Prof. Pushpak Bhattacharya: The Guru and the Visionary Paving the Way for Indic NLP

Prof. Pushpak Memorial Lecture

IndoML 2025 Conference, BITS Pilani – Hyderabad Campus

20th December 2025

By Anoop Kunchukuttan

Introduction

Good morning, everyone. I would like to thank IndoML for dedicating a memorial lecture in the honour of Prof. Pushpak Bhattacharyya. I would also like to thank the organizers for trusting me with the responsibility to deliver this memorial talk. Pushpak Sir was not just my Ph.D advisor but also my mentor and guru who enriched my life with his wisdom, guidance and support. He is no more with us, but his life and body of work inspires each one of us. In this talk, I want to share some reflections on the Pushpak Sir I knew —as a teacher, a mentor, and a builder of institutions—and reflect on how his ideas shaped multilingual NLP and Indian language technology as well as briefly look at where Indian language NLP stands today and what opportunities lie ahead.

All of us in the Indian ML community are well-aware of Pushpak Sir and his contributions to the growth of NLP and AI in India, and to Indian language NLP, in particular. He was a faculty member for 37 years from 1988 at IIT Bombay, his *karmabhoomi*. He served as the Director of IIT Patna, helping nurture the nascent institute from 2015-2020. He was also President of the Association for Computational Lingustics in the 2016 term, the first person based out of India to hold the position. He was a Fellow of National Academy of Engineering and Abdul Kalam National Fellow. He has received H.H. Mathur Research Excellence Award of IIT Bombay (2021), Manthan Award of the Ministry of IT (2009) among many other recognitions.

No doubt his career was outstanding, full of achievements and accolades – but his legacy was much more than these facts and statistics. He inspired me as well as his students, fellow researchers, colleagues and everyone else who came in touch with him with his love for science and language technology, with his kindness and empathy, and with his dedication and unwavering sense of purpose.

The Teacher

Pushpak Sir enjoyed teaching. And he made learning such a joy for his students. To me his central influence is as a teacher and mentor.

He believed deeply in teaching from *first principles and fundamentals*. He devoted much time to developing concepts. One of the very first courses I took at IIT Bombay was his **Neural Networks** course. I just took multiple machine learning courses because it seemed like a fascinating area. But I was anxious about handling ML courses – it was a totally new area. However, he put me at ease with his presentation of the course. What struck me was not just the content, but the structure of the course itself. Fundamentals were built, each idea followed naturally from the previous one, making a complex subject feel intuitive and approachable.

He was never in a hurry to move forward if the foundations were unclear. In his courses, he spent quite a lot of time on the key ideas. Learning should not be about facts, lot of details and mathematical desiderata alone – it should also be about the key insights, the big picture as well. We shouldn't miss the forest for the trees. A small but telling example of this was his choice of readings. For his NLP course, one of the suggested readings was *"The Language Instinct"* by Steven Pinker. This is not a traditional textbook. It is not filled with equations or algorithms. Rather, it is a popular science kind of book that explains the science behind language in an accessible and perceptible way.

When Sir was writing his book on machine translation, we used to have free-wheeling discussions where he used to go over existing work and discuss in detail. These used to long sessions, and Sir would branch off into great depth to various topics drawing upon his vast experience with MT. You could see the sparkle in his eyes when a key insight or topic had been nicely conveyed to his satisfaction. This was a feature of research meetings with him too – where fundamental questions would be posed, which would be food for thought.

Perhaps the most standout aspect of his role as teacher was the warmth and generosity he had for students. He was always welcoming of all students. To him every student was unique, and he gave every enthusiastic student an opportunity to learn and contribute under his guidance. Another striking aspect of his teaching was his respect for students as independent thinkers. He let students explore their strengths, weaknesses and interests and chart their own course. This can be seen in the diverse set of research questions his Ph.D students explored ranging from machine translation, multi-linguality, essay grading, computational music, sentiment and emotion analysis, eye-tracking and material science. He encouraged questions, disagreements, and alternative viewpoints. I myself took time to figure out what would be interesting and impactful for me to do, I pivoted mid-way through my Ph.D to pursue a direction I believed in - he was supportive of the same and at the time served as a well-meaning critic and guide who would make sure I make progress.

Pushpak Sir mentored and nurtured an excellent pool of NLP and AI talent and scholarship in India. He supervised around **60 Ph.D. students and more than 250 Master's students,** and inspired many more. Today, they are contributing in diverse ways across academia, industry, startups, and public institutions – inspired by his ideas of science, research, empathy, humility and perseverance.

An Institution Builder

When we speak of Professor Pushpak Bhattacharyya's contributions, it is impossible to separate the scholar from the **institution builder.** In many ways, **CFILT—the Centre for Indian Language Technology—was his life's work – the institution that he built**.

Pushpak Sir often spoke about the **need for a critical mass** to sustain meaningful research. He believed that impactful research centers cannot survive on isolated excellence; they require scale, diversity, and continuity. CFILT was his answer to that belief. It is a space that brings together **people with different skills**: linguists, engineers, data-driven researchers, systems builders, and students just discovering NLP. CFILT became a place of **rich interaction**, where diverse problems were discussed, debated, and pursued.

CFILT is a rich repository of knowledge, sustained by the strong processes that Sir insisted on. The lab maintains an excellent NLP and linguistics library, and a carefully curated archive of research outputs. Students were required to leave behind well-documented code, scripts, and datasets, ensuring long-term usability. I personally benefited from this discipline: while exploring the CFILT archives, I uncovered underutilized parallel corpora that later proved critical to my research on multilingual NLP, enabling the construction of 100 language-pair SMT models and what was then the largest publicly available English–Hindi parallel corpus.

Pushpak Sir also worked actively to build a **national network of NLP institutions across India**, collaborating closely with peers to initiate pioneering efforts such as **EILMT, ILMT, CLIA, and IndoWordNet**. They created shared agendas and a sense of collective purpose for Indian NLP at a time when the field was still taking shape.

Pushpak Sir was also very encouraging of **openness and sharing in research**. IndoWordNet was among the earliest widely available **multilingual lexical resources** for Indian languages, and its open availability played a crucial role in catalyzing research in the area. I also had first-hard experience of his approach. He was supportive of the public release of the **IIT Bombay Parallel Corpus** and it quickly became one of the most widely used early parallel corpora for Indian languages globally.

His institution-building did not stop with CFILT. At **IIT Patna**, he was the driving force behind the growth of the **AI-NLP-ML lab at IIT Patna**, mentoring its rapid development and helping establish it as an important hub for NLP research in India. Once again, his focus was on creating sustainable structures rather than short-term success.

Finally, I would add that our **AI4Bharat** initiative at IIT Madras draws direct inspiration from Pushpak Sir's ideals. Many of us who came together to build AI4Bharat had our roots in CFILT. What united us was a shared commitment to the challenges he had articulated years earlier: **building language technologies at scale for Indian languages**, grounded in linguistic understanding, openness, and social impact. In that sense, CFILT's influence extends far beyond its walls.

Contributions to Multilingual NLP and Indian Language Technology

Pushpak Sir's work spanned a lot of areas in NLP. Amongst these, his work on multilingual NLP and Indian language technology lie at the heart of his scientific legacy where he made lasting contributions across approaches, datasets, and models. It is also the area that I had the fortune of closely working with him. He was deeply fascinated by the linguistic diversity of India and the unique research challenges it posed. His research was consistently guided by the question: **how can we build effective, scalable language technologies that work for *all* Indian languages, not just the few that are well resourced?**

Language divergence and typology as first-class concerns

A central focus of his research was the systematic study of linguistic divergences, particularly between English and Indian languages. His 2002 Journal of Machine Translation article remains a seminal reference in the Indian context. Equally important was his deep engagement with Indian language typology, examining both shared structures and critical differences across languages. This balanced understanding of similarity and diversity strongly informed his work in multilingual NLP and is reflected in the rich, illustrative examples found throughout his textbooks on MT and NLP.

Resource scarcity and multilingual sharing

Pushpak Sir was acutely aware that **lack of resources is a defining reality** for Indian language NLP, and posed the research problem: how can linguistic insight and linguistic similarity be used to overcome resource scarcity?

This led him to investigate some of the fundamental research questions that defined his research.

- How can **linguistic knowledge** offset limited annotated data?

- How can **information and resources be shared across languages**?
- How can **linguistic similarity** allow richer languages to support resource-poor ones?
- How can we build **resources that serve multiple Indian languages efficiently**, while still preserving the uniqueness and identity of each language?

Many of these questions were first explored during the **rule-based era** of NLP – particularly the pioneering group at IIT Kanpur led by Prof. RMK Sinha, Prof. Rajeev Sangal and Prof. Vineet Chaitanya. Prof. Pushpak played a crucial role in **revisiting and extending them in the modern machine learning era**, ensuring that linguistic principles remained relevant even as models and methods evolved.

Across a wide range of problems, Pushpak Sir's work demonstrated how **linguistic knowledge and multilingual sharing** can lead to practical gains in Indian language technology. His work leveraged the morphological richness and script-level similarities of Indian languages, modeled structural divergences between English and Indian languages for machine translation, and developed unsupervised and multilingual approaches to transliteration and word sense disambiguation, amongst others. He also explored multilingual solutions across tasks such as machine translation, sentiment analysis, and named entity recognition, consistently showing how languages can mutually support each other under resource constraints.

My own research has also been shaped by these questions as I explored them under his guidance. As we navigate various shifts in the NLP technology landscape from early deep learning to pre-trained models to generative LLMs to reasoning models, these questions continue to be relevant and require repeated inquiry and revisiting.

IndoWordNet

Among his many contributions to Indian language NLP, **IndoWordNet** stands as one of the most influential and representative of his work. **IndoWordNet** is a **multilingual lexical knowledge resource for Indian languages**, designed to represent **word meanings (senses) and their semantic relationships** in a structured, language-linked way. It supports ~20 Indian languages, 25k-30k synsets and 150-200k word forms across all languages.

IndoWordNet is not just a lexical resource; it **reflects Pushpak Sir's vision of multilingual NLP**. It reflects his belief that:

- **Ambiguity processing lies at the center of NLP**, and that **word sense disambiguation (WSD)** plays a critical role in meaning representation.
- High-performing systems in the machine learning era require **richly annotated, high-quality data**.

- Such data can only be built at scale through **linguistic expertise, well-defined processes, and careful curation**.

He pioneered the **expansion approach** for building WordNets, enabling rapid extension to new languages by inheriting semantic relations from existing ones. This resulted in a **multilingual sense dictionary** that laid the foundation for **multilingual WSD**, including methods based on **parameter projection across languages**.

His contributions continue to shape how Indian language technologies are conceptualized and built today. More importantly, they remind us that progress in multilingual NLP is not merely about scale, but about **understanding language deeply, sharing knowledge across languages, and building resources that serve society as a whole**.

My Perspective on Indian language NLP

Because of the contributions of Prof. Pushpak and other pioneers of Indian language NLP, today we have a rich talent pool and have built strong foundations for Indian language technology. I think this is an opportune moment to reflect on where Indian language technology stands and what opportunities lie in the future. Let me make my humble attempt at that. I would like to focus a few areas that I think are critical to the advancement of Indian languages:

- Collaboration and Openness
- Data Curation and Evaluation
- Multi-linguality and relatedness
- Efficiency

Collaboration and Openness

In my view, a strong culture of openness and deep collaboration is the most essential ingredient to put Indian language technology on a trajectory of accelerated innovation, and large-scale adoption.

Openness and collaboration in AI accelerate progress by improving **reproducibility**, enabling results to be verified, shared, and extended. They enhance **education** by giving learners and practitioners access to real systems, while also driving **faster innovation** through collective problem-solving. Open approaches allow greater **customization and flexibility**, reduce development **costs**, and **lower barriers to entry** for startups, researchers, and smaller organizations. Most importantly, transparency in open and collaborative AI builds **trust**, as systems can be better understood, evaluated, and responsibly improved.

I would like to put the spotlight on a couple of aspects of collaboration and Openness:

Open data and models

Datasets and models are fundamental infrastructure for AI. They need to be open, just like the Linux operating system - which is the basis of all computing. We started slowly as a community, but over the last 5 years, there has been a multi-fold increase in the datasets and models that have been made publicly available for leading research institutions like CFILT, AI4Bharat, LTRC, amongst others. Government initiatives like AIKosh and Bhashini are also playing a role in make these resources open and accessible. We are now in a new era, where open LLM related resources are called for. In this second wave, deep-tech Indian startups, especially those supported and tasked by the IndiaAI mission to build sovereign LLMs, have an important role to play. Some of them like Soket Labs, Sarvam AI, Krutrim are already putting out some resources in the public domain. We need more of these open contributions to build a vibrant ecosystem for research and innovation in the Indian language LLM space.

I also look forward to **truly open-source LLMs which are well-documented, data, recipes and learnings are shared** – drawing inspiration from the precedent that AllenAI's Olmo models and Huggingface's SmolLM models have set. The technical reports of these models are so rich and so educational – they are catalysts for LLM research and development in the open community. We have an opportunity to do something similar in the multilingual LLM space.

Compute Access

Compute is the most premium resource today for training LLMs. We have limited compute in the country and it is fragmented amongst various Indian cloud providers. Some amount of sharing of compute resources is required to achieve our goal of building high-quality sovereign LLMs. The IndiaAI mission has come up a unique mechanism of pooling GPU resources across the country and providing GPU grants to sovereign LLMs. This is an important step in making LLM training viable.

However, one of the gaps in this approach is the duplication of effort and suboptimal overall use of compute given that each of the model trainers are training their own LLMs – albeit they have somewhat different use cases. Is there some thought towards sharing some workloads among these model builders? For instance – just a thought – most model builders would be spending a lot of compute in generating synthetic data – can this work be shared to the benefit of every company.

While startups have access to compute, **academia is significantly short on compute**. They need support to do research and innovation and tackle fundamental problems in LLM building in the Indian context. More importantly, **students need compute to be able to experiment and learn – to be able to develop high quality AI talent.** New models for making compute accessible need to be thought out.

Data Curation and Evaluation

It goes without saying that high-quality data is a fundamental ingredient to building high quality models. We have made rapid strides in data curation, but there is a long way to go. **The first wave of open resources** mainly covered fundamental language tasks like translation, transliteration, ASR, TTS amongst others. **In the second wave,** we need to curate data resource that can help power intelligent language models – chat LLMs, instruction LLMs, agents, etc. We have made a start with some web-based pre-training corpora and instruction tuning datasets like AI4Bharat's Sangraha collection, BhashaKritika from Krutrim, Updesh from Microsoft, etc. However, the scale, diversity and quality of the available corpora in different Indian languages needs to increase. Here are few concrete opportunities in this direction:

- Unlocking pdf, audio and non-web resources where we can capture data the rich literature, folklore and other cultural knowledge that are not generally represented on the web.
- High quality synthetic data providing high-density information content, various presentation formats, different levels of complexity is important to build helpful assistants.
- It is not only the data that is important, but also the data mixtures used for training are important. These are driven by usage, and methods and actual mixtures for training LLMs at various stages will be useful.
- High-quality, educational and certain domain specific datasets that are useful in mid-training.
- Preference alignment and reasoning datasets considering cultural nuances, viewpoints, local use-cases, etc. are important for advanced conversation, task completion and reasoning is Indian languages
- Building datasets for specialized areas important in the Indian context with domain experts – say tutoring systems for students, copilots for teachers in the education sector, capturing doctor/patient interactions.

Specifically, we need to develop benchmarks relevant to the Indian context – capturing use-cases relevant in India, driven by real usage and evaluating Indian trivia, cultural contexts and philosophical viewpoints.

Multilingual Learning

Multilingual learning has been made truly mainstream and transformed by deep learning, enabling powerful shared representations across languages and opening up new possibilities for Indian language technology. Multilingual training allows a single, **compact model** to serve many languages, often delivering strong **zero-shot performance** even for languages with limited data.

At the same time, significant opportunities remain. The reality is that most of the data and knowledge is present in English and knowledge transfer from English to Indian languages is still an open problem. Current systems underutilize the rich **linguistic similarities** among Indian languages, and multilingual transfer can be further extended to **preference alignment and reasoning models**. Finally, progress in this space depends on robust **multilingual evaluation**, including better automatic metrics and the use of LLMs themselves as judges to assess quality across diverse Indian languages.

Efficiency

Efficiency must become a first-class principle across every stage of LLM development, especially in the Indian context where compute, energy, and skilled manpower are constrained. Training frontier-scale models requires trillions of tokens, massive GPU clusters, and sustained expertise for high-quality data collection—costs that make unconstrained scaling impractical. Embedding efficiency into model design and training enables **sustainable research and development cycles**, rationalizes energy consumption, and allows models to be **adapted rapidly to new languages, domains, and modalities** without retraining from scratch. Yet, efficiency remains relatively underexplored in India, highlighting the need for model builders to engage deeply with efficiency-focused methods and literature. This opens up major opportunities in **data-efficient learning**, **efficient model architectures and optimization**, **capability-preserving adaptation of existing models**, and **AI-assisted data annotation**, all of which can significantly lower barriers while accelerating innovation at scale.

Last Words

As I come to the end of this talk, I realize that Pushpak Sir’s greatest legacy cannot be captured fully in papers, resources, or institutions, important as they are. It lives on in the way he taught us to think—deeply, patiently, and with humility; in the communities he built; in the talent he nurtured; and in his role in driving Indian language NLP forward. It is time for us to take that legacy forward. These are exciting times for NLP in India and for Indian language NLP. We have come a long way, but there are many exciting challenges ahead that we as a community can tackle and deliver Indian language technologies that are useful for real problems in the Indian context. I think that is the best **Gurudakshina** we could give Pushpak Sir.