Position Paper

Losing our Tail, Again: (Un)Natural Selection & Multilingual LLMs

Eva Vanmassenhove
Research Centre for Cognitive Science & Artificial Intelligence
Department of Computational Cognitive Science
Tilburg University, The Netherlands
e.o.j.vanmassenhove@tilburguniversity.edu

Abstract

Multilingual Large Language Models considerably changed how technologies influence language. While previous technologies could mediate or assist humans, there is now a tendency to *offload* the task of writing itself to these technologies, enabling models to change our languages more directly. While they provide us quick access to information and impressively fluent output, beneath their (apparent) sophistication lies a subtle, insidious threat: the gradual decline and loss of linguistic diversity. In this position paper, I explore how model collapse, with a particular focus on translation technology, can lead to the loss of linguistic forms, grammatical features, and cultural nuance. Model collapse refers to the consequences of self-consuming training loops, where automatically generated data (re-)enters the training data, leading to a gradual distortion of the data distribution and the underrepresentation of low-probability linguistic phenomena. Drawing on recent work in Computer Vision, Natural Language Processing and Machine Translation, I argue that the many *tails* of our linguistic distributions might be vanishing, and with them, the narratives and identities they carry. This paper is a call to resist linguistic flattening and to reimagine Natural Language Processing as a field that encourages, values and protects expressive multilingual diversity and creativity.

1 A Simple Evolutionary Trade-Off

A few million years ago, we lost our tails. Once useful (and presumably quite fun), they became obsolete, absorbed by evolutionary trade-offs. They disappeared through natural selection. Darwin already drew parallels between the evolution of languages and those of species noting that they follow 'curiously parallel' paths in The Descent of Man (Darwin, 1871). He went so far as to suggest that "*the survival of certain favored words in the struggle for existence is* ***natural selection***" (Darwin, 1871, p. 61), highlighting how language itself is subject to various evolutionary forces. However, it is unlikely that even Darwin could have anticipated the current path of language evolution and the latest forces that have come into play.

Today, it is not just 'natural'[1] forces but technological ones that exert strong selective pressures on languages and influence the struggle for linguistic survival on various levels. Multilingual Large Language Models (LLMs) are trained as next-word predictors over large datasets. As a result, they tend to amplify statistically likely languages and linguistic forms, whether these are words, subwords, or syntactic constructions, thereby potentially pruning away our rich, long statistical tails.

Will a technology-driven, artificial selection merely accelerate an already inevitable trajectory, or does it mark a more disruptive turn? In other words, does losing our 'language' tails imply the disappearance of obsolete, fun, and decorative elements or does it signal a more fundamental erosion of diversity and richness? In this paper, I take a position on these questions, arguing that the LLM-era marks a disruptive turn: unlike 'natural' evolution, technological selection in current mainstream LLMs (under prevailing data and training regimes) tends to reduce rather than reshape linguistic diversity.

A quick note on 'language tails': In language, as in other complex systems, many distributions are long-tailed: a few elements occur very frequently, while most are rare. These "tails" appear across multiple linguistic levels. For example, a word like 'the' appears more often than 'disperse' or 'eloquent'. Yet, these low-frequency terms are grammatical, meaningful and important within spe-

[1] While the technological forces that shape language can certainly be seen as part of the 'natural' forces influencing our linguistic ecosystem, there is a crucial shift when we move from tools that help us record, polish, or disseminate language to models that actively *generate* it. In such systems, the technology becomes not just a medium but a core driver of language change. I will return to this later in the paper.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11589–11600

cific domains or contexts. Similar long-tailed patterns can be observed across morphology, syntax, phonology, as well as at the cross-linguistic level.

2 *Natural* Language Selection

Before turning to how technology reshapes language, it is helpful to briefly consider how languages evolve 'naturally'. Evidently, this is not meant to be a comprehensive account of language evolution since doing justice to that topic requires several volumes.[2] The goal here is to highlight a few key aspects of how humans learn and transmit language, to lay the groundwork for drawing parallels with how Multilingual LLMs learn, transmit, and affect language. Broadly speaking, natural selection in languages operates at two main, related levels: *across languages* (Section 2.1), where some languages survive and thrive while others gradually disappear, and *within languages* (Section 2.2), where certain words, expressions, or grammatical structures persist over time while others fall out of use.

2.1 Across Languages

Historically, population shifts, cultural dominance, and environmental factors have repeatedly driven language extinction (Bromham et al., 2022). Nearly half of the world's roughly 6,000 languages[3] are endangered, many disappearing at an alarming rate.[4] While language loss is often viewed as something tragic, it can stem from a shared communicative goal. For instance, speakers might adopt a dominant language to improve mutual understanding. Evidently, even in these seemingly positive scenarios, something linguistically and culturally valuable is lost.

More subtly and at the same time more *universally*, we might also be witnessing a different type of loss: the loss of richness and diversity **within** languages themselves. Such a loss is less likely to be driven by collective needs like mutual intelligibility[5], but can most likely be attributed to converging pressures and constraints from language technologies. Therefore, I want to turn the attention more inwards, and focus on the erosion of diversity *within* languages.

2.2 Within Languages

Whereas the extinction of languages is often driven by external forces, changes within a language are also shaped by internal cognitive bottlenecks. The acquisition of a language, for instance, happens under rather constrained conditions: humans are exposed to sparse, ambiguous, and often noisy and imperfect input—sometimes discussed under the notion of a "poverty of stimulus" (Chomsky, 1980). Additionally, we face cognitive limitations including finite memory, a limited attention span, and bounded processing capacities. Yet, despite these many hurdles (or perhaps because of them, as suggested by DeCaro et al. (2008)) dominant languages continue to thrive.

In the following sections, I will distinguish between two types of language-internal changes: (i) the emergence of structure, briefly revisiting some foundational research on grammar and compositionality (Section 2.2.1); and (ii) the evolution of our lexicon, that is, how words are lost, gained, or transformed over time (Section 2.2.2). While in reality, these processes are to some extent intertwined, they are shaped by different pressures.

2.2.1 Towards Structure

Iterated learning (Brighton et al., 2005; Smith et al., 2003b; Ren et al., 2024a) models how linguistic structure emerges through repeated cultural transmission, where each generation of learners infers patterns from the behavior of the previous one.

To cope with limited and noisy input, humans rely on (implicit) inductive biases such as simplicity and compressibility to learn a language (Kirby et al., 2015). These structured preferences guide learning by enabling pattern extraction and structural generalization from sparse data. Research by Smith et al. (2003a) showed that when language is transmitted under cognitive and communicative pressure, structure and compositionality emerge through repeated cultural transmission cycles – processes shaped by our cognitive biases that favor more learnable and generalizable patterns. Far from being detrimental, such biases are often considered essential (DeCaro et al., 2008), allowing generalization without linguistic impoverishment.

[2] For readers interested in a broader, more comprehensive and nuanced history on, among others, the evolution of language, I recommend "The Language Puzzle" (Mithen, 2024) and "How Language Began" (Everett, 2017).

[3] Estimates vary between 3,000 and 10,000 (Crystal, 2002).

[4] https://www.unesco.org/en/articles/multilingual-education-bet-preserve-indigenous-languages-and-justice.

[5] Of course, in some specific cases or contexts, simplification can help communication across a more diverse set of speakers.

This also implies that our learning process involves *convergence*. Convergence towards systematic forms, towards structure and patterns. This convergence is not of the kind that leads to a loss of expressivity. It is a structural, higher-level convergence that supports more open-ended productivity, one where nonsense syllables can become, for instance, compositional morphemes.

While iterated learning and cultural transmission help explain how structure in language becomes more regular and learnable over time, not all aspects of language evolve in such a way. Our lexicon, the system of words and meanings, often changes in more irregular and dynamic ways. Vocabulary is more susceptible to social influence, borrowing, innovation, and drift.

2.2.2 Lexical Changes

Despite our limited memory, languages are and remain remarkably rich. Words may disappear, fall out of fashion, or decrease in usage, yet their overall vocabularies remain relatively stable, often even expanding when languages flourish driven by social needs, interactions, scientific progress, or creativity. For instance, based on over five million digitized books,[6] Michel et al. (2011) estimated that English grew by more than 70% over the past fifty years, adding on average about 8500 new words per year. A later study by Gerlach and Altmann (2013) showed that vocabulary growth largely scales with corpus size, with the rate of new-word introduction slowing but never vanishing as datasets expand. These findings can, however, be contrasted with claims of convergence towards a maximum vocabulary size (Bernhardsson et al., 2009).

The vocabulary growth reported for English in Michel et al. (2011) and Gerlach and Altmann (2013) aligns with findings showing that languages with larger speaker populations tend to have larger vocabularies (Reali et al., 2018). At the same time, widely spoken languages often exhibit simpler morphology (fewer inflections) while smaller communities may develop structurally more complex systems (Lupyan and Dale, 2010).

Word survival is frequently influenced by positive frequency-dependent selection (Pagel et al., 2019), yet frequency alone does not determine a word's fate (Altmann et al., 2011). Competition among words is also shaped by cognitive and communicative pressures that favor efficiency and clear categorization. Naming systems for colours or emotions, for example, tend to converge across cultures, reflecting shared communicative needs rather than pure frequency effects (Petersen et al., 2012). Even common words can disappear if they no longer align with prevailing cognitive, social, or technological forces. Consider the case of 'radiogram', 'Roentgenogram', and 'X-ray', once competing terms for the same concept (Petersen et al., 2012). Although 'Roentgenogram' dominated early scientific discourse, it was gradually replaced by the shorter and more efficient 'X-ray' (Piantadosi et al., 2011), a shift further reinforced by the global rise of English as the scientific lingua franca.

Vocabulary, in short, is shaped not just by usage frequency but by communicative efficiency, cognitive constraints, technological mediation, and sociocultural forces, reflecting a dynamic and context-sensitive interplay between convergence and divergence. Yet, as technological mediation increasingly shapes communication, these dynamics may come to be governed by a different kind of selection.

3 A Force of *Technology*

Darwin argued that languages evolve through a process of variation and selection, convergence and divergence, much like species. We have seen how such dynamics give rise to compositionality and structure, as well as changes in our lexicon. Broadly speaking, we have seen that language reflects a tension between convergence and divergence. However, we set out to explore what happens when the strongest selective forces are no longer human cognition or social need, but instead arise from (large-scale) technological forces.

First, we briefly examine the relationship between language and cultural technologies that predate the rise of (L)LMs (Section 3.1). This provides context for the *Descent of Language* section (Section 3.3), where I examine early signs of linguistic decline in computational models and discuss more recent work on LLMs.

3.1 Language & Cultural Technologies

It is important to note that it is not the first time that technology has played a disruptive role in shaping language. Language has never existed in isolation and (cultural) technologies have continuously influenced how language is transmitted. While it may be hard to imagine today, even *writing* was once considered a disruptive technological innovation:

[6] The 2010 analysis covered roughly 4% of all books ever published.

> *They seem to talk to you as though they were intelligent, but if you ask them anything about what they say from a desire to be instructed they go on telling just the same thing forever.*
>
> — Plato, *Phaedrus* 275d

As cognitive scientist Alison Gopnik pointed out in her 2023 ACL keynote (Gopnik, 2023), without any context, this critique of *writing* could just as well apply to LLMs today. There is a surface-level fluency that (over)confidently mimics understanding, yet lacks genuine responsiveness, real-world grounding, or explanatory depth. While we no longer worry about the *dangers* of writing, history reminds us that each new 'tool' reshapes how language is used, transmitted, and transformed. Some languages, the so-called "long-tailed" ones,[7] have been particularly impacted. Their marginalization has been reinforced by successive cultural technologies (from writing to the internet), and by their exclusion from the digital sphere. This phenomenon, described as *digital language death*, is said to affect roughly 95% of the world's languages (Kornai, 2013).

Earlier technologies also marked turning points in the evolution of language. The printing press, for instance, helped standardize spelling and grammar, elevating dominant dialects while marginalizing others (Sasaki, 2017). More recently, automatic spell-checkers boosted the 'reproductive fitness' of recognized forms at the expense of other alternatives, contributing, for example, to the rising dominance of 'colour' over 'color' (Petersen et al., 2012).[8]

While the impact of technology on language and concerns regarding its disruption are not new, the current wave of generative models operates at a different level, scale, and speed. They represent a new kind of intervention, one in which technology is not simply storing, mediating, or 'correcting' language, but actively **generating** it across different domains, tasks, and platforms. In a sense, the technology quite suddenly moved from the passenger's seat to the driver's seat.

This shift in agency may have important implications for languages and their evolution. Concerns regarding the loss of diversity *across* languages due to recent technologies, such as LLMs, have been explicitly raised in recent seminal work within the field of Natural Language Processing (NLP) by Joshi et al. (2020) and Bender et al. (2021), among others. Here, I shift the focus inward, to the internal diversity *within languages*. This dimension has received comparatively little attention, even though current multilingual LLMs exert significant influence on what is preserved, erased, or amplified within a language itself.

3.2 Large Language Models

LLMs mark a significant shift in how language is produced, accessed, and reused, and perhaps more interestingly, how the cognitive labor involved in writing and idea formulation is now often *offloaded* to technology. Earlier language technologies were built for (domain-) specific tasks; today's foundation models, however, operate across domains and applications. Large-scale models are rapidly becoming an integral part of a broad range of our everyday activities (Bick et al., 2026), allowing them to shape and influence language more directly than before. Unlike earlier tools that assisted human writing, these models now drive change by producing language themselves.

Emerging research, however, suggests that LLMs introduce subtle, yet cumulative, distortions (Shumailov et al., 2023). While initially *imperceptible*, these small changes can accumulate across multiple training cycles, leading to a phenomenon coined *model collapse*, where models trained on their own outputs progressively lose quality and diversity. In computer vision, such a collapse leads to *visible* artefacts in AI-generated images (Alemohammad et al., 2023), yet for language, the consequences remain underexplored. This is despite the fact that language shapes and possibly constrains human thought, meaning that this gradual impoverishment or distortion of linguistic output could have profound and far-reaching implications.

LLMs will likely continue to substantially influence both the content we are exposed to (images, texts, audio, etc.) and the systems that generate this content given that their output will increasingly re-enter the training cycle.[9] Therefore, at this point we can assume that interactions between models

[7] Languages that receive limited localization attention or commercial investment; this does not always correspond to speaker numbers. For instance, Bengali is the 7th most spoken language, but falls outside the top 50 most localized languages (Lionbridge, 2020).

[8] Amusingly, the tool I am writing in still nudges me away from 'colour'.

[9] A substantial portion of multilingual web content has been shown to consist of machine-translated text, indicating that language distributions are already significantly shaped by automated systems (Thompson et al., 2024).

are not hypothetical but inevitable (Martínez et al., 2023). Such interactions can occur through (partial) training on output from another LLM or a model's own output (Ren et al., 2024b). The subsequent feedback loops this creates, where models learn from their own output or that of others, accelerate concerns regarding the distortion of language.

So far, research efforts have largely focused on sustaining the benefits of training from large-scale, human-generated data scraped from the Web, summarized in this recent article as: "*LLMs' world is our word*"[10]. But perhaps more concerningly: "*Our world* might be turning into *their word*".

3.3 The *Descent* of Language

Even before terms such as 'model collapse' (Shumailov et al., 2023)[11], 'Model Autophagy Disorder' (MAD) (Alemohammad et al., 2023) or 'Habsburg AI'[12] gained traction, earlier research empirically showed how statistical language models[13] indeed amplify dominant linguistic forms while forgetting or flattening rarer, low-probability ones.

3.3.1 Precursor: Statistical and Neural Translation Models

As a precursor to this line of work, studies in Machine Translation (MT) (Vanmassenhove et al., 2019, 2021) provided empirical evidence that both statistical and neural models systematically favor frequent lexical and morphological patterns, reducing linguistic diversity. This raises two concerns: (i) technically, frequency bias diminishes lexical richness and can eliminate infrequent but grammatically necessary forms; and (ii) sociolinguistically, machine-generated translationese may, over time, influence language itself (Kranich, 2014). For instance, Vanmassenhove et al. (2021) showed that MT systems disproportionately produce masculine *président* over feminine *présidente* when translating English *president* into French, reflecting data imbalances. Under iterative training, rare forms may disappear entirely: in their experiments, the plural *présidentes* vanished. Such low-probability forms are not noise but carriers of grammatical precision, expressiveness, and social meaning. Yet, current models fail to distinguish these cases, discarding rare forms indiscriminately.

While the previous examples concerned lexical and morphological variation, similar effects arise at the morphosyntactic level. Luo et al. (2024) found MT outputs to be structurally closer to the source text than human translations, with fewer morphosyntactic divergences. Beam search biases toward "safe," high-probability constructions, reinforcing convergence and reducing syntactic diversity.

3.3.2 Generative Models

In recent work on LLMs, concerns about **model collapse** have gained traction: when models are trained on data increasingly composed of their own (or other models') output, their behavior begins to *converge*, potentially degrading over time (Shumailov et al., 2023). In computer vision, this degeneration has been made *visible*, with iterative training cycles producing recognizably distorted 'Habsburg Jaw' artifacts (Alemohammad et al., 2023). In language, such collapse is likely subtler and harder to detect, yet potentially more consequential, given the foundational role of language in human cognition and cultural evolution.

Empirical studies suggest early signs of such dynamics. McCoy et al. (2023) show that even for simple deterministic tasks, LLM behavior is sensitive to probability: GPT-4's performance drops dramatically when the correct output sequence is low-probability, revealing *embers of autoregression*. Evidence from real-world text production points in a similar direction, reflecting probability-driven effects. Kobak et al. (2024) report sharp spikes in terms such as *delve*, *crucial*, and *significant* in PubMed abstracts following the release of public LLM tools, exceeding even pandemic-related shifts (e.g. the sudden spike of *corona*). This suggests subtle, distributional nudging of academic discourse toward statistically likely phrasing.

At the distributional level, Shumailov et al. (2023) demonstrate how low-probability events disappear first in iterative training, with models gradually converging toward high-probability sequences. Extending these observations to large-scale gener-

[10] https://www.theguardian.com/technology/article/2024/sep/07/if-journalism-is-going-up-in-smoke-i-might-as-well-get-high-off-the-fumes-confessions-of-a-chatbot-helper?utm_source=chatgpt.com

[11] As recently discussed in Schaeffer et al. (2025), model collapse encompasses a broad range of phenomena; here, we explicitly focus on the distributional aspect, which relates to the erosion of long-tail linguistic diversity central to our argument.

[12] A term coined by Jathan Sadowski (https://x.com/jathansadowski/status/1625245803211272194?lang=en).

[13] Whether they are called statistical, neural or large language models, in the end, they are still all statistical models.

ative settings, Jiang et al. (2025) identify an 'Artificial Hivemind' effect, characterized not only by repetition within a single model, but also by striking homogeneity across different models when responding to diverse open-ended prompts.

Importantly, these dynamics are not limited to earlier generations of LLMs and may even be amplified by Reinforcement Learning from Human Feedback (Kirk et al., 2023) or Supervised-Fine-Tuning, where aggressively updating the model towards observed data distributions can further entrench the underrepresentation of low-probability forms. Recent work on instruction-tuned LLMs further shows that such constraints can manifest in practice as what is referred to as 'diversity collapse', with models converging toward semantically similar responses (Yun et al., 2025).[14]

Focusing more on linguistic diversity, Guo et al. (2024) leverage a comprehensive evaluation framework to assess LLM outputs and observe a significant decrease in diversity when comparing human language to LLM-generated content, especially when focusing on tasks that require creativity. They depict a rather complex pattern where instruction tuning does improve lexical diversity, but this comes at a cost, as it narrows the expressive flexibility seen by a decrease in overall syntactic and semantic diversity.

4 Thoughts & Discussions

Darwin noted the similarities between the evolution of species and that of languages. Of course, he was referring to mechanisms of gradual change through natural evolution. I could, however, not help be amused by the rather unexpected parallel that can be drawn between the evolution of our species and the effect of recent technologies on languages. It seems that, once again, our species and language are following a similar path: *We are losing our tails*. While our physical appendages became obsolete and even cumbersome, I argue that its language equivalent is anything but that, capturing the rich diversity and continuous evolution of language, shaped both by the need to articulate novel concepts and by our ongoing desire to signal belonging and distinction within social groups.

While one could argue that across languages, various natural forces, sometimes in combination with technological ones, have contributed to a significant loss of diversity across languages, leading even to *language death*, when we shift the focus to the internal diversity within languages, natural evolution has led to structure, compositionality, and a growth in terms of vocabulary, at least for thriving languages. This stands in contrast with what we observe when language is *driven* (largely or solely) by technological forces, whose internal pressures seem to lead to a reduction in expressivity, creativity, and a loss of overall lexical diversity regardless of whether the language is flourishing or not.

In the paragraphs below, I set out my reflections and formulate a position on what it means when current technological forces become the main engines of language change.

(L)LMs reduce linguistic richness and amplify biases While current models' ability to generalize over large amounts of data is one of their biggest assets, their statistical nature and the pressures that shape these models have drawbacks regarding diversity and creativity. Until recently, this might not have been a priority for our field, given that these technologies were largely regarded as domain-specific tools that were often still supervised or post-edited by humans (e.g. chatbots, MT, etc.). However, the fairly recent public release of large, general-purpose, language models raises concerns about the potential implications for language (and language technologies) in the longer term.

From the literature, it seems that indeed, (L)LMs exacerbate imbalances in the data by (i) forgetting low-probability events/words, and (ii) overgenerating high(er) probability ones (Vanmassenhove et al., 2021; Shumailov et al., 2023). This could lead to self-reinforcing loops where less frequent linguistic forms and expressions are underrepresented and risk being entirely lost in translation. Low(er)-probability events (words, subwords, etc.) contribute to the complexity and richness across and within language(s), and are important to ensure that models do not converge toward oversimplified, biased language. After all, languages are full of improbable events.

Furthermore, the effects of model collapse are likely not confined to obvious levels such as vocabulary. They may also appear at the morphosyntactic level (e.g., Guo et al. (2024)), at non-linguistically motivated subword or character levels, in the fact that (the) dominant or target language(s) structure(s) may "bleed through" (e.g., preferences for Subject-Verb-Object constructions), or in the prop-

[14] These trade-offs are often discussed under the notion of the 'alignment tax' (Bai et al., 2022; Ouyang et al., 2022).

agation of specific ideologies (e.g., representations of gender). Prior work (e.g. Cao et al. (2023) or Dokic et al. (2025)) has demonstrated how stereotypes encoded in English can "leak" into other languages in multilingual LLMs, with typologically distant languages being particularly vulnerable. This asymmetry is potentially overlooked given the dominance of English in evaluation benchmarks, which could lead to an underestimation of collapse effects in other languages. Similarly, one could hypothesize that multilingual models might start confabulating words similar to English[15] or translate English idiomatic expressions literally, even when translating into typologically distant languages, a well-known cause for errors (Karakanta et al., 2025).

And last but not least, given the unbalanced nature of language representation on the web, we can furthermore assume that minority, long-tailed and morphologically richer languages are likely affected disproportionaly. As model outputs feed back into future models, and the web increasingly becomes a space where AI systems and agents generate and exchange content with one another, the result is **a compounding distortion of language use, progressively shifting further away from diversity observed in the real world on the language-internal and cross-linguistic level**.

Methodological Blind Spots in Measuring Linguistic Diversity Current evaluation practices in NLP often involve pairwise comparisons between a single AI-generated text and a single human-written text. Because the model has been trained on massive datasets containing the writing styles, vocabularies, and linguistic patterns of millions of humans, its outputs often exhibit surface-level lexical or syntactic variety that surpasses that of an individual human writer or text (Reviriego et al., 2024). This could lead to potentially misleading conclusions, where one might start claiming that AI outputs are *overall* more diverse or more lexically rich than human texts. This way of comparing AI-written vs human-written text *overlooks people's individual variation*.

LLMs can adopt and mimic many different styles, but at inference time, they tend to converge on high-probability patterns. When many generations of outputs are compared over time, these outputs are likely increasingly uniform - a tendency that has already been observed empirically in the form of both intra- and inter-model homogeneity on open-ended questions (Jiang et al., 2025). This can be contrasted against what we observe in human language, which is inherently diverse *across* speakers, contexts, and time. If we instead evaluate diversity at scale (i.e., across many texts or over generations), human-generated language maintains a rich variation through individual idiolects, sociolects, and cultural registers shaped by our individual biases rather than a common one. **Short-term superficial lexical diversity does not guarantee long-term linguistic sustainability**. Methodological approaches that treat diversity as an isolated textual property risk drawing premature conclusions with respect to diversity in the longer term.

Aside from this methodological blind spot regarding linguistic diversity, it is worth highlighting once more that current evaluation metrics for translation (BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006), COMET (Rei et al., 2020)) are not designed to capture loss of diversity. Nor should they: diversity is not a property of a single translation, nor even necessarily of a single text. This does not mean, however, that we should ignore it: diversity emerges at the level of systems and over time. Understanding how it evolves across models and generations is important for assessing the broader impact of language technologies. In line with McCoy et al. (2023), I conclude that we should not evaluate LLMs as if they are humans but should instead treat them as a distinct type of system, one that has been shaped by its own particular set of pressures. It is thus important for metrics to be designed in order to reveal *their* idiosyncratic weaknesses.

Compositionality and Systematicity still largely elude LLM capabilities. Humans learn language under limited memory capacities; there is a critical period where we can easily learn languages, and our exposure to language is (in some ways) much more restricted than that of LLMs. Our bottlenecks, however, seem to serve as pressure mechanisms for the emergence of structure and compositionality, which allows us, among other things, to create new words that can often immediately be understood by others speaking the same language. We generalize and converge, but we do so towards a

[15] Castilho et al. (2025) showed how, just like humans, LLMs tend to resort to what they call "Lazy Gaelicisations" which involve the adaptation of English words towards the Irish ortography. This is, however, also a common strategy among Irish speakers.

productive system. Even though neural networks can behave compositionally and systematically, it is not straightforward for them.

Regarding compositionality in NMT using Transformers (Vaswani et al., 2017), recent work by Yin et al. (2024) compares the performance of different Transformer-based models (Transformer trained from scratch, pre-trained decoder-only models (BLOOMZ-7b (Muennighoff et al., 2022) and LLaMA2-13b (Touvron et al., 2023)) and a pre-trained encoder-decoder model (mT5-large (Xue et al., 2021)). They show that all these models still struggle when translating new or long compounds. Additionally, lower perplexity source sentences are more likely to be correctly translated into the target, and error rates go up when the length of the compounds increase. These findings are in line with some of the phenomena discussed in the previous section. While they do find that fine-tuned pretrained LLMs outperform Transformer models trained from scratch, they point out that this advantage could be due to pretraining exposure rather than true compositional generalizations.

In experiments similar to those conducted by Kirby et al. (2014) illustrating the effect of cultural transmission through an iterated learning framework, Kouwenhoven et al. (2025b) and Kouwenhoven et al. (2025a) empirically evaluated and compared human-human, LLM-LLM and human-LLM (artificial) language learning to compare how artificial languages differ when optimized by LLMs or humans' inductive biases.[16] Their comparisons of language learning across the three different conditions revealed that, while similar to human vocabularies, LLM languages are subtly different. The LLM optimized languages showed less diversity and variation, making them more *degenerate* in comparison to those optimized for humans. These differences were alleviated when humans and LLMs collaborated, which underscores that to achieve successful interactions between humans and machines, it is essential to optimize for communicative success since the need to be expressive in human language can prevent convergence.

More generally, Zhou et al. (2023) looked at the link between the complexity of the dataset and the ability of models to generalize. More complex datasets provide: (i) more diversity in terms of the examples the model is exposed to but also, (ii) a reduced repetition preventing the model from *non-generalizable* surface memorization. Yet, as pointed out by Dziri et al. (2023) but also by McCoy et al. (2023), these models still fail on sometimes surprisingly trivial problems and quickly decay once task complexity increases, indicating once more that these are symptoms of a more fundamental limitation.

Averaging Biases Humans are fundamentally biased. For decision-making, we rely on frugal heuristics (Gigerenzer and Goldstein, 1996) which are efficient but obviously imperfect. Again, this relates to our cognitive constraints (limited attention, processing capacity, memory). In this context, it is important to highlight that our biases are **not monolithic**. While it is true that they are partially shaped by our society, environment, and direct surroundings, we each develop a slightly different set of heuristics and biases over time. These are based on our unique experiences and contexts. Regardless of whether they are good or bad, *they are many*. Besides, some of us actively fight or question our own biases when we recognize that they could be harmful, unfair, or simply undesirable.

The biases multilingual LLMs propagate, in contrast, are averaged, dominant ones that are present in an already biased sample of training data. Rather than capturing the diversity and heterogeneity of biases in human reasoning, by letting LLMs drive language change, we risk exacerbating and normalizing dominant biases without having a critical self-reflection or self-correction component.

Invisible Gaps: The Missing Data

> *That which we ignore reveals more than what we give our attention to.*
>
> — Mimi Onuoha[17]

Finally, model collapse is as much about what is not generated as what is. Over-reliance on high-frequency or majority-culture content leads to blind spots in representation. The absence of specific linguistic structures, sociolinguistic variants, or cultural references in training data can render certain forms of expression invisible in model outputs. As models are increasingly retrained on AI-generated content, these omissions risk becoming permanent.

[16] They do not focus on behavioral biases but on the implicit inductive ones. For humans, these are biases such as preferences for compressibility, simplicity or efficiency), while for LLMs they focused on *increasingly apparent* biases of the Transformer architecture (e.g. simplicity, structure, recency)(Kouwenhoven et al., 2025a).

[17] https://github.com/MimiOnuoha/missing-datasets

5 Conclusions

A few million years ago, we lost our tails. Once functional, they became obsolete and cumbersome. I set out in this paper with the question: Do the many statistical language tails face the same fate, and if so, would we merely be losing the obsolete, fun and decorative elements?

Based on recent LLM-related research, I argue that the long statistical tails of language may indeed face a similar fate. The artificial selection driven by LLMs marks a rather disruptive shift from language being shaped by generational, cultural transmission. Unlike the evolution of language driven by humans, which despite (or because of) our cognitive constraints shaped language into a structured, productive and compositional tool with a rich vocabulary; language shaped by models tends to collapse towards what is likely, driven by statistical biases. The interplay and balance between convergence and divergence, that characterizes human behavior and communication on multiple levels, risks being lost. Unlike humans, models are not intrinsically motivated to be creative, to express belonging or differences, or to innovate. There is no capacity for critical self-reflection or self-correction, and only a limited plurality of voices.

As these systems increasingly *ingest* their own or other models' outputs, the risk of flattening linguistic diversity grows, with rare words, less-resourced languages, and culturally significant variation most at risk. Preserving the long tails of language means rethinking how we evaluate and train these systems, not just for accuracy or fluency, but for the communicative richness that makes language human. **Without our tails, we risk losing the balancing act** by converging, collapsing, and flattening the expressive multilingual linguistic, social and cultural diversity.

Limitations

The arguments presented in this paper are intended to provoke critical reflection on the trajectory of language in the era of LLMs; however, they are subject to several limitations regarding scope, empirical generalization, and my own perspective. While I draw on a synthesis of recent findings in model collapse (and more specifically 'diversity collapse'), iterated learning, and sociolinguistics, the long-term impact of LLMs on natural human language and speech remains a developing phenomenon. The limitations and concerns discussed reflect the current state-of-the-art. It is possible that future architectures, perhaps those incorporating neuro-symbolic reasoning or novel inductive biases, can mitigate certain aspects of model collapse. Our position is therefore a critique of the current trajectory of generative AI rather than an immutable law of artificial/machine intelligence.

I furthermore acknowledge an inherent selection bias in the literature reviewed and the examples provided. My perspective is situated within an academic tradition that values linguistic richness. I recognize that researchers from more functionalist or engineering-driven backgrounds might interpret the "flattening" of language as an increase in communicative efficiency or standardization rather than a loss of richness.

Acknowledgments

I thank the anonymous ARR reviewers for their thoughtful, encouraging, and constructive feedback. I am grateful to Afra Alishahi for feedback on an earlier draft, and to Dimitar Shterionov for his feedback, guidance, and many years of discussions that have shaped my thinking on this topic. I also thank my father for introducing me to Simon Kirby's work via a BBC documentary years ago, and for the steady stream of book recommendations ever since. Finally, I thank the MT Summit 2025 organizers for the opportunity to present an earlier version of these ideas as a keynote, and the audience for their insightful comments and discussions. In line with the ACL policy on AI writing assistance, AI tools were used solely to improve the English of this paper (e.g., correcting grammatical errors and suggesting alternative phrasings). No AI tools were used to generate scientific content.

References

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G Baraniuk. 2023. Self-consuming generative models go mad. *arXiv preprint arXiv:2307.01850*, 4:14.

Eduardo G Altmann, Janet B Pierrehumbert, and Adilson E Motter. 2011. Niche as a determinant of word fate in online groups. *PloS one*, 6(5):e19009.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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*, pages 610–623.

Sebastian Bernhardsson, Luis Enrique Correa da Rocha, and Petter Minnhagen. 2009. The meta book and size-dependent properties of written language. *New Journal of Physics*, 11(12):123015.

Alexander Bick, Adam Blandin, and David J Deming. 2026. The rapid adoption of generative ai. *Management Science*.

Henry Brighton, Kenny Smith, and Simon Kirby. 2005. Language as an evolutionary system. *Physics of Life Reviews*, 2(3):177–226.

Lindell Bromham, Russell Dinnage, Hedvig Skirgård, Andrew Ritchie, Marcel Cardillo, Felicity Meakins, Simon Greenhill, and Xia Hua. 2022. Global predictors of language endangerment and the future of linguistic diversity. *Nature ecology & evolution*, 6(2):163–173.

Yang Trista Cao, Anna Sotnikova, Jieyu Zhao, Linda X Zou, Rachel Rudinger, and Hal Daume III. 2023. Multilingual large language models leak human stereotypes across language boundaries. *arXiv preprint arXiv:2312.07141*.

Sheila Castilho, Zoe Fitzsimmons, Claire Holton, and Aoife Mc Donagh. 2025. Synthetic fluency: Hallucinations, confabulations, and the creation of irish words in llm-generated translations. *Proceedings of Machine Translation Summit XVII: Research Track*.

Noam Chomsky. 1980. Rules and representations. *Behavioral and brain sciences*, 3(1):1–15.

David Crystal. 2002. *Language death*. Cambridge university press.

Charles Darwin. 1871. *The Descent of Man, and Selection in Relation to Sex*, volume 1. John Murray, London.

Marci S DeCaro, Robin D Thomas, and Sian L Beilock. 2008. Individual differences in category learning: Sometimes less working memory capacity is better than more. *Cognition*, 107(1):284–294.

Kristian Dokic, Barbara Pisker, and Bojan Radisic. 2025. Mirroring cultural dominance: Disclosing large language models social values, attitudes and stereotypes. *Societies*, 15(5):142.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaïd Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Daniel L Everett. 2017. *How language began: The story of humanity's greatest invention*. Liveright Publishing.

Martin Gerlach and Eduardo G Altmann. 2013. Stochastic model for the vocabulary growth in natural languages. *Physical Review X*, 3(2):021006.

Gerd Gigerenzer and Daniel G Goldstein. 1996. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, 103(4):650.

Alison Gopnik. 2023. Large language models as cultural technologies: Imitation and innovation in children and models. https://doi.org/10.48448/8qf0-j172. Keynote lecture at ACL 2023, Toronto, July 12, 2023.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2024. Benchmarking linguistic diversity of large language models. *arXiv preprint arXiv:2412.10271*.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. 2025. Artificial hivemind: The open-ended homogeneity of language models (and beyond). *arXiv preprint arXiv:2510.22954*.

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.

Alina Karakanta, Mayra Nas, and Aletta G. Dorst. 2025. Metaphors in literary machine translation: Close but no cigar? *Proceedings of Machine Translation Summit XVII: Research Track*.

Simon Kirby, Tom Griffiths, and Kenny Smith. 2014. Iterated learning and the evolution of language. *Current opinion in neurobiology*, 28:108–114.

Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. 2015. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. 2024. Delving into chatgpt usage in academic writing through excess vocabulary. *arXiv preprint arXiv:2406.07016*.

András Kornai. 2013. Digital language death. *PloS one*, 8(10):e77056.

Tom Kouwenhoven, Max Peeperkorn, Roy de Kleijn, and Tessa Verhoef. 2025a. Shaping shared languages: Human and large language models' inductive biases in emergent communication. *arXiv preprint arXiv:2503.04395*.

Tom Kouwenhoven, Max Peeperkorn, and Tessa Verhoef. 2025b. Searching for structure: Investigating emergent communication with large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9977–9991, Abu Dhabi, UAE. Association for Computational Linguistics.

Svenja Kranich. 2014. Translations as a locus of language contact. In *Translation: A multidisciplinary approach*, pages 96–115. Springer.

Lionbridge. 2020. Localizing long-tail languages. https://www.lionbridge.com/blog/translation-localization/localizing-long-tail-languages/. Accessed: 17 April 2026.

Jiaming Luo, Colin Cherry, and George Foster. 2024. To diverge or not to diverge: A morphosyntactic perspective on machine translation vs human translation. *Transactions of the Association for Computational Linguistics*, 12:355–371.

Gary Lupyan and Rick Dale. 2010. Language structure is partly determined by social structure. *PloS one*, 5(1):e8559.

Gonzalo Martínez, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juarez, and Rik Sarkar. 2023. Towards understanding the interplay of generative artificial intelligence and the internet. In *International Workshop on Epistemic Uncertainty in Artificial Intelligence*, pages 59–73. Springer.

R Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L Griffiths. 2023. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Steven Mithen. 2024. *The language puzzle: How we talked our way out of the stone age*. Profile Books.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Mark Pagel, Mark Beaumont, Andrew Meade, Annemarie Verkerk, and Andreea Calude. 2019. Dominant words rise to the top by positive frequency-dependent selection. *Proceedings of the National Academy of Sciences*, 116(15):7397–7402.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alexander M Petersen, Joel Tenenbaum, Shlomo Havlin, and H Eugene Stanley. 2012. Statistical laws governing fluctuations in word use from word birth to word death. *Scientific reports*, 2(1):313.

Steven T Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Plato. 2002. *Phaedrus*. Oxford World Classics. Oxford University Press, Oxford.

Florencia Reali, Nick Chater, and Morten H Christiansen. 2018. Simpler grammar, larger vocabulary: How population size affects language. *Proceedings of the Royal Society B: Biological Sciences*, 285(1871):20172586.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Yi Ren, Shangmin Guo, Linlu Qiu, Bailin Wang, and Danica J Sutherland. 2024a. Bias amplification in language model evolution: An iterated learning perspective. *arXiv preprint arXiv:2404.04286*.

Yi Ren, Shangmin Guo, Linlu Qiu, Bailin Wang, and Danica J Sutherland. 2024b. Bias amplification in language model evolution: An iterated learning perspective. *arXiv preprint arXiv:2404.04286*.

Pedro Reviriego, Javier Conde, Elena Merino-Gómez, Gonzalo Martínez, and José Alberto Hernández. 2024. Playing with words: Comparing the vocabulary and lexical diversity of chatgpt and humans. *Machine Learning with Applications*, 18:100602.

Yu Sasaki. 2017. Publishing nations: Technology acquisition and language standardization for european ethnic groups. *The Journal of Economic History*, 77(4):1007–1047.

Rylan Schaeffer, Joshua Kazdan, Alvan Caleb Arulandu, and Sanmi Koyejo. 2025. Position: Model collapse does not mean what you think. *arXiv preprint arXiv:2503.03150*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Kenny Smith, Henry Brighton, and Simon Kirby. 2003a. Complex systems in language evolution: the cultural emergence of compositional structure. *Advances in complex systems*, 6(04):537–558.

Kenny Smith, Simon Kirby, and Henry Brighton. 2003b. Iterated learning: A framework for the emergence of language. *Artificial life*, 9(4):371–386.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Brian Thompson, Mehak Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. A shocking amount of the web is machine translated: Insights from multi-way parallelism. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1763–1775, Bangkok, Thailand. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 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*. Association for Computational Linguistics.

Yongjing Yin, Lian Fu, Yafu Li, and Yue Zhang. 2024. On compositional generalization of transformer-based neural machine translation. *Information Fusion*, 111:102491.

Longfei Yun, Chenyang An, Zilong Wang, Letian Peng, and Jingbo Shang. 2025. The price of format: Diversity collapse in llms.

Xiang Zhou, Yichen Jiang, and Mohit Bansal. 2023. Data factors for better compositional generalization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14549–14566.