Problems and Challenges in the Digitization of Regional Languages

Charanjiv Singh Saroa*1 and Kawaljeet Singh*

Punjabi University, Patiala, Punjab, India – 147002

ABSTRACT
In this paper, we discuss the various challenges that occur in the digitization of regional languages. We also discuss the importance of regional languages and the languages spoken in different parts of the world. A regional language is spoken in a sovereign state's region, whether a tiny area, a federated state or province, or a larger area. Regional languages can be either official state languages or minority languages. Due to diverse linguistic, technical, and socioeconomic reasons, digitizing regional languages is not easy. In this detailed article, we will look into the reasons why the digitization of regional languages is challenging, as well as the intricacies involved in the process.

Index Terms- Linguistic Complexity, Regional Language, UNICODE, Punjabi, Gurmukhi.

I INTRODUCTION
Regional languages, often known as minority or indigenous languages, contribute significantly to cultural variety and legacy. Digitization of text refers to the process of converting printed or handwritten text into a digital format that can be stored, edited, and manipulated using computers and other digital devices. This process involves using various technologies and techniques to capture the content of physical documents and transform them into electronic files. However, in the digital era, companies encounter various challenges that impede their proper digitalization and incorporation into the online world. In this essay, we will look at some of the variables that make regional language digitalization difficult. In this detailed article, we will see why the digitization of regional languages is challenging, and the kind of work involved in the process. Fonts and keyboard layouts are a major barrier to the digitization of regional languages. Every language has hundreds of typefaces. Each typeface has its unique keyboard layout in the majority of regional languages. As a result of the typeface change, the content of the matter changes, and most of the information becomes obsolete.

Digitization of text has numerous benefits, including:

Preservation: Digitizing text helps preserve historical and valuable documents, reducing the risk of physical deterioration and loss due to age, handling, or environmental factors.

Accessibility: Digitized text can be easily accessed and distributed online, making it more widely available to researchers, scholars, and the general public.

Searchability: Digitized text is searchable, enabling users to quickly locate specific information within a document or a collection of documents.
Data Analysis: Digital text can be analyzed using computational tools, allowing researchers to extract insights, trends, and patterns from large volumes of text.

Sharing and Collaboration: Digitization facilitates easy sharing and collaboration, as digital files can be easily shared, edited, and collaborated on by individuals or teams across different locations.

II REGIONAL LANGUAGES

A regional language is a language spoken in a region of a sovereign state, whether it be a small area, a federated state or province or some wider area. Regional languages may be the official languages of the state, or they may be minority languages.

Regional languages are important for several reasons. They can help to preserve cultural heritage, they can provide a sense of identity for people who speak them, and they can be a valuable tool for communication.

Here are some of the benefits of regional languages:

They can help to preserve cultural heritage. Regional languages often contain words and phrases that are unique to the region, and they can help to keep alive the traditions and customs of the people who speak them.

They can provide a sense of identity for people who speak them. Regional languages can help people to feel connected to their community and their ancestors.

They can be a valuable tool for communication. Regional languages can be used to communicate with people who do not speak the official language of the state.

Here are a few more examples of regional languages from different parts of the world:

Spain: In addition to Spanish (Castilian), Spain has several regional languages like Catalan, Basque, Galician, and Valencian, spoken in specific autonomous communities.

Italy: Italian is the official language of Italy, but there are regional languages such as Sicilian, Sardinian, Venetian, and Neapolitan, spoken in different parts of the country.

Switzerland: Switzerland is a multilingual country with German, French, Italian, and Romansh being its official languages, corresponding to different regions of the country.

Nigeria: Nigeria has over 500 languages, including regional languages like Yoruba, Hausa, Igbo, Ibibio, and Edo, spoken in various states.

Canada: Canada has two official languages, English and French, but it also recognizes several regional Indigenous languages, including Cree, Ojibwe, Inuktitut, and others.

Papua New Guinea: With incredible linguistic diversity, Papua New Guinea has hundreds of regional languages spoken throughout the country, such as Tok Pisin, Hiri Motu, and Enga.

Indonesia: Bahasa Indonesia is the official language of Indonesia, but the country is home to numerous regional languages like Javanese, Sundanese, Balinese, and many more.

These are just a few examples, and there are many more regional languages worldwide, each reflecting the unique cultural and historical aspects of the regions where they are spoken.
India: India is a linguistically diverse country with numerous regional languages, such as Marathi, Gujarati, Kannada, Telugu, Odia, Assamese, Malayalam, and many more, each predominantly spoken in different states or regions.

Over 120 languages are spoken in India, a nation with a rich linguistic diversity. Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Odia, Punjabi, Sanskrit, Santali, Sindhi, Telugu, and Urdu are among India's 22 officially recognised languages.

These tongues have official status in their respective states and are acknowledged by the Indian government. There are numerous other regional languages spoken in India in addition to these designated ones. Among the regional tongues with the greatest usage:

- Bhili, Gondi, Khandeshi, Kurukh, Tulu, Warli

Although the Indian government does not formally recognise these languages, millions of people in India nevertheless speak them. In India's cultural and linguistic landscape, regional languages are significant. They support the preservation of the distinctive cultures and customs of the nation's many geographical areas. Additionally, they also provide a sense of identity for the people who speak them.

<table>
<thead>
<tr>
<th>Language</th>
<th>Official Language in States/UTs</th>
<th>Speakers (in millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assamese</td>
<td>Assam</td>
<td>15.32</td>
</tr>
<tr>
<td>Bengali</td>
<td>West Bengal, Tripura, Assam (partly), Meghalaya (partly)</td>
<td>91.37</td>
</tr>
<tr>
<td>Bodo</td>
<td>Assam (partly)</td>
<td>1.34</td>
</tr>
<tr>
<td>Dogri</td>
<td>Jammu and Kashmir, Himachal Pradesh (partly)</td>
<td>2.31</td>
</tr>
<tr>
<td>Gujarati</td>
<td>Gujarat, Dadra and Nagar Haveli, Daman and Diu</td>
<td>53.39</td>
</tr>
<tr>
<td>Hindi</td>
<td>Hindi-speaking states and union territories</td>
<td>439.86</td>
</tr>
<tr>
<td>Kannada</td>
<td>Karnataka</td>
<td>43.99</td>
</tr>
<tr>
<td>Kashmiri</td>
<td>Jammu and Kashmir</td>
<td>4.09</td>
</tr>
<tr>
<td>Konkani</td>
<td>Goa, Karnataka (partly), Maharashtra (partly)</td>
<td>2.08</td>
</tr>
<tr>
<td>Maithili</td>
<td>Bihar</td>
<td>27.76</td>
</tr>
<tr>
<td>Malayalam</td>
<td>Kerala</td>
<td>35.8</td>
</tr>
<tr>
<td>Manipuri</td>
<td>Manipur</td>
<td>2.41</td>
</tr>
<tr>
<td>Marathi</td>
<td>Maharashtra</td>
<td>83.37</td>
</tr>
<tr>
<td>Nepali</td>
<td>Sikkim, Darjeeling district of West Bengal</td>
<td>5.54</td>
</tr>
</tbody>
</table>
III. THE DIGITIZATION OF REGIONAL LANGUAGES

The digitization of regional languages poses several challenges that hinder the smooth and comprehensive adoption of these languages in the digital era. These challenges arise from linguistic complexities, technical limitations, and socio-economic factors. Let's explore some of the major challenges in detail:

**Linguistic Complexity:**

**Script and Character Encoding:** Many regional languages have unique scripts that may not be well-supported in digital systems. Creating appropriate character encoding standards and fonts for these scripts can be challenging.
Morphological Complexity: Regional languages often have complex word formations and grammatical rules, making natural language processing (NLP) tasks more intricate and demanding.

Dialectal Variation: Regional languages may have multiple dialects and variations, which can lead to difficulties in establishing a standardized form for digitization.

Technical Challenges:

Limited Digital Resources: Regional languages may have a scarcity of digital content, fonts, and input methods, impeding the development of applications and services in these languages.

Compatibility Issues: Integrating regional language support into existing digital platforms and systems can be complex and may result in compatibility issues.

Language Technology Support: Language technology tools and resources are often initially developed for major languages, leaving regional languages with limited support.

Socio-Economic Factors:

Low Awareness and Interest: Speakers of regional languages may have limited awareness of digital technologies or may not perceive significant benefits in using their languages digitally.

Education and Literacy: Low literacy rates in some regions hinder the adoption and utilization of digital platforms, limiting the demand for digitization of regional languages.

Market Demand: Commercially, there may be less demand for products and services in regional languages compared to major languages, resulting in reduced investment in language technology development.

Language Preservation and Identity:

Language Shift: The increased use of major languages in various domains, such as education, media, and administration, can lead to a language shift, where regional languages are gradually replaced.

Threat to Endangered Languages: Digitization efforts may not be enough to save endangered languages from extinction, especially when there are small speaker populations and limited digital content available.

Quality of Language Technology:

Machine Translation Accuracy: Machine translation systems may not be as accurate for regional languages due to limited data and linguistic complexity, resulting in subpar translation quality.

Speech Recognition: Developing robust speech recognition systems for regional languages can be challenging, as these languages may have diverse accents and pronunciation variations.

Government and Policy Support:

Lack of Official Recognition: In some countries, regional languages may not have official status, leading to a lack of government support for digitization efforts.
Language in Education: Inadequate integration of regional languages into education policies can hamper the development of language technology and digital content.

To overcome these challenges, collaborative efforts from linguistic experts, governments, technology developers, and language enthusiasts are essential. Standardization of regional languages, development of language technology tools and resources, and promotion of digital literacy in these languages are crucial steps in successfully digitizing regional languages and preserving linguistic diversity in the digital age.

Current Initiatives and Case Studies

Highlight ongoing initiatives and case studies that aim to address the challenges of digitalizing regional languages. Discuss successful projects that have made strides in creating digital resources, promoting digital literacy, and overcoming linguistic and technical barriers. For example:

The "Digital India" campaign's efforts to promote digital literacy and provide digital content in regional languages (Government of India, n.d.).

The development of language-specific keyboards and input methods for regional languages, such as the Indic keyboard by Google.

Successful localization of digital services and platforms, such as e-commerce websites, in regional languages to cater to a wider user base.

IV. PUNJABI LANGUAGE

Punjabi is an Indo-Aryan language primarily spoken in the Punjab region of South Asia, which is divided between India and Pakistan. It is the native language of the Punjabi people and serves as an important aspect of their culture and identity. Punjabi is also widely spoken by the Punjabi diaspora around the world, particularly in countries with significant Indian and Pakistani communities.

Linguistic Features:

Punjabi has its own distinctive script, known as Gurmukhi, which is used to write the language in India. In Pakistan, the Shahmukhi script, a modified version of the Perso-Arabic script, is used to write Punjabi.

The language exhibits a rich phonetic and phonological system with several unique sounds and tonal variations. Punjabi is also known for its varied vocabulary, encompassing words with influences from Sanskrit, Persian, Arabic, and other regional languages.

Historical Significance:

Punjabi has a rich literary history dating back centuries. It has been used as a medium for religious scriptures, including the Guru Granth Sahib, the central religious text of Sikhism. The language has also been used in various poetic and prose works by renowned Punjabi poets and writers.

Geographical Distribution:
In India, Punjabi is an official language of the state of Punjab and is also spoken in neighboring states like Haryana, Himachal Pradesh, and Delhi. It is the most widely spoken language in the Punjab region of India.

In Pakistan, Punjabi is the most spoken language, especially in the province of Punjab, where it is the official language. However, in the country's early years, Urdu was declared the national language, which led to a decline in the use of Punjabi in official domains.

**Cultural Significance:**

Punjabi culture is deeply intertwined with the language, and Punjabi music, dance, and literature have gained popularity worldwide. Bhangra, a lively dance form, is one of the most well-known expressions of Punjabi culture.

**Challenges and Revival:**

Despite its widespread usage, Punjabi faces challenges in modern times. In both India and Pakistan, there has been a dominant influence of English and other major languages, leading to a gradual decline in the use of Punjabi in certain domains like education, media, and administration.

However, there are ongoing efforts to preserve and promote the language. Punjabi literature continues to thrive, and cultural events and institutions are instrumental in reviving the language's significance. Moreover, in recent years, there has been a resurgence of interest in using Punjabi in digital platforms, including social media and online content.

In conclusion, Punjabi is a vibrant language with a rich cultural heritage. Despite facing challenges, it continues to play a crucial role in the lives of millions of people, both in the Punjab region and in the Punjabi diaspora worldwide. Efforts to preserve and promote Punjabi are essential to ensure its continued vitality in the global linguistic landscape.

**Challenges to digitize Punjabi Text.**

**Spell check:**

All the available spellcheckers mainly work with the English language. Rarely we can find any spell checker for other languages.

**Some special Symbols:**

Some text can contain some unique types of symbols that are not available in ASCII codes and even not in the UNICODE system.

**Typing problem:**

By default, each computer contains an English (Roman) keyboard. Most of the users are not aware of UNICODE-based systems and fonts. Without the knowledge of UNICODE-based fonts users cannot type in Unicode

**Non Unicode Fonts:**

Mostly the data available in regional languages are in ASCII-based Indic scripts. we want to display that data on the website we have to upload the required font for that data and the user
first has to download and install that font, Only then the user can view that information. There are thousands of fonts are used to create data in Indian Scripts. Only Punjabi (Gurmukhi) alone has more than 225 popular fonts which are still in use for publishing books, magazines, newspapers etc. Publishers are still working with ASCII coding-based fonts. They have not used Unicode-based fonts for the following reasons:

- People resist change, due to Unicode typing issues.
- Lack of awareness of Unicode standard.
- Little support for the Unicode system in the publishing software that they are using.
- Less availability of Unicode fonts to represent text in different styles and designs.

It is always better to display information in Unicode-based fonts while displaying the information on a website. The information presently available to us is mainly in ASCII-based fonts. So we convert that ASCII-based information into Unicode-based fonts so that it can be available on the internet. The information displayed in Unicode can be seen on any computer without installing a font. Another advantage of Unicode-based fonts is that it is searchable on search engines like Google, Yahoo, Ask, Bing etc.

**Fonts & Keyboard Layouts:**

There are 100s of fonts for every language. In most of regional languages, each font has its keyboard layout. That result in changing the content of the matter with the change of font and most of the information become useless. A problem with ASCII-based fonts for Regional scripts is that there is no standardization of mapping of script characters with keyboard keys.

Popular are Punjabi fonts

<table>
<thead>
<tr>
<th>Font Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raavi</td>
<td>Official font for Gurmukhi script in Windows and macOS</td>
</tr>
<tr>
<td>AnmolLipi</td>
<td>Widely used font for Punjabi text rendering</td>
</tr>
<tr>
<td>AmrLipi</td>
<td>Popular font for Punjabi text display</td>
</tr>
<tr>
<td>Joy</td>
<td>Elegant and legible font for Punjabi script</td>
</tr>
<tr>
<td>GurbaniAkhar</td>
<td>Specifically designed for rendering Gurbani (Sikh scriptures)</td>
</tr>
<tr>
<td>Saab</td>
<td>Easily readable font with clear letterforms</td>
</tr>
<tr>
<td>Asees</td>
<td>Simple and widely used font for Punjabi script</td>
</tr>
<tr>
<td>Lohit Punjabi</td>
<td>Popular open-source font for Punjabi text</td>
</tr>
<tr>
<td>WebAkharSlim</td>
<td>Stylish and thin font variant for web applications</td>
</tr>
<tr>
<td>Sukhmani</td>
<td>Attractive and easy-to-read font for Punjabi text</td>
</tr>
</tbody>
</table>

V CONCLUSION AND FUTURE RESEARCH DIRECTION
The digitization of regional languages is a complex endeavour that involves addressing linguistic intricacies, technical hurdles, and social factors. The uniqueness of each regional language calls for tailored solutions to ensure their effective digitization and integration into the digital landscape. Collaborative efforts between governments, language enthusiasts, and technology developers are crucial to overcoming these challenges and preserving linguistic diversity in the digital age. Only through concerted actions and sustained support can regional languages flourish and thrive in the rapidly evolving digital world.

REFERENCES


[20] Online Sources
  • en.wikipedia.org/wiki/Regional_language
  • en.wikipedia.org/wiki/Languages_of_India
  • adityaupsc.wordpress.com/2013/06/27/salient-features-of-indian-society-diversity-of-india/
  • www.pointsincase.com/articles/im-the-nation-of-india-and-i-found-myself-in-america
  • https://www.myfonts.com/collections/raavi-font-microsoft-corporation
  • TTS products , https://www.understood.org
  • Text-To-Speech(TTS), http://searchmobilecomputing.techtarget.com
  • Languages of India, http://www.gutenberg.us/articles/Languages_of_India
  • Education, http://www.ncpedp.org/Education
  • Use of Computer Technology to Help Students with Special Needs, https://www.princeton.edu/futureofchildren/publications/docs/10_02_04.pdf