Tamil WordNet User guide

Table of Contents

1. Introduction

Tamil WordNet is an attempt to build a lexical network for Tamil language along the lines of the Princeton WordNet. WordNet contains information about nouns, verbs, adjective and adverbs and is organized around the notion of a synset. A synset is a set of words with the same parts-of-speech that interchangeable in certain contexts. For a detailed information on Princeton WordNet, refer (Fellbaum 1998).

2. Objectives

The main objective of the project is to build a Tamil WordNet so that it can be used as a tool for enhancing the performance of MT systems involving Tamil. In this, we have attempted to assign each word a set of all possible senses it can take. We have also captured various relationships between the words by networking the sense of these words in an appropriate manner using the relationship as a function. These word-level relations include synonymy, hypernymy, hyponymy, meronymy, holonymy and antonymy. A Machine Translation system having the source language as Tamil can effectively exploit these relations to resolve ambiguities in the text.

In this project, we have captured various semantic relations for 50,000 words in Tamil along with their senses. This is built using a database as a back-end, and a front-end user interface to view the senses and relationships.

3. Resources Consulted:

Tamil WordNet relies on Rajendran's (2001) Modern Tamil Thesaurus, which is based on Nida's (1975) Componential Analysis of Meaning. This work which is also available in the electronic form represents the ontological structure of Tamil vocabulary. Tamil vocabulary is classified into four major domains: entities, abstracts, events and relationals based on the part-of-speech categories. Other resources consulted include Technical Glossaries (English-Tamil), E-Pals Dictionary and Princeton WordNet.

4. Categories in Tamil WordNet

4.1 Nouns

Nouns in Tamil WordNet are organized into several hierarchies, each representing a unique beginner. These multiple hierarchies correspond to relatively distinct semantic fields, each with its own vocabulary. Unique beginner corresponds to a primitive semantic component in a compositional theory of lexical semantics. Noun relations are captured through lexical relations such as synonymy, hyponymy and meronymy. In case of nominal forms, corresponding verb forms are represented as a relation. The following table lists the various relations of nouns that are handled:

Relations	**Example**
Synonymy	*viiTu* ‘house’ - *illam* `house'
Hypernymy-Hyponymy	*paLLi* 'school' – *kalviccaalai* 'educational institution'
Hyponym-Hypernymy	*kalluuri* 'college' – *aracukkalluuri* `govt college'
Holonymy-Meronymy	*ndaaRkaali* 'chair' - *kaal* 'leg'
Meronymy-Holonymy	*cakkaram* 'wheel' to *vaNTi* 'cart'
Related Verb	*paTittal* ‘reading’ – *paTi* ‘read’
Coordinate terms	*kooyil* `temple' – *macuuti* 'mosque'

4.2 Verbs

The following table sums up the lexical relations that are captured in WordNet.

Relations	**Example**
Synonymy	*paTi* ‘read’ – *payilu* ‘read’
Hypernymy	*cuvai* ‘taste’ – *uNar*
Troponymy	*keeL* ‘ask’– *kenjcu* ‘plead’
Nominal	*paruku* `drink’ – *parukutal* `drinking’
Related Noun	*kaNTupiTi* `discover’ – *kaNTupiTippu* `discovery’

4.3 Adjectives and adverbs

In Tamil WordNet, only few root adjectives are listed. There exists derived adjectives and derived adverbs. The following table illustrates the relations that are assigned to adjectives and adverbs.

Relationship	**Example**
Synonyms	*caukkiyamaana* `well-being’ – *aarookkiyamaana* `well-being’
Related Nouns (adjectives)	*azhakaakaana* `beautiful’ – *azhaku* `beauty’
Related Nouns (adverbs)	*cuvaiyaaka* `tasty’ – *cuvai* `taste’

5. Database Design

Tamil WordNet's data is stored in Mysql Database. There are four tables in the database viz. *twn*, *sense, morph* and *frequency*.

twn table is the core file which contains all information pertaining to each word. There are nine fields in this table.

1. Node Index
2. Label
3. Gloss
4. Example
5. Feature
6. English
7. Index Length
8. POS

sense table has information on different senses, parts-of-speech category and its occurrence in the database. There are four fields in this table.

1. Word
2. POS
3. Number of senses
4. Index Numbers

The *morph* and *frequency* tables have information on root words and frequency of each word respectively.

6. Database File Format

Tamil WordNet data files are divided into various categories based on the Parts-of-Speech information. There are sixty-nine noun files, twelve verb files and one adjective/adverb file which are grouped based on the semantic domain. These files are used only in the early stages of WordNet development and are finally combined into two files viz.

1) twn.noun.contents
2) twn.verb.contents

The format of the two data files is given below:

/Relation/Feature/3/4/5/Index/POS/8/lexical item

There are nine fields separated by forward slash in each line. The first field is used to represent the semantic relations such as synonymy, hypernymy etc. The seond field is used for assigning features such as human and non-human. Third, fourth, fifth and eigth fields are left empty for use in providing description, example sentence etc. in future. Sixth field has Index number of each word, separated with a comma. Seventh field has Parts-of-Speech category and the last field contains concept, lexical item. Concepts with more than one word or compound words are separated with an underscore.

Following tables list information on Relation and feature along with the numbers assigned against each.

Relation	***Number***
Hypernym	0
Meronym	1
Holonym	2
Hyponym/Troponym	3
Coordinate Term	5
Nominal	7
Derived Adjective	8
Derived Adverb	9
Related Noun	10
Root Adjective	11

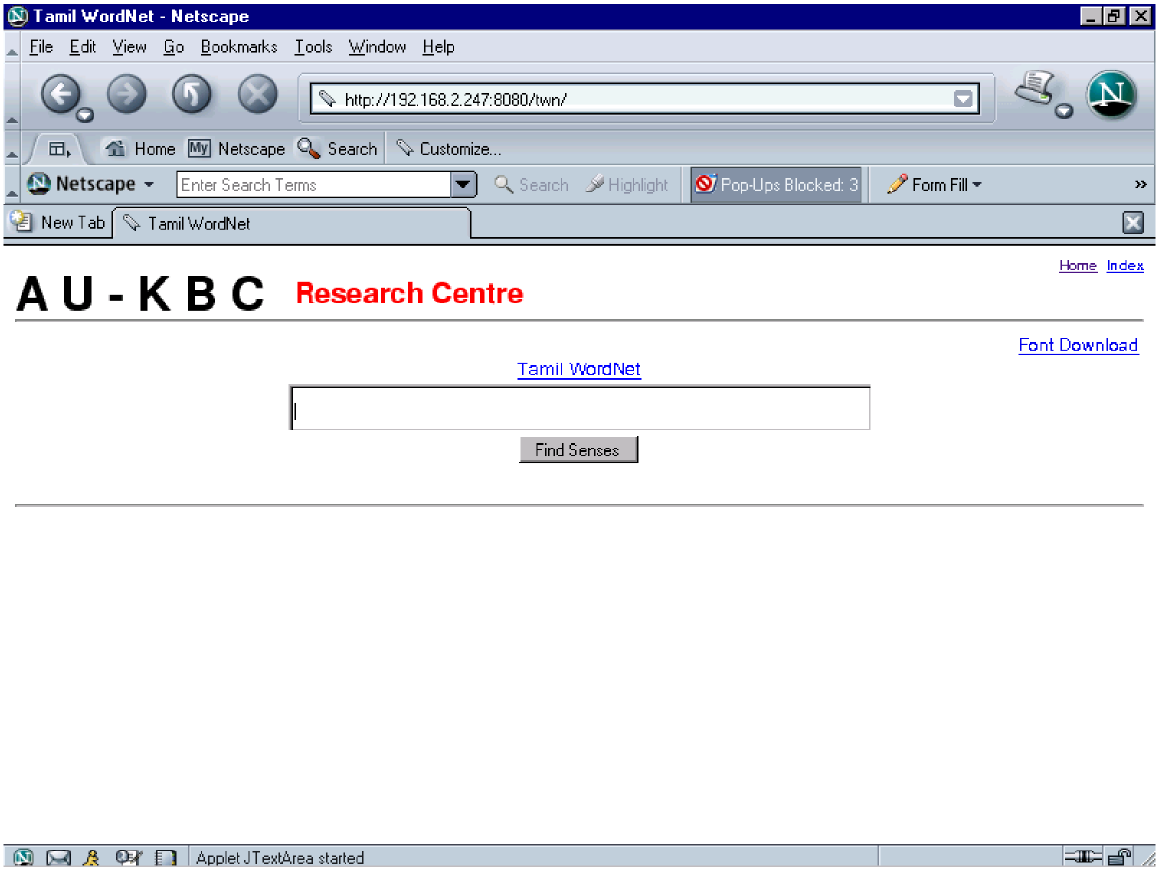
Feature	***Number***
Human	3
Non-Human	1

8. User Interface

The Web Interface has a *Search Word* box and a *Find Senses* button. Once the user submits a word, an overview of all the available information is displayed. In the *Results Display Area,* all the available senses of the submitted word are displayed. The ordering of display is based on the Parts-of-Speech category. Nouns senses appear first, followed by the Verb senses, Adjective senses appear next followed by Adverb senses. If a word is not found in the database, *No Matches Found* message appears. Under each Parts-of-Speech category list, all available senses are displayed. Below this, at the left end appears a *Button* which has *Drop-Down* menus for all available relations and a *Search* box for entering sense number search. The user can select the relation and enter the sense number in the *Sense Number* box and press *Search* button, which is available next to the

Sense Number box. Similar searches can be made for all available Parts-of-Speech categories.

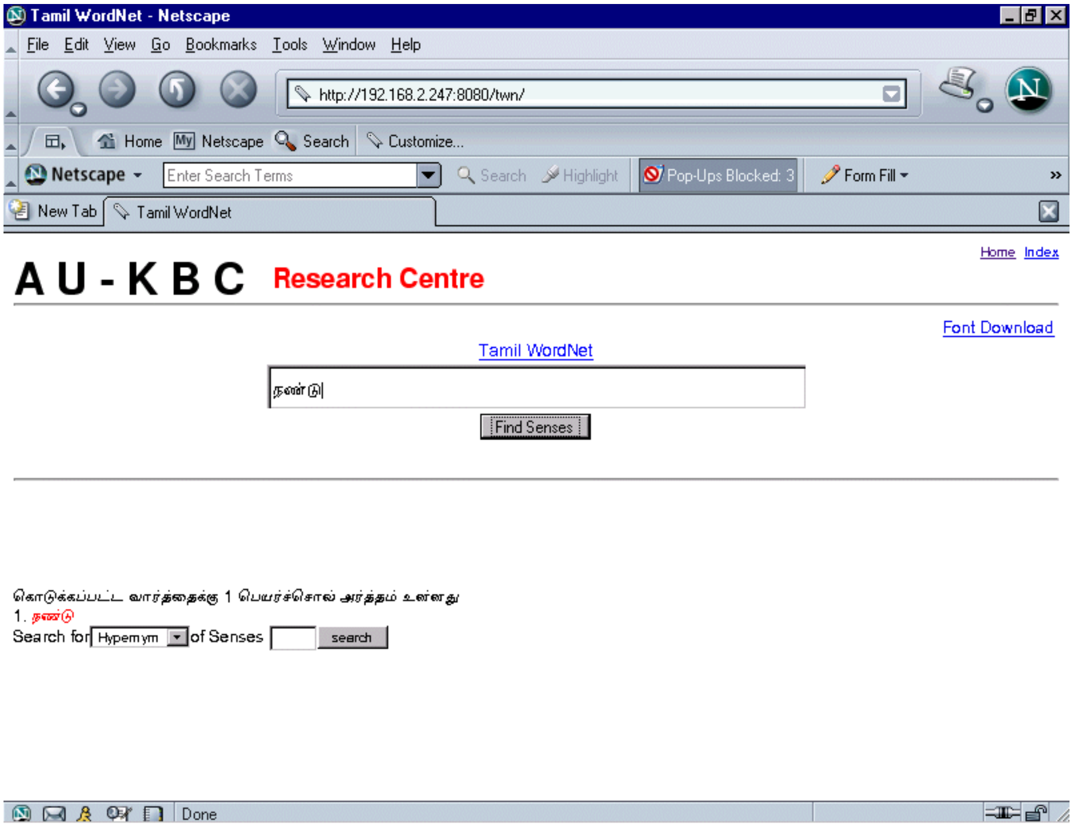
At the right top end of the Web page appears *Font Download* link for downloading fonts.



http://192.168.2.247:8080/twn/

Applet JTextArea started

Screenshot of Web Interface (Opening Page)


Tamil WordNet - Netscape
File Edit View Go Bookmarks Tools Window Help
Home Netscape Search Customize...
Netscape Enter Search Terms Search Highlight Pop-Ups Blocked: 3 Form Fill
New Tab Tamil WordNet
Home Index
Font Download
Tamil WordNet
Find Senses
search
Done

http://192.168.2.247:8080/twn/

AU-KBC Research Centre

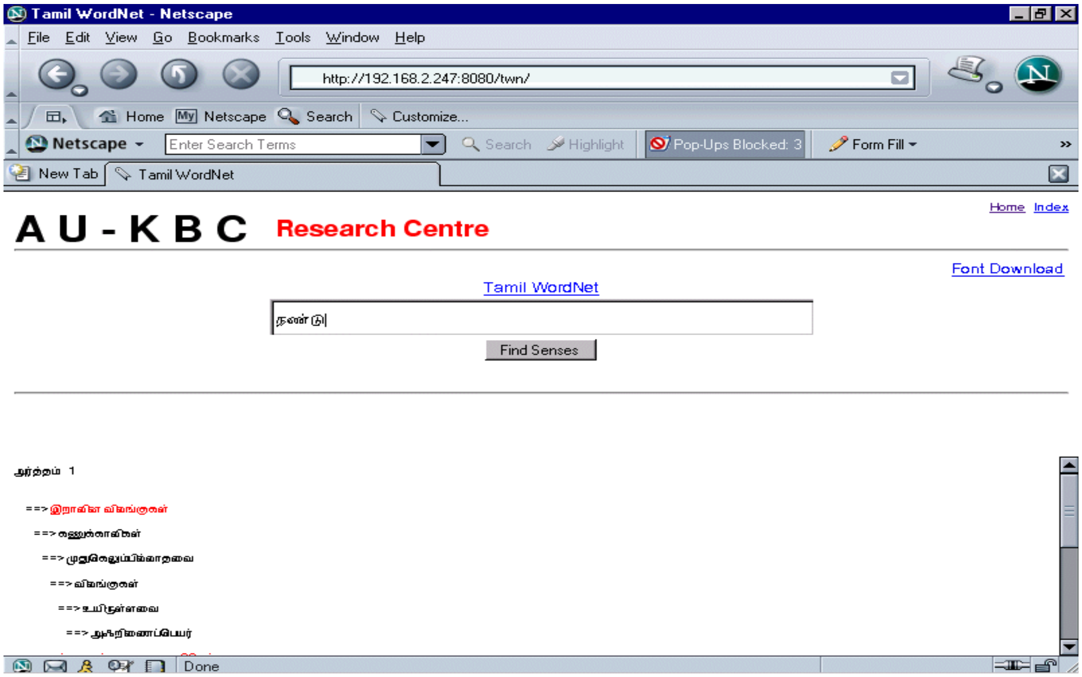
நண்டு

கொடுக்கப்பட்ட வார்த்தைக்கு 1 பெயர்ச்சொல் அர்த்தம் உள்ளது

1. நண்டு

Search for Hypernym of Senses

Screen shot of Web Interface (For the word `ndaNTu')


Tamil WordNet - Netscape
File Edit View Go Bookmarks Tools Window Help
Home Netscape Search Customize...
Netscape Enter Search Terms Search Highlight Pop-Ups Blocked: 3 Form Fill
New Tab Tamil WordNet
Home Index
Font Download
Tamil WordNet
Find Senses
Done

http://192.168.2.247:8080/twn/

AU-KBC Research Centre

நண்டு

அர்த்தம் 1

==> இறாலின விலங்குகள்
  ==> கணுக்காலிகள்
    ==> முதுகெலும்பில்லாதவை
      ==> விலங்குகள்
        ==> உயிருள்ளவை
          ==> அஃறிணைப்பெயர்

Screenshot on Hypernym search for the word 'ndaNTu'

9. Statistical Details

Parts-of-Speech	*Total Words*	*Unique Senses*
Noun	46710	37530
Verb	2881	2672
Adjective	416	388
Adverb	490	423
Total	50497	41013

Installation

(installation has been tested for windows 2000)

Please see install.txt for installation instructions.

11. Future Enhancements

The Tamil WordNet developed will be an excellent resource for researchers and public alike. However, for this to be more useful, the following improvements needs to be done.

Adding gloss and example sentences – For every sense of a given word, its gloss (meaning as found in a dictionary entry) and an example usage of that word in the particular sense. This will be extremely valuable for a casual user browsing the WordNet. This task needs to be done by a trained set of expert lexicographers with appropriate care.

Inter-linking of the words in English WordNet – The words in Tamil WordNet has to be inter-linked with those in English wordNet, which will make the WordNet useful for a variety of language processing applications involving Tamil and English such as Machine Translation, multi-lingual Question-Answering system etc.

Providing a better User-Interface – A better user-interface in the form of a dynamic 3D interactive visual interface, which can be operated by a mouse will enhance the usability of the WordNet. This will allow the users to see a word and its related terms visually (as can be seen in http://www.visualthesaurus.com).

12. References

Fellbaum, C. (ed.) 1998. WordNet: An Electronic Lexical Database.

Cambridge, MA: MIT Press.
Nida, E.A. 1975a. Compositional Analysis of Meaning: An Introduction to Semantic Structure. The Hague: Mouton
Rajendran, S 2001. *taRkaalat tamizc coRkaLanjciyam* [Thesaurus for Modern Tamil]. Thanjavur: Tamil University.
Vossen P. (eds.) 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.

URLs

Princeton English WordNet - *http://www.cogsci.princeton.edu/~wn/*
EuroWordNet - h*ttp://www.let.uva.nl/~ewn/*
Global WordNet Association - *http://www.globalwordnet.org/*