High Resource Bias in AI-Driven Neology: Structural Inequality in Lexical Innovation

Wajdi Zaghouani
Northwestern University in Qatar
wajdi.zaghouani@northwestern.edu

Abstract

Large language models (LLMs) are increasingly deployed to detect, generate, and normalize neologisms across languages. While prior work has examined their capacity to model semantic change and handle temporal drift, insufficient attention has been paid to how training data asymmetries interact with probabilistic generation mechanisms to structure lexical innovation itself. This position paper argues that AI-driven neology is shaped by systematic high resource bias that privileges dominant languages in the production, stabilization, and dissemination of new lexical items. Drawing on sociolinguistics, language political economy, lexicography, and computational modeling theory, we formalize how distributional imbalance alters innovation likelihood across languages. We introduce a taxonomy of bias types specific to AI-mediated neology, present a probabilistic account of generative reinforcement loops, and illustrate these mechanisms using documented examples from English-Arabic and English-Icelandic language pairs. We derive empirically testable predictions, outline concrete experimental protocols for their validation, and propose mitigation strategies for lexicographers, language planners, and NLP researchers.

Keywords: neology, large language models, linguistic bias, lexical innovation, high-resource languages, low-resource languages, language resources equity

1. Introduction

Neologisms emerge within speech communities through innovation, uptake, and stabilization. Historically, lexical change has been modeled as a socially distributed process observable through corpus frequency and contextual shift (Lejeune and Cartier, 2017). With the rise of large language models (LLMs), a new infrastructural actor enters this process: generative systems capable of producing plausible lexical forms largely independent of direct community grounding. These systems do not merely reflect existing language use; they actively shape the probability landscape in which new words are formed, evaluated, and propagated.

Contemporary LLMs are trained on corpora in which English and a handful of dominant languages are massively overrepresented (Joshi et al., 2020). This imbalance is not merely representational; it fundamentally reshapes the geometry of embedding spaces, the density of contextual neighborhoods, and the probability mass assigned to candidate lexical forms. As a result, generative outputs systematically favor patterns from high resource languages, creating structural asymmetries in lexical innovation that extend beyond simple performance degradation in low resource settings.

Recent empirical work has begun to document these effects. Zheng et al. (2024) demonstrate that even a single neologism can reduce machine translation quality by up to 43%, with effects more pronounced for words of non-English origin. Ármannsson et al. (2025) show reduced accuracy in morphological well-formedness judgments for Icelandic compared to English baselines. The comprehensive survey by Al-Khalifa et al. (2025) documents persistent preferences for English-derived transliterations over indigenous Arabic derivations in technical domains.

This paper advances three central claims:

1. AI-driven neology is structured by global inequalities in linguistic capital (Bourdieu, 1991).

2. Generative architectures amplify innovation originating in dominant languages while marginalizing or normalizing innovation in low resource contexts.

3. Without corrective mechanisms, LLM-integrated lexicographic practice risks reinforcing structural linguistic inequality on a global scale.

Our contribution is explicitly a position paper: it is theoretical and analytical in nature. We formalize mechanisms and derive testable predictions rather than presenting new benchmarking experiments, complementing empirical work such as NEO-BENCH (Zheng et al., 2024), the Icelandic linguistic benchmark (Ármannsson et al., 2025), and recent surveys of Arabic LLMs (Al-Khalifa et al., 2025). By focusing specifically on lexical innovation, we extend broader discussions of LLM bias (Navigli et al., 2023) to the domain of language change and resource equity. Importantly, we also outline concrete experimental protocols that would enable future empirical validation of our claims, responding to the need for actionable research directions in this emerging area.

Proceedings of NEOLOGY AND LARGE LANGUAGE MODELS, pages 16–26

2. Conceptual Delimitation

2.1. From Data Imbalance to Innovation Asymmetry

Data imbalance across languages has been widely documented in the NLP literature (Joshi et al., 2020; Navigli et al., 2023). However, the present argument concerns a distinct and previously under-theorized phenomenon: *innovation likelihood asymmetry*. Most existing discussions of imbalance focus on degraded performance in low resource languages, such as reduced translation quality or higher perplexity when encountering rare forms. Here, the focus shifts to generative dynamics.

Neology involves modeling productive morphological processes, semantic extension, compounding creativity, and lexical blending. These processes depend critically on dense distributional representations. The relevant question is therefore not only whether a language is underrepresented in training data, but whether the density of its contextual embedding space supports probabilistically plausible lexical innovation. High resource bias in neology is not reducible to general data imbalance; it concerns how imbalance actively restructures the innovation space itself.

This distinction is crucial because even morphologically rich low resource languages may exhibit suppressed indigenous creativity when mediated by current LLMs. Icelandic, with its productive compounding system, and Arabic, with its root-and-pattern morphology, both show evidence of this suppression despite their structural complexity (Ármannsson et al., 2025; Al-Khalifa et al., 2025; Wiemerslage et al., 2022).

2.2. Borrowing Versus Algorithmic Amplification

Borrowing is a natural and historically ubiquitous linguistic process. Languages routinely adopt foreign lexical material in domains of technological and cultural change. The argument here does not pathologize borrowing. Instead, it distinguishes between *organic, contact-driven borrowing* shaped by sociocultural interaction and *algorithmically amplified borrowing* driven by probabilistic reinforcement within digital infrastructures.

The issue is one of structural acceleration and asymmetry. When generative systems systematically increase the visibility and probability of dominant-language innovations, they may distort the ecological balance between borrowing and indigenous derivation. Recent lexicographic analyses document this effect in real-world dictionary compilation pipelines (Poix and Shevchenko, 2025).

2.3. LLMs as Infrastructural Mediators

This paper does not claim that LLMs autonomously create language change. Human communities remain the ultimate agents of stabilization and uptake. However, LLMs function as powerful infrastructural mediators within contemporary socio-technical networks. Gillespie (2014) argues that algorithms embedded in digital platforms increasingly determine what information is considered relevant, shaping participation in public life through procedural logics. Algorithmic systems shape visibility, salience, and circulation of linguistic forms. In generative contexts, they additionally influence which lexical candidates are more likely to be produced and repeated at scale.

As Periti and Montanelli (2024) observe in their survey of lexical semantic change through LLMs, these models fundamentally alter how we can detect, interpret, and assess meaning change over time. LLMs participate in language change not as originators but as amplifiers and redistributors of innovation probability (Navigli et al., 2023).

3. Related Work

Research on bias in large language models has grown rapidly, primarily addressing social stereotypes, toxicity, and performance disparities across demographic groups (Navigli et al., 2023; Gallegos et al., 2024). A foundational contribution by Joshi et al. (2020) documented severe underrepresentation of the majority of the world's languages in both NLP corpora and conference publications, establishing the data imbalance that underlies the present argument. Their taxonomy classified languages into six resource categories, with the vast majority falling into the lowest tiers. This imbalance reflects broader patterns of linguistic hierarchy that sociolinguists have long documented (De Swaan, 2001; Blommaert, 2010).

The political economy of large-scale language modeling has attracted increasing critical attention. Bender et al. (2021), in their influential analysis of the risks associated with ever-larger language models, highlight how training data sourced predominantly from the web systematically underrepresents marginalized communities and linguistic minorities. Conneau et al. (2020) demonstrate that while cross-lingual transfer learning can benefit low resource languages, trade-offs emerge between positive transfer and capacity dilution as model coverage expands. Similarly, Xue et al. (2021) show that massively multilingual models exhibit significant performance disparities across languages despite their broad coverage.

Subsequent empirical studies have quantified how this imbalance propagates to model behavior. Zheng et al. (2024) introduced NEO-BENCH,

a benchmark specifically designed to test LLM robustness to neologisms. Their results show that model performance is nearly halved in machine translation when a single neologism is introduced. Critically, they found that LLMs are affected differently based on the linguistic origins of words, with non-English neologisms posing greater challenges.

For morphologically complex low resource languages, Ármannsson et al. (2025) created the first manually curated linguistic benchmark for Icelandic LLMs. Native-speaker evaluation revealed markedly reduced accuracy in well-formedness judgments and morphological productivity tests compared with English baselines. Similarly, the comprehensive survey by Al-Khalifa et al. (2025) of Arabic LLMs highlights persistent challenges in handling Arabic's rich morphological system and a preference for English-derived forms in technical domains.

The Arabic case is particularly well documented in terms of resource availability. Surveys of freely available Arabic corpora have repeatedly demonstrated the imbalance between the language's massive speaker population and its comparatively limited digital resource base (Zaghouani, 2014). While substantial efforts have been made to build dialectal resources across multiple Arab countries (Zaghouani and Charfi, 2018; Bouamor et al., 2018; Charfi et al., 2019), these remain small relative to English-language resources and are concentrated in specific domains such as social media and news, leaving technical and scientific domains particularly underrepresented. This gap in domain coverage is directly relevant to our argument about neological innovation, as it is precisely in technical domains that new terminology emerges most actively.

Theoretical work on morphological productivity provides essential grounding for understanding these patterns. Bauer (2001) offers a comprehensive treatment of productivity measurement, emphasizing the scalar nature of morphological processes and the role of frequency in determining productive potential. Bybee (1995) establishes theoretical connections between token frequency and morphological representation that inform our formalization. Hamilton et al. (2016) demonstrate that statistical laws govern semantic change, with word frequency playing a key role in determining rates of meaning evolution, findings directly relevant to modeling innovation probability.

Lexicographic perspectives on AI-generated language have emerged only recently. Poix and Shevchenko (2025), in their eLex 2025 contribution, explicitly discuss the challenge of distinguishing organically occurring neologisms from synthetic LLM outputs in corpus data. They warn that AI-generated text may artificially inflate hapax legomena and distort diachronic frequency trends, raising urgent questions about the authenticity of corpus-based lexicographic evidence.

Complementary work on morphological processing by Wiemerslage et al. (2022) demonstrates that unsupervised paradigm completion for low resource languages remains fundamentally limited by sparse training signals. This limitation directly affects the capacity of LLMs to model productive morphological innovation in these languages.

The sociotechnical dynamics of algorithmic mediation have been theorized by Gillespie (2014), who argues that algorithms function as relevance-determining systems that shape public knowledge and participation. This framework illuminates how LLMs, as generative algorithms, may restructure the landscape of lexical innovation by privileging certain forms over others.

No prior publication has synthesized these threads into a unified formal account of high resource bias specifically targeting lexical innovation. The present paper fills this gap while remaining grounded in verified, peer-reviewed findings.

4. Theoretical Foundations

4.1. Neology and Lexicalization

Neologisms emerge through five primary mechanisms: morphological derivation, compounding, semantic shift, blending, and borrowing. The productivity of these mechanisms varies both synchronically and diachronically (Bauer, 2001). Successful lexicalization further requires sustained frequency growth, semantic stabilization, and eventual institutional recognition (Lejeune and Cartier, 2017). Traditional corpus linguistics treats frequency trajectories as direct evidence of community uptake, a relationship now formalized through diachronic word embeddings that reveal statistical laws governing semantic change (Hamilton et al., 2016).

In AI-mediated environments, however, synthetic generation fundamentally complicates this evidentiary basis. LLM outputs can rapidly create the appearance of frequency without corresponding human adoption (Zheng et al., 2024; Poix and Shevchenko, 2025). This raises urgent questions for lexicographers and language resource curators: How can we distinguish genuine community innovation from algorithmically amplified forms? What new methodologies are needed to track authentic lexical change in corpora increasingly contaminated by AI-generated text?

4.2. Linguistic Capital and Global Hierarchy

Language operates within a global hierarchy of symbolic power (Bourdieu, 1991). De Swaan (2001)

formalizes this hierarchy as a system in which languages occupy positions ranging from peripheral to supercentral, with English functioning as the hypercentral language connecting all others. Dominant languages accumulate institutional infrastructure, technological embedding, and cultural capital. Digital textual production mirrors and amplifies this hierarchy, a pattern that Blommaert (2010) characterizes as linguistic stratification within globalized communication systems.

Training corpora for today's LLMs are heavily skewed toward English and a small set of other high resource languages (Joshi et al., 2020). Bender et al. (2021) argue that such scale-driven approaches systematically underrepresent the linguistic diversity of the world's population, with consequences for both equity and quality. Cross-lingual representation learning demonstrates clear tradeoffs between positive transfer and capacity dilution as the number of languages increases (Conneau et al., 2020). This asymmetry is not static; it becomes operationalized in probabilistic generation. As Navigli et al. (2023) document, data selection bias in training corpora cascades into multiple forms of social and linguistic bias in model outputs. For neology specifically, this creates self-reinforcing loops that systematically disadvantage lexical creativity in low resource contexts.

4.3. Distributional Semantics and Innovation Space

The capacity of LLMs to model productive word formation depends on the density and quality of distributional representations. In high resource languages, dense contextual neighborhoods enable robust generalization to novel forms. Models can accurately predict which morphological combinations are plausible, which semantic extensions are natural, and which compounds are well-formed.

In low resource languages, sparse representations constrain these capacities. As Wiemerslage et al. (2022) demonstrate, morphological processing quality correlates strongly with training data availability. The implication for neology is that even when low resource languages possess rich productive morphological systems, LLMs may fail to model their creative potential accurately.

5. Formalizing High Resource Bias

5.1. Setup

Let L denote a language, D_L the effective training corpus size in tokens, and $P(w \mid c, L)$ the conditional token probability given context c. For high resource languages, $D_{\text{high}} \gg D_{\text{low}}$. This disparity yields more accurate estimation of conditional probabilities, denser contextual neighborhoods, and more robust modeling of morphological productivity.

5.2. Innovation Probability

Let n be a candidate neologism constructed via productive morphological processes. In a generative model,

$$P(n \mid c, L) \propto \exp(f_\theta(n, c, L))$$

where f_θ is the learned scoring function. For structurally parallel innovations across languages,

$$\mathbb{E}[P(n \mid c, L_{\text{high}})] > \mathbb{E}[P(n \mid c, L_{\text{low}})]$$

because subword representations are better optimized, productive patterns are observed at higher frequency, and contextual embeddings exhibit substantially lower uncertainty.

We note that this formalization is intended as an illustrative abstraction rather than a validated model. Its purpose is to provide a structured framework for generating testable hypotheses, which we detail in Section 8. The mathematical formulation captures the core intuition that data asymmetry translates into innovation asymmetry, and it is deliberately kept simple to highlight this relationship clearly rather than to model all relevant variables.

5.3. Morphological Productivity

Let M_L represent the modeled productivity of morphological transformations. Following Bauer (2001) and Bybee (1995) on the relationship between frequency and morphological productivity, we hypothesize:

$$M_L \propto \log(D_L)$$

As corpus size increases, the model's capacity to generalize productive transformations grows nonlinearly. Consequently, low resource languages exhibit more conservative generation behavior, reduced rates of indigenous derivation, and greater reliance on high-frequency borrowed tokens. This pattern is consistently observed in both Icelandic compounding (Ármannsson et al., 2025) and Arabic morphological systems (Al-Khalifa et al., 2025).

5.4. Generative Reinforcement Dynamics

Let $P_t(n, L)$ denote the probability of generating neologism n at time t. If n originates in a high resource language,

$$P_{t+1}(n, L_{\text{high}}) = P_t(n, L_{\text{high}}) + \alpha \cdot \text{Exposure}_t$$

Through global digital circulation, the form gains visibility. In low resource languages the update becomes

$$P_{t+1}(n, L_{\text{low}}) = P_t(n, L_{\text{low}}) + \beta \cdot \text{TranslationExposure}_t$$

where typically $\beta > \alpha$ for borrowed forms due to their higher baseline probability in the model.

This produces a feedback loop: high resource innovation $\rightarrow$ AI generation $\rightarrow$ digital uptake $\rightarrow$ corpus reintegration $\rightarrow$ increased generation probability. Such loops accelerate linguistic homogenization, as documented in broader analyses of LLM bias propagation (Navigli et al., 2023).

6. Taxonomy of Bias Types

We identify four interlocking bias types specific to AI-mediated neology:

Type 1: Distributional Bias. Unequal modeling quality resulting from corpus size disparities produces sparser representations for low resource languages (Joshi et al., 2020). This bias affects the foundational capacity to represent and manipulate lexical forms.

Type 2: Generative Amplification Bias. Disproportionate reproduction and probability boosting of dominant-language innovations during generation (Zheng et al., 2024). High resource neologisms receive higher generation probabilities even when low resource alternatives exist.

Type 3: Translational Normalization Bias. Flattening of indigenous semantic nuance when LLMs default to high resource lexical templates during translation or cross-lingual tasks. This is particularly evident in Arabic technical neology, where transliteration often supersedes productive root-and-pattern derivation (Al-Khalifa et al., 2025).

Type 4: Institutional Adoption Bias. Faster validation and lexicographic acceptance of high-visibility generative forms. AI-amplified neologisms may achieve apparent frequency thresholds for dictionary inclusion more rapidly, complicating the detection of organic innovation (Poix and Shevchenko, 2025).

These biases interact multiplicatively, producing compound effects on global lexical ecosystems. A neologism disadvantaged by distributional bias will also suffer reduced generative amplification, face stronger normalization pressure toward dominant-language equivalents, and experience slower institutional recognition.

7. Illustrative Case Studies

Table 1 summarizes selected neologisms drawn from published benchmarks. All observations are based on empirical results reported in the cited literature. While we do not introduce new experimental analyses here, these examples serve to ground our theoretical framework in documented findings, illustrating how the bias types identified in Section 6 manifest in practice across different language pairs.

7.1. Cross-Linguistic Patterns

Beyond the specific Arabic and Icelandic cases, broader patterns emerge from the empirical literature. Zheng et al. (2024) note that neologisms of different linguistic origins pose varying challenges: words borrowed into English from other languages (such as *pig butchering* from Mandarin) show compartmentalized understanding, while native English formations are more robustly represented. This asymmetry suggests that even within high resource English, the provenance of neologisms matters.

The COVID-19 pandemic provided a natural experiment in cross-linguistic neological dynamics. Technical terms like *coronavirus*, *lockdown*, and *social distancing* required rapid adaptation across languages. Observations suggest that high resource languages integrated these terms quickly and diversely, generating multiple synonyms and stylistic variants. Low resource languages, by contrast, showed slower integration and greater reliance on direct borrowing rather than calquing or indigenous derivation.

These patterns support our central claim: the generative dynamics of LLMs systematically favor high resource language innovation while constraining creativity in low resource contexts. The effects compound across the taxonomy we propose: distributional bias creates unequal starting conditions, generative amplification widens the gap, translational normalization flattens alternatives, and institutional adoption bias cements the outcomes.

7.2. English-Arabic Case Study

The English blend *doomscrolling* is densely represented in training data. NEO-BENCH demonstrates strong performance on definition tasks in high resource settings but dramatic degradation in machine translation (Zheng et al., 2024). Arabic exhibits a rich system of root-and-pattern morphology that supports productive technical neology. However, in generative outputs, transliteration or descriptive calques consistently predominate over productive derivation (Al-Khalifa et al., 2025).

Arabic's morphological system offers multiple productive mechanisms for neological derivation. The root-and-pattern system allows creation of new words through established templates: for instance, the root *k-t-b* (related to writing) generates *kitāb* (book), *kātib* (writer), *maktaba* (library), and *maktūb* (written). This system could theoretically accommodate technical neologisms through analogical extension. Similarly, Arabic possesses productive compounding mechanisms and established patterns for arabicization of foreign terms.

Despite these resources, surveys of Arabic LLMs reveal systematic preferences for transliteration (Al-Khalifa et al., 2025; Darwish et al., 2021). Technical

Neologism	Formation Type	Resource Context	Observed LLM Behavior
doomscrolling	Morphological blend	High (English)	Lower perplexity; strong definition generation; MT quality drops 43% when introduced as unknown form
pig butchering	Semantic calque (from Mandarin)	High via English	Compartmentalized knowledge; literal translations predominate over idiomatic rendering
stablecoin	Technical compound	High (English)	Accurate definition generation; successful cross-lingual transfer to related high resource languages
Icelandic compounds (e.g., *sýkingarþreyta*)	Productive compounding	Low (Icelandic)	Reduced accuracy in well-formedness judgments; lower Wug-test performance vs. English baselines
Arabic technical terms (e.g., *metaverse* equivalents)	Translational borrowing	Low (Arabic)	Strong preference for transliteration over indigenous root-and-pattern derivation
COVID-related neologisms	Multi-type	Variable	High resource languages show rapid integration; low resource languages show delayed and less diverse adaptation

Table 1: Selected neologisms illustrating high resource bias patterns. All entries are derived from empirical findings in cited published benchmarks and surveys.

terms like *internet*, *computer*, and emerging vocabulary such as *metaverse* are frequently rendered as phonetic borrowings rather than morphologically integrated forms. This pattern reflects the higher prior probability assigned to borrowed forms in training data. Even when Arabic language academies have proposed indigenous alternatives, the distributional dominance of English in training corpora biases outputs toward transliteration.

The problem is compounded by the fact that existing Arabic corpora, while growing in volume, remain concentrated in certain domains and registers. Surveys of freely available Arabic corpora have documented persistent gaps in technical and scientific writing (Zaghouani, 2014), and while large-scale dialectal corpora now cover social media registers across multiple Arab countries (Zaghouani and Charfi, 2018; Bouamor et al., 2018; Charfi et al., 2019), technical neology remains poorly represented. This domain mismatch means that LLM training data for Arabic overrepresents informal registers where borrowing is already prevalent, further reinforcing the preference for transliterated forms over indigenous derivations.

The effect creates a self-reinforcing cycle: borrowed forms dominate corpora, models learn to prefer borrowed forms, generated text contains more borrowed forms, and future training corpora inherit this bias. This dynamic threatens the productivity of Arabic's morphological system in precisely the domains, such as technology and science, where neological creativity is most needed.

7.3. English-Icelandic Case Study

Icelandic language planning has long promoted indigenous coinages through institutions such as the Árni Magnússon Institute. The language possesses extraordinarily productive compounding and derivational systems. Icelandic has historically coined native terms for modern concepts: *sími* (telephone, from an old word for thread), *tölva* (computer, from *tala* 'number' + *völva* 'prophetess'), and *sjónvarp* (television, literally 'vision-throw'). This tradition reflects deliberate language policy aimed at maintaining linguistic purity and ensuring that Icelandic remains fully functional for expressing modern concepts.

Yet LLM outputs in hybrid prompts frequently default to English technical terms or hybrid forms (Ármannsson et al., 2025). The benchmark created by these researchers specifically tests morphological productivity through tasks including well-formedness judgments, Wug tests (requiring generation of novel inflected forms), and compound interpretation. Results show that state-of-the-art models perform significantly worse on Icelandic morphological tasks compared to structurally analogous tasks in English.

This disparity is particularly striking given Icelandic's morphological regularity. The language's inflectional system, while complex, follows highly predictable patterns that should, in principle, be learnable from sufficient data. The performance gap therefore reflects not inherent difficulty but training data distribution. With approximately 350,000 native speakers, Icelandic is dwarfed in corpus representation by English's billions of speakers and massive digital footprint.

The Icelandic case reveals a fundamental ten-

sion between probabilistic modeling and institutional language policy. When LLMs consistently suggest English borrowings over indigenous Icelandic compounds, they work against decades of careful language planning. Users interacting with AI systems may increasingly encounter and adopt these borrowed forms, potentially undermining the ecosystem of indigenous neological creativity that language planners have cultivated. This represents a concrete mechanism by which AI systems may accelerate language shift even in communities with strong institutional support for linguistic maintenance.

8. Empirically Testable Predictions

Based on our formalization, we derive three specific predictions amenable to empirical validation. For each prediction, we outline a concrete experimental protocol that would enable systematic testing, responding to the need for actionable research designs that can move the field from theoretical argument to empirical investigation.

Prediction 1: Generative Diversity Hypothesis. Under symmetric prompting conditions, high resource languages will exhibit significantly higher rates of indigenous morphological innovation than low resource languages. This can be tested via controlled generation experiments extending the methodology of Zheng et al. (2024), comparing neologism generation rates across typologically similar language pairs with different resource levels.

Proposed protocol: Design a set of parallel prompts in matched language pairs (e.g., English vs. Icelandic, English vs. Arabic) that elicit neologism generation for identical novel concepts. Using multiple LLMs (both proprietary and open-weight), collect at least 100 generated responses per language per model. Annotate each generated neologism for formation type (indigenous derivation, compounding, borrowing, calque, transliteration) using trained native speaker annotators. Compute the ratio of indigenous formations to borrowed forms across languages and test for statistically significant differences using appropriate non-parametric tests given the expected non-normal distributions.

Prediction 2: Borrowing Amplification Hypothesis. The probability of borrowed forms in low resource languages will increase measurably following global exposure of dominant-language neologisms. This is testable via temporal corpus analysis comparing borrowing rates before and after major LLM deployment waves, particularly for technical vocabulary domains.

Proposed protocol: Construct time-stamped corpora for Arabic and Icelandic technical writing spanning two periods: pre-ChatGPT (2018-2022) and post-ChatGPT (2023-2026). For each period, extract neologisms related to technology, AI, and digital culture. Measure the proportion of borrowings versus indigenous formations in each period. Control for the natural increase in borrowing by comparing rates in domains where LLM-generated text is prevalent (e.g., online content) versus domains where it is rare (e.g., print publications, academic writing). A significant increase in borrowing rates disproportionately concentrated in LLM-saturated domains would support this prediction.

Prediction 3: Morphological Suppression Hypothesis. AI outputs for morphologically rich low resource languages will show lower morphological novelty compared with matched human corpora. This prediction is directly testable against the benchmarks established by Ármannsson et al. (2025) for Icelandic and the Arabic evaluation frameworks surveyed by Al-Khalifa et al. (2025).

Proposed protocol: Compile a corpus of LLM-generated text and a matched corpus of human-authored text in Arabic and Icelandic across the same domains and time periods. Measure morphological diversity using type-token ratio of morphological patterns, hapax legomena rates for derivational and compound formations, and the proportion of productive use of native morphological templates. Compare these metrics between LLM-generated and human-authored subcorpora, testing whether LLM text shows significantly reduced morphological novelty. This approach builds directly on the morphological analysis tools used in the Icelandic benchmark (Ármannsson et al., 2025) and can leverage existing Arabic morphological analyzers such as those surveyed in Darwish et al. (2021).

9. Implications and Mitigation

9.1. For Lexicography

AI-assisted corpus monitoring tools must incorporate mechanisms to distinguish organic uptake from synthetic amplification. Lexicographic workflows should adopt generative provenance tracking, flagging items that may have entered corpora through AI generation rather than community usage (Poix and Shevchenko, 2025). This may require new metadata standards for corpus annotation and revised criteria for dictionary inclusion.

9.2. For Language Policy

Language planning institutions must explicitly account for algorithmic reinforcement of borrowing. Organizations such as the Árni Magnússon Institute and Arabic language academies face new challenges in promoting indigenous terminology when probabilistic systems systematically favor borrowed forms. Promising mitigation avenues include equity-

aware fine-tuning and retrieval-augmented generation grounded in carefully curated local corpora.

9.3. For NLP Research

Evaluation metrics should incorporate cross-lingual innovation parity rather than focusing solely on aggregate performance benchmarks (Joshi et al., 2020). Concrete mitigation strategies include:

1. Balanced multilingual pre-training with explicit low resource upsampling

2. Morphology-aware tokenization schemes tailored to low resource languages (Wiemerslage et al., 2022)

3. Community-in-the-loop validation pipelines for neologism detection

4. Development of neology-specific benchmarks for low resource languages extending existing dialectal and multi-genre corpus efforts (Bouamor et al., 2018; Charfi et al., 2019)

LREC is ideally positioned to lead by developing multilingual neology resources that explicitly tag generative versus human provenance.

10. Discussion

The mechanisms formalized in this paper suggest that current LLM architectures do not merely reflect existing linguistic inequalities; they actively accelerate them within digital ecosystems. Over time, this may lead to reduced lexical diversity worldwide, with low resource languages increasingly functioning as recipients rather than co-creators of neological innovation.

The Icelandic and Arabic case studies illustrate how even languages with strong institutional support and rich morphological systems remain vulnerable. Icelandic, with its centuries-long tradition of linguistic purism and active language planning, faces pressure from AI systems that consistently prefer English borrowings. Arabic, with over 400 million speakers and a morphological system of remarkable productivity, sees its derivational potential underutilized as models default to transliteration.

10.1. Broader Consequences for Linguistic Diversity

The implications extend beyond academic concern: lexical innovation is a core mechanism of cultural expression and adaptation. Languages evolve through their speakers' creative responses to new experiences, technologies, and social configurations. When AI systems systematically suppress indigenous creativity while amplifying borrowed forms, they may contribute to broader processes of cultural homogenization, a dynamic that parallels historical patterns of language endangerment (Crystal, 2000).

Consider the domain of technology, where neological activity is most intense. If speakers of low resource languages consistently encounter AI-generated text that favors English borrowings, they may internalize these preferences. The mT5 model, despite covering 101 languages, demonstrates clear performance disparities across resource levels that reflect underlying training data imbalances (Xue et al., 2021). Over generations, this could erode the productive capacity of morphological systems that require active use to remain vital. The result would be languages that retain their grammatical structures but increasingly rely on borrowed vocabulary for modern domains, a pattern historically associated with language shift and endangerment.

10.2. Implications for Language Documentation

For endangered and low resource languages, these dynamics pose particular challenges. Language documentation efforts increasingly rely on computational tools for corpus building, lexicographic work, and language learning materials. If these tools systematically underrepresent indigenous neological patterns, documentation may inadvertently encode a biased snapshot of the language. Future revitalization efforts would then inherit these biases, potentially perpetuating reduced lexical creativity even in human-mediated contexts.

10.3. Toward Equity-Aware Language Technology

Future interdisciplinary collaboration between computational linguists, sociolinguists, lexicographers, and language communities will be essential to design equity-aware systems that preserve rather than erode global linguistic creativity. The LREC community, with its emphasis on language resources for all, is well-positioned to lead this effort. Concrete steps include developing neology-specific benchmarks for low resource languages, creating curated corpora of indigenous technical terminology, and establishing best practices for generative provenance tracking in lexicographic workflows.

10.4. Scope and Position of This Contribution

It is important to situate this paper clearly within the broader research landscape. The phenomena of

data imbalance and resource asymmetry across languages are well established in NLP (Joshi et al., 2020; Bender et al., 2021; Navigli et al., 2023). Our contribution does not claim novelty in identifying these asymmetries per se. Rather, the novelty lies in synthesizing these findings into a unified framework specifically targeting *lexical innovation*, a domain where the consequences of bias have distinct and under-explored implications for linguistic diversity, language policy, and lexicographic practice.

As a position paper, this work is designed to serve as a conceptual foundation and research agenda. The taxonomy of bias types (Section 6), the formalized predictions (Section 8), and the experimental protocols outlined therein are intended to catalyze empirical investigation. We believe that the theoretical groundwork presented here is a necessary prerequisite for principled experimental design in this area, and we invite the community to build upon it.

11. Limitations

This paper advances a theoretical and formal argument rather than presenting new empirical measurements. The probabilistic formalization and the derived hypotheses are intended as structured explanatory abstractions that guide future quantitative investigation rather than as validated models. Systematic cross-linguistic benchmarking, controlled prompting studies, and longitudinal corpus analyses will be required to validate or refine the proposed claims.

All arguments are grounded in established literature on linguistic inequality, distributional modeling, and lexical innovation. However, the paper does not provide direct experimental evidence demonstrating differential innovation likelihood across specific model architectures. The case studies draw entirely on previously published empirical results rather than new analyses, which means they illustrate rather than independently confirm the proposed framework. As such, the framework should be interpreted as a structured explanatory hypothesis rather than a definitive empirical conclusion.

We also make no claim of universal applicability across every LLM architecture, training regime, or future model generation. Differences in tokenization strategies, multilingual balancing techniques, or morphology-aware modeling may mitigate or exacerbate the effects described here. Additionally, the illustrative case studies focus on English-Arabic and English-Icelandic language pairs because they represent contrasting sociolinguistic and policy environments. Extension to other language families, especially typologically distant or endangered languages, may reveal additional bias patterns or countervailing dynamics not captured in the present analysis.

Finally, the formalization abstracts away from complex sociopolitical variables that shape language use in digital environments, including state policy, educational systems, platform moderation practices, and economic incentives. These factors interact with model design in ways that warrant dedicated interdisciplinary study.

12. Conclusion

AI-driven neology is shaped by structural asymmetries in global textual production. Through probabilistic generation and reinforcement loops, LLMs may amplify dominant-language innovation while marginalizing indigenous lexical creativity. This paper has formalized these dynamics through a taxonomy of four interlocking bias types, a probabilistic framework for innovation likelihood, and three empirically testable predictions accompanied by concrete experimental protocols. Addressing the challenge requires sustained interdisciplinary collaboration and the adoption of equity-aware design principles across the language resource pipeline.

Without deliberate intervention, generative systems risk accelerating linguistic homogenization within digital ecosystems. We call on the LREC community to operationalize the proposed taxonomy, test the derived predictions using the experimental protocols outlined in this paper, and develop concrete resources that safeguard lexical diversity for future generations.

13. Ethical Considerations

If unexamined, high resource bias in AI-driven neology may contribute to the reinforcement of existing linguistic hierarchies. Amplification of dominant-language innovation, coupled with the normalization of borrowing patterns, can accelerate processes of semantic convergence and marginalize culturally embedded lexical practices. Over time, this may contribute to diminished visibility of indigenous knowledge systems and reduced incentives for community-based lexical development.

Ethical language technology development therefore requires structural transparency and participatory governance. We advocate for full disclosure of training data composition, including language distribution and sources, to enable independent auditing of cross-linguistic representation. Open, community-governed language resources should be prioritized to ensure that local innovation is documented and accessible for both training and evaluation purposes.

Moreover, researchers working on low resource and endangered languages should be included as

equal partners in the design, evaluation, and deployment of language technologies. Collaborative models that center community expertise can help prevent extractive data practices and ensure that technological development aligns with local linguistic priorities.

Equity-aware language modeling is not only a technical objective but also an ethical commitment to sustaining global linguistic diversity in increasingly AI-mediated communication environments.

Acknowledgments

This work was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar National Research Fund (QNRF), part of the Qatar Research, Development and Innovation Council (QRDI). The author also acknowledge the Artificial Intelligence and Media Lab (AIM Lab) at Northwestern University in Qatar (NU-Q) and the MARSAD Lab for providing valuable resources and support that contributed to this research.

14. Bibliographical References

Al-Khalifa, S., Durrani, N., Al-Khalifa, H., and Alam, F. (2025). The landscape of Arabic large language models. *Communications of the ACM*, 68(10):54–61.

Ármannsson, B., Ingimundarson, F. Á., and Sigurðsson, E. F. (2025). An Icelandic linguistic benchmark for large language models. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 37–47, Tallinn, Estonia. University of Tartu Library.

Bauer, L. (2001). *Morphological Productivity*. Cambridge University Press, Cambridge.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 610–623, New York, NY. Association for Computing Machinery.

Blommaert, J. (2010). *The Sociolinguistics of Globalization*. Cambridge University Press, Cambridge.

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., and Oflazer, K. (2018). The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.

Bourdieu, P. (1991). *Language and Symbolic Power*. Harvard University Press, Cambridge, MA.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, 10(5):425–455.

Charfi, A., Zaghouani, W., Mehdi, S. H., and Mohamed, E. (2019). A fine-grained annotated multi-dialectal Arabic corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 198–204, Varna, Bulgaria. INCOMA Ltd.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Crystal, D. (2000). *Language Death*. Cambridge University Press, Cambridge.

Darwish, K., Habash, N., Abbas, M., Al-Khalifa, H., Al-Natsheh, H. T., Bouamor, H., Bouzoubaa, K., Cavalli-Sforza, V., El-Beltagy, S. R., El-Hajj, W., Jarrar, M., and Mubarak, H. (2021). A panoramic survey of natural language processing in the Arab world. *Communications of the ACM*, 64(4):72–81.

De Swaan, A. (2001). *Words of the World: The Global Language System*. Polity Press, Cambridge.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., and Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Gillespie, T. (2014). The relevance of algorithms. In Gillespie, T., Boczkowski, P. J., and Foot, K. A., editors, *Media Technologies: Essays on Communication, Materiality, and Society*, pages 167–193. MIT Press, Cambridge, MA.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (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, Online. Association for Computational Linguistics.

Lejeune, G. and Cartier, E. (2017). Character based pattern mining for neology detection. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 25–30, Copenhagen, Denmark. Association for Computational Linguistics.

Navigli, R., Conia, S., and Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Periti, F. and Montanelli, S. (2024). Lexical semantic change through large language models: A survey. *ACM Computing Surveys*, 56(11):1–38.

Poix, C. and Shevchenko, N. (2025). The challenge of AI-generated neology. In *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 Conference*, pages 318–331, Bled, Slovenia. Lexical Computing.

Wiemerslage, A., Silfverberg, M., Yang, C., McCarthy, A. D., Nicolai, G., Colunga, E., and Kann, K. (2022). Morphological processing of low-resource languages: Where we are and what's next. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 988–1007, Dublin, Ireland. Association for Computational Linguistics.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zaghouani, W. (2014). Critical survey of the freely available Arabic corpora. In *Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools, LREC 2014*, Reykjavik, Iceland.

Zaghouani, W. and Charfi, A. (2018). AraP-Tweet: A large multi-dialect Twitter corpus for gender, age and language variety identification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.

Zheng, J., Ritter, A., and Xu, W. (2024). NEOBENCH: Evaluating robustness of large language models with neologisms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13885–13906, Bangkok, Thailand. Association for Computational Linguistics.