Indigenous Writing Systems Matter: Rethinking NLP beyond Alphabetic Bias through Script-Aware Modeling

Ngoc Tan Le, Mamady Traore, Cristian Eduardo Ahumada Oliva, Fatiha Sadat
Université du Québec à Montréal (UQAM)
Montréal, QC, Canada
le.ngoc_tan@uqam.ca, traore.mamady@courrier.uqam.ca,
ahumada_oliva.cristian@courrier.uqam.ca, sadat.fatiha@uqam.ca

Abstract

Natural Language Processing (NLP) has made significant progress in recent years, largely driven by large-scale pretrained models and vast textual and multimodal corpora. However, these advances remain unevenly distributed, disproportionately benefiting high-resource languages while Indigenous and endangered languages—especially those employing diverse and less widely supported writing systems—remain underrepresented. This paper examines the role of writing system diversity in NLP, with a focus on Indigenous and endangered languages. We propose a theoretical framework that accounts for variation across writing systems and its implications for computational modeling. Specifically, we (i) provide an overview of writing system diversity, (ii) synthesize available computational resources, and (iii) present a structured analysis of challenges in modeling, tokenization, and evaluation. Our analysis shows that writing system diversity reveals structural biases embedded in current NLP pipelines. We conclude by identifying key open challenges and outlining directions for future research toward more inclusive, script-aware NLP approaches that better account for writing system variation.

1 Introduction

Recent advances in Natural Language Processing (NLP) have been largely driven by the availability of large-scale pretrained models and increasingly diverse textual and multimodal corpora (Zhang et al., 2024; Kasilingam et al., 2026). The emergence of large language models (LLMs) has further accelerated progress, enabling substantial improvements across a wide range of tasks, from machine translation to question answering (Edman et al., 2025; Scherbakov et al., 2025). However, these gains remain unevenly distributed. A growing body of work has highlighted the persistent underrepresentation of low-resource, Indigenous, and endangered languages in NLP research and infrastructure (Alam et al., 2024). At the same time, the scale and data requirements of LLMs risk amplifying these disparities, as languages with limited digital presence are less likely to be adequately represented in training data (Mishra et al., 2026). This disparity is most often explained in terms of data scarcity and limited resource availability. While these factors are central, they do not fully account for the challenges involved. An equally important but less examined dimension lies in the diversity of writing systems.

Beyond technical considerations, these challenges are closely tied to broader questions of data governance and sovereignty (Horna-Saldaña et al., 2025). For many Indigenous communities, language data is not simply a resource to be collected and processed, but a form of cultural and ancestral knowledge that is subject to principles of ownership, control, and consent (Edman et al., 2025). The development of NLP systems—particularly large-scale models trained on web-scale data—raises concerns about data extraction, lack of transparency, and the potential misuse of culturally sensitive material.

In this paper, we examine Indigenous and endangered languages by focusing their writing systems and their implications for NLP. We first provide an overview of writing system diversity, introducing key typological distinctions and representative examples with a focus on Indigenous and endangered languages. We then synthesize the current landscape of computational resources across scripts, including available corpora, tools, and benchmarks. Finally, we offer a structured analysis of challenges in tokenization, modeling, and evaluation, by proposing a theoretical framework across Indigenous writing system diversity, showing how standard assumptions break down in the presence of script variation.

Thus, the paper is organized as follows: Section 2 presents the relevant work about the writing

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 118–124

system diversity. Sections 3 presents the computational resources across Indigenous writing scripts. The structural bias in NLP pipelines and evaluations is presented in Section 4. And Section 5 addresses a theoretical framework toward script-aware NLP. Finally, Section 6 provides some conclusions and future research directions.

2 Writing system diversity: concepts and typology

Many Indigenous and endangered languages employ writing systems that diverge significantly from the alphabetic conventions that underpin most modern NLP pipelines (Fakhreldin, 2025). While most NLP research focuses on "language" as the primary unit of analysis, this study argues for a shift toward the "script" as a distinct computational entity.

2.1 What is a writing system?

In both linguistics and NLP, it is crucial to maintain a distinction between language and script (Keselj, 2009). A language refers to the spoken or signed form of communication, while a script is the visual system used to represent that language (Wierzbicka, 2014). In many Indigenous contexts, this relationship is complex; for instance, a single language may be represented by multiple scripts (*digraphia*), or a newly invented script may be central to a community's identity (Osborne, 2017; Brandt and Sohoni, 2018; Simpson, 2024). Key notions within this domain include:

- **Grapheme**: The smallest functional unit of a writing system.
- **Unit**: The level at which a script encodes information (*e.g.* phoneme, syllable, or morpheme), which directly dictates how a tokenizer processes text.
- **Orthography**: The set of standardized conventions for using a script to write a specific language. For many Indigenous languages, orthographies are often unstandardized or evolving.

Table 1 illustrates this concretely: the same Pular sentence submitted in three scripts to three frontier LLMs produces divergent language detection, inconsistent translation, and variable tokenization, despite identical semantic content.

2.2 Typology of writing systems

The structural properties of a script—such as grapheme composition and diacritics—act as computational bottlenecks (Liu et al., 2025). Indigenous scripts encompass several typological families:

- **Alphabetic**: Symbols map roughly to individual phonemes (*e.g.* N'Ko, ADLaM, Osmanya). While these are most compatible with standard NLP pipelines, they remain severely underrepresented in pretrained corpora.
- **Syllabaries**: Each symbol represents a full syllable (*e.g.* Cherokee, Vai, Canadian Aboriginal Syllabics). Subword tokenization often fails here because the symbols themselves already encode the syllabic units that Byte Pair Encoding (Sennrich et al., 2016), SentencePiece (Kudo, 2018) or WordPiece (Schuster and Nakajima, 2012; Song et al., 2021) algorithms attempt to derive statistically (Joshi et al., 2020).
- **Abugidas (Alphasyllabaries)**: Characters encode a consonant with an inherent vowel, modified by diacritics (*e.g.* Ge'ez, Tifinagh, Ol Chiki). These systems present non-trivial challenges for character decomposition, normalization, and rendering.
- **Logographic / Morphosyllabic**: Symbols represent words or morphemes (*e.g.* Mayan hieroglyphs, Dongba symbols). In these systems, segmentation is extremely challenging, and digital corpora are exceptionally sparse.
- **Semasiographic / Mixed Systems**: Systems like Nsibidi are symbolic and not strictly tied to spoken language structure. These are difficult to model with standard NLP assumptions and may require multimodal AI approaches.

2.3 Indigenous and endangered scripts

Indigenous scripts are characterized by high levels of variability and a "long march" toward digital recognition (Llanes-Ortiz et al., 2023; Manimaran et al., 2024; Agarwal and Anastasopoulos, 2024). The Atlas of Endangered Alphabets documents Indigenous and minority writing systems and highlights efforts to preserve them[1].

[1] https://www.endangeredalphabets.net/alphabets/

Many writing systems, such as the Cherokee syllabary or N'Ko, were invented in the 19th and 20th centuries as tools for cultural preservation. From both linguistic and NLP perspectives, these scripts face unique challenges:

- **Variability and Standardization**: Many communities lack a single standardized orthography. For example, the Mapuzugun community utilizes three distinct alphabets—Unificado, Ragileo, and Azümchefe—which complicates the creation of consistent datasets and models.

- **Orality and Revitalization**: Many Indigenous languages were historically oral and have only recently adopted written forms. Revitalization efforts often lead to script revival (*e.g.* Meitei Mayek), introducing further orthographic variation.

- **The "script tax"**: This structural diversity results in a systematic performance penalty (Dixit and Dixit, 2026). Models often require up to 13 times more tokens to represent the same content in Indigenous scripts compared to Latin ones, which reduces representational density and increases computational costs (Petrov et al., 2023).

3 Computational resources across Indigenous writing scripts

Resource availability for Indigenous scripts is often described as a "long march to Unicode", where digital survival depends on integration into global encoding standards (Agarwal, 2025).

- **Data Ecology**: Digital corpora are often skewed toward religious texts or news, narrowing lexical diversity. Digitization is further hampered by the lack of script-specific OCR/HTR for traditions like Wolofal (Cissé and Sadat, 2023, 2024; Le et al., 2025).

- **Infrastructure Gaps**: Formal Unicode inclusion (*e.g.* ADLaM in 2016 (Hossain, 2026)) is only a first step; practical usability requires standardized keyboard layouts and font rendering. It does not guarantee usability. And there is a persistent lack of standardized keyboard layouts and font rendering engines (Simpson, 2025).

- **Community-Led Innovation**: Projects like Amulwe Kimün and the Mapuzugun orthography converter demonstrate how minimal resources (dictionaries, grammars) can be transformed into tools for revitalization. For Mapuzugun, an orthography detector and converter handles the community's use of three distinct alphabets: Unificado, Ragileo, and Azümchefe (Ahumada et al., 2022).

4 Structural Bias in NLP pipelines

The current NLP pipeline imposes a systematic performance penalty, or "script tax", not only on non-Latin writing systems but also on underrepresented systems.

- **Tokenization Inequity**: The technique such as subword tokenization (*e.g.* Byte Pair Encoding), trained on Latin-dominant data, assumes linear sequences and stable boundaries. This tokenizer fragments Indigenous scripts into significantly more tokens—sometimes up to 13 times more than English (Asprovska and Hunter, 2024). Therefore, this causes oversegmentation in scripts with dense graphemes (*e.g.* Ethiopic), leading to higher computational costs and reduced context windows.

- **Representational Bottlenecks**: LLMs allocate disproportionate capacity to dominant scripts. This results in measurable drops in arithmetic and logical accuracy when using underrepresented scripts like N'Ko or ADLaM, even if the underlying meaning is identical to Hindu-Arabic numerals.

- **Modeling Assumptions**: Assumptions of whitespace and linearity break down for polysynthetic languages or scripts that do not use spaces.

- **Evaluation Bias**: Benchmarks, such as EXECUTE (Edman et al., 2025), often assume Latin-compatible norms and are insensitive to graphemic variation or segmentation instability. These evaluations reveal that task difficulty is shaped by writing system structure rather than character count, yet Indigenous scripts are often absent from these frameworks.

5 Toward script-aware NLP: A Theoretical Framework

To address these inequities toward script-aware NLP, inspired from (Fakhreldin, 2025), we propose a four-layer theoretical framework designed to systematically diagnose and address inequities across different writing systems of Indigenous and endangered languages. This framework shifts the unit of analysis from the language to the script.

The Four-Layer Theoretical Framework is constituted by:

(1) **Infrastructural Layer**: This foundational layer addresses the basic digital vitality of a script. Key components include Unicode allocation, the availability of standardized fonts, and keyboard input systems. Without this infrastructure, scripts remain computationally invisible regardless of their linguistic importance.

(2) **Representational Layer**: This layer focuses on modeling mechanics, specifically how scripts are segmented and stored in a model's vocabulary. It highlights issues like tokenization fragmentation, where non-Latin scripts are often over-segmented (up to 13 times more than English), and vocabulary allocation bias, which favors dominant scripts.

(3) **Functional Layer**: This layer evaluates performance on downstream NLP tasks such as Machine Translation, Named Entity Recognition, and arithmetic reasoning using script-level diagnostics to detect disparities. It identifies the "script tax", a systematic performance penalty where models underperform on tasks when using underrepresented scripts, even if the underlying meaning is identical to high-resource scripts.

(4) **Epistemic Layer**: The final layer addresses the critical and ethical framing of NLP development. It is used to reframe "low-resource" status as a product of policy neglect rather than technical intractability. It prioritizes Indigenous data sovereignty (Russ-Smith and Randell-Moon, 2025), decolonial ethics (Philip et al., 2012; Risam, 2018; Chew et al., 2023), and the reform of evaluation benchmarks that currently reproduce Western, Latin-centric standards (Bird, 2020).

6 Conclusion

In this paper, we argued that writing systems constitute a critical and underexplored axis of variation in NLP. By focusing on Indigenous and endangered languages, we examined how script diversity interacts with data availability, modeling choices, and evaluation practices.

Script diversity is not a peripheral "edge case" but a fundamental test for the next generation of NLP (Deng et al., 2024). The performance gaps identified in this study are systematically produced by engineering design choices that favor alphabetic, Latin-based norms. Progress requires a refoundation of NLP architectures to include script-aware tokenization, balanced multiscript pretraining, and evaluation metrics that respect the orthographic realities of Indigenous communities. Achieving multiscript equity is a structural precondition for a truly inclusive multilingual future (Horváth et al., 2025). Finally, we outlined future research paths and open problems for more inclusive, script-aware NLP techniques. Rather than treating non-alphabetic and low-resource writing systems as edge cases, we argued that they exposed fundamental limitations in current approaches and should therefore be central to the next generation of NLP research.

Ethical Considerations

Working with Indigenous and endangered language data involves significant risks of perpetuating colonial harms. The study of Indigenous writing systems is inextricably linked to broader questions of data governance, data sovereignty, and decolonial ethics.

Indigenous Data Sovereignty and Governance: For many Indigenous communities, language data is not merely a digital resource to be collected, but a form of cultural and ancestral knowledge subject to principles of ownership, control, and consent. Researchers must prioritize Indigenous data sovereignty, ensuring that communities remain decision-makers regarding how their scripts are documented and processed.

Decolonial Ethics and Relationality: Effective research requires building meaningful, long-term relationships with language communities rather than treating speakers as mere "information sources". This includes acknowledging the role of the researcher's positionality and involves centering the priorities of Indigenous researchers and building meaningful, long-term relationships with

language communities

Risks of Reductive Framing and Misuse: There is a significant risk in treating diverse writing systems (*e.g.* ADLaM, Vai, Tifinagh) as a monolithic group, which ignores their unique typological histories and community contexts. Furthermore, documenting technical vulnerabilities—such as the "script tax" or infrastructure gaps—could theoretically be misused to justify the continued exclusion or neglect of these scripts in global technology (Zaugg et al., 2022; Simpson, 2025).

Accessibility and Benefit-Sharing: Tools developed through these studies should be made available to the community for free, supporting revitalization and education rather than just academic advancement.

Limitations

While this study highlights critical biases, it also faces several structural and technical constraints.

Diagnostic Blind Spots: Many Indigenous scripts—including ADLaM, N'Ko, and Tifinagh—are currently absent from major tokenization and efficiency benchmarks. This omission reflects a "diagnostic blind spot" where the systems most in need of evaluation are excluded from the frameworks used to assess inequity.

Uncertainty of Scaling Laws: It remains an open question whether simply increasing the scale of models or data will reduce or exacerbate script-level disparities. Comprehensive scaling laws that account for script diversity have not yet been established.

Technical Fragmentation: Most existing script-aware interventions, such as custom tokenizers or adaptive segmentation, remain isolated experiments rather than features integrated into general-purpose multilingual models.

Dependency on Transliteration: Digitally disadvantaged languages often enter NLP pipelines through transliteration or partial tooling rather than fully native-script pathways. This often reduces the structural distinctiveness of the original writing systems in the training data.

Publication and Language Bias: The current literature relies heavily on Anglophone publication venues, which may overlook relevant scholarship published in regional or Indigenous languages that are not indexed in major databases.

Acknowledgments

This research was supported IVADO and the Canada First Research Excellence Fund. We are grateful to the anonymous reviewers for their thoughtful and valuable feedback.

Appendix

References

Milind Agarwal. 2025. *Improving Resource Creation for Low-Resource Languages Using NLP Methods*. Ph.D. thesis, George Mason University.

Milind Agarwal and Antonios Anastasopoulos. 2024. A concise survey of ocr for low-resource languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 88–102.

Cristian Ahumada, Claudio Gutierrez, and Antonios Anastasopoulos. 2022. Educational tools for mapuzugun. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 183–196.

Firoj Alam, Shammur Absar Chowdhury, Sabri Boughorbel, and Maram Hasanain. 2024. Llms for low resource languages in multilingual, multimodal and dialectal settings. In *Proceedings of the 18th conference of the European chapter of the association for computational linguistics: tutorial abstracts*, pages 27–33.

Marijana Asprovska and Nathan Hunter. 2024. The tokenization problem: Understanding generative ai's computational language bias. *Ubiquity Proceedings*, 4(1).

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th international conference on computational linguistics*, pages 3504–3519.

Carmen Brandt and Pushkar Sohoni. 2018. Script and identity–the politics of writing in south asia: an introduction. *South Asian History and Culture*, 9(1):1–15.

Kari AB Chew, Wesley Y Leonard, and Daisy Rosenblum. 2023. Decolonizing indigenous language pedagogies: Additional language learning and teaching. *Handbook of languages and linguistics of North*, pages 767–788.

Thierno Ibrahima Cissé and Fatiha Sadat. 2023. Automatic spell checker and correction for under-represented spoken languages: Case study on wolof. In *Proceedings of the Fourth Workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 1–10.

Thierno Ibrahima Cissé and Fatiha Sadat. 2024. Advancing language diversity and inclusion: Towards a neural network-based spell checker and correction

Model	ISO 639-3 Detected	Script Found	Translation	Text Domain	Tokens	Chars.
Latin: Sabu salli Bilaali fodaama malal						
GPT-5.4	fuf	Latin	incorrect	religion	5	32
Gemini 3	ful	Latin	incorrect	proverb	8	32
Claude 4.6	fuf	Latin	incorrect	religion	5	34
ADLaM: 𞤅𞤁𞤅𞤁𞤖 𞤁𞤖𞤑𞤭𞤄 𞤔𞤅𞤑𞤅𞤉𞤦 𞤔𞤅𞤑𞤁𞤉 𞤖𞤦𞤁𞤔						
GPT-5.4	fuf	Adlam	incorrect	religion	5	32
Gemini 3	ful	Adlam	incorrect	religion	18	32
Claude 4.6	fuf	Adlam	incorrect	religion	5	36
Ajami: سَبُ صَلِّ بِلَالِ فُدَامَ مَلَلْ						
GPT-5.4	und	Arabic	unknown	unknown	5	33
Gemini 3	ful	Ajami	incorrect	religion	11	29
Claude 4.6	fuf	Ajami	incorrect	religion	5	33

Table 1: Illustrative probe: the same Pular (`fuf`) sentence—meaning *"Because the adhan, Bilal has been promised eternal happiness"*—submitted in three scripts to three frontier LLMs. No model produced a correct translation in any script. `ful` denotes the Fula macrolanguage rather than the specific `fuf` (Guinea Pular) code. Gemini's token count inflates from 8 (Latin) to 18 (ADLaM), a 2.25× increase illustrating script-dependent tokenization fragmentation. Character-count disagreement across models for the same raw string further shows that even basic Unicode/code-point handling is inconsistent. None of the four major MT systems (NLLB-200, AfriNLLB, MADLAD-400, SeamlessM4T v2) supports the `fuf` ISO code, rendering Pular computationally invisible at the infrastructure level.

for wolof. In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages@ LREC-COLING 2024*, pages 140–151.

Yinlin Deng, Chunqiu Steven Xia, Chenyuan Yang, Shizhuo Dylan Zhang, Shujing Yang, and Lingming Zhang. 2024. Large language models are edge-case generators: Crafting unusual programs for fuzzing deep learning libraries. In *Proceedings of the 46th IEEE/ACM international conference on software engineering*, pages 1–13.

Aradhya Dixit and Shreem Dixit. 2026. The script tax: Measuring tokenization-driven efficiency and latency disparities in multilingual language models. *arXiv preprint arXiv:2602.11174*.

Lukas Edman, Helmut Schmid, and Alexander Fraser. 2025. Execute: A multilingual benchmark for llm token understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1878–1887.

Mohammed Fakhreldin. 2025. Developing a comprehensive nlp framework for indigenous dialect documentation and revitalization. *International Journal of Advanced Computer Science & Applications*, 16(4).

Cesar Jhonnatan Horna-Saldaña, Juan Ernesto Perez Perez, and Manuel Leonardo Toro Galeano. 2025. Artificial intelligence in the preservation of native languages and bridging the information access gap for indigenous peoples. *Journal of Enabling Technologies*, 19(1):63–75.

Alíz Horváth, Cosima Wagner, David Joseph Wrisley, Joanne Bernardi, Aanya Chadha, Danielle Garcia, Till Grallert, Sharanya Ghosh, Yuri Ishida, and Aleksandra Kaye. 2025. Who are the users in multilingual digital humanities? *Digital Scholarship in the Humanities*, page 1.

Anushah Hossain. 2026. Unsettled scripts: Script invention and the bind of digital recognition. *Modern Languages Open*, 1(1).

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*, pages 6282–6293.

N Kasilingam, Vaishnavi S Kedhlaya, Saakshi S Moolya, and Manvith N Poojary. 2026. Multilingual nlp models: A review of techniques, tools, and real-world use cases. In *2026 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, pages 1–5. IEEE.

Vlado Keselj. 2009. Book review: Speech and language processing by daniel jurafsky and james h. martin. *Computational Linguistics*, 35(3).

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75.

Ngoc Tan Le, Ali Mijiyawa, Abdoulahat Leye, and Fatiha Sadat. 2025. The best of both worlds: Exploring wolofal in the context of nlp. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 1–6.

Yihong Liu, Haotian Ye, Chunlan Ma, Mingyang Wang, and Hinrich Schütze. 2025. Langsamp: Language-script aware multilingual pretraining. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1743–1770.

Genner Llanes-Ortiz and 1 others. 2023. *Digital initiatives for indigenous languages*. UNESCO Publishing.

A Manimaran, Mohammad Haider Syed, M Siva Kumar, S Selvanayaki, Gurram Sunitha, and Asmita Manna. 2024. Enhancing asian indigenous language processing through deep learning-based handwriting recognition and optimization techniques. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(8):1–20.

Yash Mishra, Suyash Mishra, and Kedarnath Senapati. 2026. Attention amplification in multilingual llms: Why script representation matters. *DOI: 10.21203/rs.3.rs-8959575/v1*.

Henry S Osborne. 2017. *Indigenous Use of Scripts as a Response to Colonialism*. Ph.D. thesis, University of Oregon.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. *Advances in neural information processing systems*, 36:36963–36990.

Kavita Philip, Lilly Irani, and Paul Dourish. 2012. Postcolonial computing: A tactical survey. *Science, Technology, & Human Values*, 37(1):3–29.

Roopika Risam. 2018. Decolonizing the digital humanities in theory and practice. In *The Routledge companion to media studies and digital humanities*, pages 78–86. Routledge.

Jessica Russ-Smith and Holly Randell-Moon. 2025. Ai and indigenous data sovereignty: Knowing, engaging, and learning in new data contexts. *Somatechnics*, 15(3):287–295.

Dmitry Scherbakov, Nina Hubig, Vinita Jansari, Alexander Bakumenko, and Leslie A Lenert. 2025. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *Journal of the American Medical Informatics Association*, 32(6):1071–1086.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

L Simpson. 2025. *Modern Indigenous writing systems: From inception to Unicode*. Ph.D. thesis, Queen Mary University of London.

Logan Simpson. 2024. From icons to identities: Analysing visual cultural elements in emerging scripts. *Visible Language*, 58(2):42–81.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2021. Fast wordpiece tokenization. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 2089–2103.

Anna Wierzbicka. 2014. Language and cultural scripts. In *The Routledge handbook of language and culture*, pages 339–356. Routledge.

Isabelle A Zaugg, Anushah Hossain, and Brendan Molloy. 2022. Digitally-disadvantaged languages. *Internet Policy Review*, 11(2):1–11.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12401–12430.