Online Multilingual *Amarakośa*: the relational lexical database

Girish Nath Jha
Special Center for Sanskrit Studies
Jawaharlal Nehru University, New Delhi
girishjha@gmail.com

R. Chandrashekar
Special Center for Sanskrit Studies
Jawaharlal Nehru University, New Delhi
ramaswamy.chandrashekar@gmail.com

Umesh Kumar Singh
Special Center for Sanskrit Studies
Jawaharlal Nehru University, New Delhi
umeshvaidik@gmail.com

Vibhuti Nath Jha
Special Center for Sanskrit Studies
Jawaharlal Nehru University, New Delhi
vibhutijha158@gmail.com

Satyendra Pandey
Special Center for Sanskrit Studies
Jawaharlal Nehru University, New Delhi
pandeysatyendra80@gmail.com

Surjit Kumar Singh
Special Center for Sanskrit Studies
Jawaharlal Nehru University, New Delhi
surjit.jnu@gmail.com

Mukesh Kumar Mishra
Special Center for Sanskrit Studies
Jawaharlal Nehru University, New Delhi
mukeshscssjnu@gmail.com

Abstract

This paper outlines the ongoing research project called "Online Multilingual Amarakośa (OMA)" for a multilingual lexical resource for Amarakośa (AK) which not only lets users store ontological equivalents of Sanskrit concepts in their languages, but also lets them search and edit. The work has tremendous applications in Word Sense Disambiguation (WSD) process of Machine Translation Systems, in Knowledge Representation, and in language pedagogy.

This system hosted at http://sanskrit.jnu.ac.in has a Java front-end and relational database server as back-end. At present, it has the following features –

- facility for online multilingual data entry in Sanskrit, Hindi, Kannada and English.
- data storage in multi-scriptural Indian langauge unicode
- stores up to 50 synonyms with grammatical information and detailed glosses
- cross-referencing among synonyms
- ontology display
- search capability in the supported Indian language (Jha et al, 2005)

Near future enhancements include

- selective data export in the user defined format
- display of the scanned imges of the pages from authorative text
- smarter search engine
- text processing based on AK

The system is intended to be used in the following domains –

- Multilingual concept acquisition
- Word Sense Disambiguation
- Machine Translation among Indian languages by way of Sanskrit
- As a model for building multilingual online systems for other seminal texts of Indian intellectual tradition
- Sanskrit wordnet (Jha et al, 2006)

1 Introduction

India has seen amazing strides in Information and Communication Technology (ICT) applications for Indian languages in general and for Sanskrit in particular. Since Machine Translation from Sanskrit to other Indian languages is often the desired goal, traditional Sanskrit lexicography has attracted a lot of attention of ICT and Computational Linguistics community. While several attempts are being made to build word-nets on traditional Indian epistemological and logical principles, the need for generating a multilingual lexical resource for AK, the Sanskrit lexicon built on ontological principles, has been largely ignored. AK, the 4th CE lexicon developed by Amarasimha has influenced modern lex-

icographic techniques in quite the same way as Pāṇini and Chomsky have done to generative linguistics. There have been some efforts in building a bilingual Sanskrit electronic dictionary. Huet (2004) is working on creating a Sanskrit – French electronic dictionary. Bontes (2005) had built a standalone system of Monier Williams dictionary. The Cologne Digital Sanskrit Lexicon contains Monier-Williams Sanskrit-English Dictionary has approximately 1,60,000 main entries. It has an online search facility in both Sanskrit and English. Capeller's Sanskrit-English Dictionary has been converted to a digital format similar to the Cologne project and has online search facility in both Sanskrit and English. It has only 50,000 entries. Apte Sanskrit Dictionary Search is a web Sanskrit dictionary based on the famous work of V. S. Apte - *The Practical Sanskrit-English Dictionary*. Andre Signoret's French-Sanskrit dictionary is freely downloadable from the net. The BhāratīyaBhāṣā multilingual dictionary built by Central Hindi Directorate, New Delhi under TDIL, Govt. of India funding consists of nearly 5000 common words in 14 different languages. It is available for download from the TDIL site. The Sanskrit Dictionary-Database being prepared by Jong-cheol Lee (2005), Academy of Korean Studies, Seoul, Korea will include mappings among Sanskrit, Tibetan, Chinese and Korean. Mohanty et. Al. (2004) has done some work on representing ontologies of Sanskrit words using Navya Nyāya methodology. As we can see, none of these works focuses on the AK and its rich semantic ontologies. In terms of search, the Cologne and Capeller's works are comparable.

As a text, AK has three *kāṇḍa* (chapters), each subdivided into *varga* (classes). The first and second *kāṇḍa* have 10 *vargas* each. The third *kāṇḍa* has 5 *vargas*. *Figure1* illustrates the structure of AK

There have been attempts to put the text of AK online or in digital formats. But there has been no attempt to create a version of this work which not only allows the users interactively build a database of AK but also search and test.

2 The AK system

The online system is being developed using Java servlets as front-end hosted on Apache-Tomcat platform and MS SQL server with multilingual Unicode data as backend.

The system has the following components –

- the relational database
- the data entry component
- search component
- detail search component
- data editing component

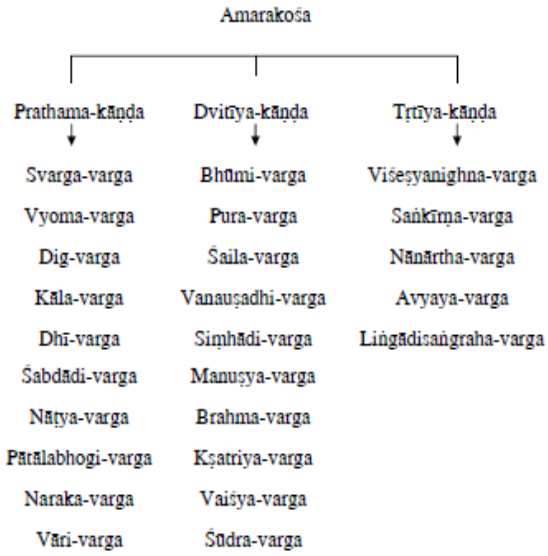

Figure 1: Structure of AK

2.1 The AK Database

The AK database is a relational database designed using MS SQL server objects and procedures. The database includes intricate relationships between the base words, synonyms and multilingual glosses.

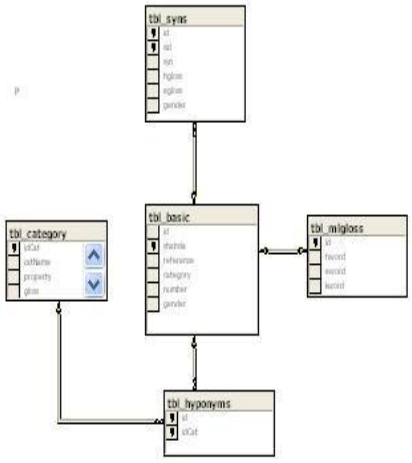

Figure 2: AK database-diagram

2.2 Data entry component

This component allows users to enter data in their language by way of user validation by password checking. Users can select the basic Sanskrit entry and provide information in their language (at present 8 Indian languages including English are supported). The screen capture for this page is given below -

Figure 3: Basic data entry page

After the basic data, the synonyms (up to a maximum of 50) can be entered with other relevant information.

2.3 Data Search

The search facility procided is of three kinds – direct search, alphabetical search and search by AK structure. Search can be done in the base word (Sanskrit), multilingual glosses and synonyms. There is scope for more languages in future. For example, the alphabet search of 'छ' displays all the words starting with this letter as:

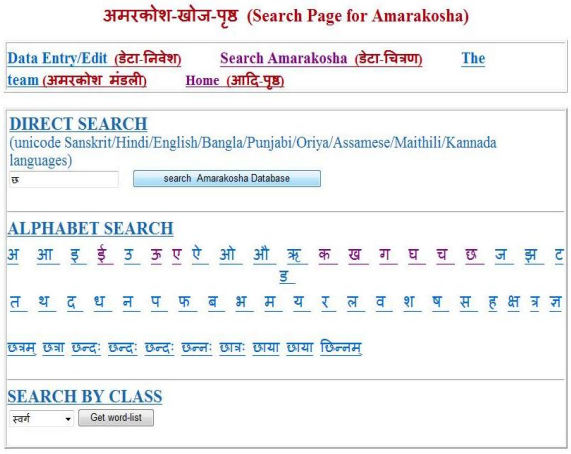

Figure 4: Direct-search page

The direct search can be done by typing the word in Devnagari (using an inbuilt Unicode keyboard for the iTrans scheme) or by clicking on the word list obtained by alphabet search.

Figure 5: Alphabet-search page

A successful direct search displays the basic information including the multilingual glosses and all the synonyms associated with the search string –

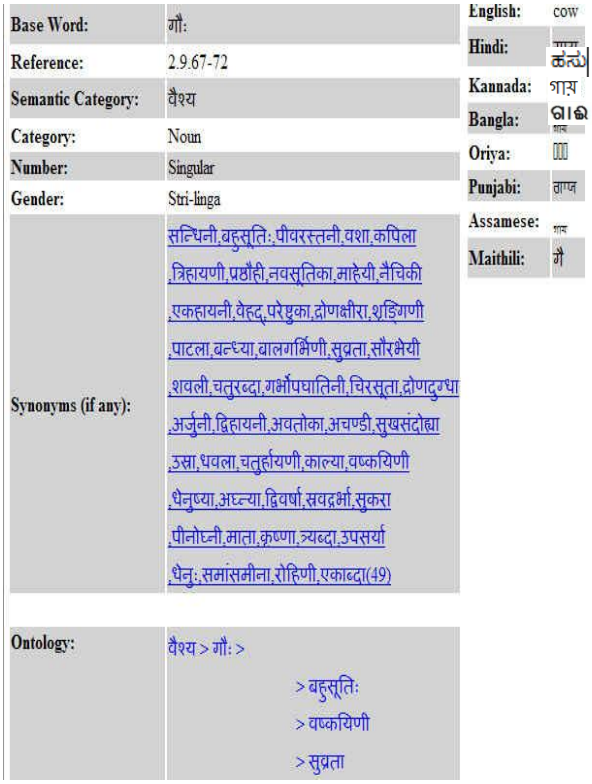

Figure 6: Search-result page

The search by structure (semantic classes) can be done by selecting a semantic class from the drop down box as follows –

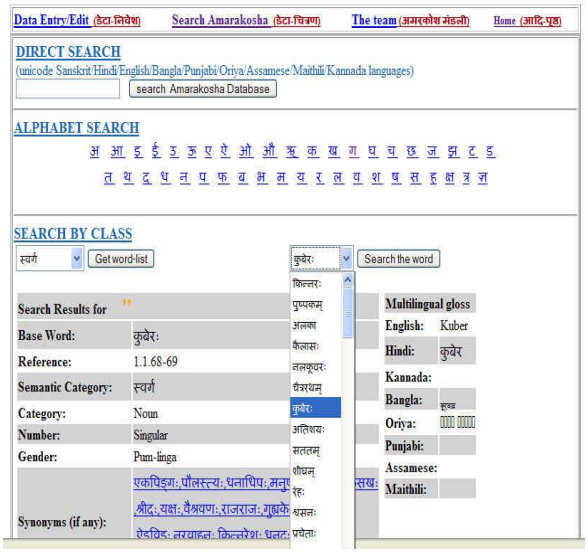

Figure 6: Search by semantic class

Clicking on the synonym link takes to the details page where each synonym entry is explained according to AK. The image that follows shows the synonym details for the English search string 'cow' –

Amarakosha (अमरकोश) Synonym details page

Syn#	Synonym	Gender	English Gloss	Hindi Gloss
1.	सन्धिनी	पुल्लिंग	cow which has co-habitated with a bull	साँड़ के साथ संगम की हुई गाय
2.	बहुसूतिः	पुल्लिंग	cow which has delivered several calves	सूधी अर्थात् न मारने वाली गाय
3.	पीवरस्तनी	पुल्लिंग	cow with thick udder	मोटे स्तनों (थनों) वाली गाय
4.	वशा	पुल्लिंग	sterile cow	बाँझ अर्थात् बच्चा नहीं पैदा करने वाली गाय
5.	कपिला	पुल्लिंग	Yellowish-brown cow	कपिल (पीली-भूरी) गाय
6.	त्रिहायणी	पुल्लिंग	Three year old cow	तीन वर्ष की उम्र वाली गाय
7.	प्रष्ठौही	पुल्लिंग	first time pregnant cow	आँकर अर्थात् पहले-पहल गर्भ धारण की हुई गाय

Figure 7: Search-result-detail page

3.4 Data Edit

This module lets language experts (by login only) to edit a wrongly entered data. The experts are sought through a registration module connected to the OMA website at http://sanskrit.jnu.ac.in/user/register.jsp which stores the user information in a database and determines who to give access for editing AK database directly on the server. The following screen allows the logged-in experts to edit AK data on the server -

अमरकोश डेटा-परिवर्तन-पृष्ठ (Amarakosha Data-Edit Page)

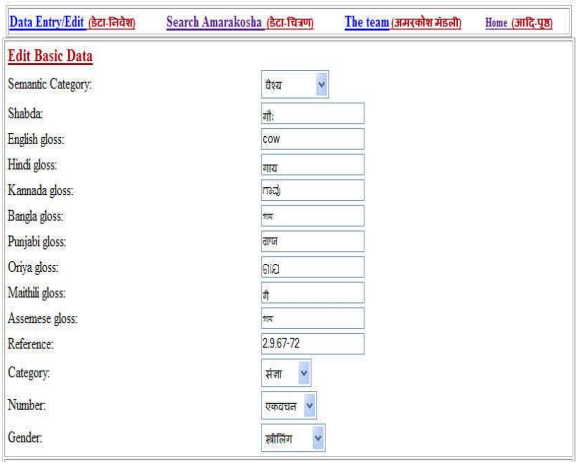

Figure 8: Data edit page

4 Structure of lexical data

The current structure of AK is a change from the earlier database structure mentioned above. While the text edition being followed is the same (*Rāmāśramī ṭīkā (RT)*), the data organization has undergone significant changes. We have also allowed POS annotation with two tagsets – the JNU tagset (Chandrashekar, 2007), and the Sanskrit Consortia tag-set (2009). The abstract data structure in the base table is as follows –

- *kāṇḍa>varga>śloka*
- reference of AK according to *RT*
- Word id
- Word in the AK according to *RT*
- Nominal base
- Word type
- Gender
- Number
- JNU POS Tag
- Sanskrit Consortium POS Tag
- AK gloss
- RT / Amarapadavivṛti gloss
- Interpolation
- Variant reading
- Other reading
- Inferred reading
- Pāṇini sūtra

The relations table stores the relations of the words with only the ids of the words related as follows –

- id of the first word
- id of the second word
- relation between the two words

Currently, we have stored the lexical items obtained from AK verses as per the RT, but later on we may also display respective *śloka* (verse) corresponding to the searched word.

5 Limitations and future enhancements

We would like to include as many langauges as possible depending on the resources at our hand. A Sanskrit textual analysis based on AK database is being carried out at the level of Ph.D. research. An analytical research (at the level of M. Phil.) is being carried out on the homonyms used in AK. These components will be added in future. The system presented here is the first Sanskrit *śāstra* made available in this format. We aim to include more and more texts in future. Other enhancements include improved display of information in terms of showing the *ślokas* with scanned image of the page where the word occurs. We are also developing a data download module where data export in customized formats will be supported.

6 Credits

The OMA project was started with the first batch (2002) of Ph.D. students–Sudhir, Chandrashekar, Sharda, Nagesh, Asha Shahi, Manju Pandit, Ashok Tiwari, Devendra Singh, Uma. The idea of an online interactive multilingual Sanskrit dictionary was liked by the BBC when they interviewed Dr. Girish Nath Jha in their *aaj ke din* Hindi program. Subsequently the project got funded under a generous grant from J.N.U. under the UPOE scheme. Currently, the editing and formatting of data is in process under a funding from Dept. of Information Technology for the Sanskrit Consortium. The authors of the paper would like to thank the above mentioned students and institutions as well as the following students and staff who contributed data and gave valuable suggestions – Sangeeta, Priti Bhowmick, Subash, Manji Bhadra, Muktanand Agrawal, Sachin Kumar, Diwakar Mani, Diwakar Mishra, Surjit Kumar Singh, Sureshwar Meher, Debashis Ghosh, Satyamudita Snehi and Vishav Bandhu.

References

Andre Signoret's French-Sanskrit dictionary (2005), http://asignoret.free.fr/index.html (accessed : 23 March 2005)

Apte Sanskrit Dictionary Search, 2009 http://aa2411s.aa.tufs.ac.jp/~tjun/sktdic/

BhāratiyaBhāṣā multilingual dictionary,2005, TDIL website (accessed : 23 March 2005)

Bontes Louis, 2005, Monier William Digital Dictionary http://members.ams.chello.nl/l.bontes/. (accessed : 23 March 2005)

Capeller's Sanskrit-English Dictionary,2005 http://www.uni-koeln.de/phil-fak/indologie/tamil/cap_search.html (accessed: 23 March 2005)

Chandrashekar R., 2007, *POS tagging for Sanskrit*, Ph.D. thesis, JNU

Cologne Digital Sanskrit Lexicon,2005, http://www.uni-koeln.de/phil-fak/indologie/tamil/mwd_search.html (accessed : 23 March 2005)

Dadhimatha, Pandita Sivadatta, 1929, The *Nāmaliṅgānuśāsana (Amarakośa)* of Amar asimha with the commentary (*Vyākhyāsudhā* or *Rāmāśramī*) of Bhanuji Dikshit, Nirṇaya Sāgar, Bombay

Huet Gerard, 2005, Sanskrit –French dictionary, http://pauillac.inria.fr/~huet/SKT/indo.html (accessed : 23 March 2005)

Jha Girish Nath et al, 2005, *Information technology applications for Sanskrit lexicography: case of Amarakośa*, procs of the 4th AsiaLex conference organized by NUS, Singapore

Jha Girish Nath et al, 2006, *Computational lexicography and Amarakośa : an online RDBMS approach*, Presented at the *National Seminar of Language and Interface*, Deptt of Linguistics, Delhi University

Lee Jong-cheol, 2005, Sanskrit Dictionary-Database, Academy of Korean Studies, Seoul, Korea http://www.hm.tyg.jp/~acmuller/ebti/dictionaries/sanskritdb.htm (accessed : 23 March 2005)

Mohanty et al, 2004, *Ontological analysis in Sanskrit wordnet*, *Procs of ICSLT-O-COCOSDA*, New Delhi, 2004

Ramanathan, A.A. 1978, Amarakosa with the unpublished South Indian commentaries, Vol. 1- 3, The Adyar Library and Research Centre.