Digital dictionaries offer access to vanishing languages

A screenshot of some of the literary works available on the Bharatavani website.

The word for sunlight or sunshine in Angami — a language spoken by around 130,000 people in the North East — is *niakikezie*. In the Ao-language of Nagaland, it is *anüpu oranüsangwa*. And this reporter in far away Bengaluru could look up these words and many more from several Indian languages, thanks to digital dictionaries available on the Bharatavani website.

Most cities in India have infrastructure to teach many foreign languages . But how many look inwards to tap the domestic cultural motherlode of more than 1,500 Indian languages? It is this question that spurred Bharatavani, an online Indian Languages platform hosted by the Central Institute for Indian Languages (CIIL), Mysuru, to not only publish content in 121 Indian languages, but work towards starting online classes.

What is particularly causing ripples of excitement among linguists and researchers is the compilation of digitised searchable dictionaries. In a little over a year since its inception, the portal offers 262 unilingual and multilingual dictionaries in 50 Indian languages — all of them in a searchable format on android platforms — which can be accessed on Bharatavani's free Android app.

The number of languages covered will soon cross a hundred, said Beluru Sudarshana, consultant with CIIL. "Bharatavani is not publishing new works, but we are for the first time digitising available dictionaries in smaller languages, to bring it to a wider audience," he said. Malto-English-Hindi, Odia-Ho, English-Ao and Lepcha-English are some of the dictionaries on offer — most of them available in a searchable format and not as cumbersome PDF files.

Prof. Panchanan Mohanty, Dean, School of Humanities, University of Hyderabad and an expert in Eastern Language research, who is also on the Bharatavani committee, likened Bharatavani to Project Tiger, arguing for conservation of India's fast-depleting language heritage. But more significantly, the digitised database of dictionaries is a goldmine for linguistic research in the country, he said.

These dictionaries can now be linked to create a large database of words across various languages, using English, Hindi or regional languages as the source words. With over seven lakh source words at present, the potential of the database is immense. For instance, the use of Odia source words will result in an Odia-English-Ho-Munda-Khadia-Kui-Oraon-Saura dictionary, integrating a family of Austroasiatic languages spoken in central-eastern India. The integration of these dictionaries is still a work in progress.

Linguist G.N. Devy, who spearheaded the People's Linguistic Survey of India, believes this resource will help speed up socio-linguistic research and not just along themes of structure and genealogy, thereby ensuring better development planning.

"One serious challenge is that children from communities speaking non-scheduled languages are pushed out of schools leading to development deprivation. For an imaginative user, content on Bharatavani may help in designing a curriculum in these languages," he said, adding that starting from scheduled languages, Bharatavani has now broadened its scope to smaller languages that have over 10,000 speakers. "But there are several languages with fewer than 10,000 speakers, which Bharatavani needs to work on in its second phase."

This undertaking is not without it challenges. For one, Optical Character Recognition (OCR) is still

in a primitive stage even for major Indian languages. Thus constructing digitised databases for smaller languages will be a problem as their script cannot be scanned and converted into text format. Tedious desktop publishing is the only viable option.

Another hurdle is that unicode script input drivers are available in only recognised scripts. Incidentally, the Bharatavani portal will soon provide a virtual keyboard, integrating all available Unicode drivers of India languages for users to search for words by typing in language of their choice.

The bigger problem, however, is proofreading, said Mr. Sudarshana. "Ideally, for a multi-lingual digital dictionary we need to carry out a collaborative online proof-reading process, each expert looking at their language of expertise. In most of these smaller languages, it's tough to even get language experts. Most are old and not equipped to proofread online. We have opted for assisted online proof reading, where a person reads out the text to the expert and makes suitable changes in the database on the expert's recommendation, which is time consuming," he said,

Bharhavani is steering in uncharted terrain, but researchers and linguists on board this project are optimistic that it will unveil India's landscape of languages to its citizens despite the many challenges.

Struggling to meet the demand after production delays, Apple supplier Foxconn has reportedly employed thousands of students to assemble iPhone X who