NLP Subfield Survey: Impoverished Languages

Andres Gutierrez
College of Computing, Data Science & Society
University of California, Berkeley
andres_gutierrez@berkeley.edu

Abstract

The development of linguistic techniques in the course of NLP has been bench marked by their success in mostly high-resource language settings. As a consequence, research practices in the field have become biased and skewed away from supporting the vast majority of the Earth's estimated 7000 naturally languages. Our understanding of these languages, what I call the *impoverished languages* is limited by the pace by which high-resource practices move. In this paper, we will discuss the development of the impoverished languages and how their study adds linguistic value to the field of NLP.

1 Introduction

An estimated 7000 natural languages are spoken around the Globe. Despite this diversity, mainstream discussion and developments in the field of NLP revolve around techniques and applications supporting less than 2% of the 7000 languages. Historically, this is not too surprising. Westernization of the world has made it so that a select few languages have become the ones of convenience and utility such that local tongues carry less social weight. In consequence, the language focus of NLP itself has become westernized–those considered *high-resource languages*.

On the opposite end, we have what I collectively call the **impoverished languages**: low-resource languages, unknown scripts, and constructed language (or conlangs). Thematically, this set of language types have been glossed over in research but add deep linguistic value.

Within this paper, it is my hope to synthesize the papers definitive to the discussion of impoverished languages that either: **(a)** promote the visibility of each language type, **(b)** discuss application of existing methods to a language type, **(c)** develop a method more conscious to the nature of a language type not accounted in common-place study as a consequence of biases from working with higher-resource languages.

Word count: 2196

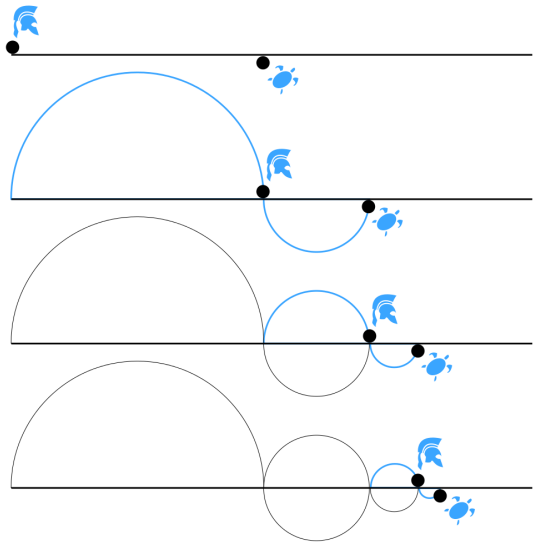

Figure 1: **Zeno's Paradox: Achilles and the Tortoise**. From Grandjean (2014), licensed under CC BY 3.0 CH.

2 Low-Resource Languages

In a footrace between Achilles and a tortoise, Achilles will never surpass the tortoise if it had a head start. Each point Achilles passes, the tortoise has already done so and has moved up to further point Achilles must again cover (Figure 1).

As Nigatu et al. argue, low-resource languages take the position of Achilles, limited by the pace at which the tortoise, or high-resource languages, moves. The analogy of Achilles extends to unknown scripts and conlangs granted their impoverished qualities.

Part of the lack of representation for low-resource languages stems from a general disunity as to what exactly "low-resource" refers when it

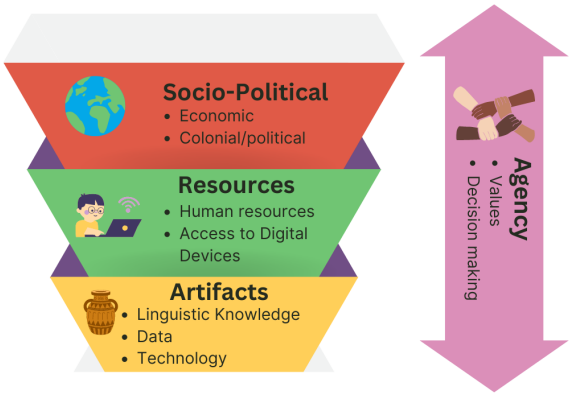

Figure 2: **Four overarching aspects that contribute to a language being classified as low-resource**. From The Zeno's Paradox of Low-Resource Languages (Nigatu et al., 2024).

appears in different papers. Nigatu et al. set forth official terminology having analyzed 150 papers that feature the term. A language, they define, can be "low-resourced" across four different axioms: socio-political, resources, artifacts, and agency (Figure 2).

2.1 Preservation Efforts

Standard LLMs are trained on corpora like OSACR, CommonCrawl, and PILE which are skewed heavily to latin scripts. This is evident when evaluating models in common NLP tasks like NER and POS tagging where these multilingual LLMs perform considerably better on a language if it's presented in its romanzied form (Adilazuarda et al., 2025). Ultimately, this contributes to the erasure of a language and is entirely more problematic if a script is not unicode-supported.

NUSAAKSARA, a benchmark dataset collected from 8 of scripts from Indonesia's 700 languages, or *askara*, was developed specifically with preservation in mind. The benchmark, annotated by native speakers and experts, tested on state-of-the-art models like GPT-4o and Gemini consistently report poor ChrF++ scores in tasks of translation (11.-20.) and transliteration (>1 CER) (Adilazuarda et al., 2025).

The lackluster performance stresses the further need for local script integration in training of common-place models and the support of unicode formatting.

Against High-Resource Practice Bias Byte-pair encoding is a common method of tokenization by which models ingest language, but its function is often over-reductive of morphologically complex languages. MOVXOC (Teklehaymanot et al., 2025), a novel tokenizer for Ge'ez scripts, uses a blended vocabulary of BPE and a morpheme boundary cap, that scores well for intrinsic MorphScores that reflect adherence to the morphological nuances of Ge'ez scripts that current BPE systems fail to capture.

Teklehaymanot et al., Chang and Basit, and others have made it clear that tokenizers can largely shape LLM reasoning. This can largely impede the use of LLMs in the work of decipherment when it comes to certain glyph-based scripts, like Mandombe, that stumble upon mistaken interpretation (Figure 3). This gap promotes the development of glpyh-based tokenizers (Shih et al., 2025).

2.2 Model Interventions

The curse of multilingualism describes the tendency for LLMs to fail in their unerstanding of a single target language despite being trained on a diverse set of languages.

Early techniques to surmount this issue looked into modifying the models themselves. *Meta-Learning*, as it's structured in Nooralahzadeh et al. (2020), fine tunes in a **zero-shot** or **few-shot** setting[1] of tasks selected from a pool of auxiliary languages which shows the greatest benefit in Question-Answering tasks with greater linguistic diversity.

We see in comparison efforts to modularize the architecture of models. The MAD-X approach (Pfeiffer et al., 2020) introduces plug-in

[1] "*X-shot*" presented in recent NLP study is commonly associated with prompt-interventions in the sense of in-context learning; however Nooralahzadeh et al. use the term to unconventionally refer to selection of down-stream tasks.

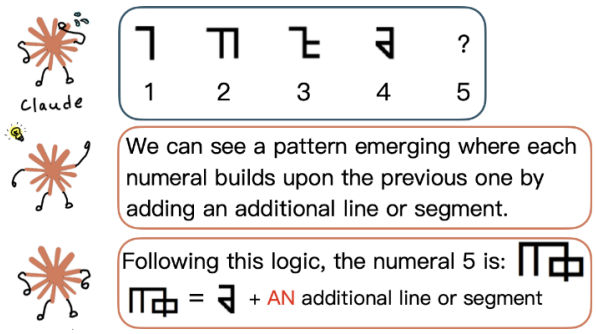

Figure 3: Example incorrect LVLM geometric interpretations on Mandombe Script. From Shih et al. (2025), licensed under GFDL.

"adapters"–language, task, and invertible. Each one is finetuned to a target language & corresponding task with the intent of being swapped in upon query. However, this approach, unlike the work in Nooralahzadeh et al.'s meta-learning approach, requires explicit finetuning on the target language.

Data Scale The success of a model–within and outside the LLM Space–is commonly understood to depend on the volume accessible data. Translation quality seems to show improvements at roughly 25,000 samples in training (Chang and Basit, 2025) .

2.3 Prompt Interventions

As models have grown beyond architectural comprehension, it has become more convenient to perform tweaks via in-context learning. MWE (*multi-word expressions*) are of particular research interest in low-resource settings as they require out-of-context comprehension which often does not appear in training-sets. Dimakis et al. prompt Greek-to-English translation while providing the "glosses" of the MWE. GPT-4 with this aided translation is rated higher by Greek speakers than those proposed by the No Language Left Behind multilingual model (NLLB) and GPT-4 base model, warranting some merit to the dictionary-approach.

Similar research with the dictionary-prompt approach on Inuktitut finds lackluster performance with BLEU scores as the metric (Elsner and Needle, 2023). However, further tinkering finds in-context grammar rules and chain-of-thought prompts encourage more methodical dissection of presented polysynthetic words. While still not producing promising results, it leads to another relevant question: are LLMs able to extrapolate and apply in-context learned grammar at all?

Zhang et al. pose in-aided translation as a process of two steps: retrieval and application. A "rule-by-rule" approach, instead of a full in-context grammar book, prompts the LLM iteratively presenting a sampled rule and a context sentence asking if they are related thus framing the task as binary classification. Once a correct rule is identified it is provided for in-context translation. All of the models they used in their experiments showed improvements of their BLEU translation scores over an unaided baseline. Interestingly, reformatting the textual linguistic rules to that of code elicited the model's math and logic strength which showed considerable translation improvement which showed greater benefit as the complexity of the rule expanded (Zhang et al., 2025).

3 Unknown Scripts

3.1 Acquiring Character Inventories

Before a script can be deciphered, it is essential to have a character inventory describing the basic units of the writing system and the corresponding sounds for speech production.

Early approaches have fitted a character(c)-sound(s) pairs as a transition between observable and unobservable states to map entire sequences of sound or script, i.e $P(s_1, s_2, ...s_n|c_1, c_2, ..., c_n)$[2]. The proper inventories can be built using an Expectation-Maximization (EM) algorithm until a convergence criteria has been met (Knight and Yamada, 1999).

Alternatively, if a script has been identified to have linguistic roots with known languages, cognates become a powerful tool for vocabulary building. Tamburini (2023) introduces flexible encoding between characters of a known script, K, and an unknown script, L, in an simulated-annealing optimization setting to provide concrete cognate identification within Aegan scripts.

Character Alignment: Sequence Ciphers Sequence-to-Sequence ciphers have been contended as a basis for 1:1 decipherment without knowledge of the plain text. Mimicking a neural machine translation (**NMT**) approach, transformers augmented with **frequency analysis** of the enciphered keys helps align meaning for transformers (Aldarrab and May, 2021). Other work has elaborated this sequence-to-sequence approach attempting to line up cognate pairs across languages with a latent flow variable, $\mathcal{F}$, under a minimum-cost-flow approach again optimized by EM to encourage non-lazy translations (Luo et al., 2019).

Character Individuality Commonly languages rely on stylistic variation in the form of allographs adding another nuance to building character inventories: is a character we come across new or a variation of one we already know? To answer this, deep-clustering on VAE-based encoders has proven effective in producing novel character clusters in Crypto-Greek (Born et al., 2023).

[2]Although not cited as so in (Knight and Yamada, 1999), this is effectively a Hidden-Markov-Model that assumes the importance of sequential appearance of states while the traditional EM algorithm does not account for the order of point appearance.

3.2 Frames of an Unknown Script

A look into cross-lingual transfer work done by Tufa et al. makes it evident that some researchers consider the mere absence of the target language in pre-training to make a language an unknown-script. Prior work discussed on dictionary-aided translation for low-resource languages then has implications for understanding unknown-scripts.

Unknown, Hoax-ed, or Down-Right Synthetic? Acquired by Wilfred Michael Voynich from a Jesuit Villa near Rome, the now famously named "Voynich manuscript" has puzzled cryptographers, some viewing it as no more than a hoax. Layfield et al. sees potential in its linguistic value. Using statistical context-free probabilities coupled with what they call a *uniqueness point* of each word, they find the regression pattern to strongly match the behavior of common languages like English, Korean, Hebrew, etc . . . These findings suggest the manuscript to be written in a natural language–or one that is synthetically designed carefully enough to mimic the intuitions behind the grammar and vocabulary base of other natural languages.

Though not the aim of the paper, they introduce a method by which a candidate unknown script or constructed language can be said to be "*natural*."

4 Constructed Languages

The development of constructed languages in the NLP space has proved useful in two regards. First, it has helped us understand the bounds of language; where a language fits typologically if it does at all. Second, synthetic generation of languages that are "human-like" can compensate in low-resource or unknown script settings.

4.1 The Continuum of Languages

A popular view in the NLP space is that LLMs are capable of learning any language and thus cannot be seen as informative tools for linguistic understanding (Chomsky, 2023). This has prompted definitions on the realms of language spanning from possible to impossible. GPT-2 trained on a series of impossible languages via English permutations of SHUFFLE, REVERSE, or WORDHOP reports higher perplexities over a baseline English (Kallini et al., 2024). Researchers view this a clear evidence LLMs are not equally capable of understanding all languages and thus are aware of some linguistic nuance. However, this research has faced criticism, specifically from Hunter who argue

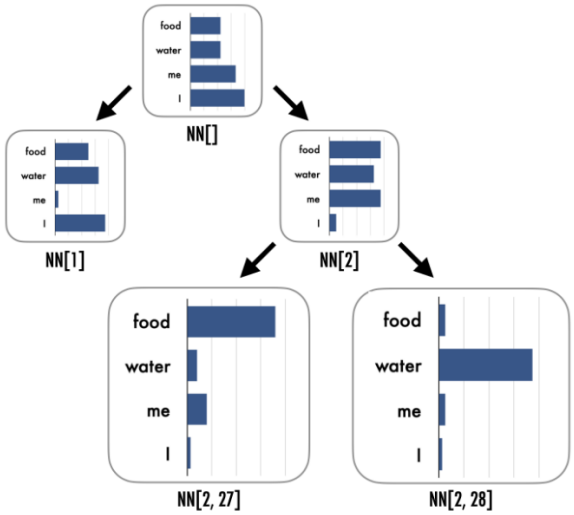

Figure 4: **An example hierarchical Pitman-Yor process**. From Towards More Natural Artificial Languages (Hopkins, 2022).

impossible-languages should not be considered in the conversation of constituency-based languages as they do not fall in the confines of possible–and thus productive-to-study–languages.

The question on what LLMs internalize or prefer in language has opened research into investigating a model's linguistic biases. When asked to conjugate synthetic verbs presented in training data, GPT-2 performed poorly as the number of distinct verb classes grew suggesting a preference for simplicity (Tosolini and Blevins, 2025). Additionally, presented unseen verbs in one-shot settings accuracy of conjugations tanked suggesting the model to be memorizing a linguistic understanding rather than generalizing from it.

To match the natural structure of languages, researchers have experimented with context-free grammars (CFG) which govern sentence validity. One such approach follows **selectional preference** using dependency bigrams to enforce head-dependent relationships (Hopkins, 2022). The Pitman-Yor process, visualized in Figure 4, demonstrates how this can be done through linked distributions between children and parent tokens.

4.2 Playing by the Linguistic Rules

Construction of a new language is no smaller a task even with a clean understanding on how a language should be designed. Esperanto, originally developed by L. L. Zamenhof in 1887, is an pinnacle example of a successful naturally oriented conlang. In school settings, synthetic Esperanto sentences

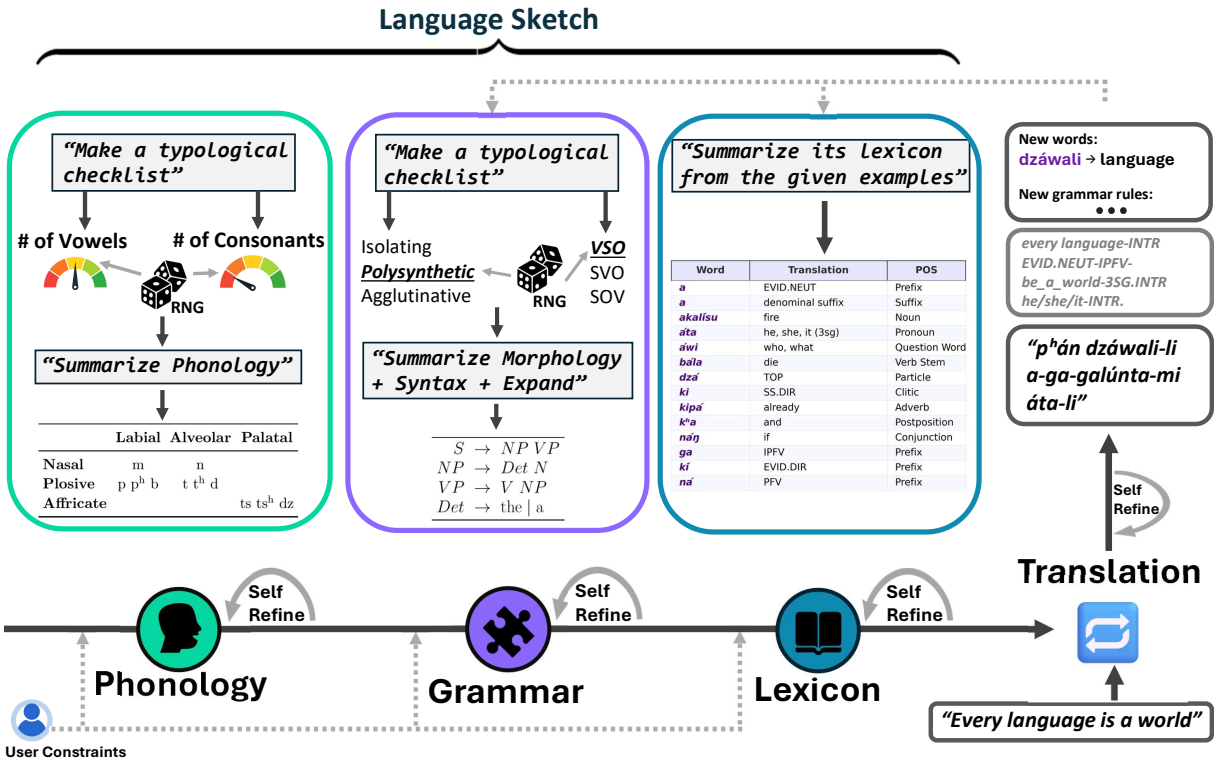

Figure 5: The ConlangCrafter pipeline for modular constructed language generation, decomposed into sequential stages of phonology, grammar, and lexicon construction with self-refinement at each step, culminating in translation. From Alper et al. (2026).

have been generated following a valency framework with their coherence graded sentences by a weighted combination of their bigram, trigram, and **depgram** probabilities (Bick, 2019).

The Galactic Dependencies Treebank (Wang and Eisner, 2016) is a modern project that has developed nearly 50,000 synthetic languages. Starting from a **superstrate language**'s dependency tree found in the Universal Dependencies framework, a permutation of the parts-of-speech distribution generates a new language. In "single-source transfer" for dependency parsing of low-resource languages, their "unearthly" languages prove useful once added to the source pool of languages[3].

Consistency of a Language Alternatively, LLMs have been used to generate synthetic languages. As LLMs are prone to hallucination, this presents issues of self-consistency in the construction of a conlang. Iterative prompts have helped ensure a degree of self-consistency by not over-extending the context window (Taguchi and Sproat, 2026). *ConlangCrafter* described in Figure 5, instead, leverages the hallucination as a strength generating languages that are diverse yet maintain consistent translation quality through a process of self-refinement from a sub-critic and sub-editor model. (Alper et al., 2026). The capability to generate consistent language from minimal description suggests LLMs' linguistic understanding can aid translation from following an annotated set of transformations that take a high-resource language to the target low-resource language.

5 Concluding Remarks

The diversity of languages is a feature that arguably makes linguistic study attractive. Historically, however, NLP as a field has been dedicated to a select few languages biasing methods to how higher-resource languages function. In this paper, we have discussed the waves of research that has pushed visibility of **impoverished languages** and structured how common-practice methods can be adapted or made more conscious to the stylistic differences in low-resource, unknown, or constructed settings.

Acknowledgments

This subfield survey was written for assignment as part of U.C. Berkeley's INFO 159, Natural Language Processing, Spring 2026 semester under the

[3]This effect is pronounced with the when the language is *natural*, stressing the importance of linguistic diversity to a model's training set.

instruction and guidance of Professor David Bamman. My synthesis of the subfield of what I define as "impoverished languages' is not intended to appear in any publication. Nor is this paper intended to described all papers in the space but those that are most definitive to the discussion to equip a reader with adequate knowledge for future readings and contribution to the field.

References

Muhammad Farid Adilazuarda, Musa Izzanardi Wijanarko, Lucky Susanto, Khumaisa Nur'aini, Derry Tanti Wijaya, and Alham Fikri Aji. 2025. NusaAksara: A multimodal and multilingual benchmark for preserving Indonesian indigenous scripts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28371–28401, Vienna, Austria. Association for Computational Linguistics.

Nada Aldarrab and Jonathan May. 2021. Can sequence-to-sequence models crack substitution ciphers? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7226–7235, Online. Association for Computational Linguistics.

Morris Alper, Moran Yanuka, Raja Giryes, and Gašper Beguš. 2026. Conlangcrafter: Constructing languages with a multi-hop llm pipeline. *Preprint*, arXiv:2508.06094.

Eckhard Bick. 2019. Automatic generation and semantic grading of Esperanto sentences in a teaching context. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, pages 10–19, Turku, Finland. LiU Electronic Press.

Logan Born, M. Willis Monroe, Kathryn Kelley, and Anoop Sarkar. 2023. Learning the character inventories of undeciphered scripts using unsupervised deep clustering. In *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*, pages 92–104, Toronto, Canada. Association for Computational Linguistics.

Emily Chang and Nada Basit. 2025. How many words does it take to understand a low-resource language? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 207–224, Albuquerque, USA. Association for Computational Linguistics.

Noam Chomsky. 2023. Noam Chomsky on Language, Left Libertarianism, and Progress. Conversations with Tyler, Episode 182: Noam Chomsky.

Antonios Dimakis, Stella Markantonatou, and Antonios Anastasopoulos. 2024. Dictionary-aided translation for handling multi-word expressions in low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2588–2595, Bangkok, Thailand. Association for Computational Linguistics.

Micha Elsner and Jordan Needle. 2023. Translating a low-resource language using GPT-3 and a human-readable dictionary. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–13, Toronto, Canada. Association for Computational Linguistics.

Martin Grandjean. 2014. Henri bergson et les paradoxes de Zénon : Achille battu par la tortue ? Blog post. Licensed under CC BY 3.0 CH.

Mark Hopkins. 2022. Towards more natural artificial languages. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 85–94, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tim Hunter. 2025. Kallini et al. (2024) do not compare impossible languages with constituency-based ones. *Computational Linguistics*, 51:641–650.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14691–14714, Bangkok, Thailand. Association for Computational Linguistics.

Kevin Knight and Kenji Yamada. 1999. A computational approach to deciphering unknown scripts. In *Unsupervised Learning in Natural Language Processing*.

Colin Layfield, Lonneke van der Plas, Michael Rosner, and John Abela. 2020. Word probability findings in the Voynich manuscript. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 74–78, Marseille, France. European Language Resources Association (ELRA).

Jiaming Luo, Yuan Cao, and Regina Barzilay. 2019. Neural decipherment via minimum-cost flow: From Ugaritic to Linear B. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3146–3155, Florence, Italy. Association for Computational Linguistics.

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The Zeno's paradox of 'low-resource' languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Yu-Fei Shih, Zheng-Lin Lin, and Shu-Kai Hsieh. 2025. Reasoning over the glyphs: Evaluation of llm's decipherment of rare scripts. *Preprint*, arXiv:2501.17785.

Chihiro Taguchi and Richard Sproat. 2026. Creating conlangs to probe the metalinguistic grammatical knowledge of llms. *Preprint*, arXiv:2510.07591.

Fabio Tamburini. 2023. Decipherment of lost ancient scripts as combinatorial optimisation using coupled simulated annealing. In *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*, pages 82–91, Toronto, Canada. Association for Computational Linguistics.

Hailay Kidu Teklehaymanot, Dren Fazlija, and Wolfgang Nejdl. 2025. MoVoC: Morphology-aware subword construction for Ge'ez script languages. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13131–13144, Suzhou, China. Association for Computational Linguistics.

Alessio Tosolini and Terra Blevins. 2025. Analyzing the linguistic priors of language models with synthetic languages. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 7–15, Vienna, Austria. Association for Computational Linguistics.

Wondimagegnhue Tufa, Ilia Markov, and Piek Vossen. 2024. Unknown script: Impact of script on cross-lingual transfer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 124–129, Mexico City, Mexico. Association for Computational Linguistics.

Dingquan Wang and Jason Eisner. 2016. The galactic dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4:491–505.

Chen Zhang, Jiuheng Lin, Xiao Liu, Zekai Zhang, and Yansong Feng. 2025. Read it in two steps: Translating extremely low-resource languages with code-augmented grammar books. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3977–3997, Vienna, Austria. Association for Computational Linguistics.