Natural Language Processing in the Era of Large Language Models: Foundations, Integration, and Low-Resource Frontiers

Monisha Gottam *

Posted Date: 6 March 2026

doi: 10.20944/preprints202603.0570.v1

Keywords: large language models; natural language processing; low-resource languages; multilingual NLP; transfer learning; tokenization; transformer architecture; cross-lingual transfer

 doi:10.20944/preprints202603.0570.v1

Article

Natural Language Processing in the Era of Large Language Models: Foundations, Integration, and Low-Resource Frontiers

Monisha Gottam

Department of Applied Data Science, San Jose State University, San Jose, CA, USA; monisha.gottam@sjsu.edu

Abstract

Large Language Models (LLMs) have fundamentally transformed the landscape of Natural Language Processing (NLP), subsuming and redefining tasks that were once addressed by specialized, modular pipelines. This paper surveys the role of classical and contemporary NLP within modern LLM architectures, examining how foundational techniques — tokenization, syntactic parsing, semantic representation, and discourse modeling — have been absorbed into, and continue to inform, the pre-training and fine-tuning paradigms of transformer-based models. We further investigate the critical challenge of linguistic inclusivity, focusing on low-resource and morphologically complex languages that remain underserved by dominant English-centric corpora. Drawing on recent advances in cross-lingual transfer learning, multilingual pre-training, and data augmentation, we assess the progress and persistent gaps in extending LLM capabilities to such languages. Case studies on Southeast Asian, African, and indigenous language NLP toolkits illustrate practical strategies and remaining bottlenecks. We conclude by outlining open research directions at the intersection of structural NLP and generative AI.

Keywords: large language models; natural language processing; low-resource languages; multilingual NLP; transfer learning; tokenization; transformer architecture; cross-lingual transfer

1. Introduction

The relationship between Natural Language Processing and Large Language Models is one of both inheritance and transformation. NLP, as a field, developed over decades a rich repertoire of techniques: part-of-speech tagging, named entity recognition, syntactic parsing, coreference resolution, and semantic role labeling, among others [1,2]. These were crafted largely as modular components in processing pipelines, each consuming linguistically annotated data and producing structured representations of language.

The arrival of the transformer architecture [3] and its subsequent scaling into billion-parameter models such as GPT [4], BERT [5], and T5 [6] did not render classical NLP obsolete; rather, it radically changed its role. Pre-trained LLMs implicitly encode rich linguistic knowledge that emerges from self-supervised objectives over massive corpora, capturing morphological, syntactic, and semantic regularities without explicit annotation [7]. Yet the theoretical frameworks of NLP remain indispensable for understanding, evaluating, and auditing LLM behavior — and for building systems in data-scarce settings where brute-force scaling is not a viable strategy.

This preprint is organized as follows. Section 2 revisits foundational NLP components and traces their integration into LLM architectures. Section 3 examines the multilingual challenge, focusing on low-resource settings. Section 4 surveys state-of-the-art approaches and benchmarks. Section 5 presents targeted case studies. Section 6 discusses open problems and future directions.

 doi:10.20944/preprints202603.0570.v1

2. NLP Foundations Inside LLM Architectures

2.1. Tokenization and Subword Modeling

Tokenization is the first — and arguably most consequential — NLP decision embedded in every LLM. Classical approaches relied on whitespace segmentation or hand-crafted rules; modern LLMs universally employ subword tokenization algorithms such as Byte-Pair Encoding (BPE) [8], WordPiece [9], and SentencePiece [10]. These methods balance vocabulary coverage against sequence length, enabling models to handle morphologically rich languages with limited out-of-vocabulary exposure.

Nevertheless, tokenization remains an Achilles' heel for linguistic diversity. Languages with complex morphology — such as Turkish, Finnish, or Tagalog — are systematically over-segmented by vocabularies optimized for English, yielding longer token sequences, higher inference cost, and diminished representational fidelity [11]. Recent work by Rust et al. [12] demonstrates that vocabulary mismatch is a primary driver of performance degradation in multilingual BERT variants for low-resource languages.

2.2. Syntactic and Semantic Representations

Probing studies have extensively demonstrated that transformer hidden states encode syntactic information, including dependency parses and constituency structures, despite receiving no explicit syntactic supervision [13,14]. Work by Tenney et al. [14] showed that lower layers of BERT capture POS and chunking information while higher layers encode semantic role labels and coreference — mirroring the classical NLP pipeline in depth.

This emergent syntactic structure has practical implications: it suggests that LLMs can serve as general-purpose NLP backends, reducing the need for task-specific annotated datasets. However, this implicit encoding is fragile under distribution shift, performing poorly on non-standard dialects, code-switched text, and languages not represented in pre-training [15].

2.3. Discourse and Pragmatics

Beyond sentence-level processing, LLMs have absorbed capabilities traditionally associated with discourse NLP: coherence modeling, anaphora resolution, and pragmatic inference. The attention mechanism, with its capacity to relate arbitrary token pairs across long contexts [3], provides a structural basis for discourse-level reasoning. Long-context models such as Longformer [16] and extended context variants have further pushed these boundaries, enabling document-level summarization and multi-turn dialogue management.

3. The Low-Resource Language Challenge

3.1. The Data Imbalance Problem

Of the approximately 7,000 living languages documented globally, fewer than 100 have meaningful representation in the corpora used to train contemporary LLMs [17]. The Common Crawl, which underlies much of modern pre-training data, is estimated to be approximately 46% English, with the remaining share distributed heavily toward high-resource European languages [18]. Languages such as Tagalog, Yoruba, Swahili, Amharic, and dozens of others have token counts orders of magnitude smaller, producing severely undertrained representations.

This disparity has practical consequences: LLMs deployed in multilingual settings exhibit substantial performance gaps between high- and low-resource languages on benchmarks including XNLI [19], TyDiQA [20], and XTREME [21]. These gaps are not merely quantitative; they reflect genuine failures of understanding in morphology, pragmatics, and cultural knowledge that scale alone cannot easily remedy.

 doi:10.20944/preprints202603.0570.v1

3.2. Cross-Lingual Transfer and Multilingual Pre-Training

Multilingual pre-trained models such as mBERT [5], XLM-R [22], and mT5 [23] represent the dominant strategy for extending NLP capabilities to low-resource languages through shared multilingual representations. XLM-R, trained on 100 languages with a two-trillion-token corpus, demonstrated that sufficiently large multilingual models achieve competitive performance even on languages with limited training data, via cross-lingual transfer [22].

Zero-shot and few-shot cross-lingual transfer — the ability to fine-tune on high-resource language data and generalize to low-resource targets — has emerged as the primary practical strategy for NLP development in resource-scarce contexts [24]. Instruction-tuned LLMs such as mT0 [25] and BLOOMZ [25] further extend this paradigm, demonstrating that multilingual instruction following can be achieved with surprisingly few target-language examples.

3.3. Data Augmentation and Synthetic Data Strategies

In the absence of large labeled corpora, data augmentation has proven essential for low-resource NLP. Techniques include back-translation [26], cross-lingual data projection, template-based generation, and LLM-assisted corpus synthesis. Back-translation, originally developed for neural machine translation, has been repurposed to generate pseudo-labeled training data for classification and named entity recognition tasks [26].

More recently, LLMs themselves have been used as data generators for low-resource NLP, prompting high-resource-language models to produce training instances that are then projected or translated into target languages [27]. While promising, this approach risks propagating hallucinations and culturally inappropriate content, underscoring the need for human-in-the-loop validation, especially for safety-sensitive applications.

4. State-of-the-Art Approaches and Benchmarks

4.1. Instruction Tuning and RLHF for Multilingual Settings

Instruction tuning — fine-tuning LLMs on diverse task descriptions paired with target outputs — has become a standard technique for eliciting zero-shot generalization [28]. When applied in multilingual settings, instruction tuning has been shown to substantially improve performance on low-resource languages, even when the instruction dataset itself is predominantly English [25]. This phenomenon, termed 'multilingual task arithmetic,' suggests that instruction-following capabilities transfer across languages more readily than raw language modeling performance.

Reinforcement Learning from Human Feedback (RLHF) [29], while predominantly studied in English, presents both opportunity and risk for low-resource languages. Human preference data is expensive to collect for every language, and crowd-sourced preferences may not reflect culturally specific notions of helpfulness or harmlessness. Approaches such as preference data projection and culturally-aware RLHF are active areas of inquiry [30].

4.2. Benchmark Landscape

Evaluating multilingual NLP progress requires benchmarks that go beyond high-resource languages. Key benchmarks include XNLI (cross-lingual natural language inference, 15 languages) [19], TyDiQA (typologically diverse question answering, 11 languages) [20], XTREME and XTREME-R (cross-lingual transfer, 40+ languages) [21], and AfriSenti [31] for African language sentiment analysis. These benchmarks reveal consistent performance hierarchies strongly correlated with training data volume, emphasizing the need for targeted low-resource interventions beyond simply scaling general-purpose models.

 doi:10.20944/preprints202603.0570.v1

5. Case Studies in Low-Resource NLP Toolkits

5.1. CalamanCy: A Tagalog NLP Toolkit

Tagalog, spoken by over 90 million people as a first or second language in the Philippines, has historically been severely underrepresented in NLP research. CalamanCy [32,35] is a spaCy-based NLP toolkit developed to address this gap, providing production-ready pipelines for Tagalog including tokenization, POS tagging, morphological analysis, dependency parsing, and named entity recognition. The toolkit leverages transformer backbones pre-trained on Filipino web corpora and demonstrates that focused, community-led NLP development can produce high-quality tools for low-resource languages outside the standard English-centric ecosystem.

A subsequent implementation and evaluation presented at the 2025 IEEE International Conference on Industrial Technology & Computer Engineering (ICITCE) [35] further validated CalamanCy's practical utility in industrial and academic settings, confirming its robustness across diverse Tagalog text domains. CalamanCy's design philosophy emphasizes interoperability with existing NLP infrastructure, enabling downstream integration with modern LLM pipelines. The toolkit highlights both the potential and the effort required for community-driven low-resource NLP: achieving competitive performance requires careful curation of training data, attention to morphological complexity, and sustained community engagement [32,35].

5.2. AfroNLP and African Language Models

The AfroNLP initiative and associated models — including AfriBERTa [33] and Afro-XLMR — represent coordinated efforts to build NLP infrastructure for African languages at scale. Africa hosts over 2,000 languages, the vast majority of which have no digital NLP resources whatsoever. AfriBERTa, trained on 11 African languages using a compact corpus aggregated from Common Crawl and curated sources, demonstrates that even modest pre-training can yield meaningful NLP capabilities when combined with careful multilingual fine-tuning [33]. These efforts underscore the importance of community-centered data collection and the limitations of relying solely on web-scraped corpora for languages with limited digital footprints.

5.3. Lessons Across Case Studies

Across these case studies, several common themes emerge: (1) community and domain expertise is irreplaceable in corpus curation for low-resource languages; (2) subword tokenization must be adapted to the morphological typology of the target language; (3) cross-lingual transfer from high-resource relatives provides a strong baseline but cannot substitute for native-language pre-training data; and (4) evaluation benchmarks must reflect the specific linguistic properties and use cases of the target community rather than simply translating English benchmarks.

6. Open Problems and Future Directions

Despite significant progress, fundamental challenges remain at the intersection of NLP and LLMs, particularly for linguistically diverse settings. We identify five priority research directions:

Linguistically-Informed Tokenization: Next-generation tokenizers should incorporate morphological analysis to better serve agglutinative and polysynthetic languages, moving beyond frequency-based BPE toward linguistically motivated segmentation.

Efficient Low-Resource Adaptation: Parameter-efficient fine-tuning methods — including LoRA [34], prefix tuning, and adapter layers — offer promising paths for adapting large multilingual models to low-resource languages with minimal data and computation.

Culturally-Aware Alignment: RLHF and instruction tuning methodologies must be adapted to account for cultural variation in language use, pragmatics, and user expectations, requiring diverse human feedback collection beyond English-speaking populations.

 doi:10.20944/preprints202603.0570.v1

Evaluation Ecosystem: There is an urgent need for evaluation benchmarks that cover a substantially broader set of languages, particularly for underrepresented language families in sub-Saharan Africa, the Pacific, and indigenous communities globally.

Interpretability and Linguistic Grounding: As LLMs become de facto NLP engines, interpretability research must maintain rigorous connection to linguistic theory, enabling diagnosis and correction of model failures rooted in genuine linguistic misunderstanding rather than superficial pattern matching.

7. Conclusion

This paper has examined the deep and evolving relationship between Natural Language Processing and Large Language Models. LLMs have absorbed and extended core NLP capabilities — from tokenization to discourse modeling — through scale and self-supervision, while classical NLP frameworks remain essential for evaluation, linguistic grounding, and low-resource development. The challenge of linguistic inclusivity is both an ethical imperative and a scientific frontier: the billions of speakers of low-resource languages deserve NLP tools of comparable quality to those available in English, and building such tools will require novel approaches to data, modeling, and community engagement. We hope this survey serves as a useful reference for researchers and practitioners working to make language technology genuinely universal.

Submitted as a preprint. This work has not undergone peer review. Comments welcome.

References

1. Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.
2. Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed.). Draft. https://web.stanford.edu/~jurafsky/slp3/
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., et al. (2020). Language models are few-shot learners. NeurIPS, 33, 1877-1901.
5. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proc. NAACL-HLT, 4171-4186.
6. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR, 21(140), 1-67.
7. Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. Transactions of the ACL, 8, 842-866.
8. Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. Proc. ACL, 1715-1725.
9. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144.
10. Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. Proc. EMNLP, 66-71.
11. Ács, J. (2019). Exploring BERT's vocabulary. Budapest University of Technology and Economics Blog.
12. Rust, P., Pfeiffer, J., Vulic, I., Ruder, S., & Gurevych, I. (2021). How good is your tokenizer? On the monolingual performance of multilingual language models. Proc. ACL-IJCNLP, 3118-3135.
13. Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. Proc. NAACL-HLT, 4129-4138.
14. Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. Proc. ACL, 4593-4601.
15. Blasi, D. E., Anastasopoulos, A., & Neubig, G. (2022). Systematic inequalities in language technology performance across the world's languages. Proc. ACL, 1407-1423.
16. Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv:2004.05150.

 doi:10.20944/preprints202603.0570.v1

17. Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. Proc. ACL, 6282-6293.
18. Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., et al. (2022). Quality at a glance: An audit of web-crawled multilingual datasets. Transactions of the ACL, 10, 50-72.
19. Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., & Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. Proc. EMNLP, 2475-2485.
20. Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., & Palomaki, J. (2020). TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. Transactions of the ACL, 8, 454-470.
21. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. Proc. ICML, 4411-4421.
22. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., et al. (2020). Unsupervised cross-lingual representation learning at scale. Proc. ACL, 8440-8451.
23. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. Proc. NAACL-HLT, 483-498.
24. Lauscher, A., Ravishankar, V., Vulic, I., & Glavas, G. (2020). From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. Proc. EMNLP, 4483-4499.
25. Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., et al. (2023). Crosslingual generalization through multitask finetuning. Proc. ACL, 15991-16111.
26. Sennrich, R., Haddow, B., & Birch, A. (2016). Improving neural machine translation models with monolingual data. Proc. ACL, 86-96.
27. Moller, A. G., Dalsgaard, J. A., Pera, A., & Aiello, L. M. (2023). Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks. arXiv:2304.13861.
28. Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). Finetuned language models are zero-shot learners. Proc. ICLR.
29. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., et al. (2022). Training language models to follow instructions with human feedback. NeurIPS, 35.
30. Santurkar, S., Durmus, E., Ladd, F., Lee, C., Ganguli, D., & Hashimoto, T. (2023). Whose opinions do language models reflect? Proc. ICML, PMLR 202.
31. Muhammad, S. H., Yimam, S. M., Ahmad, I. S., et al. (2023). AfriSenti: A Twitter sentiment analysis benchmark for African languages. Proc. EMNLP, 13968-13981.
32. Miranda, L. J. (2023). CalamanCy: A Tagalog natural language processing toolkit based on spaCy. Proc. 3rd Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), ACL, 1-7.
33. Ogueji, K., Zhu, Y., & Lin, J. (2021). Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages. Proc. 1st Workshop on MRL, 116-126.
34. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. Proc. ICLR.
35. V. Parupally, "CalamanCy: A Tagalog Natural Language Processing Toolkit," 2025 IEEE International Conference on Industrial Technology & Computer Engineering (ICITCE), Penang, Malaysia, 2025, pp. 45-51, doi: 10.1109/ICITCE65255.2025.11210765.