Report of the Workshop on
WordNet for Dravidian Languages
2-3 June 2003

Organised jointly by

AU-KBC Research Centre, MIT Campus of Anna University, Chennai
Central Institute of Indian Languages, Mysore
Department of Linguistics, Tamil University, Thanjavur

WordNets are being globally established for a number of languages, an example of which is EuroWordNet for European languages. Analogous to EuroWordNet, building an IndoWordNet needs individual WordNets for all the major Indian languages. Currently, WordNets are being developed at IIT-Bombay (Hindi and Marathi), MS University, Baroda (Gujarathi), IIT-Kharagpur (Bengali), Utkal University, Bhubaneswar (Oriya) and the AU-KBC Research Centre, MIT Campus of Anna University, Chennai (Tamil).

Among Indian languages, Dravidian languages such as Tamil, Telugu, Kannada and Malayalam have a close association in many linguistic features such as morphology and semantics. They also share a lot of culture-specific features. Building a common WordNet for this family of languages will make easier the task of a linked IndoWordNet.

The present workshop aimed to explore the possibility of developing such a Dravidian WordNet.

The objectives of the workshop were:

- to present ongoing work on WordNets in Dravidian languages
- to explore the common features of Dravidian languages that are useful in a WordNet
- to exchange ideas, experience, tools and resources in building WordNets
- to evolve a program for the development of WordNets in the four major Dravidian languages.

Faculty for the Workshop

The faculty for the workshop were:

Jayaram, B.D., Central Institute of Indian Languages, Mysore
Pushpak Bhattacharya, IIT-Bombay, Mumbai
Rajendran, S., Department of Linguistics, Tamil University
Uma Maheswar Rao, G., Centre for A.L.T.S., University of Hyderabad

Participants for the Workshop

In all, about 30 participants attended the two-day workshop. Participants came from the following institutions:

Kannada: Central Institute of Indian Languages, Mysore
Kuvempu University, Shimoga

Malayalam: SCERT & ER&DCI, Thiruvananthapuram

Telugu: Rashtriya Sanskrit Vidyapeet, Tirupati & University of Hyderabad

Hindi: IIT-Bombay, Mumbai

Tamil: AU-KBC Research Centre, MIT Campus
Dept of Computer Science, Anna University
Tamil University, Thanjavur
University of Belize, Central America

Details about the workshop

The first day started with S. Rajendran's (Tamil University) talk on `Remarks on WordNet'. He gave a detailed presentation on the English WordNet, including the concepts of synsets and the semantic relations that are dealt within a WordNet. He also gave examples in Tamil.

The second talk was by Pushpak Bhattacharya (IIT-Bombay) on `WordNet for Hindi and towards IndoWordNet'. He explained the various techniques that were followed in developing a Hindi WordNet. He also discussed about three principles, viz. Minimality, coverage and replaceability in constructing a synset. Further, he explained how the Hindi WordNet's synsets and glosses were useful in constructing a Marathi WordNet.

In the afternoon session, G. Uma Maheswar Rao (University of Hyderabad) in his lecture on `Some issues in building Dravidian WordNet' discussed the problems and issues involved in building a Dravidian WordNet. He supplemented his talk with examples from EuroWordNet.

B. D. Jayaram (CIIL) gave a brief account on how corpus can be utilized in building WordNet. The title of his paper was `Using corpus for building WordNet'. He also explained how they extracted unique words from the corpus.

The last presentation was on `A Tamil WordNet-Construction Domain' by S.Arulmozi (AU-KBC). During the demonstration, the participants were shown how wordlists from all the major dravidian languages could be

incorporated in a hierarchical tree.

On the morning session of the second day, all the participants were given a list of 500 words from the domain of `body-parts'. During this session the participants were asked to prepare a list of synsets in their respective languages. After this, the participating groups exchanged notes on an exercise that was carried out on a list of 900 words taken from the broad semantic domain – `Construction' - that was provided by AU-KBC Research Centre two months before the workshop. The participating groups discussed the problems faced by them while giving equivalents to this wordlist.

The post-lunch session began with a discussion on building ontologies. S.Rajendran reviewed the various ontological classifications. G.Uma Maheswar Rao elaborated on ontological classifications. Ravisankar Nair (SERCT) stressed the need for a common format for ontologies in Dravidian languages.

Panel Session: Dravidian WordNet

Dr. M. Ponnavaiko (Director, Tamil Virtual University) chaired the panel session. The panelists were B. D. Jayaram, S. Rajendran and G. Uma Maheswar Rao .

G. Uma Maheswar Rao started the session by making a point on EuroWordNet's "Expand and Merge" Model. As it seemed that the *Expand* model is best suited for dravidian languages, using the Tamil ontological classification (developed by S. Rajendran for his electronic thesaurus) and extending it to other dravidian languages would be the starting point for this effort. L.Halemane of C.I.I.L pointed out that the universal entities and language specific entities should be kept in mind while developing the dravidian WordNet. G.Uma Maheswar Rao responded by saying that individual WordNets would fill the lexical gaps. Ilangovan Padmanaban (University of Belize) stressed the need for replaceability and pointed out that finding synsets among the synonymous words is a huge task.

Ramasree (Rastriya Sanskrit Vidyapeet, Tirupati) gave a brief note on the application of WordNets. Arti Sharma (IIT-Bombay) pointed out the usefulness and relevance of the Hindi WordNet in building the Marathi WordNet.

Dr.C.N.Krishnan (Director, AU-KBC Research Centre) stressed the need for a programme for the future by coming up with a joint proposal for funding. He pointed out that the Ministry of Information Technology and European Union could be approched in this regard. G.Uma Maheswar Rao felt that such a proposal should comprise of the following items: 1. Time Frame, 2. Size of the WordNet, and 3. Funding. B.D.Jayaram said that the language-specific WordNets can be developed in different institutions and then can be

linked later on. He said that there should be a common format made available for building WordNets individually.

Action Plan:

1. S. Rajendran (Tamil University) will coordinate the efforts for building a Dravidian WordNet

2. Language-specific WordNets will be coordinated by the following institutions:

Kannada - Central Institute of Indian Languages, Mysore
Malayalam - SCERT, Thiruvananthapuram
Telugu - University of Hyderabad
Tamil - AU-KBC Research Centre, Chennai

3. Each language-group will work on the Nouns and AU-KBC Research Centre will send to the participating groups the following:

i) Ontological classification from Rajendran's electronic thesaurus
ii) Domain-wise noun list

4. By the end of six months, each participating group will complete the development of a WordNet for nouns.

Co-ordinators:

B. D. Jayaram
C.I.I.L
Manasagangotri
Mysore 570006
jayaram@ciil.stpmy.soft.net

S.Rajendran
Dept. of Linguistics
Tamil University
Thanjavur 613 005
raj_ushush@yahoo.com

S.Arulmozi
AU-KBC Research Centre
MIT Campus of Anna University
Chennai 600044
arulmozi@au-kbc.org