable English to Sinhala Machine Translation, ICIA , IEEE , 2006
15. Xia Wenhong, Reading Based on Metaphor Theory in Cognitive Linguistics, Second International Conference on education and training, 2009
16. Zou Qin. "The Old Man and The Sea" in the "biblical" Metaphors [J]. Of foreign literature, 1993 (4) :16-24.

Vaasagi – The Story Analyzer for Tamil

Lavanya.P, Siva Shankari.M , Subhatra Priyadarshini.R and Mahalakshmi G.S.
Department of Computer Science and Engineering,
Anna University, Chennai-600025
mahalakshmi@cs.annauniv.edu, gs_maha@yahoo.co.in

Abstract - *In this paper we present the design and implementation of Vaasagi-The Story Analyzer, which analyzes input Tamil short stories and generates a suitable title, moral and the type of the story. The title generated contains either the description of the main character in the story or a meaningful phrase related to the story. Moral or theme of the story is expressed by proverbs. The story is made attractive by including an appropriate picture from the image corpus. The keywords in the story are extracted for analysis and thereby the essence of the story is accessed.*

Keywords: Natural Language Analysis, Natural Language Understanding, Intelligent Information Extraction, Intelligent Agent

1. Introduction

Vaasagi - The Story Analyzer assists the user in analyzing the theme of the story. The system automates the process of story analysis by saving the user's time spent in reading the entire story. The entire process consists of four modules namely story type identification, title generation, and moral generation and image insertion. The first module categorizes the given story under a predefined story type. The second module generates an appropriate title for the story. The third module generates the moral conveyed in the story. The fourth module inserts a picture related to the story to make it more interesting. The entire system is supported by domain corpus containing domain words, the associated story types, morals and relevant images. The TAB-Tamil is used for encoding and the font used in developing Vaasagi is TAB-Inaimathi. The Syntactic Tamil Parser has been used to identify the parts of speech and Tamil Morphological Analyzer has been used to analyze the root words of the parsed output. Bigram and Trigram algorithms have been designed and implemented for title generation.

Even though, much advancement has already been achieved on story analysis for English [2], the research on Tamil story analysis is still a novel field. Like other languages, analyzing Tamil stories require Named Entity Recognizing algorithms, which is currently under research. In this paper, we

discuss the model and implementation of Tamil Story Analyser for short stories. In this system we have used Vaanavil Parser [3] for Tamil. It accepts a sentence as an input and generates the syntactical tree representing the structure of the sentence. It tackles both simple and complex sentences by making use of phrase structure grammar and look-ahead to handle free word order. This tool uses the Atcharam - morphological analyzer [1] to obtain the root word. Apart from implemented results of Vaasagi, we also discuss certain issues in applying our system for long stories and novels.

2. Design of Vaasagi

This section explains the overall flow of the design of Vaasagi, the Tamil short story analyzer (Fig 1). The name 'Vaasagi' means 'a female reader' in Tamil. The system attempts to analyse the input stories like how a reader would analyse and therefore, the name Vaasagi. Following sections discuss about the module-wise functionalities and operating constraints of the system. The system framework logically has four sections to deal with namely Story Type Identification, Story Title Generation, Moral Generation and Image Insertion.

2.1 Domain Dictionary

After an extensive study of a large number of Tamil short stories [3,4] a Domain Dictionary has been built for Vaasagi. The dictionary has the following contents: Keywords, Story Types, Morals and Pictures. Keywords are a large set of significant words, which commonly occur, in any Tamil short story. The words in the story are compared with these keywords which help in choosing the appropriate story type, moral and picture.

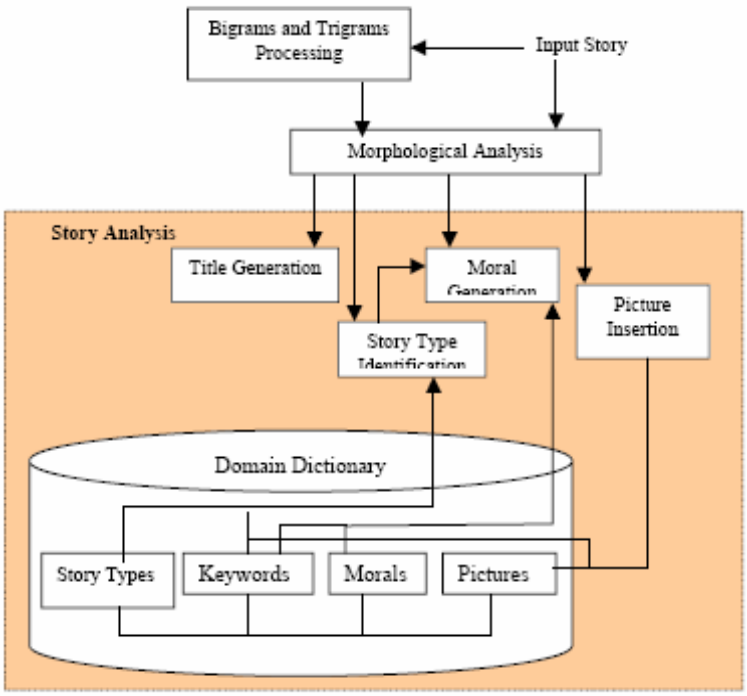


Figure 1. Design of Tamil Short Story Analyser, Vaasagi

The Domain Dictionary has a set of predefined story types under which the given story is categorized. The keywords associated with each type are also specified in the dictionary. The dictionary also has a set of predefined morals in the form of Tamil proverbs. The story type and the keywords associated with each moral are also specified in the dictionary. A large collection of pictures for inserting into the story is maintained in the dictionary. The keyword associated with each picture is also specified.

2.2 Title Generation

While constructing a title for a short story it is appropriate to have only two to three words in the title. For a two-word title to be meaningful it must have either the 'Adjective + Noun' structure or the 'Noun + Noun' structure. For the three word title to be meaningful it must have the 'Noun + Relative Participle + Noun' structure. In Vaasagi, we have considered only the two structures 'Adjective + Noun' and the 'Noun + Relative Participle + Noun' structure. For generating the title Bigrams and Trigrams algorithms are involved.

2.2.1 **Bi-gram Algorithm.** The probability of occurrence of the consecutive pair of words (bi-grams) is calculated and the meaningful pair in the story is taken as the title of the story.

```
Bigram (Story S)
1.Preprocess S to remove the punctuation marks
2.For each pair of adjacent words(w1,w2)
        b = w1 + w2                          * b - bigram
        Add b to B
3.For each b in B
        Find f, the frequency of occurrence of b
4.Frequency of highly repeated b, max = Maximum(f)
5.Find fb with frequency max                 *fb – bigrams with highest frequency max
6.For each fb
If (Analyse(fb) == 'Adjective + Noun')
Title = fb
```

2.2.2 Tri-gram Algorithm. Like Bi-grams, meaningful consecutive occurrence of three words is identified. The algorithm for Tri-gram Identification is given below.

```
Trigram(Story S)
1.Preprocess S to remove the punctuation marks
2.For each set of three adjacent words (w1,w2,w3)
t = w1 + w2 + w3                             * t - trigram
Add t to T
3.For each t in T
Find f, the frequency of occurrence of t
4.Frequency of highly repeated t,
max = Maximum (f)
5.Find ft with frequency max                 *ft – trigrams with highest frequency max
6.For each ft
        If (Analyze (ft) == 'Noun + Relative Participle + Noun')
        Title = ft
```

In Bigrams algorithm the most frequently occurring adjacent pair of words with the structure "Adjective+ Noun" is displayed as the title. In Trigrams algorithm the most frequently occurring set of three adjacent words with the structure "Noun + Relative Participle + Noun" is displayed as the title. The probability for the repeated bigram or trigram to be meaningful is higher in Tamil compared to English. Implementing the algorithm for Tamil stories is therefore more advantageous than implementing the same for English stories.

2.3 Story Type Identification

The Domain Dictionary has a set of predefined story types and each type has a set of associated keywords. The story is fed into the Morphological Analyzer and all the root words are matched with the keywords. The story type with the highest number of matches selected at the output (refer Figure 5).

2.4 Moral Generation

The Domain Dictionary has a set of predefined morals in the form of proverbs. Each moral has a set of associated keywords and story type. The keywords and the story type are identified and the moral related to them is chosen.

2.5 Picture Insertion

The parsed output of the story is analyzed using the Morphological analyzer and the root words of the nouns are collected. The collected words are compared with the keywords in the database. The picture for the noun with the highest frequency of matches is inserted in the story.

3 Results

Figure 4 Input Story with Picture Insertion

Vaasagi is implemented using Java. A large number of Tamil short stories [4][5] were studied to build the domain dictionary. After an extensive study of the titles of the existing stories a conclusion was made that the title was either of the two forms - 'Noun + Relative Participle + Noun' or 'Adjective + Noun'.

The input Tamil story (Fig. 4) is about the 'The Crane and the Fox' which is more familiar among kids. For the Tamil Story in Fig.4, the story type and moral generated can be seen from Figure 6. The intermediate process taking part is shown in the figure 5.

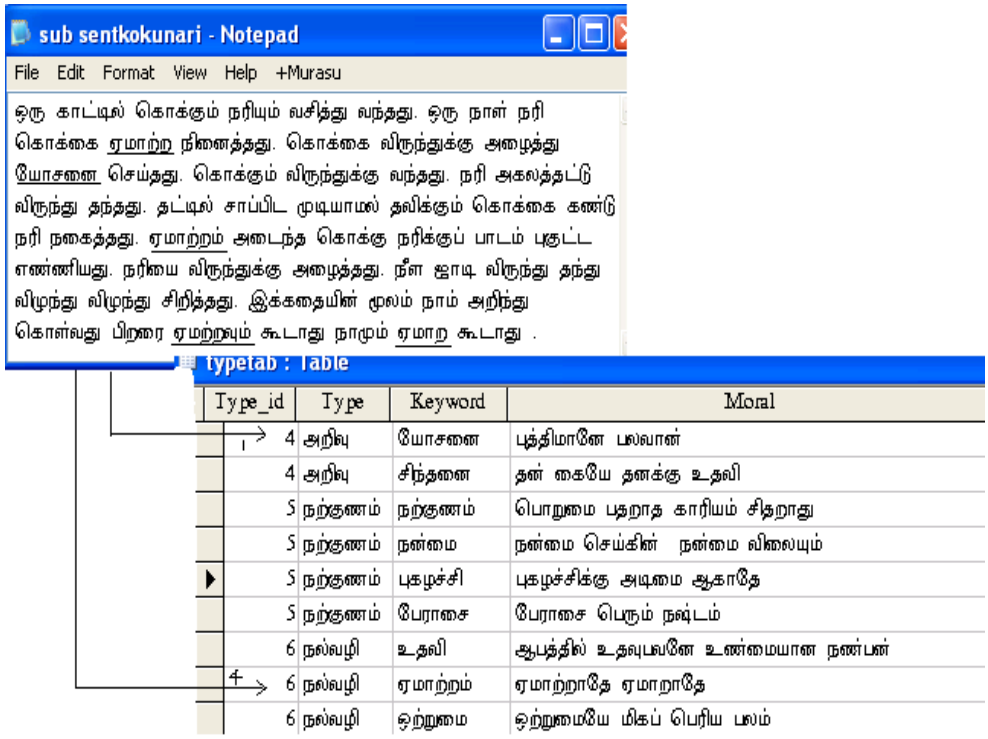


Figure 5 Story type and Moral Generation

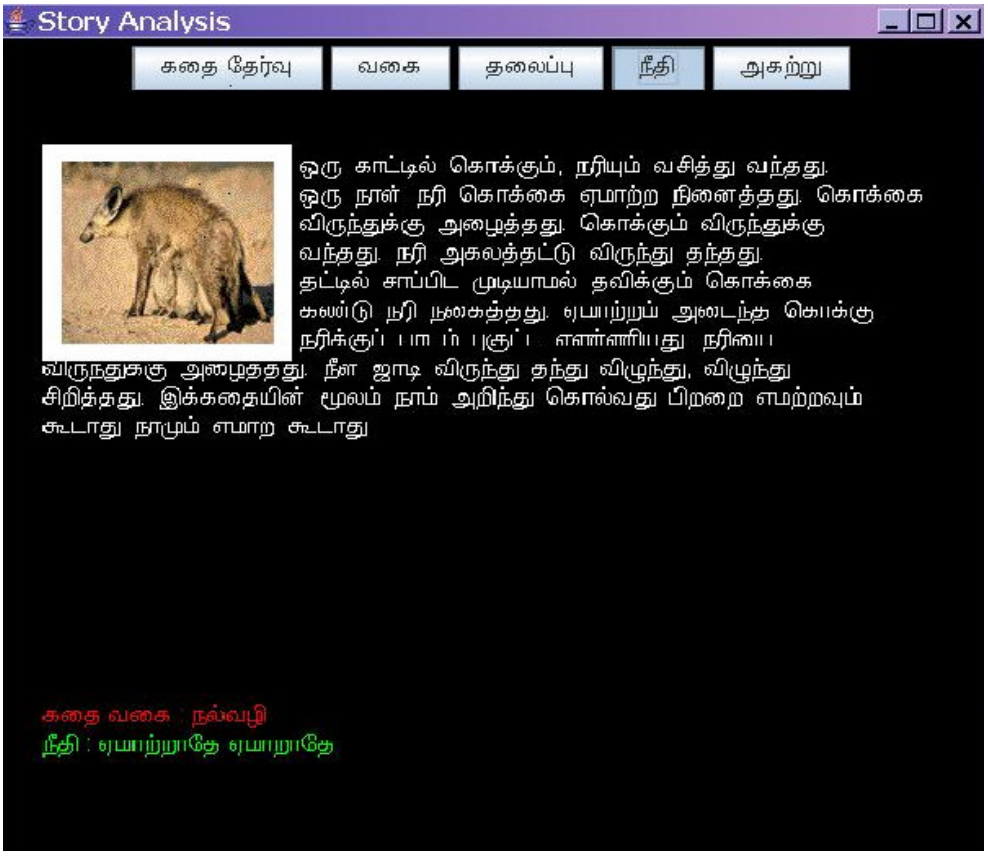


Figure 6 Story Type and Moral Output

The story type and moral are generated by consulting with the domain dictionary of Vaasagi. For the given input story the title generated by Vaasagi is shown in the figure 7.

Figure 7 Title of the story identified by Vaasagi

4. Limitations of Vaasagi

Vaasagi does not generate the title in the format "Noun+Noun". This is because difficulty arises in including the conjunctions in between two nouns by analyzing the story semantically. Named Entity Recognition algorithm has to be implemented which can be used for identifying the place that can be included in the title. However, Title generation using bi-grams and n-grams algorithm, which is implemented in Tamil stories may not be meaningful for English sto-

ries, since the algorithm does not check for semantics.

5. Future Enhancements

To make Vaasagi applicable for novel analysis, the domain dictionary must be extended to include more words belonging to different slangs and different periods of use. For example to analyze a Tamil novel belonging to the ancient Sangam period the dictionary must have those ancient Tamil words which may not be in use now. And for Tamil poems the different styles of poem writing must be studied well because poems do not take grammatical syntax into account. Moreover, given a character name in the story, generating its description can be added as an enhancement. This can be done by picking out the adjectives associated with the character name and framing the description in a few sentences. Also, Named Entity Recognition algorithm has to be used to identify the noun as a character name or a place name. This can also be used to generate the title of the story when there is also a chance of including the place name into the title. In our current system the story is analyzed and its type is found and given in a single word. Further the theme of the story can be found and expressed in a few sentences, which is left as a future work. Tamil Story can be converted in to speech using text to speech conversion tool in the future.

In Vaasagi, we have used the Bi-gram and Tri-gram algorithm for generating the story's title. For a short story it appears good to have a title in two or three words only. So we have not implemented up to N-gram algorithm. The actual n-gram algorithm takes each unigram, assigns a weight to it and checks for the frequently occurring meaningful n-grams. In our system we have simplified the algorithm by finding out the more frequently occurring bi and tri-grams and passing them through the morphological analyzer to verify the grammatical and syntactic correctness of the title. The semantic correctness of the title should be verified and considered in future. Also when the bi-gram or tri-gram does not form a grammatically correct sense a few stop words can be included between the keywords to give a meaningful title.

One Picture for each story is inserted as only short stories are considered here. As a part of enhancement when large stories and novels are taken into account a relevant picture for each page can be inserted. Though our system works well for all Tamil short stories, when the same is extended for English stories a few problems arise. The bi-gram and tri-gram structure that we have assumed for a Tamil title does not hold good for an English title. Moreover in English the probability for the repeated n-gram to be meaningful is very less when compared to Tamil.

6. Conclusion

We have discussed in this paper about the design and development of Vaasagi – the Story Analyzer that greatly assists the user in story analysis and automates the processes of title generation, moral generation, and type identification and picture insertion. In our current system only short, simple Tamil stories are analyzed and the outputs are generated. This can be extended for analyzing novels and poems in the future. The major challenge we expect is in applying efficient Named Entity Recognition algorithms and the collection and usage of period dictionaries for any morphologically rich language like Tamil.

7. Acknowledgements

The paper was extensively discussed to develop a non-existing project Tamil story analyzer. The first three authors extend their special thanks to Ms. G.S. Mahalakshmi, Anna University and Mr. Lakshmanapandian for their enthusiasm, active participation and support. We would like to thank RCILTS-Tamil, Anna University, Chennai for providing the Vaanavil Tamil

Parser and the Atcharam Morphological analyzer to carry out the project Vaasagi succesfully.

8. References

[1] P Anandan, K. Saravanan, Ranjani Parthasarathi and T.V.Geetha, Morphological Analyzer for Tamil , International Conference on Natural Language Processing (ICON) , Mumbai, India. December 18-21, 2002.

[2] Harry Halpin, Johanna D. Moore and Judy Robertson, Towards Automated Story Analysis Using Participatory Design, Proceedings of the ACM Workshop on Story Representation, Mechanism and Context, ACM Multimedia, NY, NY, October 15, 2004.

[3] K. Saravanan, Ranjani Parthasarathi and T.V. Geetha, Syntactic Parser for Tamil, Tamil Internet Conference, Chennai, India, December 2003.

[4] Talk Tamil – Stories and Rhymes for Children, http://tamilforkids.tripod.com/index.html

[5] N. Udayakumar, Tamil Kutti Kathaigal, http://tamil-kutti-kathaikal.blogspot.com

OFFLINE TAMIL HANDWRITING RECOGNITION FROM DOCUMENT IMAGES

S.Abirami, B.Raghavardhini, V.Sasireka and R.Sutha,
Department of Information Science and Technology, Anna University, Chennai—25.
abirami_mr@yahoo.com, raghavardhini@gmail.com, reka.cse2198@gmail.com,
mailtosutha@yahoo.com

Abstract

Handwriting Recognition is the ability of a computer to receive and interpret handwritten input, which translates the document into the digital data used by the computer. This paper tries to analyze a quick survey of Offline Handwriting Techniques.

1. Introduction

Handwriting Recognition entails Optical Character Recognition. It handles formatting and performs correct segmentation into characters. It may take place in one of two ways, either by scanning of written text or by writing directly on to a peripheral input device. Handwriting Recognition has active community of academics. Active Areas includes,

*Online Handwriting Recognition.

*Offline Handwriting Recognition.

Online Handwriting Recognition deals with the movement of pen tips which is sensed and recognized. In offline Handwriting Recognition , the image of written text is scanned from the piece of paper. It falls in the area of Optical Character Recognition (OCR), where only the static information is available.

Offline Character Recognition usually known as Optical Character Recognition because the image of writing is converted to bit pattern by optically digitizing device such as scanners. OCR usually requires users to scan the documents containing handwritten text and then extracts individual characters from scanned document to letter code. Offline Handwriting Recognition involves four steps as stated below:

Preprocessing step involves removal of noise, skew correction, slant correction, binarization. The major step is to segment the document which is scanned into lines, words and characters. In Feature Selection step different people use different model to extract the characters. Feature Extraction deals with how the particular character is extracted from the documented image. Some uses graph techniques to extract the feature. In Feature Recognition step the feature which was extracted in previous step is recognized by using HMM(Hidden Markov Model).

1.1 Importance of Offline over Online

The main disadvantage of Online Handwriting Recognition is that it is used to the one, only who has the ability to be connected to the internet, but offline is used to the one, who is not currently con-

nected to it.

Offline technology is widely used in business which processes lots of handwritten documents, like insurance companies.

In Online Handwriting Recognition, the characters should be recognized at that instance, but in Offline, the character can be recognized at anytime after documentation.

2. Related Works

Many authors researched about Offline Handwritten Recognition and came up with several advantages and disadvantages. They are discussed in this paper one by one:

Author JAGADESH [1] [2][3] used Radial Bass Function Neural Network (RBFNN) to retrieve the actual character, where the accuracy of OCR is increased. The advantages was over efficiency and slant correction but as he used only the basic Preprocessing steps, it resulted in finding only 7 Tamil characters.

Author SUTHA [4] used forward Multi Layered Perception (MLP) network to recognize handwritten tamil characters. Though had advantages over accuracy, they lacked in large sized characters. They recognized standard sizes of 2x2, 8x8, 32x32, 48x48, 64x64 of totally 34 samples.

Another work published by AKSHAY APTE [5] identifies the character as the one which it is closest to in terms of Euclidean distance. Hough Transform is also used ,non-uniform thickness of the character strokes and false branching also removed. Accuracy is based only on font size 18, 20 & 22. This lead to major disadvantages while using characters less than 18.

ALEX GRAVES [6] did Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks which uses HMM's Neural Network Hybrid model. He came up with testing of large fonted text, but the major disadvantage is that the input character must meet the assumptions made by HMM (Hidden Markov Model).

In other paper[7], YANIKOGLU separator algorithm is used where there are 8 fixed angles. Here we choose a subset of 4 of the separator angles to segment the word. He came up with large data samples with multiple writers and resulted in high performance. He also met up with disadvantage, where he lacked in feature recognition step where HMM's performance drowned while recognizing large fonted text.

3. Conclusion

Thus the paper concludes that many authors researched many papers, but they recognized only alphabetical characters and omitted punctuations, numbers and special symbols. Few authors succeeded in recognizing all font sized characters but they lacked, since they observed a small number of data samples. Though they were solved by other authors, they lacked in large sized fonts and misfigured characters.

Thus to overcome the above problems,more research has to be done especially for Tamil characters, where all 247 characters, consonants, vowels, punctuations, all sized fonts has to be extracted and recognized with accuracy and efficiency.

4. References

[1] Jagadeesh Kannan .R, Prabhakar. R and Suresh R.M. (2008), "Off-Line Cursive Handwritten Tamil Character Recognition", WSEAS Transactions on Signal Processing, Vol 4, No.6, pp.351-360.

[2] Jagadeesh Kannan .R, Prabhakar. R and Suresh R.M. (2008), "An improved Handwritten Tamil Character Recognition System using Octal Graph", Journal of Computer Science, Vol 4, No.7, pp.509-516.

[3] Jagadeesh Kannan .R, Prabhakar. R and Suresh R.M. (2009), "A Comparative Study of Optical Character Recognition for Tamil Script" , European Journal of Scientific Research, Vol.35, No.4, pp.570-582.

[4] Sutha.J and Ramaraj. N (2007), " Neural Network Based Offline Tamil Handwritten Character Recognition System", Proceedings of the International Conference on Computational Intelligence and Multimedia Applications, Vol.2, pp.446-450.

[5] Akshay.A and Harshad. G, "Tamil Character Recognition Using Structural Features".

[6] Alex. G and Jurgen.S, "Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks".

[7] Yanikoglu. B.A., Sandon. P.A (1994), "Recognizing Off-Line Cursive Handwriting", Proceedings of the Computer Vision and Pattern Recognition, pp.397-403.

கணிணியியலில் தமிழ்ப் பயன்பாடு

இலக்குவனார் திருவள்ளுவன்
7/1, மாவு ஆலை முதல் தெரு,, மயிலாப்பூர், சென்னை 600 004
பேசி: 98844 81652 / 044 6499 3317

எத்துறையாயினும் அத்துறையறிவு தாய்மொழியில் வெளிப்படுத்தப் பட்டால்தான் அம்மொழியினருக்கு முழுப் பயன்பாடு கிட்டும்; அத்துறையும் சிறப்பான வளர்ச்சியை எட்டும். அந்த வகையில் கணிணியியலில் முழுமையும் தமிழ் மொழி பயன்படுத்தப்பட்டால்தான் கணிணியியல் முழு வளர்ச்சியடைந்ததாகும். இப் பொழுது அந்நிலை இன்மையால் அதனை வலியுறுத்துவதே இக்கட்டுரையின் நோக்கம்.

கட்டுரையாளர்களும் நூலாளர்களும் இதழாளர்களும் தமிழில் கணிணியியலை விளக்குவதில் பெரும் ஆர்வம் காட்டி வருகின்றனர். ஆனால், அவ்வாறு விளக்குவதில் உள்ள ஆர்வம் தமிழைப் பயன்படுத்துவதில் இல்லை. கணிணிக் கலைச் சொற்களாக நல்ல தமிழ்ச் சொற்கள் இருப்பினும் அதைப் பயன்படுத்தாதவர்களும் உளர்; தமிழ்க் கலைச் சொற்கள் இன்மையால் அயல் மொழிச் சொற்களைத் தமிழ் வரிவடிவில் எழுதுவோரும் உளர். எனவே, இந்நிலை மறைந்து இன்னிலை தோன்றப் பின்வரும் செயற்பாடுகளில் ஈடுபட கணிணியியலாளர்கள் முன் வரவேண்டும்.

1. இருக்கின்ற தமிழ்க்கலைச் சொற்களைப் பயன்படுத்தல்.
2. இல்லாதவற்றிற்குப் புதிய கலைச் சொற்களை உருவாக்கல்.
3. நேர் பெயர்ப்புச் சொற்களைத் தவிர்த்தல்
4. ஒலி பெயர்ப்புச் சொற்களை விலக்குதல்
5. நடைமுறையில் பொருந்தாச் சொற்கள் இருப்பின் தக்க கலைச் சொற்களை உருவாக்கல்

6. சொற்சுருக்க எழுத்துகளைத் தமிழில் குறிப்பிடல்.
7. தலைப்பெழுத்துச் சொற்களைத் தமிழில் குறிப்பிடல்.
8. விசைச் சொற்களைத் தமிழில் குறிப்பிடல்.
9. கணிமொழிக் கட்டளைகளைத் தமிழில் அமைத்தல்.
10. கணிப்பொறியின் பகுதிகளைத்தமிழிலேயே குறித்தல்

சுருக்கமாகச் சொல்வதாயின், தமிழை மட்டுமே தெரிந்த ஒருவர் கணிணியியலை நன்கு புரிந்து கொள்ளும் அளவிற்குத் தமிழை மட்டுமே பயன்படுத்திக் கணிணியியலை விளக்கும் காலம் விரைவில் வர வேண்டும்.

தமிழ் ஆர்வலரான கட்டுரையாளர் சிலர், தத்தம் படைப்புகளில் நல்ல தமிழ்ச் சொற்களைக் கையாண்டுவரினும், கணிணித்தமிழ் அறிஞர் சிலர் நல்ல தமிழ்ச் சொற்களைத் தொகுத்து அகராதிகள் வழங்கியிருப்பினும், அவற்றை அறியும் தேடுதல்-வேட்கையின்றியும், அல்லது அறிந்தாலும், அத் தமிழ்ச்சொற்களைப் பயன்படுத்த வேண்டும் என்ற கடப்பாட்டு உணர்வு இல்லாமலும், கணிணித்துறையினர் ஆங்கிலச் சொற்களையே கையாண்டு கணித்தமிழ் வளர்ச்சிக்குத் தடையாக இருக்கின்றனர். எனவே, அறிமுகப் படுத்தப்பட்ட கலைச் சொற்களைப் பயன்படுத்த வேண்டும் என்ற உணர்வு படைப்பாளர்களுக்கு வர வேண்டும்.

கலைச் சொல்பெருக்கத்திற்குத் தடையாக இருப்பது சொல்லைப் புரிந்து கொண்டு படைக்காமல், 'சொல்லுக்குச் சொல்' என்ற நேர்முறையில் ஆக்கப்படும் கலைச்சொற்களும் தமிழ்ச் சொற்களைக் கையாளாமல் ஒலிபெயர்ப்புச் சொற்களாக மூலச் சொற்களைக் கையாளுமாகும். இவற்றை உணர்ந்து, புத்தம்புதுக் கலைச் சொற்களை நாளும் உருவாக்கவும், உருவாக்கப்பட்ட கலைச் சொற்களைப் பயன்படுத்தவும் நாம் முன்வர வேண்டும். கலைச் சொற்கள் சுருங்கியனவாகவும், அவற்றின் அடிப்படையில் மேலும் புதிய கலைச் சொற்களை ஆக்க வாயிலாகவும் அமைய வேண்டும்.

அறிவியல் துறைகளைப் புரிய வைப்பதற்கும் அறிந்து கொள்வதற்கும் கையாளப்படும் கலைச்சொற்கள் தன்-விளக்கமாயும் எளிமையாயும் அமைய வேண்டும். அவ்வாறு இல்லாச் சூழலில், தவறாகப் புரிந்து கொள்ளவோ விளங்காமல் குழப்பம் அடையவோ வாய்ப்புகள் ஏற்படுகின்றன. எனவே, விரைந்து வளரும் கணிணியியலில் துறைவளர்ச்சிக்கேற்ற கலைச்சொல் பெருக்கமும் அமைய வேண்டும். கலைச்சொற்கள் பெருகுவதற்கான தடைகளை நீக்க,

1. ஒவ்வொருவர் ஒவ்வொரு வகையாகக் கையாளுதல்.
2. சில நேரங்களில் ஒருவரே வெவ்வேறு வகையாகக் கையாளும் நேர்வும் உள்ளது.
3. நடைமுறைக்கு நல்லசொற்கள் வந்துவிட்டபின்னும் கொச்சையாகக் கையாளுதல்.
4. சுருங்கிய கலைச்சொல்லாக இல்லாமல், விளக்கச் சொற்றொடராகக் கையாளுதல்
5. பொருள்விளக்கமான கலைச்சொல்லைக் கையாளாமல், நேருக்குநேர் மொழி பெயர்த்துக் கையாளுதல்.
6. தவறான சொல்லாக்கத்தைக் கையாளுதல்
7. சொல்லும் அதன் பயன்பாட்டுக் காலத்திற்கு ஏற்பப் பொருள்மாற்றம் அடைகிறது. எனவே, இதுதான் இச்சொல்லுக்குப் பொருள் என்னும் பிடிவாதம் இன்றிச் சூழ க்கேற்ற பொருள் விளக்கத்தைக் கையாளாமை.

ஆகியவற்றை அறவே நீக்குதல் வேண்டும்.

ஆதலின், நடைமுறையில் இல்லாத கலைச்சொற்களுக்குத் தேடுதல் வேட்கையுடன் புதிய கலைச் சொற்களை உருவாக்க வேண்டும். கலைச்சொல் புனையும் ஈடுபாடும் ஆர்வமும் இல்லாதவர்கள் கலைச் சொல் வல்லுநர்கள் மூலம் புதிய கலைச்சொற்களைப் படைக்கத் தூண்டுதலாய் இருத்தல் வேண்டும். அறிமுகமாகியுள்ள கலைச் சொற்கள் உரிய பொருள்தராதனவாகவும் தொடர்போன்றும் அமைந்து இருப்பின் அவற்றிற்கும் உரிய பொருத்தமான சுருக்கமான கலைச் சொற்களை உருவாக்கிப் பயன்படுத்த வேண்டும். ஒவ்வொருவர் ஒவ்வொரு வகையான கலைச் சொற்களைக் கையாளுதலும் ஒருவரே வெவ்வேறிடத்தில் வெவ்வேறு வகையான சொற்களைக் கையாளுதலும் படிப்பவர்களிடையே குழப்பத்தை ஏற்படுத்தி எதிர் விளைவுகளை உருவாக்கும். எனவே, நிலைத்து விட்ட நல்ல சொற்களை மாற்றும் முயற்சியைக் கைவிட வேண்டும். அதே நேரம் நடைமுறையிலுள்ள சொல்லைவிடப் பொருத்தமான கலைச்சொல் அறிமுகப்படுத்தப்பட்டால் பிடிவாதத்துடன் முந்தைய சொல்லையே கையாளாமல் புதிய கலைச் சொற்களைக் கையாளும் மனப்பக்குவமும் வேண்டும். ஓரிடத்தில் பொருத்தமாக உள்ள கலைச் சொல் வேறிடத்தில் உரிய பொருளைத் தராமல் பொருந்தாமல் நிற்கும். எனவே, சொல் இடத்திற்கேற்ற பொருளைப் பெரும் என்பதை உணர்ந்து சூழலுக்கேற்ற கலைச் சொல்லையே பயன்படுத்த வேண்டும். தேவையான இடங்களில் அடைப்பிற்குள் ஆங்கிலச் சொல்லையோ நடைமுறையில் உள்ள சொல்லையோ குறிப்பிடத் தயங்கக் கூடாது.

மூலச் சொற்களையும் தலைப்பெழுத்துச் சொற்களையும் சுருக்க அமைப்புச் சொற்களையும் ஆங்கிலத்திலேயே குறிப்பிட்டால் தவறல்ல என்னும் மனப் போக்கு பெரும்பாலாரிடம் உள்ளது. இதுவும் தவறான நிலைப்பாடாகும். இவையும் தமிழில் இருக்கும்பொழுது கணினியறிவியல் மேலும் எளிமையாகத் திகழும். ஐ.நா. என்பது போன்ற தமிழ்ச் சுருக்கக் குறியீடுகள் பொருளை விளங்க வைக்க உதவுவதை எடுத்துக்காட்டாகக் கூறலாம். தமிழில் இருந்தால் புரியாது என்று சொல்வதெல்லாம் மேலோட்டச் சிந்தனையே!

கணினியியலில் ஆங்கில ஒலிபெயர்ப்பிலேயே கலைச் சொற்களும் தலைப்பெழுத்துச் சொற்களும் எண்ணிலடங்கா அளவு கையாளப்பட்டுத் தமிழ் மொழி சிதைந்து வருவதைப் பலரும் உணரவில்லை. 'மணிப்பிரவாளம்' என்ற பெயரில் மொழிக்கொலை புரிந்து பாழ்பட்ட நிலையிலிருந்து அண்மைக் காலத்தில் மீண்டுவரும் வேளையில் ஆங்கிலக்கலப்பு விளைவிக்கும் தீங்கைப் பெரும்பான்மையர் புரிந்து கொள்ளவில்லை. பிற அறிவியல் துறைகளில் நிகழும் சொல்லாக்கத் தவறுகள்தாம் கணினியியலிலும் நடைபெறுகின்றன. ஆனால், பிற துறைகளுடன் ஒப்பிட முடியாத அளவு கணினியியலில்தான் ஆங்கில ஒலி பெயர்ப்புச் சொற்கள் மிகுதியாகக் கையாளப்படுகின்றன. இவை முற்றிலும் உடனடியாகக் களையப்பட வேண்டும். சுருக்கக் குறியீடுகள், தலைப்பெழுத்துகள் என எந்த வடிவிலும் ஆங்கிலத்தைப் பயன்படுத்தாமல் கீன மொழியிலேயே குறிக்க வேண்டும் எனச் சீன அரசு ஆணை பிறப்பித்து நடைமுறைப்படுத்தி வருகிறது. இது போல் தமிழ்நாட்டரசும் ஆணை பிறப்பித்து நடைமுறைப்படுத்த வேண்டும்.

சொல்லில் உயர்வு தமிழ்ச் சொல்லே என்னும் பாரதியாரின் பொன்மொழியை உணர்ந்துதமிழில் எண்ணித் தமிழிலேயே எழுதத் தொடங்கினால் அரிய கலைச் சொறகளைக் கூட அழகு தமிழில் அருமையாகக் கூற இயலும். தமிழ் எழுத்துகளில் அமைந்தன மட்டுமே தமிழ் என்பது நம் முன்னோர் கூற்று. ஆகவே, தமிழ்ப்படைப்புகளில் அயற் சொற்களும் கிரந்த எழுத்து முதலான அயல் எழுத்துகளும் பயன்படுத்தக்கூடா. இவற்றை ஊக்கப்படுத்துவதற்காக அரசு, தமிழ்க்கலைச் சொற்களைப் பயன்படுத்தும் நூல்களை மட்டுமே பாட நூல்களாக வைக்க வேண்டும்; கலப்பு நடையைக் கைவிட்டு நல்ல தமிழில் எழுதப்படும் நூல்களுக்கு மட்டுமே பரிசுகள் வழங்க வேண்டும். தமிழ்ப்பகைவர்களுக்குப் பட்டங்களும் விருதுகளும் பொற்கிழிகளும் வழங்கி மொழி இன அழிப்பிற்குத் துணை போகாமல் தமிழ் அன்பர்களை மதித்துப் போற்ற வேண்டும்.

கலைச் சொற்களை மட்டும் தமிழில் வழங்கினால் போதுமா? கணிக்கட்டளைகளையும் தமிழிலேயே அமைத்தல் வேண்டும். அதற்கு முதற்கட்டமாகக் கணிணிச் செயற்பாட்டுக் கட்டளைகளைக் குறிப்பிடும் விசைகளின் பெயர்கள் தமிழில் இருக்க வேண்டும்.(கணி விசைப் பெயர்கள்)

Enter Key - புகுவி விசை
Control Key - யாப்பு விசை
Alternate Key - வினை விசை
Delete Key - நீக்கி விசை
Escape Key - விலக்கி விசை
Home Key - ஆதி விசை
End Key - அற்றவிசை
Shift Key - முறைமை விசை
Tab Key - பெயர்த்தி விசை
Number Lock key- எண்தாழ் விசை
Scroll Lock Key - சுருணை விசை
Insert Key - செருகி விசை
Page up Key - ஏற்றி விசை
Page down Key - இறக்கி விசை
Pause Key - நிறுத்தி விசை
Print Screen Key - பதிப்பி விசை
Up Arrow Key - மேலம்பு விசை
Down Arrow Key - கீழம்பு விசை
Left Arrow Key - இட அம்பு விசை
Right Arrow Key - வல அம்பு விசை
Back Space Key - முன்னிட விசை
Functional Keys - செயல் விசைகள்
User Keys - பயனர் விசை
Caps.lock key - முறைமைத் தாழ் விசை

இவை போன்று கட்டளைச் சொற்களையும் தமிழில் அமைத்து இம்முயற்சியை விரைவுபடுத்த வேண்டும் கணிப்பொறியின் பகுதிகளைத் தமிழிலேயே குறித்தல் வேண்டும் அப்பொழுதுதான் கணியியல் குறித்த முழுமையான தமிழ் நூல்களைப் படைக்க இயலும்.

இவையனைத்தையும் தமிழில் அமைக்கக் கணிணியியலாளர்கள் முன்வரின் கணிணியியலில் தமிழ் தலைமையுற்றுத் திகழும். தமிழ்வழியாகக் கல்வி அமையாமையாலேயே நம் நாட்டில் புதிய புனையும் அறிஞர்களும் கண்டுபிடிப்பாளர்களும் உருவாகவில்லை என்பார் செந்தமிழ்ச் செம்மல் பேராசிரியர் சி.இலக்குவனார். கணிணி உலகில் நாளும் அறிஞர்கள் பெருக வாழும் மொழியாம் தமிழில் முழுமையாய் கணிணியறிவியல் அமைய வேண்டும்.

செயல் செய்வாய் தமிழுக்குத்

துறைதோறும் துறைதோறும் சீறிவந்தே என்னும் பாவேந்தர் பாரதிதாசன் கட்டளைக்கிணங்க நாம் கணிணியறிவியலிலும் தமிழ்ப் பயன்பாட்டை முழுமையாகக் கொண்டு வர வேண்டும். அதுவே நாம் செய்யும் எப்பணிக்கும் முதற்பணியாய் அமைதல் வேண்டும்.

உத்தமம்