Keynote Address

8.1 Lexicographic Tradition in India
(With Reference to Lexical Resources)

N. Raj Shekharan Nair, Annamalai University

I. INTRODUCTION

INDIA has a long and ancient lexicographic and grammatical tradition. Ya:ska's Nirukta, the earliest etymological work of Sanskrit, Amarkosh of Amarasimha a traditional verse dictionary, and Nighantus belong to the dictionary tradition of India. Koshas and Nighantus were mostly lists of synonyms and homonyms. In addition, the ancient grammatical works like Paaini's Aṣṭa:dhya:yi: in Sanskrit belonging to the 5th B.C. and Tolka:ppiyar's Tolka:ppiyam of Tamil of 3rd C.B.C. provide lots of lexical information. We can say that these two and other traditional grammars were partial lexical resources. Aṣṭa:dhya:yi: has accessory texts called *dhatupa:ṭha,* verb roots, *gaṇapa:ṭha*, lists of nominals, *linga:nu sa:sana*, gender of Sanskrit words, *phit sutras* rules on accentuation of words. In Tamil Tolka:ppiyam list of classes of words are included as part of the text itself. It devotes one whole chapter for describing the meanings of lexemes or content words. This chapter can be considered to contain the seeds of later lexicographic tradition of Tamil. This tradition combined with the Sanskrit Kosha tradition lead to the compilation of numerous verse Nighaṇṭus in Tamil. The dictionary tradition in India owes itself to the Western Christian missionaries, which started in the 17th century. This was later followed by the native scholars. These are the beginnings of lexical resources in Indian languages.

As a linguist, my thrust in my address will be to outline the different types of lexical information that should find a place in the lexical resources to be developed for Indian languages. It will be pointed out that importance must be given to the special grammatical and semantic features of Indian languages. While grammars of languages will emphasize the structural aspects like syntactic patterns, word formation, etc., which are common to classes or groups of words, lexical resources must concentrate on the formal (structural) and semantic aspects of individual words. Formal aspects will include the marking of the parts of speech (form class) of the words like noun, verb, adjective, adverb, etc., sub-categorization on the basis of inflectional classes (e.g., conjugation classes of verbs), gender marking, etc. A special feature of Indian languages is the case inflexion of nouns which is equivalent to the prepositional phrases of English. Here the case selection or rection of verbs is very important for Indian languages. For example a class of involuntary physical and mental state verbs having the meanings *'be hungry', know, understand,* etc. do not take nominative case or subject in a sentence. They take only dative or fourth case, e.g., Hindi, *muje bhu:k lakti:hai*, Tamil *enakku pacikki* r̲atu, Telugu, na:ku-a:kaliga:undi. 'I am hungry' (lit. 'It is hungry to me') Malayalam: enikku visakkannu. Providing of case selection features of verbs for Indian languages is a must.

At the semantic level meanings of polysemous and homonymous words, synonyms, antonyms, class inclusion (hyponymy), part of a whole (meronymy), etc. must be indicated. In addition collocational restrictions (e.g., dog barks; horse neighs; lion roars) must also be indicated.

Providing of both grammatical and semantic information must be on the basis of exhaustive study of vast corpus of each language. For this corpus development of multiple natures, is essential. Lexical resources can also study dialect variation, both regional and social (i.e., caste and religious group based). Recording of different occupational vocabularies of Indian languages is an urgent need as international terms are fast replacing the native words. This is essential to preserve the lexical wealth of Indian languages.

Lexical resources will be of great use for language learning-teaching, creative writing, translation between Indian languages, creation of technical terminology, and for sociological, cultural and historical study of the different Indian language groups. The effort of C-DAC will go a long way in fulfilling the needs of different types of language users and create awareness among the common public of the multiple uses of lexical resources, apart from coordinating the works of different agencies already involved in corpora and lexical resource development.

II. LANGUAGE CORPORA

India is a multilingual country. As per the census of India there are 1652 mother tongues. There are many languages, especially tribal languages, which are at the verge of extinction due to various reasons. So it is necessary to generate corpora for every Indian language. Various institutions are engaged in the preparation of corpus. The corpus developed by Shastri (1988) Shivaji University is the first corpus for Indian English called 'Kolhapur Corpus of Indian English' (KCIE).

The following is a list given by Niladri Sekhar Dash (Indian Linguistics 64: 2003) on the corpora generation projects involving Indian languages.

N. Rajasekharan Nair is with CAS in Linguistics, Annamalai University, Annamalai Nagar – 608 002 (e-mail: rajasekharan245@yahoo.co.in)

Agencies	Indian languages corpora
Indian Institute of Technology, Kanpur	Hindi, Nepali
Indian Institute of Technology, Mumbai	Marathi, Konkani
Indian Institute of Technology, Guwahati	Assamese, Manipuri
Indian Institute of Sciences, Bangalore	Kannada, Sanskrit
Indian Statistical Institute, Kolkata	Dangla
Jawaharlal Nehru University, New Delhi	Sanskrit
University of Hyderabad, Hyderabad	Telugu
Anna University, Chennai	Tamil
MS University, Baroda	Gujarati
Utkal University, Bhubaneswar	Oriya
Thapar Institute of Engg. and Tech., Patiala	Punjabi
ER&DCI, Tridendrum	Malayalam
CDAC, Pune	Urudu, Sansk rit, Kashmiri

The above studies show the significance of such ventures.

III. SPEECH CORPORA

The scholars in many universities in foreign countries are showing keen interest to create the speech corpora for Indian languages. This seems to be a modern trend in creation of lexical resources. Recently Prof. Peter Juel Henrichsen, Centre for computational modeling of Language (CMOL), Department of Computational Linguistics, Copenhagen Business School, Denmark visited my Department (Centre of Advanced study in Linguistics, Annamalai University) and had discussions with all the faculty members. During the interaction he informed us their plan to create the speech corpora for Indian languages, especially for Tamil and a few Dravidian Tribal Languages. A colleague of mine visited the Copenhagen Business School later by invitation and interacted with the faculty members there. Now we are exploring the possibilities to have a MoU with them on the creation of the speech corpora. Niladri Sekhar Dash has highlighted the importance of speech corpora in his article on 'Speech Corpora and Text Corpora' (Indian Linguistics: 67: 2006).

Prof. Peri Bhaskara Rao, Chair, Institute for the study of Languages and Cultures of Asia and Africa, Tokyo University of Foreign studies, Tokyo, Japan has been collecting linguistic data from the Toda Tribal people of Nilgiris, Tamilnadu for the past few years.

Dr. Christiane PILOT – Raichoor, a renowned scholar in Dravidian Linguistics, LACITO-CNRS, France has been engaged in the digitization of Dravidian Tribal languages of Wayanad District, Kerala. I have been involved in the collaborative project. I have recorded the folk songs of two Dravidian Tribal communities, namely, Mullakurumar and Kurichyar. The digitization process was initiated at LACITO during November 2006 when I visited there. This will be a speech corpus, with other relevant details incorporated. This digital corpus will be a document for the scholars to do future research. The collection of speech data from other tribal communities of Wayanad district will be continued and the data will be digitized. I would like to mention here that these speech varieties are at the verge of endangerment. David Crystal in his book 'Language Death' points out that a language will die if people of that community are not using that language. The mother tongue use of many tribal communities is decreasing. Mostly in home domain they use their mother tongue. The older people who use their languages are dying due to illness, etc. The younger generation, who study in schools and colleges in Malayalam language medium, which is the major language of Kerala, use their mother tongue less frequently. In this circumstance there is all possibility that these tribal tongues will be lost soon. Hence we are recording the languages along with folk songs and digitizing them.

IV. DICTIONARIES

The lexicographic as well as computational lexicographical activities are going on in Indian Universities and Institutions in an appreciable manner. The trained linguists, lexicographers and computational lexicographers are involved in the compilation and also in the research. At the Centre of Advanced Study in Linguistics, Annamalai University I have completed along with Dr. S. Raja a UGC Major Research Project entitled 'A study of Dialect Variation in Tamil Agricultural Vocabulary'. Data were collected from all over Tamilnadu – mainly from the five major dialect regions of Tamilnadu. Both the lexical variations and phonological variations are identified. The results of this survey are presented in the form of a dialect dictionary, showing the variation. The same recorded cassettes can be used to generate speech corpora.

Dr. Christiane PILOT – Raichoor, who is a Dravidologist, has been working on Badaga language spoken in Nilgiri district of Tamilnadu. She has compiled A Badaga – English dictionary along with Paul Hockings (1992) which is applauded by scholars. Since Badaga has no script, the head entries are given in roman script

Similarly, I along with Dr. S. Raja, completed a Tamil-Malayalam Translator's Dictionary. A Malayalam-Tamil Dictionary is also being compiled by us. The Centre for Applied Linguistics and Translation Studies, Hyderabad has a dictionary project. Under that they have taken up the works on Telugu-Hindi Nishantarvu, Telugu-Oriya nikhantuvu, Telugu-Kannada, Nighantuvu and Telugu-Malayalam Nighantuvu. Here the same strategy is employed. For one Telugu main entry there will be two or three target language equivalents, which are given in an order of frequency. I have been collaborating with the Telugu-Malayalam Nighantuvu. Prof. Umamaheswar Rao is

the Chief Editor for all the volumes. In addition to this he is engaged in projects concerning to online lexicography.

Tamil Lexicon was published by the University of Madras in 6 volumes. It has a supplementary volume also. Now Prof. Jeyadevan and the team in the University of Madras are engaged in the revision of the Tamil Lexicon by incorporating new entries. Similarly the compilation of Malayalam Lexicon in 11 volumes was entrusted to the University of Kerala, Trivandrum. 7 volumes have been published. The first volume was published in 1956. The works in connection with the volumes 8, 9 and 10 are going on simultaneously. In addition to this there are many good monolingual and bilingual dictionaries available in Malayalam. A Pedagogical dictionary in Malayalam (Sabda Surabhi 2 volumes, 2005) compiled by the former Malayalam Lexicon Chief Editor Dr. B.C. Balakrishnan is noteworthy. In this he has incorporated all the Malayalam words available in the Malayalam books up to 12th Std. with details and meanings. This is a good lexical resource for future researches also. Dr. Shanavas of the Department of Linguistics, University of Kerala is engaged in the preparation of Malayalam corpus using modern technology.

'Preparation of Generative Lexicon using MRDs' is an ongoing project undertaken by Prof. S. Rajdendran, Tamil University. Another project undertaken by him is 'Indian Languages to Indian Languages' (Machine Translation System). A bilingual mapping dictionary of Tamil-Malayalam and Malayalam-Tamil will be the byproduct of this project.

While talking about the 'Desiderata for a Modern English Dictionary of Tamil' Harold F. Schiffman opines that "The non-Tamils who learn an Indian language other than Sanskrit or Hindi is immediately aware of the problem of lack of adequate materials for learning the language, and especially the lack of decent dictionaries" (South Asian Language Review Vol.V, Jan. 1995. No.1).

Gregory James, the author of 'A History of Tamil Dictionaries' in one of his articles opines that "The nikantu remained the principal style of dictionary composition in Tamil from the earliest times until at least the 16th C. A.D". Tiva:karam with 9,500 entries 8th C.AD. is the first nikantu for Tamil language. Then pinkalam with 14,700 entries came into existence between 8th C and 13 C AD. Another popular nikantu called cu:ta:mani nikantu by Mantalapurutar, the first metrical lexicon was published during 1520 AD.

Dr. L. Shobha, Au-KBC Research Centre has undertaken a few projects the results of which will be useful for the creation of lexical resources for Indian languages. The ontological study is a must for information retrieval. Her work an 'Tamil-Hindi Transfer Lexicon' is useful in many ways. In the other important research project entitled 'Named entity' she has incorporated more than thirty thousand proper names of Tamil with the details. It seems this is the first attempt in any Indian language. An electronic thesaurus of Tamil has been prepared by the Department of Linguistics of Tamil University in collaboration with the Department of Computer Science of Tamil University. The CALTS, Hyderabad is engaged in the preparation of multilingual thesaurus. It has to be noted that IIIT, Hyderabad is engaged in preparing lexical resources for Indian Languages. A team of scholars are involved in it.

Dr. T.V. Geetha of Anna University of Chennai has done a good amount of research on the Tamil lexical resources under the Technical development of Indian Languages. Her online pedagogical dictionary (English-Tamil and Tamil-English) can be used easily by any body. As a whole the present scenario is bright on the preparation of lexical resources for Indian Languages.

V. WORDNET

"WordNet was originally conceived and developed as a lexical database for English on the basis of psycholinguistic properties. The major lexical categories like Nouns, verbs, adjectives and adverbs are organized in terms of sets of synonyms (synsets), each representing a lexical concept." (Dravidian WordNet, AU-KBC Research Centre, Chennai). The first version of Tamil WordNet has been prepared by AU-KBC and Tamil University with the financial support of Tamil Virtual University. It is going to be launched as an open source material in on-line.

Developing WordNet is an important and basic need for lexical resources. The Euro WordNet developed for European languages is really useful for various kinds of research. Similarly if an India it will be a very good resource. The scholars who work on this field are aware of this phenomenon and hence, lot of such studies are in progress in India. It is to be mentioned that IIT, Mumbai has been engaged in the preparation of Hindi WordNet. It is available in the internet in its full-fledged form. This has to be appreciated since it is the first one from India.

I take this opportunity to appreciate the organizers for conducting this National Seminar on 'Creation of Lexical Resources for Indian Language computing and processing' with the clear cut objectives which are spelt out in the call letter itself. An added attraction of this seminar, which is interdisciplinary in nature, is that linguists, computational linguistics, lexicographers, computational lexicographers, NLP/NLT researchers/ developers, computer scientists, etc. are involved in the deliberations. I am sure there will be a very good outcome in the field by the deliberations for all the three days.

ACKNOWLEDGEMENT

I thank Prof. K. Balasubramanian and Prof. S. Rajendran who helped me in the preparation of this paper.

REFERENCES

[1] Cruse, D.A. 1986. Lexical Semantics. Cambridge: Cambridge University Press.
[2] Cuypers, I & G. Adraens. 1997. Periscope: the EWN Viewer. EuroWordNet Project LE4003. Deliverable D008d012. Amsterdam: University of Amsterdam.
[3] Fellbaum, C. 1990. "English Verbs as a Semantic Net". International Journal of Lexicography, Vol. 3, No.4, 278-301.
[4] - - - - -. 1998. "A Semantic Network of English Verbs". In: Fellbaum, C. (ed.). WordNet: An Electronic Lexical Database. Cambridge: MIT Press.
[5] Fellbaum, C. (ed.) 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.
[6] Gross, D. and K.J. Miller 1990. "Adjectives in Wordnet". International Journal of Lexicography, Vol. 3, No.4, 265-277.
[7] James, G. 2002. *colporul*: History of Tamil Lexicography. Chennai: CreA.
[8] Louw, M. 1997. *The Polaris User Manual*. Internal Report, Lernout & Hauspie.
[9] Lyons, J. 1977. Semantics (Vol.1). Cambridge: Cambridge University press.
[10] Miller, G.A. 1990. "Nouns in WordNet: a lexical inheritance system". International Journal of Lexicography Vol 3, No. 4, 245-264.
[11] Miller, G.A. 1991. Science of Words. New York: Scientific American Library.
[12] - - - -. 1998. "Nouns in WordNet". In: Fellbaum, C. (ed.). 1998. WordNet: An Electronic Lexical Database. Cambridge: MIT Press.
[13] Miller G.A., R. Beckwith, C. Fellbaum, D. Gross, K.J. Miller. 1990. "Introduction to WordNet: An On-line Lexical Database". International Journal of Lexicography, Vol. 3, No.4, 235-244.
[14] Miller, K.J. 1998. "Modifiers in WordNet". In: Fellbaum, C. (ed.). 1998. WordNet: An Electronic Lexical Database. Cambridge: MIT Press.
[15] Nida, E.A. 1975a. Compositional Analysis of Meaning: An Introduction to Semantic Structure. The Hague: Mounton.
[16] - - - -. 1975.b. Exploring Semantic Structure. The Hague: Mouton.
[17] Pustejovsky, J. 1995. The Generative Lexicon. Cambridge: MIT Press.
[18] Rajendran. S, 1978. Syntax and Semantics of Tamil Verbs. Ph.D. Thesis. Poona: University of Poona.
[19] - - - -. 1983. Semantics of Tamil Vocabulary. (Report of the UGC sponsored Postdoctoral Work in Manuscript). Poona: Deccan College Post-Doctoral Research Institute.
[20] - - - -. 1995. 'Towards a Compilation of a Thesaurus for Modern Tamil.' South Asian Language Review 5.1: 62-99.
[21] - - - - -. 2001. *taRkaalat tamizc coRkaLanjciyam* [Thesaurus for Modern Tamil]. Thanjavur: Tamil University.
[22] - - - -. 2002. 'Preliminaries to the preparation of a Word Net for Tamil.' Language in India 2:1.
[23] - - - -. 2003. 'Pre-requisite for the Preparation of an Electronic Thesaurus for a Tex Processor in Indian Languages'. Language in India 3:1.
[24] Rajendran, S., S. Arulmozi, B. Kumara Shanmugam, S. Baskaran, and S. Thiagarajan. 2002. "Tamil WordNet." Proceedings fo the First International Global WordNet Conference. Mysore: CIIL, 271-274.
[25] Rajasekharan Nair, N. and S. Raja. 2006. A study of dialect variation in Tamil Agricultural vocabulary (mimeo).
[26] S. Sundarabalu. 2006. Semantic field and Lexical structure in Modern Tamil: Ph.D. dissertation.
[27] Tengi, R.I. 1998. "Design and Implementation of the WordNet Lexical Database and Searching Software". In: Fellbaum, C. (ed.) 1998. WordNet: An Electronic Lexical Database. Cambridge: MIT Press.
[28] Uma Maheswar Rao, G. 2003. 'Some Issues in building Dravidian WordNet', Paper presented in the *Workshop on WordNet for Dravidian Languages*, Chennai: AU-KBC Research Centre, 2-3 June 2003.
[29] Vossen P. (eds.) 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.
[30] Vossen, P. 2001. 'Condensed Meaning in EuroWordNet', In: P. Bouillon & F. Busa (eds.). 2001. *The Language of Word Meaning*. Cambridge: Cambridge University Press, pp. 363-383.
[31] Princeton English WordNet–http://www.cogsci.princeton.edu/~wn/
[32] EuroWordNet – http://www.let.uva.nl/~ewn/
[33] Global WordNet Association – http://www.globalwordnet.org/
[34] http://framenet.icsi.berkeley.edu/
[35] http://wordnet.princeton.edu/
[36] http://dictionary.cambridge.org/

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.