Beyond Multilinguality: Typological Limitations in Multilingual Models for Meitei Language

Badal Nyalang
MWire Labs
Shillong, India
nyalang@mwirelabs.com

Abstract

We present MeiteiRoBERTa, the first publicly available monolingual RoBERTa-based language model for Meitei (Manipuri), a low-resource language spoken by over 1.8 million people in Northeast India. Trained from scratch on 76 million words of Meitei text in Bengali script, our model achieves a perplexity of 65.89, representing a 5.2× improvement over multilingual baselines mBERT (341.56) and MuRIL (355.65). We argue that Meitei's agglutinative morphology and complex word formation present typological challenges that multilingual models with broad language coverage fail to capture effectively. Through comprehensive evaluation on perplexity, tokenization efficiency, and semantic representation quality, we demonstrate that domain-specific pretraining significantly outperforms general-purpose multilingual models for low-resource languages. Our model exhibits superior semantic understanding with 0.769 similarity separation compared to 0.035 for mBERT and near-zero for MuRIL, despite MuRIL's better tokenization efficiency (fertility: 3.29 vs. 4.65). We publicly release the model, training code, and datasets to accelerate NLP research for Meitei and other underrepresented Northeast Indian languages.

1 Introduction

Meitei (also known as Manipuri) is an endangered Tibeto-Burman language spoken by approximately 1.8 million people, primarily in Manipur, India, and parts of Bangladesh and Myanmar. Despite being recognized as one of India's 22 scheduled languages and having a rich literary tradition spanning centuries, Meitei remains critically underrepresented in natural language processing research. The language exhibits complex agglutinative morphology, subject-object-verb word order, and can be written in both Meitei Mayek (indigenous script) and Bengali script, with the latter being more prevalent in digital contexts.

From a typological perspective, Meitei presents specific challenges for multilingual language models. As an agglutinative language, Meitei forms words through extensive suffixation and compounding, resulting in long, morphologically complex word forms that strain tokenization strategies optimized for isolating or mildly inflecting languages. Its SOV word order and predominantly head-final phrase structure differ from the SVO patterns dominant in many high-resource languages represented in multilingual models. These typological characteristics-shared with other Tibeto-Burman languages of the region-suggest that multilingual models trained predominantly on typologically distant languages may allocate insufficient capacity to capture Meitei's linguistic structure, motivating our investigation of monolingual alternatives.

Recent advances in transformer-based language models have revolutionized NLP across high-resource languages (Devlin et al., 2019; Liu et al., 2019). However, these benefits remain largely inaccessible to low-resource languages like Meitei due to their underrepresentation in multilingual models (Joshi et al., 2020; Ponti et al., 2020). While multilingual model such as mBERT (Devlin et al., 2019) and Indic-specific models like MuRIL (Khanuja et al., 2021) and IndicBERT (Kakwani et al., 2020) have attempted to address linguistic diversity, they often allocate insufficient capacity to individual low-resource languages, resulting in suboptimal performance (Conneau et al., 2020; Lauscher et al., 2020).

The critical challenge for endangered languages is not merely technological but existential: without adequate digital infrastructure and NLP tools, these languages risk accelerated decline in the digital age (Moseley, 2010) . Recent work has demonstrated that language-specific models trained from scratch can significantly outperform multilingual alternatives for low-resource languages, even with limited data (de Vries et al., 2019; Virtanen et al., 2019;

Proceedings of the 8th Workshop on Research in Computational Linguistic Typology and Multilingual NLP, pages 32–38

Martin et al., 2020). This finding is particularly relevant for Northeast Indian languages, which collectively represent over 220 distinct languages but remain marginalized in mainstream NLP research.

In this work, we present MeiteiRoBERTa, a monolingual RoBERTa-based encoder model trained from scratch on 76 million words of Meitei text. Our contributions are threefold: (1) We release the first publicly available transformer-based language model specifically designed for Meitei in Bengali script, (2) We demonstrate through rigorous evaluation that our model achieves 5.2× better perplexity than multilingual baselines while exhibiting superior semantic understanding, and (3) We provide comprehensive comparative analysis against state-of-the-art multilingual models, offering insights for future low-resource language modeling efforts.

2 Related Work

2.1 Multilingual Language Models

The development of multilingual BERT (mBERT) (Devlin et al., 2019) marked a significant milestone in cross-lingual NLP, demonstrating that a single model could capture linguistic patterns across 104 languages. Subsequent work on XLM-R (Conneau et al., 2020) extended this to 100 languages with improved performance through larger-scale training. For Indian languages specifically, MuRIL (Khanuja et al., 2021) and IndicBERT (Kakwani et al., 2020) were introduced as specialized multilingual models covering 17 and 12 Indian languages respectively. However, recent studies have shown that these multilingual models suffer from the "curse of multilinguality", where increased language coverage leads to decreased per-language performance, particularly for low-resource languages (Pfeiffer et al., 2020; Üstün et al., 2020).

2.2 Low-Resource Language Models

Growing evidence suggests that monolingual models trained from scratch can outperform multilingual alternatives for low-resource languages (de Vries et al., 2019; Virtanen et al., 2019). Recent work on BanglaBERT (Bhattacharjee et al., 2022), AraBERT (Antoun et al., 2020), and CamemBERT (Martin et al., 2020) has demonstrated significant performance gains through language-specific pretraining. For Northeast Indian languages, preliminary efforts include work on Assamese (Nath et al., 2023) and limited experiments with Manipuri NER systems (Singh and Bandyopadhyay, 2009), but no comprehensive pre-trained language model for Meitei has been publicly released prior to this work.

2.3 RoBERTa Architecture

RoBERTa (Liu et al., 2019) introduced key improvements over BERT including dynamic masking, removal of next sentence prediction (NSP), larger batch sizes, and longer training sequences. These modifications have consistently demonstrated superior performance across various benchmarks (Clark et al., 2020; Lan et al., 2020). Recent adaptations of RoBERTa for low-resource languages (Nguyen and Nguyen, 2020; Agerri et al., 2020) have shown that the architecture's efficiency makes it particularly suitable for scenarios with limited computational resources and smaller corpora, motivating our choice of RoBERTa over alternative architectures.

3 Methodology

3.1 Model Architecture

MeiteiRoBERTa follows the RoBERTa-base architecture with 12 transformer layers, 12 attention heads, and a hidden dimension of 768, totaling 125 million parameters. We trained a custom Byte-Pair Encoding (BPE) tokenizer with a vocabulary size of 52,000 tokens, optimized specifically for Meitei morphology in Bengali script. The tokenizer was trained on the full corpus to minimize out-of-vocabulary rates and handle the language's agglutinative morphological structure efficiently.

3.2 Training Data and Preprocessing

Our training corpus comprises 76 million words of Meitei text, representing a curated aggregate of the IndicCorp v2 subset (Gala et al., 2023) and independent crawls of local news archives, government documents, digitized literary collections and web content. The corpus underwent rigorous preprocessing including deduplication, language identification filtering, and quality assessment. Data sources include publicly available digital archives, news portals, and literary collections. The data was chunked into 353,123 blocks of 512 tokens for efficient batch processing during training.

The dataset is publicly available at `https://huggingface.co/datasets/MWirelabs/meitei-monolingual-corpus`.

3.3 Training Configuration

The model was trained from random initialization using masked language modeling (MLM) with a 15% masking probability. We employed an effective batch size of 256, a learning rate of 6×10^{-4} with linear warmup and decay, and trained for 3 epochs. Training was conducted on NVIDIA A40 GPUs. The final training loss converged to 4.1855, indicating successful learning of Meitei language patterns.

4 Evaluation

4.1 Baselines

We compare MeiteiRoBERTa against three multilingual baselines: (1) mBERT (Devlin et al., 2019): 110M parameters covering 104 languages and (2) MuRIL (Khanuja et al., 2021): 235M parameters optimized for 17 Indian languages including transliterated text. These models represent the current state-of-the-art for multilingual NLP and are commonly used for low-resource Indian languages.

4.2 Evaluation Metrics

Perplexity. We evaluate language modeling capability using perplexity on a held-out validation set of 19,038 samples, calculated as PPL = exp(average loss), where the loss is computed through masked language modeling prediction accuracy.

Tokenization Efficiency. We measure subword fertility (average number of subword tokens per word) on a test set of 10 diverse Meitei sentences. Lower fertility indicates more efficient tokenization and better vocabulary alignment with the language's morphology.

Semantic Representation Quality. We assess the quality of learned representations using semantic similarity tests on 4 manually curated sentence pairs (2 semantically similar, 2 dissimilar) drawn from our test set. Sentence pairs were constructed to test the model's ability to distinguish between paraphrases with shared semantic content versus unrelated sentences from different domains. We compute cosine similarity between [CLS] token embeddings, which serve as sentence-level representations in BERT-style models, following standard practice for semantic similarity evaluation (Devlin et al., 2019). For each pair labeled as semantically similar or dissimilar, we measure the model's ability to separate semantically related from unrelated content through the separation score (average high similarity - average low similarity). The choice of [CLS] serves as a typological probe; if a model cannot distinguish basic sentence-level meaning in its pooling layer, it indicates a fundamental failure to map the language's syntax and morphology into a coherent semantic manifold

Model	Parameters	Perplexity
mBERT	110M	341.56
MuRIL	235M	355.65
MeiteiRoBERTa	125M	65.89

Table 1: Perplexity comparison on Meitei validation set (19,038 samples). Lower is better.

Model	Vocab Size	Fertility
mBERT	119K	4.79
MuRIL	197K	3.29
MeiteiRoBERTa	52K	4.65

Table 2: Tokenization efficiency on 10 diverse Meitei sentences. Lower fertility is better.

5 Results and Analysis

5.1 Perplexity Comparison

Table 1 presents the perplexity results on the Meitei validation set. MeiteiRoBERTa achieves a perplexity of 65.89, substantially outperforming mBERT (341.56) and MuRIL (355.65) by factors of 5.2× and 5.4× respectively. This dramatic improvement demonstrates that monolingual pre-training with language-specific tokenization provides superior language modeling capabilities compared to general-purpose multilingual models, even when the latter have significantly more parameters.

5.2 Tokenization Efficiency

Table 2 shows tokenization efficiency measured by subword fertility. MuRIL achieves the best fertility score of 3.29, followed by mBERT (4.79) and MeiteiRoBERTa (4.65). While MuRIL's extensive vocabulary (197K tokens) enables more efficient tokenization, our custom BPE tokenizer with 52K tokens achieves competitive performance while maintaining a more compact model size. The trade-off between vocabulary size and model compactness is favorable for resource-constrained deployment scenarios.

5.3 Semantic Representation Quality

Table 3 presents semantic similarity evaluation results. MeiteiRoBERTa demonstrates exceptional

Model	High Sim	Low Sim	Sep.
mBERT	0.983	0.948	0.035
MuRIL	0.993	0.993	0.000
MeiteiRoBERTa	0.968	0.199	0.769

Table 3: Semantic representation quality on curated Meitei sentence pairs. Higher separation indicates better semantic understanding.

semantic understanding with a separation score of 0.769, vastly outperforming mBERT (0.035) and MuRIL (0.000). The near-perfect similarity scores (>0.95) from multilingual models for both related and unrelated sentence pairs indicate their failure to capture fine-grained semantic distinctions in Meitei. In contrast, MeiteiRoBERTa shows high similarity (0.968) for semantically related pairs while correctly assigning low similarity (0.199) to unrelated pairs, demonstrating genuine semantic comprehension.

The 0.000 separation score for MuRIL may indicate limited semantic differentiation under the current probing setup, where the model's pre-training on Indo-Aryan languages fails to provide the structural handles necessary to differentiate Meitei's Tibeto-Burman semantic features.

6 Discussion

Our results provide strong empirical evidence for the superiority of monolingual language models over multilingual alternatives for low-resource languages. The 5.2× perplexity improvement and dramatically better semantic separation (0.769 vs. 0.035) demonstrate that dedicated models can capture language-specific nuances that multilingual models miss, even when the latter have 2-3× more parameters.

The stark contrast in semantic representation quality is particularly noteworthy. MuRIL's near-zero separation score (0.000) suggests that despite its optimization for Indian languages, it shows limited ability to differentiate semantic contrasts in Meitei under our evaluation setting, assigning uniformly high similarity scores regardless of semantic content. This may indicate representation collapse or insufficient training signal for Meitei, aligning with recent findings on catastrophic interference in massively multilingual models (Artetxe et al., 2020), where low-resource languages receive insufficient training signal to develop meaningful representations.

While MuRIL demonstrates superior tokenization efficiency (fertility: 3.29 vs. our 4.65), reflecting the benefits of its larger vocabulary (197K tokens) optimized across multiple Indic languages, this advantage does not translate to better language understanding in our evaluation. This finding suggests a more nuanced relationship between tokenization efficiency and semantic representation quality than previously assumed. Efficient tokenization may be necessary but not sufficient for low-resource language support-dedicated model capacity and language-specific pre-training appear essential for developing robust linguistic representations. The trade-off between vocabulary size and model compactness remains an important consideration, particularly for deployment in resource-constrained environments where MeiteiRoBERTa's smaller vocabulary offers practical advantages.

The success of MeiteiRoBERTa with only 76 million words of training data (compared to billions used by multilingual models) has important implications for other low-resource languages. It demonstrates that even modest-sized corpora can yield high-quality language models when combined with appropriate architecture and training strategies. This is particularly encouraging for the hundreds of endangered languages worldwide that face similar resource constraints.

6.1 Implications for Underserved Communities

Our work directly addresses the digital divide affecting Northeast Indian linguistic communities. By providing the first comprehensive language model for Meitei, we enable potential applications in education, digital governance, cultural preservation, and content moderation. The model can support machine translation, sentiment analysis, and information retrieval systems that were previously unavailable for Meitei speakers. This technological infrastructure is crucial for preventing language endangerment in the digital age and ensuring equitable access to AI technologies.

From a typological perspective, our results highlight systematic limitations in how multilingual models handle morphologically complex, agglutinative languages. The dramatic performance gap suggests that current multilingual pretraining approaches-which distribute model capacity across typologically diverse languages-may inherently disadvantage languages with rich morphology and head-final syntax. This has broader implications for the approximately 220 languages of Northeast

India and hundreds of other agglutinative languages worldwide that share similar typological profiles with Meitei.

7 Conclusion and Future Work

We have presented MeiteiRoBERTa, the first publicly available transformer-based language model for Meitei, achieving state-of-the-art performance with 5.2× better perplexity than multilingual baselines and superior semantic understanding (0.769 separation vs. 0.035 for mBERT). Our comprehensive evaluation demonstrates that dedicated monolingual models remain the most effective approach for low-resource language processing, even in the era of massively multilingual models.

Future work will focus on several directions: (1) extending support to Meitei Mayek script through script-agnostic representations or dual-script training, (2) fine-tuning for downstream tasks including named entity recognition, sentiment analysis, and machine translation, (3) exploring few-shot learning capabilities for related Tibeto-Burman languages, and (4) investigating cross-lingual transfer learning between Bengali-script Meitei and other languages using the same script. Most importantly, we aim to establish partnerships with Meitei language communities to ensure our future work aligns with community priorities and contributes meaningfully to language preservation efforts.

The model (https://huggingface.co/MWirelabs/meitei-roberta) and evaluation datasets (https://huggingface.co/datasets/MWirelabs/meitei-monolingual-corpus) are publicly available to support further research on Meitei and other low-resource Northeast Indian languages.

These findings suggest potential typological limitations in multilingual models when applied to morphologically complex, agglutinative languages such as Meitei, with possible implications for other underrepresented Tibeto-Burman and related agglutinative languages.

Limitations

While our work represents a significant advance for Meitei NLP, several limitations warrant discussion:

Script Coverage: Our model only supports Bengali script, excluding the indigenous Meitei Mayek script still used in some contexts. Future work should explore multi-script models or script-agnostic representations.

Evaluation Scope: We primarily evaluate intrinsic metrics (perplexity, tokenization, semantic similarity). Downstream task evaluation on named entity recognition, sentiment analysis, and machine translation would provide additional insights into practical utility.

Corpus Limitations: Our 76M word corpus, while substantial for a low-resource language, may not capture the full linguistic diversity of Meitei, including dialectal variations and specialized domains. The corpus is also biased toward formal written text from news and government sources.

Community Involvement: Data collection and model development occurred without extensive consultation with Meitei language communities. Future efforts should employ participatory methods to ensure alignment with community needs and values.

Potential Biases: The model may inherit biases present in the training corpus, including underrepresentation of certain demographics, topics, or perspectives. We have not yet conducted comprehensive bias audits.

Computational Resources: While more efficient than training multilingual models, our approach still requires substantial computational resources (GPU training), which may limit reproducibility for researchers with limited access.

Ethical Considerations

Data Sourcing: All training data was collected from publicly available sources. We employed language identification and quality filtering to ensure corpus integrity, but acknowledge that web-scraped data may contain errors, biases, or copyrighted material despite our filtering efforts.

Cultural Sensitivity: Language technology for endangered languages carries responsibility for cultural preservation. While our model enables digital applications, we recognize that technology alone cannot address underlying sociolinguistic challenges. Community-led language revitalization efforts remain paramount.

Dual Use: Like all language models, MeiteiRoBERTa could potentially be misused for generating misinformation, impersonation, or surveillance. We advocate for responsible deployment and recommend implementing appropriate safeguards in downstream applications.

Access and Equity: By releasing our model, code, and data publicly, we aim to democratize ac-

cess to NLP tools for Meitei. However, meaningful access requires not just open models but also computational resources, technical expertise, and internet connectivity; resources unevenly distributed in Northeast India.

Acknowledgements

We thank the SIGTYP 2026 reviewers for their constructive feedback, which significantly improved the typological framing and clarity of this work. The author would like to acknowledge the **AI4Bharat** team for the **IndicCorp v2** initiative (Gala et al., 2023), which provided a vital foundational dataset for this study. This research was conducted under the research and development initiative at **MWire Labs**, Shillong. We also thank the digital archivists of Northeast India whose efforts in preserving online Meitei texts provided the primary signal for our monolingual pre-training.

References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for basque. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 4781–4788.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*, pages 9–15.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7375–7388.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations (ICLR)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Jay Gala, Pranjal A. Chitale, A. K. Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M., Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages. *Transactions on Machine Learning Research*. Survey of IndicCorp v2 coverage.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6282–6293.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. MuRIL: Multilingual representations for Indian languages. *arXiv preprint arXiv:2103.10730*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations (ICLR)*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4483–4499.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: A tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7203–7219.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*, 3rd edition. UNESCO Publishing.

Abhijnan Nath, Sheikh Mannan, and Nikhil Krishnaswamy. 2023. AxomiyaBERTa: A phonologically-aware transformer model for Assamese. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11629–11646, Toronto, Canada. Association for Computational Linguistics.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2009. Named entity recognition for Manipuri using support vector machine. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 811–818.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly universal dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.