Lexical database of the Torwali Dictionary

Inam Ullah

A paper presented at The Asia Lexicography Conference,
Chiangmai, Thailand
24th – 26th May, 2004

ABSTRACT

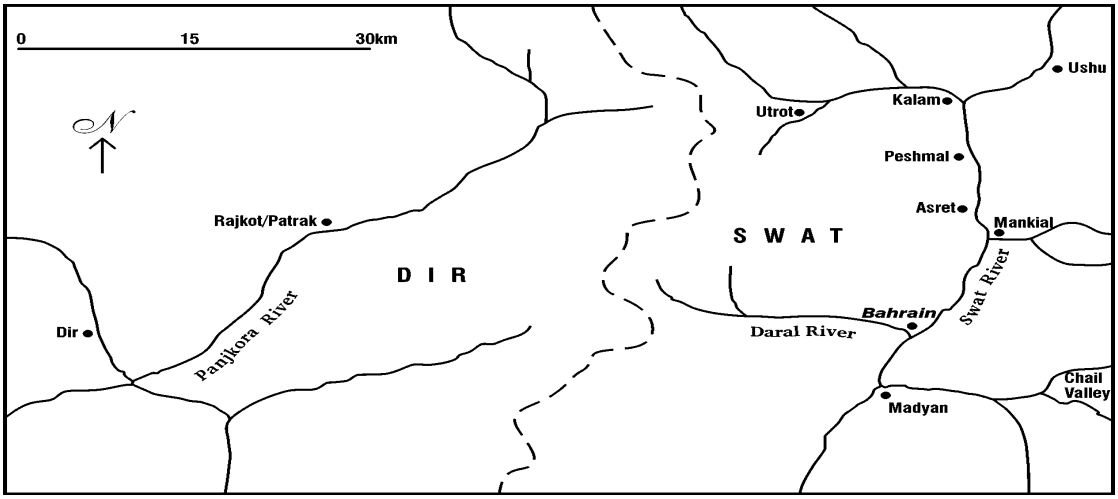
Torwali *is one of at least twenty four-lesser known languages of northern Pakistan that have been remained unwritten and previously less exposed to the academic community of the world. According to L.R Turner, 'these languages may not be significant politically, but historically they are of great importance'. Yet, there is hardly an institution in Pakistan that supports such activities as preservation and documentation of these lesser languages. Smaller groups or individuals strive to preserve and document their mother tongues, but due to the lack of a national policy on languages and low priority given to them, only foreigners take pain to explore and analyze these languages. The* ***Torwali-English Urdu-Dictionary Project*** *is an example of an indigenous initiative, that of a non-linguist lexicographer from within the speech community without any institutional support.*

In the paper that I propose, I will deal with several aspects of the lexical database of the ***Torwali*** *dictionary. My objective is to highlight the beginning of the dictionary project by providing some background on the geographic setting and genetic affiliation of the* ***Torwali*** *language as well as the methodology used for collecting and compiling lexical material and the problem of deciding orthography for a previously unwritten language. I shall also talk about the linguistic software 'shoebox' which is a wonderful lexicographic tool for entering, parsing and interlinearizing text which I adapted for the compilation of the dictionary at a later stage. I will outline the various lexical fields used in the database and the items, which have been included in different files, such as the main database, clans, plants, trees, proverbs, affixes, doubtful entries, etc. I will then mention my involvement with the University of Chicago for recording sound files of the example sentences in the database and its inclusion among the Less Commonly Taught Languages of Pakistan on the World Wide Web. I will conclude my discussion by describing some of the problems presently faced as well as some comments regarding the future of this project.*

GEOGRAPHIC OCCURRENCE OF THE LANGUAGE

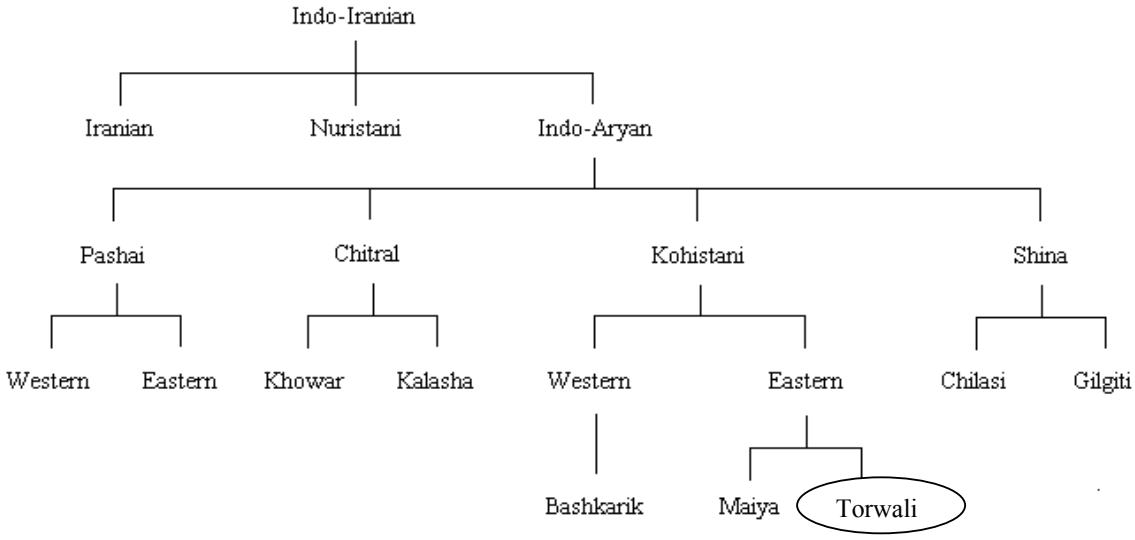
Torwali belongs to 'Kohistani' sub-group of 'Dardic' branch that constitutes Indo-Aryan family of languages. The speech community calls itself 'Kohistani' and the name 'Torwali' is given by the neighboring speech communities, like Kalam Kohistani, Indus Kohistani etc. It is spoken by 80,000 to 90,000 people in the northern hilly areas of Pakistan, more than one third of which have been migrated to the bigger cities in other parts of the country.

There are two dialects of Torwali, the *'Bahrain Dialect'*, which is the bigger one, and the *'Chail dialect'*. According to a socio-linguistic survey, there is 89% similarity between these two dialects. '*Bahrain Dialect'* is spoken along the 35 Km stretch of Swat River, north of Madyan to Asret and the '*Chail Dialect'* is spoken in the 8 Km long Chail valley, in the west of Madyan, (See map below). According to some researchers, Torwali was once a major language covering the whole area of today's District Swat.



GENETIC AFFILIATION OF THE LANGUAGE

Here is a figure showing genetic affiliation of the Torwali language with other Indo-Aryan languages, according to Richard Strand. See his website, http://users.sedona.net/~IngIndex0.html



MOTIVATION FOR THE PROJECT.

Being a mother-tongue speaker of Torwali, the author was always curious to know anything written about his language while a student at the University of Peshawar. For the first time he found materials on Torwali in the 'Linguistic Survey of India' by Sir George Grierson. Later, he studied other books on the languages of north Pakistan in which he found Torwali text with a lot of phonetic and semantic mistakes. This motivated him to work on his native language in order to present it more accurately to the academic community. In the meantime, fortunately, he was introduced with Dr. Joan Baart of SIL International (formerly Summer Institute of Linguistics), who was working on the neighboring Kalami language, also known as Gawri. It was Dr. Baart who encouraged

this author to write a dictionary for the Torwali language. Thus, in August 1996, the author started working on the dictionary compilation just picking up paper and pencil with the help of Platt's Dictionary of Urdu, Classical Hindi and English.

NEED OF THE PROJECT

For language learning and instruction few resources are more crucial than dictionaries. A good dictionary, specially of a previously unwritten language serves many purposes at one time. It meets the issue of orthography, documentation and preservation of the language, particularly since it is endangered.

Is Torwali an Endangered Language?

The author's view is a resounding 'yes'. Though the degree of endangerment may be debatable but the author considers Torwali an Endangered Language on the following grounds.

- Torwali is a small language of less than 0.1 m people
- Close to half of its speaker have migrated permanently to the bigger cities of Pakistan where use of their mother tongue is limited to the elderly people only
- Torwali speaking area is a tourist spot which makes the language exposed to linguistic encroachments
- Having no writing tradition in the language, it is vulnerable to extinction

The author is fortunate by enjoying the help of a group of young Torwali people who give their input on the issues of orthography and the spelling system. They acknowledge benefits of the Torwali Dictionary Project as the following.

Benefits Of The Project

- To bring the Torwali community into contact with the national and international mainstream.
- To contribute to the promotion of mother tongue education in the Torwali speaking people through increasing their opportunities for learning English and Urdu.
- To make linguistic material available to the academic community of the world for further research on the language.
- To record, document, and preserve an unwritten language of Swat Kohistan and to save it from extinction.

Orthography For The Unwritten Language

The issue of orthography has been resolved with the help of aforementioned Torwali people. There were four new consonants and one vowel which were to be represented by specially modified Perso-Arabic characters.

METHODOLOGY

There was no idea of proper methodology in the author's mind for dictionary writing. When a wordlist of about 1200 items prepared on paper, Dr. Baart introduced the author to the 'Shoebox method' by using index cards (3x5 inch) for writing each entry word

with Torwali (in Perso-Arabic script), English gloss, Part of Speech, Urdu gloss and paradigm etc. (because the author had no access to the computer technology) So, the database was transferred from a notebook to index cards and put into real shoe boxes alphabetically. About 2000 cards were written with pencil and arranged in the boxes.

For collecting lexical material the author maintained personal contacts with many Torwali speaking people, including family members, relatives, friends, students and language activists belonging to different locations of the Torwali speaking area. This enabled the author to find out variants and synonyms for many lexical items. In order to recall and collect maximum number of items the author used the idea of 'semantic domains' (though, at that time the author was unaware of this term). Students and friends were given particular domains to gather as many items as they can. For example, they were told to collect items on a variety of domains as 'agriculture', 'bullocks' and 'watermill' etc.

The Helpful 'Shoebox'

In the meantime, Dr. Baart introduced this author to the useful computer software, 'Shoebox' and spent many days teaching how to use 'Shoebox'. After some time, all the 2000 words data was transferred from the real shoe boxes to the computer 'Shoebox'.
The use of the Shoebox made a radical change in the mode of entering and managing the database. Shoebox proved to be highly useful for the author. Previously, it was a tedious job to search for a card containing data. It took several minutes to find a specific card for making important changes. But, to find a required record file on the computer is now a matter of seconds. The author found this software most useful for the following features.

- While entering a record, shoebox recognizes duplicate or multiple matches
- Having need of different fonts for a trilingual dictionary the typesetting problem now has been solved
- Right-to-Left and Left-to-Right script can be typed in the same record
- Shoebox can sort the database in any order
- Filtering is a very strong feature a field linguist can use in a number of ways for different purposes
- Browse, search and jump features are very useful
- Shoebox can open different databases at the same time with all the copying, cutting and pasting features available
- Above all, Shoebox has the export feature and the still more wonderful MDF tool

In short, this author has found the software very useful except a few drawbacks like,

- It is a complex software
- There is no easy solution for spell checks

Work done so far…

- ***Main Entries or Headwords = 5756***
- ***Subentries = 1688***

Each with maximum relevant information under respective lexical fields, at least with the following.

- Gloss Vernacular - Gloss English
- Reversal English - Definition in English
- Gloss National - Part of Speech
- Paradigm set - Paradigm Gloss etc. etc….

- About 2700 English definitions with character style codes for special typesetting, such as:
  - fv: for vernacular fonts
  - uc: for underline characters
  - ub: for underline bold
  - ui: for underline italic

- Etymology of about 400 entries
  - With English gloss of the Sanskrit
  - Reference from the Turner's Comparative Dictionary of Indo-Aryan languages

- About 75 names of tribes and clans
  - With genealogical information
  - Location of the tribe and clan

- Around 3000 example sentences of Torwali
  - Translated into English
  - Recorded on a DAT recorder
  - Edited and copied to CD

- 187 plants and trees names
  - With English names
  - About 75 entries with botanical/scientific names
- 39 Torwali proverbs with English translation
- 47 notes under Encyclopedic information as myths, historical and cultural information

Work Still To Be Done…

- Expansion of the ***database*** up to 7000 head words and sub entries
- Addition of the ***tonal description*** to all the entries and sub entries
- Work on the ***etymology*** of as many Torwali words as possible
- Inclusion of ***example sentences*** up to 5000 with English translation
- Completion of ***scientific names*** of all the flora and fauna mentioned in the database
- To complete addition of the ***character style codes*** throughout the database
- Addition of the ***special words and variants*** relating to Chail dialect of Torwali
- Addition of Torwali ***idioms and phrases*** under the proper headword
- ***Editing and spell checking*** thoroughly

Lexical Fields Used in the Electronic Database of Torwali and supported by MDF

No	Field Marker	Field Name	Field Uses	Fonts Used
1	\lx	Lexical entry	Lexeme or headword of the lexical entry; actually, phonemic form of the word.	Standard Orientalist
2	\va	Variant	Variant forms in the same or different dialects of Torwali	Standard Orientalist
3	\hm	Homonym Number	Lexemes that sound or are spelled the Same but have no semantic relationship	English numerals
4	\ph	Phonetic Form	Basically for phonetic form but, used for tone of a few hundred words, marked by L, H, LH and HL.	English
5	\gv	Vernacular Gloss	Torwali word	Modified Perso-Arabic fonts
6	\ge	English Gloss	Short English gloss	English
7	\re	Reversal (English)	Used for an English index of the database	English
8	\sn	Sense Number	Sense number under a sub entry	English numerals
9	\de	English Definition	A longer English definition	English
10	\gn	National Gloss	Urdu gloss for the headword	Urdu
11	\dn	Definition in National Language	MOSTLY UNFILLED	Urdu
12	\ps	Part of Speech	Part of speech of a headword or a sub entry. This field is the only one used as range set for consistency in the database	English
13	\pd	Paradigm Set	Used primarily for inflated singular or plural forms of a word	Standard Orientalist
14	\pde	Paradigm Gloss (English)	Used for English gloss of the paradigm sets	English
15	\cf	Cross Reference	Cross Reference of another headword	Standard Orientalist
16	\se	Subentry	Used for derivative words of the headword with the normal range sets	Standard Orientalist
17	\lt	Literal Meaning	Literal meaning of a word or phrase within a record	English
18	\et	Etymology	This field only refers to an original Sanskritic word or phrase from which the Torwali word has been evolved.	Standard Orientalist
19	\eg	Etymology Gloss	Gloss of the original Sanskritic word or phrase	English
20	\es	Etymology Source	Up till now the source has been noted for a few hundred Torwali entries as T-1235, T-11640, T-9202 etc., referring to a headword in Turner's Dictionary of Indo-Aryan Languages for an original Sanskrit word	English
21	\bw	Borrowed Word	Used for a borrowed language as 'Pash' for Pashto or 'Urd' for Urdu etc.	English
22	\mn	Main Entry	Used to refer a minimal entry back to its main entry	Standard Orientalist
23	\sy	Synonym	Used to refer a Torwali synonym of the headword	Standard Orientalist
24	\ue	Usage (English)	This field has been used to mention a restriction in usage or referring a Torwali	English

			dialect, such as, 'Chail Dialect' of Torwali	
25	\xv	Example (Vernacular)	Used for a Torwali Example sentence in a local or cultural background of the speech community	Standard Orientalist
26	\xe	Example (English)	English Translation of the example sentence	English
27	\ee	Encyclopedic Information	Extra cultural and historical information about a lexical item such as a myth, tribe and clan etc.	English
28	\sc	Scientific Name	Used mainly for Botanical and Zoological names of plants and animals respectively	English

Lexical fields not supported by MDF but used by author's only.

1	\nt	Note	Simple working notes, questions etc relating to the . Not meant for printing	English
2	\addNotes	Additional Notes	Working notes worth further investigation.	English
3	\inf	Informant	Name and relevant information about the informant to check authenticity	English

Association with the University of Chicago

In 1999, Dr. James H. Nye., director South Asia Language Center developed and submitted a proposal under the tittle of "Digital Dictionaries of South Asia" before the US department of Education. The project aimed at making high-quality dictionaries of the languages of south Asia universally available in digital format. The main objective of the project has been to make electronic dictionaries accessible to the international community via the World Wide Web. Dr. Nye. contacted this author and invited to include his lexical database after examining a few sample pages of the first print out. This author agreed in principal that the deal will be a non-exclusive one, meaning that, he will be free to publish the database in print anywhere and also to sell the electronic file again to someone else if he wants to. Similarly, the University of Chicago will be bound to present the data on the web on a not-for-profit basis and will honor the intellectual property rights of the dictionary compilers. The database will be available on http://dsal.uchicago.edu/dictionaries

Digital Dictionaries of Less Commonly Taught Languages of Pakistan.

Based on the above, a smaller project developed under the title of "Digital Dictionaries of the Less Commonly Languages of Pakistan" that included three languages of the North Pakistan, that is, Pashto, Khowar and Torwali. This project is significant because the development of instructional materials and programs in the smaller languages spoken in Pakistan has lagged far behind the development of materials in other Less Commonly Taught Languages. This project also includes the addition of digital audio files into the web edition of this dictionary which is an innovative feature of this project. Digital audio

files are to be linked to the on-line dictionary so the students will be able to hear the pronunciation of the entry words and selected sentences as examples of word usage. In the existing database, some 3000 example sentences have been added and recorded so far.

Audio Strategy

Once example sentences of Torwali are created and translated into English, recording takes place in the sound proof environment at the Allama Iqbal Open University Islamabad. There are appropriate pauses between two successive sets of entry words and example sentences. Each session of the recording is separated from others by putting a filter marker in the lexical file of the database. In order to ensure high a quality recording, a DAT recorder has been purchased by the University of Chicago and located at the American Institute of Pakistan Studies, AIPS. After the recording, a print out of the entry word and example sentences are carefully marked with the pauses made during the recording so as to facilitate the editing process.

Offer from the National Language Authority (Muqtadra Qaumi Zuban)
The National Language Authority of Pakistan has formally offered this author to re-arrange the existing database for the development of a Torwali-Urdu glossary, as part of their policy to relate regional and local languages with the national language of Pakistan. This would exclude the English part of the database, but the work on this project has not yet started due to logistic problems.

The problems ahead….

- One of the main problems is that the project had not been planned and designed from the very beginning. The author is basically a teacher in the government school system. He started compiling this dictionary as a hobby, but it is a time-consuming and laborious job, though, he enjoys the advantage of being a mother-tongue speaker.
- There has been the lack of proper professional and technical expertise required for the work of dictionary writing. Thanks to Mr. Wayne Lunsford, director, Frontier Language Institute (FLI), who has been giving linguistic training to many language activists of the area, including this author. This will, hopefully, put this project on the right tract, but the author feels that a short course on a topic like *Dictionary Project Management* would enhance his capacity.
- The author is reluctant whether to publish the existing database in the form of a wordlist or search for subventions to publish it as a complete dictionary.
- The author also has personal ambitions to earn an academic degree in lexicography (MA or Ph.D) and train language workers of his country who want to compile dictionaries of their mother tongues.

Let’s hope for the best and thank you all!