Digital Dictionary Development for Torwali, a less-studied language: Process and Challenges

Inam Ullah
Torwali Research Forum
Bahrain, Swat, Khyber Pakhtunkhwa
Pakistan
torwalpk@yahoo.com

Abstract

Torwali is an endangered and less-studied language spoken in the north of Pakistan. Recently, the community celebrated publication of the first ever Torwali dictionary both in print and an online version. This paper discusses issues and challenges regarding lexicography of a previously non-written language; from data collection by the native speakers having no set goals and training or institutional support, to organization and presentation of the data for producing multiple versions of the dictionary. The first section describes the process of developing the database using the methods of wordlists and semantic domains. The proceeding sections describe the technical development of its printed and online version in detail, and discuss orthographical, technical, computational and social concerns of the project. The paper concludes with recommendations for future dimensions of the present work and for similar projects with special consideration to lexicographical work on non-written languages.

1. Introduction

1.1 Torwali language

Torwali belongs to the Kohistani sub-group of the Indo-Aryan Dardic languages, spoken in the upper reaches of district Swat of northern Pakistan. It has two dialects (the Bahrain and Chail dialects), with a total of approximately 90,000 to 100,000 speakers. Close to half of the population has migrated to bigger cities where language shift is a common phenomenon.

1.2 Motivation or need for the project

The project initiator, a mother-tongue speaker of Torwali, when studied the written materials on the language for the first time, found many semantic and phonetic errors. This initially motivated him to work on his native language in order to present it more accurately to the academic community. Later, after receiving encouragement from the community elders, he decided to compile a dictionary of Torwali based on the idea that dictionaries can be a crucial resource for language learning and instruction, particularly with regard to endangered languages. A good dictionary can address issues of orthography, documentation and language preservation. Previously, the locals found it difficult to write Torwali language using the alphabets of neighbouring regional languages or the national language as some of its peculiar sounds had no representation in their alphabets. The main goals were, therefore, to record, document and preserve a hitherto unwritten language of Swat Kohistan and thus, to safeguard it for the future generations.

1.3 Intended audience of the dictionary

Initially, the intended audience was the academic community. The aim was to provide them with error-free material of Torwali for further research. However, later, in view of the interest of the Torwali community, it was decided that the intended audience would include both the academic and the speech communities. During the compilation process, numerous difficulties

Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages: Vol. 2 Extended Abstracts, pages 11–16,
Honolulu, Hawai'i, USA, February 26–27, 2019.

emerged regarding decisions to present the data in a way that would benefit both the communities equally. As a result, it was decided that the primary audience would include those Torwali-speaking Torwalis and Torwali-learning Torwalis whose preference is the socio-cultural information like clans, place names, medicinal plants, cultural items, myths and oral traditions. Thus, the final product of the database is intended for students, Torwali speakers across the globe, tourists, and researchers.

1.4 The selection of dialect

While compiling data for Torwali dictionary the 'Bahrain dialect' was decided to be the standard dialect because: (i) it is spoken by a larger number of Torwali speakers; (ii) Bahrain is the cultural, political and administrative center of Torwali community; and, (iii) the compiler of the data speaks Bahrain dialect of Torwali.

Despite the above-mentioned decision, some words peculiar to the Chail dialect were added to the database with the tag of 'Chail dialect' However, it was not possible to enter Chail variation of every Torwali word due to space issues.

1.5 Previous literature

Several western researchers have worked on this language. In 1880, John Biddulph published *Tribes of the Hindoo Koosh*, which contained the first linguistic description of the Torwali language. The most extensive work on the language was carried out by Sir George Grierson which is known as *Torwali: A Dardic language of Swat Kohistan* (1929). In the late 1980s, SIL International carried out a sociolinguistic survey in northern Pakistan which included the Torwali community. Wayne Lunsford's work, *An Overview of Linguistic Structures in Torwali, A Language of Northern Pakistan* (2001) is another major work on Torwali after Grierson's.

2 The undecided project goals

There were no set goals at the beginning of the project. It was all about 'writing a dictionary of Torwali'. The compiler, being a government school teacher, had no previous knowledge or training of lexicography. He worked on the project as a hobbyist and therefore, did not time-frame it. However, he stored the database in an electronic format to serve the academic purposes of research.

3 Methodology: Printed Version

Major part of the database was developed over the past one and half decade by the active help of the author's students, colleagues, friends and relatives. Data was collected on index cards and paper and was entered in to the computer program called 'Shoebox'. 'Shoebox' was replaced with the improved version 'Toolbox'

3.1 Data collection and verification

Both wordlist and semantic domain methods were used for the data collection. Being bilingual, the author used Urdu wordlists for recalling Torwali words. But specific cultural items, plants and animal names about which the author himself was unaware could not be recorded using this method. He, therefore, adopted the method of semantic domains. He asked his Torwali students, friends and family members to make lists of words relating to a specific semantic domain or sub-domain. For example, a group of students was asked to bring a labeled sketch of the interior of a watermill.

To ensure correctness and completeness, cross-checking and verification of the data was conducted through multiple sources within the community, such as, various people living in different localities (valleys and side-valleys of the indigenous area as well as in different urban centers).

3.2 Expansion of the database

Printouts of the existing database were distributed among the Torwali speakers for verification and further addition of lexical items that had been left out. At the same time the author made a partial use of lists of *Semantic Domains* prepared by SIL International under The Dictionary Development Process (DDP) http://www-01.sil.org/computing/ddp/ddp_wordcoll.htm?. This process facilitates lexicographers to collect words for the development of dictionaries of minority languages. It helped increase the database from 5200 to 8000 lexical entries by including names and related information of places, plants and clans as well as idioms, proverbs and, words related to Chail dialect.

3.3 Verification and refinement of specific semantic domains

Lists containing words of special domains were verified by the Torwali practitioners of the field concerned. For example, Torwali words for diseases and ailments were verified by and discussed with qualified medical practitioners living in the indigenous area. Similarly, items relating to forests, botany, watermills or agriculture were verified by the Torwali speakers working in the respective fields.

3.4 Consistency Checks

Consistency checks were employed both automatically and manually, depending on the availability of the features in the Toolbox program in which the database was stored and handled. For example, Toolbox consistency checks support parts of speech but not spell checks, particularly in the national and source languages. Thus, spell checks were carried out manually.

3.5 The use of lexique pro

Toolbox file opened in Lexique Pro and was exported to Microsoft Word in a standard dictionary format. Thus final export was made through Lexique Pro instead of Toolbox due to repeated problems faced while exporting. But there were lots of formatting issues which had to be fixed manually. For example, various Toolbox field markers, such as, \ue, \vr, \lt, \va and \ps were changed into Urdu script to meet the practical needs of developing Torwali-Urdu print version as they were not supported by Lexique Pro during export process. Proofreading of the entire exported file was done page by page. Arising technical issues were resolved through the 'trial and error' principle.

4 Methodology: Online Torwali Dictionary (OTD)

4.1 Overall architecture (OTD)

To develop the online version of the dictionary, the lexicon data was contained in a Toolbox database file. The Toolbox database file was taken to Lexique Pro to be converted into an xml formatted file. The corresponding xml file was exported, using MDF (Multi- Dictionary Formater). The xml file was used as input to the dictionary's website application which transformed the word detail into html formatted content for users.

4.2 Microstructure and macrostructure of the OTD

Lexique Pro generated xml file of lexicon data in LIFT (Lexicon Interchange FormaT) language. The detailed xml encoding format of LIFT formed the microstructure. The method of accessing information by the user formed the macrostructure.

Macrostructure and microstructure were navigable to some extent. Words were linked on the basis of parts of speech, and synonyms were cross referenced as navigable links deeming OTD to be termed as hyperlexica (Gibbon, 1999).To ensure that user accessed the dictionary information with ease, the following navigational ways were made available on the website:

- Torwali alphabet was enlisted on the website. Through this, user could have the list of words starting with the selected Torwali letter.
- User could navigate the words by syntactic categories, affixes or phrases. These indices would display a corresponding wordlist making it easier to find the word. Native users could ponder upon the list and find out if some words were missing. It could help teachers make a lesson plan to teach a certain category to Torwali-learners.
- Users could search any Torwali word by entering it using onscreen Torwali keyboard provided on the website. If a word existed in the lexicon, then the result would either be a single word or more than one word if its homonyms existed.
- Through any of the above three ways a user could reach a word or word list. The word detail would be displayed by clicking on the desired word. It would contain all the information fields already displayed in the hard copy of the dictionary except the usage field. ('Usage' depicts the geographical usage, obsolescence and vulgarity of the word.)
- Reverse lookup; Urdu to Torwali navigation for words is also one of the

features of the online Torwali dictionary. Urdu word list would help user traverse back to Torwali equivalent.

4.3 Orthography Issues

Since Torwali was an oral language until the start of the lexical compilation, the orthography issues constituted a major source of problem during the process of compilation.

4.3.1 Decision on the script

In order to be in cultural and historical harmony with the regional language Pashto and national language Urdu, it was decided with the help of the community activists that Perso-Arabic script would be used for writing Torwali, because both the Pashto and Urdu languages are written in Perso-Arabic script.

4.3.2 Characters for peculiar sounds

There are five distinct sounds in Torwali which are absent in both Pashto and Urdu. Grierson identified and mentioned these sounds in his work (Grierson, 1929). A Dutch phonetician (Baart, 1999) worked on the neighboring Kalami (or Gawri) language in the 1990's and presented them to the community people for approval. After a detailed discussion all of them recognized the five sounds (figure 3).

4.3.3 Standard spellings

During serial workshops organized for discussion on language issues among Torwali language activists, it was noticed that many words were being written with different spellings. This issue was also evident in the database where non-unique words were quite often found for the same lexical item. The participants decided to adopt the spellings which occurred with high frequency.

4.4 Technical Issues

4.4.1 Conversion of legacy fonts into Unicode fonts

After using 'Shoebox' program for many years, Torwali lexical data had to be shifted to its newer version 'Toolbox'. Hence, all the legacy fonts needed to be converted to Unicode supported fonts. Almost all the characters representing Torwali sounds were assigned Unicode positions except the Voiced Retroflex Affricate.

4.4.2 Torwali support in Nafees Pakistani Web Naskh

Center for Language Engineering, CLE (formerly CRULP) developed the Burushaski-Torwali-Khowar (BTK) font which is a character-based Nafees Pakistani Web Naskh Open Type Font in 2009. It was an extension of Nafees Web Naskh supporting several regional languages including Torwali in addition to Urdu.

4.4.3 Torwali keyboard development

In order to support Torwali characters to be typed easily along with Urdu characters, Torwali keyboard was developed by Center for Language Engineering, CLE (formerly CRULP). This keyboard was based on and similar to the Urdu Phonetic Keyboard so that the additional characters for Torwali Language could be typed easily along with the regular Urdu characters.

4.5 Issues relating to XML file

4.5.1 Some of the Torwali examples were not exported to xml format by Lexique Pro

According to LIFT, xv field in word entry contained examples in vernacular language and xe, xn, xr fields contained examples in English, national and regional languages. If xv did not exist, other example fields could not be exported to xml format. In case of Torwali dictionary, xv field contained examples in the form of IPA symbols and xr contained examples in Torwali language. In some cases xv field did not exist or in other words pronunciation of example sentences did not occur. Therefore dummy xv was inserted where xv was empty by using Toolbox, so that xml element corresponding to xr could be generated and thus displayed to the user. As xv-value was not to be displayed in hard copy or on website therefore dummy value could be used to save time and insert remaining pronunciations of Torwali example sentences afterwards.

4.5.2 Text formatting for hard copy dictionary

Toolbox was used to compile and manipulate the lexicon. However, its export features did not work well for publishing hard copy of the dictionary. For

this purpose Lexique Pro was used. There were default formatting styles (known as Multiple entry style) for each of the fields in a word entry. These styles were used by Lexique Pro during the process of export to HTML or WORD format.

4.5.3 Sorting of non-written languages

As Torwali was not a written language therefore collation sequence was not readily available for it. Though, collation rules had been explicitly mentioned in Toolbox, diacritics were not handled as ignorable characters due to which sorting was interrupted. The presence of diacritics caused the word to be processed in sub-sequences. Therefore, the hard copy of the dictionary was not properly sorted. This discrepancy was later removed and headwords were displayed in a proper sequence in the online dictionary version.

4.5.4 Gloss field and reverse lookup

In gloss field, semicolon is used to separate the multiple gloss terms. In Urdu gloss \gn, Urdu semicolon was used but was not recognized as a separator. Therefore, all the gloss field content was handled as single gloss term.

4.5.5 Encoding

Word detail of the Torwali-Urdu was in XML format therefore Unicode (that is, utf-8) encoding had to be used by the website. Secondly, Urdu and Torwali characters could not be presented by ASCII encoding. This is because web application configuration is set for languages using Unicode, otherwise the characters appear illegible on the interface.

4.6 Social and Other Issues

Like every living language Torwali has also many taboo and slang words. Torwali natives differed in their treatment of these words in the dictionary of their language. Some suggested that these words must be avoided as they may create wrong impression among the children and 'outsiders' about the community. Others said that these words were a part of their language and had to be recorded. After long sessions of discussions with community activists and elders it was decided that obscene slangs were to be avoided but words with offending connotations could be tagged with 'offending'. Similarly, some clans with shady histories did not want their historical information to become a part of the dictionary. Therefore, their names were included but not their history.

5 Future Work

Based on the existing database several enhancements can be made to enrich the practical uses of the dictionary.

Indexing on the basis of semantic domains can be incorporated in the interface.

The Torwali grammatical and collocation information can be enhanced to form a useful resource for Torwali to Urdu translational work leading to localization. When translating a sentence from source to target language, the context of the word sometimes changes the choice of word in target language (Saleem, 2007). Such constructions can be resolved when proper collocations and grammar of words are given.

The interactive interface for users can be added to contribute linguistic information of a new word or to an existing word. After the linguistic verification of the contributed information, it can either be approved and added to dictionary or disapproved.

Torwali to English dictionary can be developed and the corresponding online interface can be merged to the existing one - OTD.

Talking Torwali Dictionary can be produced to help community members living away from the indigenous area to learn the language of their ancestors.

Example sentences are good resource for better explanation of a word in a dictionary. There are only 1200 example sentences which need to be expanded under each lexical entry.

Specific information needs to be added for plants and animals rather than the generic formation as "a kind of…".

6 Recommendations

- The 'Usage' field should be exported by Lexique Pro. Currently, LIFT stated that <usage> element corresponding to \et exists but it is not exported by Lexique Pro.
- Collation sequence of non-written and less taught languages should be included by collation consortia, so that these can be readily available to linguistic tools. Collation sequence leads to properly sorted language data which is more efficient to navigate and manipulate.
- For non-written or less taught languages, data collection is quite a difficult task. Therefore, instead of top-down approach, bottom-up approach (Carr, 1997) is more helpful. Especially through emails or forums or a dictionary website page dedicated to user contribution are the easy and fast ways to collect the data. These small contributions can be of great benefit to all.
- Many cultural items, whose precise alternatives are not available in target languages, are best explained with the help of drawings and pictures.
- Idioms and proverbs embody the essence of a language. There are about 600 idioms and proverbs in the database which can fairly be expanded to thousands.

7 Acknowledgments

We want to thank all those who contributed in making possible the creation of Torwali dictionary: especially, the Torwali community who very persistently supported the compiler of this work, both practically and morally; National Geographic (http://www.nationalgeographic.com) who supported the completion and editing of the Torwali lexicon; the International Development Research Center (IDRC), Ottawa, Canada who funded the project for bringing the Torwali lexical data online; and the University of Chicago who initially supported the development of the dictionary content.

8 References

Baart, Joan L.G. 1999b. *'A Sketch of Kalam Kohistani Grammar'. Islamabad: National Institute of Pakistan Studies and Summer Institute of Linguistics. (Studies in Languages of Northern Pakistan Vol. 5).*

Carr, M. 1997. *'Internet Dictionaries and Lexicography.' Internal Journal of Lexicography, vol. 10 No. 1. 1 Feb 2011. http://ijl.oxfordjournals.org/content/10/3/209.full.pdf+html.*

Gibbon, D. 2000. *'Computational Lexicography.' In F. van Eynde and D. Gibbon (eds.), Lexicon Development for Speech and Language Processing. Dordrecht: ELSNET, Kluwer Academic Publishers, 1-42.*

Grierson, George. A. 1929. *Torwali: An Account of a Dardic Language of the Swat Kohistan. London: Royal Asiatic Society.*

Lunsford, Wayne. 2001. *An Overview of Linguistic Structures in Torwali, A Language of Northern Pakistan. M.A. Thesis, University of Texas at Arlington.*

Martin, J. B. & Mauldin, *M. M. 1997. 'Practical and Ethical Issues in Lexicography: Examples From The Creek Dictionary Project.' In C.Pye (eds.), 1996 Mid-America Linguistics Conference Papers, 565-573.*

Saleem, M. I. 2007. *'Bilingual Lexicography: Some Issues with Modern English Urdu Lexicography – a User's Perspective.' Linguistik online. 1 Feb 2011. http://www.linguistik-online.com/31_07/saleem.pdf/.*

Ullah, Inam. 2004. *'Lexical Database of the Torwali Dictionary.' In The Asia lexicography conference. Chiangmai: Payap University.*