Lyceum India Journal of Social Sciences || ISSN (Online): 3048-6513
Volume: 3 Issue: 6 || April 2026 (Special Issue 2) || DOI: 10.5281/zenodo.19956266

Preserving Low-Resource Indian Languages Through Natural Language Processing: Challenges and Imperatives

Dr. Annie Rajan[1]

Abstract:

The importance of conserving languages, any language in general, using Natural language processing (NLP). This can be done by building corpus that can be accessed by computer systems, so it becomes feasible to build linguistic tools that can help in the growth of the language. The objective of the work is to demonstrate, that in the present era the creation and enrichment of existing corpora and tools for any low resource language is a fundamental requirement for conserving the language, as well as a key facilitator.

NLP resources for many Indian languages are not keeping pace with the growing digitization of the economy and of public life. The potential for NLP to cater to digital markets as well as social media flourishing in the Indian hinterland is growing rapidly, but the availability of corpora and NLP tools is lagging behind. The list of deprived languages includes not only the 18 officially recognized regional languages in India, but many more spoken languages of small communities or tribal populations.

Automatic NLP of Indian languages of small communities like has been limited by the scarcity of the language-specific digital linguistic resources. Raw textual data are available in ample quantities, but they are not in a form suited for NLP. Among the different kinds of data are newspaper articles, literary works and poems. Such data sets need to be collected into corpora, and further, annotated in different ways according to the intended application.

NLP techniques are a challenge for low resource languages because NLP techniques require linguistic knowledge that can be only developed by experts and some by speakers of that language, and they require a lot of labelled data which is again expensive to generate.

Keywords: Indian Languages, Natural Language, Culture, Society

[1] Associate Professor & HOD, DCTs Dhempe College (Autonomous), Miramar, Goa

Introduction

Boundaries of communication has been shrinking due to the research exploration in Natural Language Processing (NLP). Digital environment is the requirement of this developing world. Digital resources and tools in NLP are a vital need; at the same time, it is a challenging task taken by researchers in the field of computational linguistic. For a global communication there is a need for sufficient enough of resources that can help in the development of tools required for a digital presence of the language. Communication, information retrieval, speech recognition, is a result of interaction between the disciplines of linguistics and artificial intelligence. Large number of languages across the globe are having limited or no resources for building NLP tools, such languages need contribution from linguist and computational tools need to be developed so that these languages have a digital presence. India is a country having 22 languages in the Eighth Schedule of Indian Constitution [1], these languages are challenging because resources for some of the languages are less or next to nil.

Language reflects embedded cultural perceptions, prejudices, values and power structures. The continuing sustenance of one's distinct cultural identity depends on a number of factors, among which language is prominent. Resorting to standardization has always posed a distinct threat to marginalized languages, realized in our own time by the use of English as the common language of expression in the digital world. While many countries have dual language websites, pan-Indian content is often restricted to English for the sake of convenience. Websites in Hindi and major regional languages like Marathi, Tamil, and Bengali are available, but their content is often inadequate. The state of affairs is even more dire for languages of small communities, such as Konkani and Dogri. This situation cannot be overcome unless there is collaboration between various stakeholders, including language scholars, linguists and computer scientists. This paper is an attempt to address the question of enhancing the presence of marginalized languages in the digital domain, and the pressing need for collaboration among specialists. It looks at the role of contemporary research in natural language processing (NLP) for marginalized languages and the real threat of elimination that these languages face if they cannot make a place for themselves in the digital world. It is not just a medium of communication that would be lost with these languages, but the rich cultural heritage they carry within them.

While the number and scope of digital transactions is rapidly increasing all over the world, low-resource languages like Konkani (State language of Goa, India) are deprived of automatic NLP. Moreover, the number of native Konkani speakers is getting reduced [1]. Creating a comprehensive set of well-performing NLP tools for Konkani written in Devanagari, its official script [2], is a necessity for enhancing the presence

of the language in this digital era. The language is getting diminished otherwise. Such a development has its own toll, like the loss of a culture, its folk songs and its literature [3]. The challenges in the NLP of Konkani are unique due to the multiplicity of scripts and dialects, and the resulting lack of standardization of both language and script. More generally, the Konkani language, its evolution and history been affected by a variety of deep historical, regional, and cultural factors [3]. Compared to the other languages with a similar number of native speakers, automation efforts in Konkani are delayed and slower to develop [5-7].

Konkani and Goa have a special political and cultural significance, for several reasons. Goa is a very popular tourist destination, and not least, Goa has had a long history of being a colonial possession of a European power. It was the first in South Asia to go under western colonial control in 1,510 and the last to emerge from it in 1961. As a result, Goa is regarded as a region, and Konkani a language, that experienced a uniquely intense interaction with Portuguese culture [8]. Devanagari was declared as the official script for Konkani, though a multiplicity of other scripts is used to this day, for reasons dictated by geography as well as history.

According to the 1971 census of India, about 1.5 million people declared Konkani to be their mother tongue, while the 1981 census shows the total population of Goa itself to be only 1.08 million. In the absence of complete and up to date breakdown of figures, the present number as per 2011 census, of those who speak Konkani as the mother tongue is estimated to be about 2.5 million, while a total of about 3.5 million can speak Konkani either as a first or second language. The 2011 census indicates that the number of Konkani speakers has actually reduced, even as the population of the region as a whole has increased. This apparent paradox is be-cause newer generations of Konkani speakers have begun to adopt the dominant language of their region as their mother tongue. This is a sufficient reason why NLP researchers need to pay attention to Konkani, and to similar small-population, low-resource languages.

Konkani is an Indo-Aryan language belonging to Indo-European family that originated in India. It is the southernmost of the Indo-Aryan languages. In relation to modern Indian Languages, it is closer to Gujarati and Marathi. Konkani is one of the twenty-two languages included in the Eighth schedule of Indian constitution. Albeit the official language of Goa, it is also spoken in various parts of Karnataka, Maharashtra, Kerala, Kenya, Uganda, Pakistan, Persian Gulf and Portugal. Devanagari is the most widely accepted script for Konkani, which is declared as the official regional language of Goa. Konkani is the medium of instruction in some primary schools and is taught as a main language and optional language at the high school and college level. Organizations like Konkani Bhasha Mandal, Konkani Sahitya Parishad,

Goa Konkani Akademi, Thomas Stephen Konkani Kendr and other similar institutions are working to promote the language and encourage Konkani language.

Computational linguistic tools for Konkani

When we started the research work on Konkani language, there were no computational linguistic tools other than Konkanverter. This online tool (Figure 1), known as the 'Konkanverter,' is a script conversion utility designed and created by the World Institute of Konkani Language at the World Konkani Centre in Mangalore. It facilitates the conversion of Konkani text from one script to another. It supports conversions such as: a) Devanagari to Kannada, Roman, and Malayalam, b) Kannada to Devanagari, Roman, and Malayalam, c) Roman to Kannada, Devanagari, and Malayalam, d) Malayalam to Kannada, Devanagari, and Roman.

We also have a YouTube link on terminologies of Konkani and a website *www.audiopustakam.com* (Figure 2) for elementary learning of this language. There is also standalone device developed of Audiopustakam so that schools can use it without the need of an internet connection. These tools were developed to help and promote learning Konkani. AudioPustakam's has created audio data resource of Std 1, 2, 3, 4 for Konkani text book of SCERT followed in the state of Goa. The documentation of audio data will help to educational institutions for intervention learning, slow learners of language, language learners, speech corpus of speech data for researchers. Standalone low-cost audio player with pre-recorded audio has been developed. Tools of morphological analysis (Figure 3), spell checking [9] (Figure 4), stemming (Figure 5), Named Entity Recognition [10] and Part of Speech (PoS) [11-12] tagging (Figure 6) for Konkani has also been developed using Python Django framework. Among these tools, only the PoS Tagger makes use of neural network based trained model, whereas the other tools make use of dictionary lookup, rule-based analysis and traditional machine learning approaches.

Fig. 1

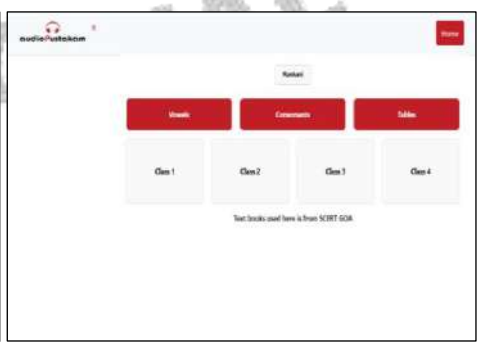

Fig. 2

Fig. 3

Fig. 4

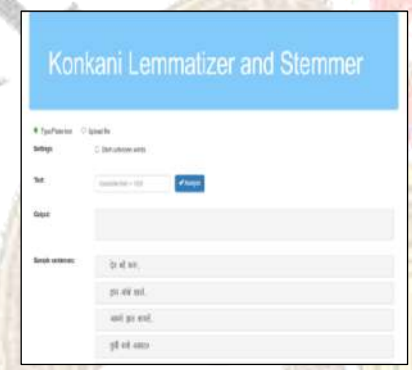

Fig. 5

Fig. 6

Conclusion

While the volume and scope of digital transactions in regional languages is rapidly increasing all over the world, low-resource languages like Konkani are deprived of automatic NLP. Moreover, the number of native Konkani speakers is getting reduced. Creating a comprehensive set of well-performing NLP tools for Konkani written in Devanagari, its official script, is a necessity for enhancing the presence and continuance of the language in this digital era. The language would get diminished otherwise. Such a downward trend has its own toll, like the loss of a culture, its folk songs and its literature.

The challenges in the NLP of Konkani are unique due to the multiplicity of scripts and dialects, and the resulting lack of standardization of both language and script. More generally, the Konkani language, its evolution and history have been affected by a variety of deep historical, regional, and cultural factors. Compared to the other languages with a similar number of native speakers, automation

efforts in Konkani are delayed and slower to develop. The facilities created through this research would attract more researchers towards Konkani NLP.

Automatic natural language processing (NLP) of Indian languages of small communities has been limited by the scarcity of the language-specific digital linguistic resources. Creating and enriching the existing corpora for low resource languages is important and providing annotations are fundamental requirements.

References

The Goa, Daman and Diu Official Langauge Act, 1987 (PDF), U.T. Administration of Daman and Diu, 19th December 1987. Retrived 26th December 2014.

Mathew Almedida. S.J., " A Description of Konkani." (1989), Thomas Stephens konkani Kendr, Miramar, Panaji. Goa.

Sanjana Manerkar, et. al., "Experiences in Building the Konkani WordNet Using the Expansion Approach" (2010), IIT Bombay, Mumbai. Maharashtra.

Shilpa N Desai, et. al., "Insights on the Konkani WordNet Development Process" (2017), The WordNet in Indian Languages, pp. 101–117.

Sanjana Manerkar, et. al., "Konkani WordNet: Corpus-Based Enhancement using Crowdsourcing" (2022), Transactions on Asian and Low-Resource Language Information Processing, vol. 21(4).

Rocky V. Miranda., "The Status of Konkani During the Portuguese Era" (1982), South Asian Review, vol. 6(3).

Annie Rajan, Nehal Kalita, Ambuja Salgaonkar., "Spell Checker for Low-resource Konkani Language" (2025), Journal of Information and Communications Technology: Algorithms, Systems and Applications, vol. 1(3).

Annie Rajan, Ambuja Salgaonkar., "Named Entity Recognizer for Konkani Text" (2021), ICT with Intelligent Applications: Proceedings of ICTIS 2021, pp. 687–702.

Annie Rajan, Ambuja Salgaonkar., "Part of speech (PoS) tagging for Konkani language using HMM" (2022), ICT Systems and Sustainability: Proceedings of ICT4SD 2021, pp. 601–609.

Annie Rajan, Ambuja Salgaonkar, Arshad Shaikh., "Deep Learning for Part of Speech (PoS) Tagging: Konkani" (2022), Smart Trends in Computing and Communications: Proceedings of SmartCom 2022, pp. 337–346.

https://en.wikipedia.org/wiki/Konkani_language – Accessed 11/04/2026