Documenting Languages in Danger

Inam Ullah

Chairperson,
Mother Tongue and Heritage
for Education and Research (MOTHER)

"Languages are not only tools of communication; they also reflect a view of the world. Languages are vehicles of value systems and cultural expressions and are an essential component of the living heritage of humanity. Yet, many of them are in danger of disappearing."

UNESCO on Language Endangerment

We will discuss about….

- What is Language Endangerment?
- Languages of Pakistan
- Factors of Language Endangerment
- Why should we care?
- What is Language Documentation?
- Some popular tools of documentation
- Organizations involved in Language Documentation
  - Efforts on International Level
  - Efforts on National Level
  - Efforts on Regional Level
  - Efforts on Community/Local Level
- Process and Steps of Language Documentation
- What Language Technology can do for saving Endangered Languages?
- Computational Resources needed for Pakistani Languages
- Related activities

What is Language Endangerment?

An **endangered language** is a language that is at risk of falling out of use as its speakers die out or shift to speaking another language.

Language loss occurs when the language has no more native speakers, and becomes a "**dead language**". If eventually no one speaks the language at all, it becomes an "**extinct language**".

Languages are currently disappearing at an accelerated rate due to the processes of globalization and neo-colonialism, where the economically and politically powerful languages dominate other languages.

The top 20 languages spoken by more than 50 million speakers each are spoken by 50% of the world's population, whereas many of the other languages are spoken by small communities, most of them with less than 10,000 speakers.

Number of languages is not precise

No precise number of contemporary languages in the world is known, and it is not well defined what constitutes a separate language rather than a dialect. Estimates vary depending on the extent and means of the research undertaken, and the definition of a distinct language and the current state of knowledge of remote and isolated language communities. The number of known languages varies over time as some of them become extinct and others are newly discovered.

Language Position at world level

- ***Total Number of Languages = 7000***
  - ***Definitely endangered = 646***
  - ***Severely endangered = 527***

Languages of Pakistan

(ethnologue.com)

Aer	Farsi, Eastern	Kamviri	Pashto, Central
Badeshi	Gawar-Bati	Kashmiri	Pashto, Northern
Bagri	Ghera	Kati	Pashto, Southern
Balochi, Eastern	Goaria	Khetrani	Phalura
Balochi, Southern	Gowro	Khowar	Sansi
Balochi, Western	Gujarati	Kohistani, Indus	Savi
Balti	Gujari	Koli, Kachi	Shina
Baluchi	Gurgula	Lahnda	Shina, Kohistani
Bateri	Hazaragi	Lasi	Sindhi
Bhaya	Hindko, Northern	Loarki	Sindhi Bhil
Brahui	Hindko, Southern	Marwari	Torwali
Burushaski	Jadgali	Memoni	Urdu
Chilisso	Jandavra	Od	Ushojo
Dameli	Kabutra	Ormuri	Vaghri
Dehwari	Kachchi	Pahari-Potwari	Wakhi
Dhatki	Kalami	Pakistan Sign	Waneci
Domaaki	Kalasha	Language	Yidgha
English	Kalkoti	Panjabi, Western	

Position at National Level

(UNESCO atlas of Endangered Languages of Pakistan)

- ***Definitely endangered languages:***

Gawri, Bateri, Gwarbati, Kati, Kundal Shahi, Ormuri, Palula, Torwali, Ushojo, Wakhi, Yidgha

- ***Severely endangered languages:***

Chilisso, Dameli, Domaaki, Gawro, Kalasha, Kalkoti

Factors of Language Endangerment

UNESCO Experts propose 9 factors of vitality and endangerment in measuring the level of endangerment of the world's languages. These are:

- 1. Intergenerational language transmission;
- 2. Absolute numbers of speakers;
- 3. Proportion of speakers within the total population;
- 4. Loss of existing language domains;
- 5. Response to new domains and media;
- 6. Materials for language education and literacy;
- 7. Governmental and institutional language attitudes and policies;
- 8. Community members' attitudes towards their own language; and
- 9. Amount and quality of documentation.

Factor 1

Degree of Endangerment	Grade	Speaker Population
Safe	5	The language is used by all ages, from children up.
Unsafe	4	The language is used by some children in all domains; it is used by all children in limited domains.
Definitively endangered	3	The language is used mostly by the parental generation and up.
Severely endangered	2	The language is used mostly by the grandparental generation and up.
Critically endangered	1	The language is used mostly by very few speakers, of great-grandparental generation.
Extinct	0	There exists no speaker.

Factor 2-3

Degree of Endangerment	Grade	Proportion of Speakers Within the Total Reference Population
Safe	5	All speak the language.
Unsafe	4	Nearly all speak the language.
Definitively endangered	3	A majority speak the language.
Severely endangered	2	A minority speak the language.
Critically endangered	1	Very few speak the language.
Extinct	0	None speak the language.

Factor- 4

Degree of Endangerment	Grade	Domains and Functions
Universal use	5	The language is used in all domains and for all functions.
Multilingual parity	4	Two or more languages may be used in most social domains and for most functions.
Dwindling domains	3	The language is in home domains and for many functions, but the dominant language begins to penetrate even home domains.
Limited or formal domains	2	The language is used in limited social domains and for several functions.
Highly limited domains	1	The language is used only in a very restricted domains and for a very few functions.
Extinct	0	The language is not used in any domain and for any function.

Factor-5

Degree of Endangerment	Grade	New Domains and Media Accepted by the Endangered Language
Dynamic	5	The language is used in all new domains.
Robust/active	4	The language is used in most new domains.
Receptive	3	The language is used in many domains.
Coping	2	The language is used in some new domains.
Minimal	1	The language is used only in a few new domains.
Inactive	0	The language is not used in any new domains.

Factor-6

Grade	Accessibility of Written Materials
5	There is an established orthography, literacy tradition with grammars, dictionaries, texts, literature, and everyday media. Writing in the language is used in administration and education.
4	Written materials exist, and at school, children are developing literacy in the language. Writing in the language is not used in administration.
3	Written materials exist and children may be exposed to the written form at school. Literacy is not promoted through print media.
2	Written materials exist, but they may only be useful for some members of the community; and for others, they may have a symbolic significance. Literacy education in the language is not a part of the school curriculum.
1	A practical orthography is known to the community and some material is being written.
0	No orthography available to the community.

Factor-7

Degree of Support	Grade	Official Attitudes Toward Language
Equal support	5	All languages are protected.
Differentiated support	4	Minority languages are protected primarily as the language of the private domains. The use of the language is prestigious.
Passive assimilation	3	No explicit policy exists for minority languages; the dominant language prevails in the public domain.
Active assimilation	2	Government encourages assimilation to the dominant language. There is no protection for minority languages.
Forced assimilation	1	The dominant language is the sole official language, while non-dominant languages are neither recognized or protected.
Prohibition	0	Minority languages are prohibited.

Factor-8

Grade	Community Members' Attitudes toward Language
5	*All* members value their language and wish to see it promoted.
4	*Most* members support language maintenance.
3	*Many* members support language maintenance; others are indifferent or may even support language loss.
2	*Some* members support language maintenance; others are indifferent or may even support language loss.
1	Only *a few* members support language maintenance; others are indifferent or may even support language loss.
0	*No one* cares if the language is lost; all prefer to use a dominant language.

Factor-9

Nature of Documentation	Grade	Language Documentation
Superlative	5	There are comprehensive grammars and dictionaries, extensive texts; constant flow of language materials. Abundant annotated high-quality audio and video recordings exist.
Good	4	There is one good grammar and a number of adequate grammars, dictionaries, texts, literature, and occasionally-updated everyday media; adequate annotated high-quality audio and video recordings.
Fair	3	There may be an adequate grammar or sufficient amount of grammars, dictionaries, and texts, but no everyday media; audio and video recordings may exist in varying quality or degree of annotation.
Fragmentary	2	There are some grammatical sketches, word-lists, and texts useful for limited linguistic research but with inadequate coverage. Audio and video recordings may exist in varying quality, with or without any annotation.
Inadequate	1	Only a few grammatical sketches, short word-lists, and fragmentary texts. Audio and video recordings do not exist, are of unusable quality, or are completely un-annotated.
Undocumented	0	No material exists.

Levels of Endangerment in Languages

UNESCO *distinguishes* ***four*** *levels of endangerment in languages, based on intergenerational transfer.*

- **Vulnerable**: Most children speak the language, but it may be restricted to certain domains (e.g., home).
- **Definitely endangered**: Children no longer learn the language as mother tongue in the home.
- **Severely endangered**: Language is spoken by grandparents and older generations; while the parent generation may understand it, they do not speak it to children or among themselves.
- **Critically endangered**: The youngest speakers are grandparents and older, and they speak the language partially and infrequently.

Why should we care?

Language extinction is not new—languages have been dying since ancient times. However, languages are becoming extinct today at an alarming rate. Of the nearly 7,000 languages in the world today, some 3,000 (43%) are endangered

Why should we care?

Experts have predicted that in the worst-case scenario 90% of all languages will be extinct within 100 years; in the best-case scenario, only 50% will survive, and just 10% are considered safe during the next century (see Krauss 1992).

Why should we care?

Languages not being learned by children are not just ***endangered***, they are ***doomed***.

Can we apply this statement to some Pakistani Languages?

Why should we care?

We should be concerned over the crisis of language loss, for compelling reasons.

(1) Human concerns : Languages are treasure houses of information on literature, history, philosophy, and art. Their stories, ideas, and words help us make sense of our lives and the world around us.

(2) Lost knowledge : Specific knowledge is often held by the smaller speech communities of the world—knowledge of medicinal plants and cures, identification of plants and animals yet unknown scientifically, new crops, etc.

(3) Scientific understanding of human language: Linguists have the goal of understanding what is possible and impossible in human languages, and through the study of human language capacity, of advancing knowledge of how the human mind works. For these goals, language extinction is a disaster.

(4) Human rights: Language loss is often not voluntary; it frequently involves violations of human rights, with oppression or repression of speakers of minority languages. It is a matter of injustice when people are forced to give up their languages by repressive regimes or prejudiced dominant societies.

What is Language Documentation?

Language documentation is the documentation in writing and audio-visual recording of grammar, vocabulary, and oral traditions (e.g. stories, songs, religious texts) of a language. It entails producing descriptive grammars, collections of texts and dictionaries of the languages.

Major Steps for Language Documentation:

- **Recording** *stories, narratives, personal histories, poetry, songs etc.*
- **Transcribing** *the recorded material (writing down phonetically).*
- **Translating** *into language(s) of wider communication.*
- **Analyzing** *the material to uncover the structure and function of the language.*
- **Compiling** *of corpus, glossary, dictionary and grammar of the language.*
- **Disseminating** *or sharing the material with academic and speech community.*
- **Archiving** *the recorded and processed material for later use and future generations.*

Priorities of documentation

- *to create a range of high quality materials to support description of a variety of language phenomena*
- *to enable the recovery of knowledge of the language even if all other sources are lost*
- *to generate resources in support of language maintenance and/or learning*

Media of Language Documentation

- video
- audio
- images
- written (e.g. transcription, description/analysis)
- metadata (structured data about materials, typically in written form)

Metadata in documentation

Metadata is "data about data" – structured cataloging information describing characteristics of events, participants, recordings, and details of other data files. Metadata provides the keys for understanding data.

Data formats in Documentation

Choosing the best formats for data can be complex, and advice about formats tends to change as technologies evolve.
It is important to have a basic awareness of the following:

- *character encoding*: how characters are represented, e.g. Unicode, ASCII
- *data encoding*: how meaningful structures in the data are marked (using, for example, XML, Toolbox, MSWord tables, spreadsheet columns, labels etc)
- *file encoding*: how all the data is packaged into a file (e.g. plain-text, MSWord, PDF)
- *physical storage medium*: the physical form used to store the file (e.g. hard disk, compact flash cards, CD, etc)

Archiving

- Archiving is for the benefit of depositors, the language community, and other researchers or interested people in the future. It provides long-term security for the materials.
- There are many archive facilities available for the recorded materials of endangered languages around the world. Endangered Language Archive (ELAR) at SOAS, London is one of them.

Some popular Tools of Documentation

- Toolbox
- Transcriber
- ELAN
- PRAAT
- Speech Analyzer

Organizations involved in Language Documentation

- Google's Endangered Languages Project (http://www.endangeredlanguages.com)
- The Linguist List (http://linguistlist.org)
- Catalogue of Endangered Languages (ELCat), led by University of Hawaii at Manoa.
- SIL International (www.sil.org)
- The Hans Rausing Endangered Languages Project (www.hrelp.org)
- Foundation for Endangered Languages
- Endangered Languages Fund
- Dokumentation Bedrohter Sprachen (DoBeS) 'Documentation of Endangered Languages'
- Research Centre for Linguistic Typology
- Linguistic Data Consortium
- World Oral Literature Project, Voices of Vanishing Worlds
- National Science Foundation (USA) under DEL program

Few words about SIL International

SIL International (formerly known as Summer Institute of Linguistics) most organized and well-funded faith-based organization committed to serving language communities worldwide.

SIL does this primarily through research, translation, training and materials development.

Founded in 1934, SIL has a staff of over 5,500 coming from over 60 countries. SIL's linguistic investigation exceeds 2,590 languages in nearly 100 countries.

Efforts on National Level

- Department of Pakistani Languages, AIOU, Islamabad
- Forum for Language Initiatives (FLI), Islamabad
- Institute of Applied Linguistics (IAL), Karachi

Efforts on Regional level

- Khyber Pakhtunkhwa Regional Languages Authority
- Mother Tongue and Heritage for Education and Research (MOTHER)

Efforts on Community/Local level

- Anjuman Taraqi-e-Khowar
- Burushaski Research Academy
- Gawri Cultural Society
- Gandhara Hindko Board
- Shina Language and Cultural Society

What NLP community can do for saving endangered languages?

"Seen from a computational point of view, all languages pose the same challenges..There is no reason to treat endangered and non-endangered differently"

(Trond Trosterud, 2008)

- *Obstacles before digital documentation and preservation… the issues of portability.*
- *A lot of linguistic documentation is stored in Microsoft formats, secret binary formats that cannot be opened 5-10 years later.*

Without the help from Language Technology we'll turn the endangered languages into endangered data!!!

A possible answer is *Unicode* encodings and *XML* markup. These international, open standards will ensure that the zeros and ones in the digital storage can always be decoded.

Open Language Archives Community (OLAC) is an international network of 25 digital archives in six countries, currently holding over 30,000 items: dictionaries, grammars, field-notes, text collections and recordings. The network is growing rapidly, as more linguists collect and upload their materials.

Can we have a repository of Pakistani languages?

Computational Resources needed for Pakistani Languages

- Unicode supported fonts for all Pakistani languages
- Keyboards
- Orthography standards
- Spell-checkers
- Parallel Corpus for as many Pakistani languages as possible along with Urdu
- Online/digital Dictionaries
- Machine Translation

References

http://ww2.cs.mu.oz.au/~sb/TELIA.html
http://www.hrelp.org/documentation/whatisit/
http://www.sil.org/sociolx/ndg-lg-cahill.html
http://www.ethnologue.com/show_country.asp?name=pk
http://en.wikipedia.org/wiki/Endangered_language
http://cmucelt.org/index.html
http://www.endangeredlanguages.com/about/
"Documentary and descriptive linguistics", Nikolaus P. Himmelmann (1998).

Thank You