



# South Asia Language Resource Center

## SALRC Grant Letter of Intent 2005-2006

Proposal title

Sindhi Digital Dictionary

Language(s)

Sindhi

Project period (Earliest start date is 8/15/2005;  
project must be completed by 8/14/2006)

8/15/2005 to 8/14/2006

Is this proposal an extension or expansion of a  
previously SALRC funded project? Yes

Faculty PI

Dr. Sarmad Hussain (Consultant)

Department

South Asia Language Resource Center  
(SALRC)

Office address

Judd Hall 207  
5835 S. Kimbark Avenue  
Chicago, IL 60637

Have you previously received SALRC funding  
for an unrelated project?

No

Requested funding amount

\$ 14,400

City

State

Zip code

Institution

The University of Chicago

Office telephone

(773) 834 3399

FAX

E-Mail

sarmad.hussain@nu.edu.pk



# South Asia Language Resource Center

      

**Description of proposal** (Please do not exceed three pages.):

You must respond to the points covered in the RFP including technology, timeline, budget, pedagogical approach and the unique aspect of this project compared to existing materials.

**Sindhi Digital Dictionary**  
*A proposal submitted to the SALRC*  
Sarmad Hussain  
[sarmad.hussain@nu.edu.pk](mailto:sarmad.hussain@nu.edu.pk)

**Overview.** We propose to construct a digital dictionary for Sindhi to address research and pedagogical needs related to the Sindhi language. The proposed dictionary project will be based on Mewaram Parmanand's *Sindhi-English Dictionary* (1866-1938, non-copyrighted), and lexicographic information contained there will be entered into a structured database (HTML format) suitable for use with a variety of computing applications. The digital dictionary will be searchable based on a wide range of content features, including grammatical features, meaning, Perso-Arabic text encoding, and Roman transcription. The project will be designed to be done



## South Asia Language Resource Center

by Dr. Hussain in collaboration with Dr. Jennifer Cole at University of Illinois Urbana-Champaign, and the data design and entry work will be carried out by Dr. Hussain.

**Statement of Need.** The immediate need for a Sindhi digital dictionary is for use with *Online Sindhi*, the web-course Sindhi language course currently being created by Dr. Cole with support from the SALRC. Online Sindhi offers a self-guided introduction to spoken and written Sindhi, and includes audio content in addition to Sindhi text in the modified Perso-Arabic script and its Roman transliteration. Students using Online Sindhi will need a searchable electronic dictionary in HTML format, where dictionary entries can be directly linked to lesson and script pages where new vocabulary is introduced. It is our goal that the digital Sindhi dictionary can also be used as a basic reference for scholars and students who want to access Sindhi literature on the web. For instance, the *Shah-jo-risalo*, a classic compilation of Sindhi sufi poetry, has recently been uploaded in searchable Unicode text at <http://www.freewebs.com/majidbhurgri/shah.htm>.

The Sindhi digital dictionary will be offered for inclusion in the electronic dictionary materials of the Digital Dictionaries of South Asia (DDSA) project at the University of Chicago. The DDSA project has plans to scan Mewaram's Sindhi-English dictionary, the same dictionary targeted for use in the project we propose. The primary advantage of our proposed Sindhi digital dictionary over the scanned image is that it will be in Unicode text format, and will therefore be searchable on the basis of the Sindhi-Arabic or Roman text encoding, as well as by other grammatical and usage features that are encoded in the data structure we will adopt.

Motivating the current proposal is the idea that the dictionary created to serve Online Sindhi should be constructed in a format that will allow its future integration with other electronic Sindhi materials. Towards this goal, we propose to adopt the dictionary format developed by Dr. Hussain for Urdu, which has been designed with the explicit goal of portability across a variety of electronic language and speech applications, including spell-checking, machine translation, text-to-speech, and other technologies for spoken and written Sindhi. The dictionary will also be valuable for linguistic research on Sindhi morphology, phonology and phonetics.

**Plan of Work.** Dr. Hussain, in consultation with Dr. Cole, will construct a word list for the project that will include approximately 2,000 word bases including all of the core vocabulary covered in the first-year Online Sindhi course. The dictionary information for each word in this list will be taken from Mewaram Parmanand's *Sindhi-English Dictionary* (1866-1938, non-copyrighted), which presents Sindhi words in the Sindhi-Arabic script. We envision a second and third phase of the dictionary project in which the coverage will be extended to 5,000 words, and then further to achieve complete coverage of Mewaram's dictionary, which contains approximately 20,000-25,000 entries. (Each entry is a morphologically distinct word. A single root morpheme may be the base for multiple entries which differ in derivational or inflectional morphemes.) Mewaram's dictionary is currently the most comprehensive Sindhi-English dictionary available. The same author has produced a multi-volume Sindhi-Sindhi dictionary unsurpassed in its coverage and which is widely viewed by Sindhi literary scholars as an outstanding lexicographic reference.

Dr. Hussain will produce (i) a 2,000 word lexicon of Sindhi with XML tags, (ii) GUIs to present the word data (1 GUI per word plus general GUIs), (iii) word-based search for searching through GUIs. The lexical entries will encode information such as spelling, part-of-



## South Asia Language Resource Center

speech, meaning and usage, as presented in Mewaram's dictionary. Dr. Cole will also provide input towards the XML tagset and for the GUI design. Drs. Cole and Hussain will be in regular email communication to oversee progress on the project and to deal with questions as they arise. We plan that future work on the project, undertaken in Phase II, will develop advanced search tools (e.g., to support wildcard search or search on other XML fields).

Comments of reviewers have also been received and more discussion on word grouping will be taken up and addressed during the design process. The corpus will not be accumulated in the first phase, but will be addressed directly in subsequent phases of the project.

**Budget Justification.** Our estimate of the budget needed for this project is \$14,400 for the first phase. This estimate is based on the known costs of the Urdu dictionary project recently completed by Dr. Hussain and his staff. The budget covers consultancy cost for Dr. Hussain, who will be undertaking the lexical development process. The work done by Dr. Cole at UIUC will be conducted as part of her regular research activities. Integration of the Sindhi digital dictionary with the Online Sindhi web-course will be done by Dr. Cole's current assistant on the Online Sindhi project, as part of the planned work for that project and will not require any new assistant salaries or other expenses. The proposed work will be done over a 12-month period.

Because the work will be carried out in Pakistan rather than in the US, there will be an enormous savings to the project. Furthermore, the experience of Dr. Hussain brings a savings in that he has recent experience with a similar project of larger scale (the Urdu dictionary project funded by Ministry of IT, Government of Pakistan), and will be able to use methodologies developed for that project.