MLO Modern Languages Open

“No Language Left Behind?” Predictive Text, Generative AI and Dilemmas of Digital Inclusion for Marginalized Languages

SPECIAL COLLECTION: THE DIGITAL LIVES OF GREATER SOUTH LANGUAGES

ARTICLES – DIGITAL MODERN LANGUAGES

PETER CHONKA
STEPHANIE DIEPEVEEN
YIDNEKACHEW HAILE

*Author affiliations can be found in the back matter of this article

LIVERPOOL UNIVERSITY PRESS

ABSTRACT

Ongoing developments in natural language processing (NLP) and natural language generation (NLG) raise critical questions about linguistic bias and inequalities in datasets. Problematizing the concept of digital “inclusion” for linguistic vitality, this article addresses an adjacent research gap concerning dilemmas that are emerging from the increasing operability of globally marginalized languages with NLP and NLG technologies. We draw on our previous work on how Google Search “Autocomplete” algorithms interact with three languages indigenous to East Africa – Amharic, Kiswahili and Somali – each with their own historical, political and orthographic features that condition this automated interaction in complex and unpredictable ways. Highlighting different forms of harm that might result from the operation of predictive text in each language, we argue for the necessity of situating the algorithmic experiences of marginalized languages within specific historical, cultural and political contexts. From here, we consider the ways that the predictive logics of autocomplete have been expanding through the rapid intrusion of generative AI into the results of mainstream search engines. We grapple with the global implications of this rapidly changing AI information landscape and consider the potential impacts of such developments on linguistic contexts in East Africa and the dynamics of multi-scalar language inequality within and beyond the region.

CORRESPONDING AUTHOR:
Peter Chonka
King’s College London, UK
peter.chonka@kcl.ac.uk

TO CITE THIS ARTICLE:
Chonka, Peter, Stephanie Diepeveen and Yidnekachew Haile 2026 “No Language Left Behind?” Predictive Text, Generative AI and Dilemmas of Digital Inclusion for Marginalized Languages. *Modern Languages Open*, 2026(1): 6 pp. 1–16. DOI: https://doi.org/10.3828/mlo.v0i0.465

INTRODUCTION

In September 2020, if a user typed the Somali word *gabadh* (girl) into Google Search, the top "Autocomplete" suggestion that they would be presented with would be the Somali word for "naked" (*qaawan*). Testing with a research browser and across different locations (to minimize the impact of an individual user's previous search habits), there was significant consistency here, with the autocomplete algorithm presenting users with a sexualized word as a likely search query even before they had completed the search and viewed the results. If the user was to complete the search using this "predicted" additional keyword, this would become search pattern data, potentially making it more likely that this same sexualized autocomplete suggestion would reappear for other users. Contemporaneously, entry of the English word "girl" into Google Search did not generate similar problematic autocomplete suggestions. If algorithmic autocomplete predictions are based on factors that include the search behaviour of other users, did the appearance of *qaawan* in the Somali example indicate that searches by users of this language contained more sexualized content than those using English? This conclusion is untenable given the proportion of global Internet traffic taken up by pornography. A more plausible explanation would be that autocomplete for the English search keyword "girl" was subject to a range of filtering techniques that would remove harmful predictions. This degree of filtering has not been in operation for Somali.

In both 2020 and at the time of writing (2024), a Google search for Somali politicians (using the Somali-language spelling of their name) would frequently generate a prediction that invites users to query the "clan" identity of that individual. Given Somalia's history of conflict that has been shaped by this social category, written public discussion of specific clan identities is a taboo in formal Somali-language media. Today, in a Somali interaction with ChatGPT, this generative artificial intelligence (AI) tool will provide this information directly to a user. Considering the cultural and political complexities of clan-based discourse in the Somali context, questions around content moderation/user protections for predictive/generative AI are arguably more ambiguous than in the previous gendered/sexualized example. What is clear is that predictive autocomplete and generative text AI tools are putting contentious language into the digital public in ways that contrast with mainstream local journalistic practice. Also striking is the pace at which generative language technologies are increasing their capabilities across a wide spectrum of global languages. Just in the period during which this article was under review, ChatGPT went from being a tool that "unofficially" worked in a limited fashion with Somali language inputs, to "officially" incorporating this language and becoming adopted by companies in Somalia looking, for instance, to automate their online customer support (Hormuud).

In our previous work (Chonka et al.) we analysed results from tests of predictive autocomplete suggestions for gendered and political keywords in three East African languages through the conceptual lens of digital colonialism (Couldry and Mejias; Kwet; Madianou). To take into account varying orthographic and spelling features, as well as the political context of each language, we developed specific sets of keywords for each test (e.g. Somali spellings of politicians' names), while also testing some common terms explored across all three (e.g. girl/woman, boy/man). We have argued that digital colonialism manifests in the unequal exposure of different groups of language users to algorithmic harms. In light of subsequent (and rapid) developments in the field of generative AI, this article considers the implications of marginalized global languages interacting with predictive text algorithms beyond search engine autocomplete and in relation to the potential impacts of these technologies on languages themselves and the cultures they are situated in. We are also concerned with multi-scalar global/local linguistic power imbalances, since data-driven predictive/generative text is already "perpetuating cycles in which both advantage and disadvantage [for different languages] are magnified" (Zaugg, "Language Justice" 262).

Thinking beyond search engine autocomplete also means considering how predictive/generative text functionality is now penetrating (or colonizing?) an increasingly broader range of digital platforms, ranging from email and online writing tools to business operation platforms. It is necessary to examine how natural language processing (NLP) and natural language generation (NLG) work, particularly given that the latter gained widespread public attention through the release of powerful AI-driven tools that simulate meaningful linguistic interaction with users.

While search keyword autocomplete and NLG-based technologies differ, both are underpinned by probabilistic prediction and extraction from massive online corpora/datasets. Furthermore, the boundaries between these technologies are now inextricably blurred. Google is now clearly enacting its vision of the future of information retrieval through automated conversation agents that provide users with synthesized answers to queries, as opposed to directing them towards sources for their own perusal (Shah and Bender). Both Microsoft and Google have now incorporated their respective generative AI models as prominent features of their search engines, offering unsolicited answers to user queries. Since Google has not "fixed" the issue of obviously problematic African-language autocomplete suggestions that we pointed out in our earlier research (Chonka et al.), it is essential to revisit these questions in light of generative AI developments of search engines themselves. Here, it is remarkable that for all of Google Gemini's current sophistication in Somali language generation and interaction, Google Search Autocomplete still does not filter out the suggestion of the Somali word for "f*ck", even for innocuous search keywords, and presents the user with troubling unsolicited content (e.g. non-consensual sexual image sharing) (Chonka 30).

Issues of linguistic bias in NLP development – and the dominance of English in datasets – have long been critiqued (Bender). Building on Zaugg's discussions of the "complexity of inclusion" ("Language Justice" 245) and language justice in digital spaces, we look beyond the incorporation of marginalized languages in the work of NLP development communities (Joshi et al.). We examine specific dilemmas that emerge as a result of intentional or unintentional linguistic inclusion, and the interaction of the languages with various types of predictive language-processing or generative tools. Can we automatically assume that digital inclusion here is essential for the "vitality" of marginalized languages (defined by Zaugg, "Language Justice" 250 in terms of both digital linguistic supports and use) in a global context where linguistic diversity appears to be increasingly threatened (Griffiths)? This question is inherently political. Languages rarely just "die" and historically they have been more often "killed" through dynamics of colonial violence and/or other forms of marginalization. Therefore, what role does automated language processing, prediction and generation play for "minority" languages – cumulatively spoken by the global majority – that have hitherto been under-represented in the development processes of these technologies?

We are interested in features of different languages that make them interact in particular ways with predictive text algorithms, which have been primarily designed for English-language users and content. We approach these questions as (broadly) qualitative social scientists who work in East African linguistic contexts to understand the social, cultural and political aspects of digital transformations in this diverse part of the world. We consider how linguistic features are tied to specific histories of language development, and the ways these histories are connected to global and local power dynamics. While we recognize the Anglophone dominance of the global Internet, the English/non-English Internet binary needs to be problematized, not least because languages other than English are hardly homogeneous. Obvious and subtle differences between African languages – in their scripts, transliterated spellings, their relationship to colonial languages, the status of their dialects, their role as regional lingua franca, and the scope and location of diaspora users – *all* potentially influence how their modern forms interact with externally designed algorithms, often in unexpected ways.

Linguistic marginalization is also multi-scalar and must be understood in relative terms. The languages that we discuss in this article (Amharic, Kiswahili and Somali) could all be considered *globally* "under-resourced" in terms of the availability of labelled datasets for machine-learning training and engagement with them by NLP practitioner communities (Joshi et al.). Nonetheless, these are dynamic digital languages with large communities of users generating significant quantities of (often unlabelled) natural language data online. At a *regional* level these languages also predominate over the multiple other languages spoken by different groups in this part of the continent (see Zaugg, "Language Justice" for a wider picture of relative digital vitality of the many languages of Ethiopia). Amharic, Kiswahili and Somali's existing digital presence allows language models to be built from them, which in turn enables their interaction with predictive text algorithms. The simultaneous development of "multilingual language models" (which apply patterns identified from globally dominant languages to inform outputs in lower-resourced languages) are also increasing the actual and potential interoperability of lower-resourced language with predictive/generative tools (Nicolas and Bhatia). These interactions

"in the wild" with emerging technologies could be seen as helping to sustain (some) African languages in digital spaces. Equally, the unsupervised nature of this interaction might lead to particular types of harm, such as users' everyday exposure to inappropriate predictions that conflict with local cultural norms, or the longer-term ingraining of forms of bias. Furthermore, if extractive "surveillance capitalism" or "data colonialism" (Couldry and Mejias; Zuboff) is increasingly based on natural language processing, the machine-readability (and generate-ability) of marginalized languages is also linked to global power inequalities, which might themselves feed back into dynamics of linguistic vitality or decline.

This article extrapolates a new set of questions from our research on predictive text and three African indigenous languages, contributing to current debates around bias and inequality in predictive/generative text-based technologies. We suggest that our findings on language and predictive text have implications for how to approach the study of the effects of predictive and generative text-based technologies on marginalized languages. We first review a selection of literature that links the interrogation of the power of search engines to directions in NLP/NLG research, and (hitherto limited) engagement with how knowledge of African languages fits into these increasingly intertwined fields. We then recap our research findings on Google Search Autocomplete for three East African languages. Here we provide new reflections on how analysis of specific linguistic features (tied to the languages' cultural, orthographic and political histories) can help us understand how they interact with predictive text algorithms. We conclude by outlining how this approach might be useful for the study of the algorithmic and digital future(s) of specific marginalized languages in a broad, complex and rapidly developing field of research, as well as the study of everyday practice around predictive and generative NLP and NLG technologies.

POWERS OF PREDICTION: SEARCH ENGINES, NATURAL LANGUAGE PROCESSING AND GLOBAL LANGUAGE INEQUALITIES

Outputs of predictive text mining have long been visible to everyday Internet users through search engines that "autocomplete" queries entered into them. These "predictions" to complete or add keywords are based – according to companies such as Google – on what a user enters, their search history (if logged in) and what others are searching for (Google). While it is common knowledge that Google Search results are heavily influenced by companies bidding to have their links positioned highest in searches for specific keywords (Machill et al.), brand names and logos are also increasingly visible in its autocomplete predictions. This indicates the influence of the platform's commercial logics on this part of the information-retrieval process too, even before a user has completed a search query. Our work on how predictive autocomplete algorithms interact with African languages has built on a range of scholarship interrogating the cultural, political and economic power of search in digital societies. Baker and Potts analysed how autocomplete predictions reproduced racist, biased and stereotyped discourses in English-language searches for group-based keywords. Focusing on search results, Safiya Umoja Noble's influential research revealed how Google's commercial imperatives exposed marginalized groups to the harms of algorithmic bias, for instance in the propensity of keywords such as "Black girl" to generate misogynistic, "pornified" and racist results.

Problematizing the semantic distinction between search "predictions" (what Google says it provides for users) and "suggestions" (an external influence on a user's information-seeking behaviour) is helpful in conceptualizing the power of technologies that "think" one step ahead of humans in their choices around language use. Miller and Record's examination of autocomplete highlights its potential influence on users, and the difficulties of relying on user accountability or automated filtering and appropriation for reporting and removing problematic predictions. They note that popular searches (reflected in autocomplete) are not necessarily helpful for the searcher; that feedback loops can occur when autocomplete raises the visibility of certain keywords, making them more likely to be searched for again; and that users are unable to turn off these suggestions (Miller and Record 1950).

Noting that there was little or no direct engagement in this critical scholarship with how predictive search algorithms interact with globally non-dominant and non-European languages, our initial work on Google Autocomplete results for Amharic, Somali and Swahili explored and highlighted possible algorithmic harms (Chonka et al.). We conceptualize these in relation to

the dynamics of digital colonialism, the multiple ways in which global patterns of information and communication technology development reproduce and reinforce historical power imbalances, and violent extractive logics of imperialism (Kwet; Couldry and Mejias; Madianou). This manifests in historical continuities in control over physical digital infrastructure (Mainwaring and Aldrich); exploitative and harmful digital labour practices that enable platforms to function (e.g. content moderation) and underpin developments in AI (Anwar and Graham); as well as data extractivism from African populations viewed as vital growth markets by companies such as Google (Oyedemi).

Mpofu and Salawu's study of the linguistic localization of Google Search in Zimbabwe highlights the dynamics of linguistic colonialism in the Anglo-centrism of cyberspace. The localization they critique gestures towards global multilingual inclusivity, but is ultimately redundant in the wider context of language policies and practices that privilege English globally. In Mpofu and Salawu's case study, linguistic colonialism manifests in a lack of online content or functionality for African languages. However, for our analysis of linguistic contexts on the African continent, the problem is not so much that indigenous language material is lacking, but instead how search engine algorithms (irrespective of interface translations) interact with the African-language keywords that users input to access content in those languages. As the vignette that prefaced this article highlighted, we have conceptualized digital colonialism in search predictions in relation to the unequal exposure of users of globally marginalized languages to algorithmic harms, for instance misogynistic suggested keywords that are not filtered from searches in the same way that they are in English. This work did not call for a technical fix to the problem; given the diversity and nuances of global language cultures, automated mechanisms could never be exhaustive in their protection or moderation capabilities. Rather, we started to build a case for radical transparency in predictive search technologies, whereby companies should be responsible for making data on manual and automated predictive keyword filtering accessible and auditable to relevant linguistic communities.

In this article we think through the potential impact of similar technologies on marginalized languages themselves, as well as beyond search engine autocomplete. Indurkhya's review of emerging directions in predictive text mining indicates the proliferation and increasing scope of these technologies, and highlights that "the boundary between the fields of Text Mining and Natural Language Processing is becoming increasingly blurry" (155). The former has been conventionally understood as the process by which structured data is extracted from (unstructured) language-based corpora, whereas the latter is intended to simulate the human capacity for understanding and generating language. Both of these (interrelated) sets of machine-learning techniques have underpinned the development of predictive and generative language applications, for instance, from the evolution of systems designed to predict users' word input on a limited device interface (e.g. a basic mobile phone number pad), through to the emergence of full smartphone alphanumeric keypads and the consolidation of predictive text as a ubiquitous innovation for timesaving and convenience.

Another shift is visible in the development of predictive text beyond suggestions of single words to longer phrases, and its spread beyond mobile phones to a range of applications such as email platforms and grammar/writing assistants. More recently, public attention has shifted to the increasing capacity of the natural language generating subfield of AI technology, which can produce longer and more complex texts. These either augment human writing or engage users in what appears to be dialogue, as in the case of the high-profile release of OpenAI's ChatGPT in November 2022. This type of "chatbot" is a different technology to a search engine autocomplete algorithm, but both work through the extraction and processing of large text-based corpora derived from the wider Internet, and both operate in a predictive fashion. Writing about applications based on earlier iterations of the GPT language model, Perrotta et al. explain that when

> faced with a prompt, the [language model] system takes into consideration the relationships between words in a sentence – with the process of self-attention allowing the algorithm to estimate how words interact with other words "in context". The system therefore "gambles" on those associations in a probabilistic fashion and produces an output that respects those "rules of self-attention". (Perrotta et al. 6; see also Vaswani et al.)

Google Search predictions and developments in natural language processing/generation are increasingly intertwined because the techniques of the latter are increasingly shaping the operation of the former. Shah and Bender critically analyse Google's proposals (now enacted) to "reimagine search" through large language models that would enable automated "conversational agents" to answer user queries directly, rather than directing users to sources that contain the information sought. They argue that speed and convenience should not be the primary goal of an information-retrieval tool and that "searching is a socially and contextually situated activity with diverse set of goals and needs for support that must not be boiled down to a combination of text matching and text generating algorithms" (Shah and Bender 221). Reducing user interactions with a search engine (and external sources of information) through automated query responses has significant implications for transparency and accountability regarding the information provided. Of particular interest to us are the ways in which power imbalances between users and platforms would be compounded by global inequalities around language and representation, and the ways in which various algorithmic biases affect different communities.

In outlining various sources of bias that affect existing language models (namely in datasets, annotation processes, input representations, models and research design), Hovy and Prabhumoye point out that "most NLP research is still in and on English" and that despite the abundance of data available from other languages, "most NLP tools skew towards English" (Hovy and Prabhumoye n.p.). However, this doesn't mean that applications such as ChatGPT cannot interact with marginalized global languages. Later in this article we discuss examples of this phenomenon that indicate that even limited engagement by NLP development communities with particular languages is not preventing language models drawing from available digital corpora and enabling their functionality. Joshi et al. note that a very small proportion of the over 7000 languages of the world are "represented in the rapidly evolving language technologies and applications" (6282). They quantitatively examine the relationship between language features, language resources, and the representation of languages at NLP conferences to conceptualize a typology of languages that are currently "winning" or "losing". Their findings are presented to encourage these professional fora "to prioritise the resolution of the predicaments [...] so that no language is left behind" (6282). Their paper starts by explicitly highlighting Somali as an example of a language on the disadvantaged side of this digital divide, in terms of its unusual linguistic features, lack of mentions at NLP conferences, and the (then) poor quality of existing machine-translation tools. And yet our work has shown that Somali has a vibrant digital life; it interacts with autocomplete algorithms, and can be engaged directly by an AI text generator such as ChatGPT (even before OpenAI identified Somali as a language it formally supported).

This interoperability might be explained by the development of multilingual language models. These have been used to overcome differences in the scale and quality of digital corpora across languages. By training models on multiple languages simultaneously, multilingual models are designed to identify patterns in higher-resource languages and apply these to other lower-resourced languages. However, these emerging approaches to training across languages have generated concerns about errors, lack of attention to the context of local language speakers, and difficulties in identifying and correcting problems (Nicolas and Bhatia).

As with our Google Search Autocomplete case study, our question in this article is less about a language's lack of operability with such predictive tools, but rather the (non-)existence of safeguards against harms from certain types of probabilistic language generation and potential accountability. This is where the limited practitioner engagement described by Joshi et al. might accentuate risks for certain communities of language users. Their framing of the technical field of NLP in terms of some languages "winning" and others being "left behind" implies that the ability of languages to interact with predictive NLP technologies is inherently good, for instance in terms of their survival in the digital age. But what exactly does inclusion in the field of predictive language technologies actually entail?

We can interrogate this question by examining research (which has almost exclusively focused on English) on the impacts of predictive text applications on languages themselves and their use. The rise of SMS mobile messaging prompted popular concerns that predictive input functions would make "textisms" proliferate, prevent children from learning to spell, and thus degrade

the quality of the English language. However, empirical research with school and college students has largely debunked this perceived impact on literacy (Waldron et al.). The wider semantic impacts of predictive text interventions on how people engage with digital platforms has received surprisingly little research attention. Nonetheless, Arnold et al. ("Predictive Text") have demonstrated through tests with human writers that in certain contexts (e.g. image caption writing) the use of predictive tools might influence how people write. Those who were given suggestions produced shorter captions with fewer words compared to those who were not presented with predictions/suggestions.

Again, through controlled experimentation, Arnold et al. have also demonstrated that manipulating the sentiment of text predictions for certain writing tasks (e.g. restaurant reviews) can bias users towards that sentiment in their eventual input. They also point to a potential recursive "chain" where "biases in training data cause biases in system behaviour, which in turn cause biased human-generated products" ("Sentiment Bias" 1). As predictive language technologies become more pervasive, greater quantities of text are generated by machines and become accessible to those same machine-learning processes to further inform and refine their future outputs. This raises important questions about ingrained feedback loops that we explore in relation to digitally "under-resourced" global languages later in this article. Arguably, the impacts of recursive cycles whereby AI-generated data are increasingly training AI-driven processes are under-theorized, even in relation to English-language-dominated NLP, let alone marginalized global languages. This is a complex research gap, but it is a general area for which we wish to explore a range of new questions about the implications of the evolving and expanding practices of predictive text within global linguistic inequalities. These are informed by our previous work on East African languages and the nuances of their unpredictable interactions with Google Search Autocomplete.

PREDICTION PREDICAMENTS: GOOGLE SEARCH AUTOCOMPLETE AND THREE EAST AFRICAN LANGUAGES

To interrogate the interactions between Google Search Autocomplete and users of different language communities, we empirically analysed its functionality in three African indigenous languages. Autocomplete works across Google's offering of applications and platforms, and has been a core feature of Google Search, not least since a user cannot turn it off. Google has outlined policies specifying that its autocomplete predictions are limited to exclude content that is violent, dangerous, sexually explicit, hateful against groups or disparaging to individuals. Here we show how user–algorithm interactions are mediated by political and cultural-linguistic contexts, creating sometimes quite jarring predictions/suggestions.

Amharic, Somali and Kiswahili have distinct orthographic features and user communities. Amharic is the second most frequently spoken Semitic language after Arabic. It is derived from the ancient language Ge'ez whose abugida-based script gave rise to several Afro-Asiatic and Nilo-Saharan languages. These include more than twenty Ethio-semitic languages in Ethiopia and Eritrea, including Amharic, Tigrinya and Gurage (Abebaw et al.). Amharic has 33 base characters with six derivations of each. It has seven vowels and seven forms (orders) that match these. They represent a combined consonant-vowel and constitute a syllabic system. Overall, Amharic therefore currently comprises around 231 characters including core characters, labialized characters, punctuation marks and numerals (sources differ on the exact number given the dynamism of the language and blurred distinctions between Amharic and other Ethiopian languages that use the same base writing system but include additional and new characters for distinctive sounds). Its user community is predominantly in Ethiopia and global diasporic communities, with at least one-third of Ethiopia's population speaking the language. Amharic has the status of a constitutionally recognized national language in a multi-ethnic country, and is regarded as a national lingua franca.

Somali is a Cushitic language spoken across multiple countries in the region: Somalia, Djibouti and parts of Ethiopia and Kenya. "Standard" Somali combines multiple (mostly "northern") dialects (Lamberti), and was established as an official lingua franca in the 1970s by Somalia's then military regime. It was at the same time that a formalized Latin-based orthography was standardized by the government and came into wider use (Laitin et al.). Contemporary debates focused on whether Somali should use a Latin alphabet, an Arabic-based script or a new and

indigenously developed writing system. Somalia has long been perceived to be characterized by a higher level of linguistic homogeneity by comparison with neighbouring Ethiopia and Kenya, where ethnic diversity is more frequently associated with the use of different languages. While this a problematic oversimplification (there is both linguistic and ethnic diversity in Somalia and a history of marginalization of minoritized groups), the standardization of Somali has contributed to the predominance of this language across most of the politically dominant clans. "Standard" Somali is also shared across boundaries in the context of political fragmentation. For example, while the (internationally unrecognized) Republic of Somaliland has claimed its independence from Somalia since 1991, this split is conceived of primarily in political and not ethno-linguistic terms (Chonka and Healy), notwithstanding some north/south Somali dialect differences that we discuss below.

Like Somali, Kiswahili is spoken across multiple countries in East Africa. It is a national language in the DRC, Kenya, Tanzania and Uganda. Unlike Somali and Amharic, a greater number of Swahili speakers speak it as a second language. Kiswahili is a Bantu language and includes loanwords from English and Arabic, reflecting its history as a language used for trade and the colonial penetration of East Africa. Its written form uses a Latin orthography, though in the past it has been written using Arabic script. Currently, there is little difference in the spelling of proper names in Kiswahili and English.

Noting different features and contexts of use for each language, we chose not to conduct a systematic comparative test, but instead selected a sample of politicized and gender-related keywords specific to each. Testing was completed in the UK using private browser and logged-out user settings, with an additional Somali test in Somalia to consider the effect of user location. We did not complete the searches or click through to any of the results so as not to influence subsequent search results. The full results are published in Chonka et al. and the underlying datasets are available online.[1]

Our results indicate how Google Search Autocomplete predictions interact with the orthographic features of each language, as well as the political and cultural-linguistic contexts in which the languages are used. This results in variation in *how* the language is included in the algorithmic system. Small differences between the languages appeared to have a big impact on the types of predictions that users would see – for instance in terms of the frequency of predictions in those same languages (versus predicted English terms). In turn, this had an impact on the likelihood of the appearance of problematic or potentially harmful content to users.

The presence of filtering varied across the languages, with predictions for Amharic, a language with its own script and fewer Google users using that script, appearing to be hardly moderated at all. For all three languages, we tested Autocomplete results for keywords related to gender, including "woman", "girl" and "man". Our results revealed that at the time of inquiry, Autocomplete did not filter out sexualized and violent content in two of the three languages: Amharic and Somali. In Amharic, almost all of the Autocomplete predictions associated with "woman", "women" and "girl" were sexually explicit. For example, predicted results for the keyword "woman" in Amharic included "to satisfy a woman" and "to f**k a woman". Sexualized results also appeared in the Somali search for girl, for example, "virgin" and "naked". In contrast, there were no sexually explicit predictions for the Kiswahili search.

Sexualized predictions were not the only potentially problematic content that emerged, and this varied across the languages. A more complicated and context-specific case of questionable predictions materialized in the search involving politicized keywords in Somali. We tested search predictions for the names of ten male politicians in Somalia. Of these, six returned clan-related keywords, most often the Somali noun for "clan" (*qabiil*), and in some cases the names of specific clans. Such predictions could potentially influence a user's search and increase the likelihood that they would use this suggested keyword to find information about a political figure's clan identity. Clan divisions have played a role in state-collapse and conflict in Somalia since the 1980s (Kapteijns), and have a longer history of manipulation and instrumentalization by colonial powers, and in post-independence state-building periods (Samatar). Local news

1 Amharic: https://docs.google.com/spreadsheets/d/164SmYiFQW8mIp4F2rjU3nLdZXYAAC1ahO2iiBtqSJDU/edit?usp=sharing; Kiswahili: https://docs.google.com/spreadsheets/d/12-T32irGX99TEBqvXtOWhXYFSS2HZkFDG1LOtUbSE6I/edit?usp=sharing; Somali: https://docs.google.com/spreadsheets/d/14lm7Q0HfLMmOQ3xKHAu3_L1i8Vevj8eJjt3pnQVXCxE/edit?usp=sharing.

media tend to avoid explicitly discussing clan identities in broadcasts and open discussion of clan politics remains something of a taboo – even if these are semi-institutionalized in modern political power-sharing agreements, and are very much open for debate on social media. Clan has been understood as an "invisible" identity marker (Lewis) because clan differences are generally not characterized by clear linguistic or cultural/religious variance (unlike for many ethnic groups in neighbouring countries).

The fact that Somali names are spelled differently in Somali (as opposed to their English transliterations) helped to determine the language of the Autocomplete predictions. Predictions for names of high-profile figures whose names had a similar spelling in Somali and English (e.g., the nickname of the former President "Farmajo") tended to be dominated by English keywords, also indicating their higher international profile and reflecting the search patterns of non-Somali-speaking users.

The visibility of "clan" in the Somali Autocomplete predictions contrasts with the Kiswahili results around a similar theme. In Kiswahili, we similarly tested using the names of prominent male Kenyan politicians. In Kenya, ethnic identities are intertwined with political contestation, and, at times, this has manifested in violence (Mueller; Kanyinga). Under colonial rule, legal and administrative instruments divided people arbitrarily along tribal lines, and ethnic identity has continued to be used in post-independence politics to manipulate group belonging and division. There were no explicit references to ethnicity in the Kiswahili search results for the names of prominent Kenyan politicians. This could be due to a number of factors related to the language and the user base, which contrast with the Somali political and linguistic context of "clan". While talking explicitly about ethnicity in Kenya has been increasingly critiqued, especially following post-election violence in 2007, ethnicity is arguably more "visible" than (Somali) clan identity, in that it is often characterized by language difference. The overlap between the English and Kiswahili spelling of names might also be significant. External English-language searches for prominent politicians might result in fewer Kiswahili predicted keywords. Additionally, the position of English as an official language in Kenya (unlike Somalia) means that much of the mainstream media space is taken up by English content, and this might be reflected in search predictions and the language data that underpin them.

Looking across the three languages, dilemmas emerge between the prospects for their digital vitality through "unsupervised" interactions with platform algorithms, and potentials for bias and user exposure to harm through automated text generation. Different forms of harmful content emerged through the user–algorithm interaction. In Amharic, sexualized and misogynistic search results indicate a lack of attention to (and lack of sophistication of) algorithmic filtering in the Amharic script. In Somali, we see an additional potential area of politically problematic suggestions/predictions that emerges in relation to Autocomplete predictions about "clan". The presence of such markers indicates how local political contexts are part of the creation of Autocomplete suggestions, within the underlying dataset and alongside individual search behaviours. A lack of moderation of (or informed community engagement with) automated language generation in complex and conflicted contexts can allow for the unexpected emergence of different forms of harmful predictions. Some, as with the gendered results, might be received with a relatively strong social consensus that such content is inappropriate. Others, such as "clan", might be more ambiguous and raise difficult questions about the productive effects of predictions, especially where algorithms generate political content that might recursively influence discourses in the digital public.

While harmful predictions in our testing of Kiswahili were less apparent, its "inclusion" within predictive algorithms has other potentially negative effects on its digital vitality. Kiswahili search results were more intertwined with predictions in other languages. This probably reflects multiple factors, some tied to the language itself, which is written in Latin script and shares words with English and Arabic. Also factors linked to the user community could help to explain this difference, given that (outside of Tanzania) many Kiswahili speakers use it as a second language. The result in Kiswahili was a set of predictions that suggested to users an opportunity to switch to another language. Some predictions took the form of a "correction" to the Kiswahili keyword, assuming that a user intended to type a different word. For example, within the gender-based keywords, Autocomplete prediction mistook a Kiswahili common noun (*mke*, wife) for an indexed proper noun (Mike Tyson, or the Turkish football team, MKE Akaragucu).

For politicians' names, when a name was combined with a conjunction (e.g., "Kenyatta na", "Kenyatta and"), the conjunction was translated into English in some predictions. Here results for Kiswahili users blurred with other languages in a multilingual digital search ecosystem, with Autocomplete predictions directing users potentially away from cultural-linguistically specific results.

BEYOND "INCLUSION": POLITICIZING AND HISTORICIZING THE DIGITAL FUTURES OF GLOBALLY MARGINALIZED LANGUAGES

Regarding predictive Autocomplete, the experiences of these three East African languages vary in important ways. The challenges and opportunities for the different languages cannot be considered only through the lens of inclusion in algorithmic learning. This section argues how other factors related to the orthographic history of the language and user context come into play in our results, and point to the need to re-politicize how linguistic inclusion is considered and analysed in relation to predictive text algorithms. We outline this position by discussing three factors that inform and shape our results: orthography, user communities and datasets/corpora. We suggest that interrogating the political and social dimensions of each factor, specific to each language, can help to make sense of predictive text outcomes and effects. This is intended as a starting point for illustrating how the operation and development of predictive text algorithms are tied to political and cultural-linguistic contexts. In this, we shift attention away from simple questions of linguistic inclusion to debates over the algorithms themselves, and how they reflect, intersect with and reinforce forms of power and inequality.

LANGUAGE FEATURES, ORTHOGRAPHY, SPELLINGS, DIALECTS

Our results point to the heterogeneous ways in which "neglected" languages – those that tend not to inform most NLP development and/or have smaller corpora of high-quality data for training models – interact with predictive text algorithms, with varying outcomes related to algorithmic harms and linguistic vitality. Amharic and Somali seemed to show the least applied content moderation, evident primarily around the number of sexually explicit predictions. Kiswahili diverged: with less clearly problematic content, but also potentially less relevant results for a Kiswahili-speaking user community. All three languages examined have been considered to be "relatively ignored" in papers presented at NLP conferences. Joshi et al. categorize Amharic as one of the least represented languages with nine structural/semantic features that are not discussed in most conference papers. Kiswahili is identified as having eight "ignored" features, but with a slightly lower rate of error in the translation of English into Kiswahili (45.64%), versus into Amharic (60.71%). However, while both Kiswahili and Amharic have relatively similar levels of "ignored" features, and relatively similar numbers of estimated speakers (18 million versus 22 million, respectively [Joshi et al.]), filtering of predictions – especially of gender-related terms – differed. While rates of error in translation might play a role, this difference nonetheless indicates the importance of interrogating more closely if and how specific orthographic features pose problems for NLP models, and what this means for languages' digital vitality.

Features of the modern written scripts of these three languages are tied to their specific histories of development, change and contestation. For Somali, for example, the decision in the 1970s to officially adopt a Latin alphabet (over an Arabic-based script or a new indigenous orthography) was highly controversial, but portentous for the language's later presence in digital spaces, given that the Latin alphabet was the "native" script of early digital technology developments. Diversity and survival of dialects or sub-national languages (themselves often tied to geographical or political divisions) might be influenced by the ways in which predictive algorithms amplify biases in language corpora, making some more dominant in digital publics. This is potentially of great significance in some of the multilingual and multi-ethnic East African contexts we examine, with their histories of violent fracture and ongoing state contestation or reconstruction in which ethnicity and language have been frequent fault-lines. Dialect diversity can also have a profound but very unpredictable impact on how predictive algorithms respond to user inputs. The test that we conducted with the Somali word for "girl" used the northern dialect spelling (*gabadh*). The same word using the southern dialect spelling (*gabar*) produces an entirely different set of Autocomplete predictions, none of which are in Somali

(Figures 1 and 2). It is important to note here that while the "naked" predicted keyword has (at the time of writing) disappeared from the *gabadh* predictions (possibly because of user behaviour, or content moderation applied after we drew attention to this in previous research), a whole host of different but equally problematic suggestions were still being offered to the user at the time of writing (2024), including "girl [who was] raped" and "girl f**ked Facebook".

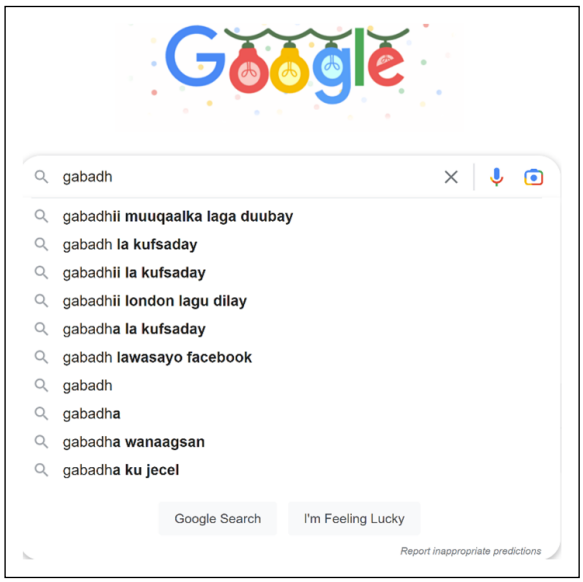


Figure 1 Google Search Autocomplete suggestions for Somali word *gabadh*.

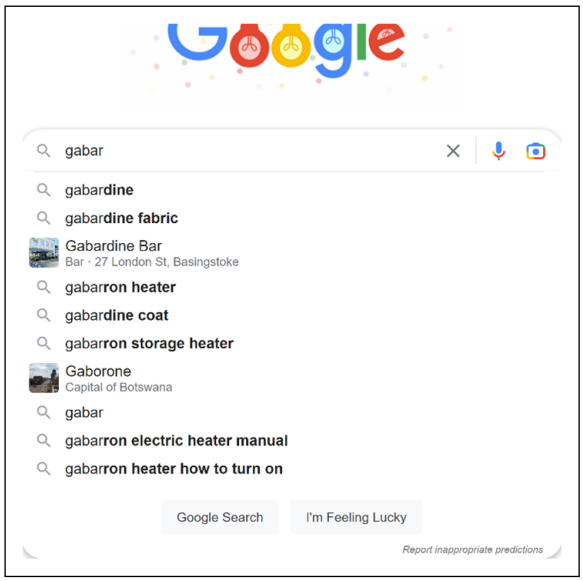


Figure 2 Google Search Autocomplete suggestions for Somali word *gabar*.

Both of these spellings (*gabadh*/*gabar*) are "correct", found in Somali dictionaries, and are equally common. The former's use of the Somali alphabet letter *dh* (a voiced retroflex plosive sound), which is less common in English, skews those results towards Somali predictions, whereas *gabar* resembles a wider range of words from different languages. We (and probably the NLP community) have barely begun to scratch the surface of how small dialectical variations might influence predictive algorithms. This includes uncertainty about how these algorithms might be shaping the ways in which dialects are used (or become less used) in digital publics, and how this maps on to political contestation over language use in territories struggling with reconciling diversity (e.g., Somalia or Ethiopia).

WHO IS USING THESE LANGUAGES? IN WHICH (DIGITAL) SPACES? FROM WHERE?

While all three languages are national languages for specific countries in the region, they are also spoken by significant global diasporic communities. We examined the impact of location on Google Search, comparing prediction results from Somali-language tests done in the UK and in Somalia. Here we found no significant differences. This indicates a lack of location-based

individualization, potentially suggesting greater independence from algorithmic moderation but also greater geographical reach of problematic results. It is possible that a significant amount of Somali-language use of Google Search is conducted by Somali speakers in the diaspora. Research has examined how geographically dispersed communities might contribute to conflict in "homelands" through their online engagement in hate speech, incitement or active mobilization for violence (or, conversely, peacebuilding efforts) (Osman). The online traces of these behaviours might influence digital publics in the Horn of Africa, as they are reflected through algorithmic language predictions that become visible to users there. This raises new questions for digital (conflict) diaspora researchers, but also for linguists around the transnational aspects of the global use of marginalized languages and their impact on the language models that underpin NLP.

The geographical distribution and patterns of use of Kiswahili as a regional lingua franca warrant a different set of considerations about the relative influence of various groups of Kiswahili speakers on Autocomplete predictions. Kiswahili is the native language of Swahili people along the coastal region of East Africa. However, being spoken by the majority in wider East Africa as a second language, and interspersed with loanwords from other languages, it is not surprising that Kiswahili Autocomplete results are not only in Kiswahili. This also raises questions about which issues or topics different user communities are likely to discuss in their native, versus a second, language. The lack of geographical/contextual specificity in the Kiswahili results makes it difficult to identify which user communities might have a greater influence on predictions. Irrespective of this, the results of the Autocomplete algorithm do not encourage the use of the language as clearly as in either Somali or Amharic. Instead, some of the predictions seem to reflect the use of Kiswahili in combination with other languages. This could have implications for the user communities for whom Kiswahili is a primary or native language, as well as for the vitality of Kiswahili as a digital language.

CORPORA FOR LANGUAGE MODELS

The functioning of Autocomplete in the three languages of our study indicates that there are digital data in these languages which are being drawn upon to train predictive algorithms. For instance, as individuals are using Google Search in the Amharic script, this contributes to algorithmic-based predictions for new queries. This is evident in the very fact that the Autocomplete algorithm produces results: it has data to draw on to make predictions. Autocomplete algorithms rest on assumptions about the underlying sources of data from past and ongoing queries, and their relationship to an individual's future query. As has been much more widely critiqued (Eubanks), large datasets retain biases that reflect inequalities in the social, cultural and political contexts in which they have been constructed. An added complication in our results is that corpora for marginalized languages are probably smaller or less visible, which could result in the over-representation of particular types of text or types of user. This is especially a concern given the diverse geographical distribution of these language-speakers, and in contexts where political/societal conflict have fractured states, often along ethno-linguistic lines.

Technology companies have responded to the under-representation of certain languages by incentivizing the development of non-English datasets,[2] for which, of course, they also have a commercial incentive. However, this distracts from the more fundamental challenge: that any dataset premised on prior human interactions will reflect societal biases. If searching is a contextually situated activity (Shah and Bender), the datasets that are generated through search behaviour also reflect specific contexts, both in who/what is present in the dataset, and who/what is excluded or unrecognized. The languages we examined might not only be limited by relatively smaller underlying corpora. These corpora might also reflect multiple biases – both general biases that are present in any dataset based on human behaviour, and specific biases tied to the potential over-/under-representation of different language or dialect speakers in different settings. This potential is evident in the varied results across languages. However, when considered within the sociopolitical contexts of language use, exactly what constitutes these datasets and how they relate to results remain unclear, particularly given the opacity of proprietary predictive algorithms.

2 For example https://lacunafund.org/about/ (accessed 24 September 2025).

QUESTIONING THE FUTURE OF AI NATURAL LANGUAGE GENERATION FOR MARGINALIZED LANGUAGES

Our previous research explored Google Search Autocomplete because this – at the time – appeared to be the most accessible and immediately visible example of machine-learning processes predicting and generating text in East African indigenous languages. In the relatively short space of time since we conducted the tests discussed above, more technologies have become available for public use that can interact with these languages. Prominent among these have been "conversational" AI models such as ChatGPT – technologies with the ability to automatically predict user information retrieval needs and synthesize answers for them out of available online data and LLMs, as opposed to directing those users to specific sources. ChatGPT garnered significant international attention and generated much debate about the extent to which it represented a step-change in the capabilities of natural language generating technology. Since then, multiple companies have launched their own LLMs, including X, Meta, Google, Baidu and Alibaba, as well as companies specifically working on generative AI (e.g. Anthropic). While more systematic experimentation with ChatGPT and other LLMs is necessary, our initial interactions have shown the former to be capable of generating grammatically correct and semantically coherent responses to questions posed to it in, for example, the Somali language. This capacity has developed significantly in the period in which this article was under review, and Somali is now officially listed as a supported ChatGPT language. Similarly, our recent engagement with ChatGPT in Amharic also illustrates a significant improvement from our initial testing, which had generated content almost totally unrelated to the prompts. Currently it can produce fairly meaningful and convincing responses, albeit often including some scrambled texts and misspellings. As we noted above, while search autocomplete algorithms and AI natural language generation are different technologies, both are based on extraction from available natural language text data online and both have predictive features. Furthermore, artificial conversational agents are being positioned by companies such as Google as the future of searching. The implications of this shift (made all the more imaginable by the uncanny capacities of tools such as ChatGPT to respond to human questions) are both potentially profound and concerning, as Shah and Bender outline.

Given the rapid development of NLG technologies, it is perhaps unsurprising that there is not yet sufficient research on their impact on marginalized languages and their global communities of users – or research that steps beyond analysis of their "inclusion" in the development processes of the technologies. There is often an implicit assumption that in order for marginalized languages to survive in the digital age, they will need to interact with NLP/NLG technologies. This could occur through the conscious efforts of designers to incorporate these languages in their models, or take place "in the wild" as unsupervised machine learning engages with the online natural language data that is generated by these communities. Furthermore, multilingual language models are increasing the capacity of generative AI to draw on data across different linguistic datasets.

NLG is rapidly becoming the future of information retrieval (and evolutions of multiple other productive sectors) and it is increasingly difficult for users to opt out of exposure to automatically generated content. Further, language translation seems to be a central and emerging use of generative AI. A 2025 Microsoft-led study on conversations between US-based users and their generative AI system suggests that generative AI is most applicable to fulfilling the role of a translator or interpreter, versus other occupations (Tomlinson et al.). It is therefore perhaps obvious that there is a need for globally non-dominant languages to develop a presence and visibility through these applications.

Nonetheless, a future research agenda could interrogate whether the proliferation of NLG technologies might actually reinforce linguistic inequalities and the dominance of English (and other over-represented digital languages) at the expense of "minority" languages. Considering minoritization at both global and regional scales, this article has shown that there is a huge range of variables – linguistic, cultural, political, historical and geographical – that might shape how different globally marginalized languages can interact with predictive/generative text algorithms, beyond the activity and activism of those developer communities who might wish to include them in language modelling. We have emphasized the heterogeneity of a selection of African languages (from just one sub-region of the continent) to illustrate important

differences that relate to these factors and that might determine the nature of their presence and visibility in language models.

An even broader research agenda could interrogate the desirability of inclusion in NLG from a perspective that foregrounds global and local power imbalances and how these are reflected through language. Here it might be instructive to look at the evolution of discussions about inclusion in datasets for other applications of machine learning. For example, scholars have long pointed out that the efficacy of machine-learning-based biometric/facial recognition technologies has been compromised by the whiteness of the image-based datasets that train the algorithms (Buolamwini and Gebru). Responses to this form of algorithmic bias have often focused on increasing data "inclusivity", that is, adding more Black faces to image datasets, to improve the functionality of systems. But if this technology can be considered to be inherently oppressive – as exemplified by its deployment by security services in ways that undermine individual privacy and autonomy, or enable ever more sophisticated forms of racial profiling – is "inclusion" here a good thing? This is a live question globally, for example, in African contexts, as Chinese companies (world leaders in facial recognition) strike deals with governments to provide digital security infrastructure in exchange for access to the "training data" this will capture (Jili).

Can linguistic inclusion in language models be considered an analogous example? Given the precarity of many at-risk global minority languages (and the possible benefits of digitization), a clear answer to this question might prove elusive. Nonetheless, some scholars are beginning to encourage investigation from this more critical starting point. In her discussion of digitization for Amharic, Zaugg points out that "while the benefits of these supports for digitally-disadvantaged language communities are clear, the reality is that standardised script use, standardised spelling, and NLP systems in particular increase a language community's legibility for digital surveillance" ("Digital Surveillance" 227). Such surveillance can be undertaken by military/state security, corporate and humanitarian actors, but there is also a potential trade-off in relation to machine readability for effective content moderation. This is a key theme that emerged from our research on Autocomplete predictions and possible protections against harm. Elsewhere (Chonka et al.) we have indicated our scepticism that the relevant tech companies can guarantee that their predictive language algorithms will not make unexpected and potentially harmful interventions into digital publics in contexts affected by instability and ongoing conflict. We also doubt that they would or could enable full transparency to relevant linguistic communities around how moderation/filtering is taking place in relation to complex and ambiguous content in specific cultural contexts. Here we speculate about how trajectories of unsupervised NLP/NLG machine learning might be developing "in the wild", and in ways that might not easily be rolled back. Consideration of the multiple factors that influence how certain languages come to be engaged by these rapidly proliferating technologies will be vital for understanding the impacts on these languages themselves, their worldwide user communities, and what can and should be done to protect global linguistic diversity.

AUTHOR AFFILIATIONS

Peter Chonka orcid.org/0000-0001-9871-1377
King's College London, UK

Stephanie Diepeveen orcid.org/0000-0002-5468-9312
King's College London, UK; University of Cambridge, UK

Yidnekachew Haile orcid.org/0000-0002-7927-7357
Royal Holloway University of London, UK; Teesside University London, UK

REFERENCES

Abebaw, Zeleke, et al. "Transliterating Latin to Amharic Scripts Using User-defined Rules and Character Mappings." *International Journal of Digital Libraries*, vol. 24, no. 1, 2023, pp. 63–75. DOI: https://doi.org/10.1007/s00799-023-00346-5

Anwar, Mohammad Amir, and Mark Graham. *The Digital Continent: Placing Africa in Planetary Networks of Work*. Oxford University Press, 2022. DOI: https://doi.org/10.1093/oso/9780198840800.003.0007

Arnold, Kenneth C., et al. "Predictive Text Encourages Predictable Writing." In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, 2020, pp. 128–38. DOI: https://doi.org/10.1145/3377325.3377523

Arnold, Kenneth C., et al. "Sentiment Bias in Predictive Text Recommendations Results in Biased Writing." In *Proceedings of the 44th Graphics Interface Conference*. Canadian Human–Computer Communications Society, 2018, pp. 42–49.

Baker, Paul, and Amanda Potts. "'Why Do White People Have Thin Lips?' Google and the Perpetuation of Stereotypes via Auto-Complete Search Forms." *Critical Discourse Studies*, vol. 10, no. 2, 2013, pp. 187–204. DOI: https://doi.org/10.1080/17405904.2012.744320

Bender, Emily M. "On Achieving and Evaluating Language-independence in NLP." *Linguistic Issues in Language Technology*, vol. 6, no. 3, 2011, pp. 1–26. DOI: https://doi.org/10.33011/lilt.v6i.1239

Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 2018, pp. 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html. Accessed 24 September 2025.

Chonka, Peter. "Digital Governance and Security in the Horn of Africa." *Rift Valley Institute/XCEPT*, 2025. https://riftvalley.net/wp-content/uploads/2025/03/RVI-250328-XCEPT-Chonka-Digital-Governance_Final.pdf. Accessed 9 October 2025.

Chonka, Peter, and Sally Healy. "Self-determination and a Shattered Star: Statehood and National Identity in the Somali Horn of Africa." *Nations and Nationalism*, vol. 27, no. 1, 2021, pp. 61–79. DOI: https://doi.org/10.1111/nana.12646

Chonka, Peter, Stephanie Diepeveen and Yidnekachew Haile. "Algorithmic Power and African Indigenous Languages: Search Engine Autocomplete and the Global Multilingual Internet." *Media, Culture & Society*, vol. 45, no. 2, 2023, pp. 246–65. DOI: https://doi.org/10.1177/01634437221104705

Couldry, Nick, and Ulises A. Mejias. *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford University Press, 2019. DOI: https://doi.org/10.1515/9781503609754

Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St Martin's Publishing Group, 2018.

Google. "How Google Autocomplete Predictions Work." https://support.google.com/websearch/answer/7368877. Accessed 30 May 2021.

Griffiths, James. *Speak Not: Empire, Identity and the Politics of Language*. Zed Books, 2021. DOI: https://doi.org/10.5040/9781350237865

Hormuud Telecom. "Garaad macmaal." (AI) X post/promotional video, 16 May 2024. https://x.com/Hormuud/status/1791153645284716796

Hovy, Dirk, and Shrimai Prabhumoye. "Five Sources of Bias in Natural Language Processing." *Language and Linguistics Compass*, vol. 15, no. 8, 2021, e12432. DOI: https://doi.org/10.1111/lnc3.12432

Indurkhya, Nitin. "Emerging Directions in Predictive Text Mining." *WIREs Data Mining and Knowledge Discovery*, vol. 5, no. 4, 2015, pp. 155–64. DOI: https://doi.org/10.1002/widm.1154

Jili, Bulelani. "The Rise of Chinese Surveillance Technology in Africa." *Electronic Privacy Information Centre*, 31 May 2022. https://epic.org/the-rise-of-chinese-surveillance-technology-in-africa/. Accessed 18 December 2022.

Joshi, Pratik, et al. "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*. Association for Computational Linguistics, 2020, pp. 6282–93. DOI: https://doi.org/10.18653/v1/2020.acl-main.560

Kanyinga, Karuti. "The Legacy of the White Highlands: Land Rights, Ethnicity and the Post-2007 Election Violence in Kenya." *Journal of Contemporary African Studies*, vol. 27, no. 3, 2009, pp. 325–44. DOI: https://doi.org/10.1080/02589000903154834

Kapteijns, Lidwien. *Clan Cleansing in Somalia: The Ruinous Legacy of 1991*. University of Pennsylvania Press, 2012. DOI: https://doi.org/10.9783/9780812207583

Kwet, Michael. "Digital Colonialism: US Empire and the New Imperialism in the Global South." *Race & Class*, vol. 60, no. 4, 2019, pp. 3–26. DOI: https://doi.org/10.1177/0306396818823172

Laitin, David D., et al. *Politics, Language, and Thought: The Somali Experience*. University of Chicago Press, 1977.

Lamberti, Marcello. *Die Somali-Dialekte: Eine Vergleichende Untersuchung*. Buske Verlag, 1986.

Lewis, Ioan. "Visible and Invisible Differences: The Somali Paradox." *Africa*, vol. 74, no. 4, 2004, pp. 489–515. DOI: https://doi.org/10.3366/afr.2004.74.4.489

Machill, Marcel, et al. "Search-Engine Research: A European-American Overview and Systematization of an Interdisciplinary and International Research Field." *Media, Culture & Society*, vol. 30, no. 5, 2008, pp. 591–608. DOI: https://doi.org/10.1177/0163443708094010

Madianou, Mirca. "Technocolonialism: Digital Innovation and Data Practices in the Humanitarian Response to Refugee Crises." *Social Media + Society*, vol. 5, no. 3, 2019, pp. 1–13. DOI: https://doi.org/10.1177/2056305119863146

Mainwaring, Sarah, and Richard J. Aldrich. "The Secret Empire of Signals Intelligence: GCHQ and the Persistence of the Colonial Presence." *The International History Review*, vol. 43, no. 1, 2021, pp. 54–71. DOI: https://doi.org/10.1080/07075332.2019.1675082

Miller, Boaz, and Isaac Record. "Responsible Epistemic Technologies: A Social-Epistemological Analysis of Autocompleted Web Search." *New Media & Society*, vol. 19, no. 12, 2017, pp. 1945–63. DOI: https://doi.org/10.1177/1461444816644805

Mpofu, Phillip, and Abiodun Salawu. "African Language Use in the Digital Public Sphere: Functionality of the Localised Google Webpage in Zimbabwe." *South African Journal of African Languages*, vol. 40, no. 1, 2020, pp. 76–84. DOI: https://doi.org/10.1080/02572117.2020.1733833

Mueller, Susanne D. "The Political Economy of Kenya's Crisis." *Journal of Eastern African Studies*, vol. 2, no. 2, 2008, pp. 185–210. DOI: https://doi.org/10.1080/17531050802058302

Nicolas, Gabriel, and Bhatia Aliya. "Lost in Translation: Large Language Models in Non-English Content Analysis." *Centre for Democracy and Technology*, 23 May 2023. https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/. Accessed 16 December 2023.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018.

Osman, Idil. *Media, Diaspora and the Somali Conflict*. Springer, 2017. DOI: https://doi.org/10.1007/978-3-319-57792-0

Oyedemi, Toks Dele. "Digital Coloniality and 'Next Billion Users': The Political Economy of Google Station in Nigeria." *Information, Communication & Society*, vol. 24, no. 3, 2021, pp. 329–43. DOI: https://doi.org/10.1080/1369118X.2020.1804982

Perrotta, Carlo, et al. "Artificial Intelligence and the Affective Labour of Understanding: The Intimate Moderation of a Language Model." *New Media & Society*, vol. 26, no. 3, 2022, pp. 1585–1609. DOI: https://doi.org/10.1177/14614448221075296

Samatar, Abdi Ismail. "Destruction of State and Society in Somalia: Beyond the Tribal Convention." *The Journal of Modern African Studies*, vol. 30, no. 4, 1992, pp. 625–41. DOI: https://doi.org/10.1017/S0022278X00011083

Shah, Chirag, and Emily M. Bender. "Situating Search." In *ACM SIGIR Conference on Human Information Interaction and Retrieval*. Association for Computing Machinery, 2022, pp. 221–32. DOI: https://doi.org/10.1145/3498366.3505816

Tomlinson, Kiran, et al. "Working with AI: Measuring the Occupational Implications of Generative AI." arXiv preprint, 10 July 2025. DOI: https://doi.org/10.48550/arXiv.2507.0793

Vaswani, Ashish, et al. "Attention is All You Need." *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, 2017. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. Accessed 24 September 2025.

Waldron, Sam, et al. "Use of Predictive Text in Text Messaging over the Course of a Year and Its Relationship with Spelling, Orthographic Processing and Grammar." *Journal of Research in Reading*, vol. 40, no. 4, 2017, pp. 384–402. DOI: https://doi.org/10.1111/1467-9817.12073

Zaugg, Isabelle. "Digital Surveillance and Digitally-Disadvantaged Language Communities." In *Proceedings of the Language Technologies for All (LT4All)*. UNESCO, 2019, pp. 227–30.

Zaugg, Isabelle. "Language Justice in the Digital Sphere." In *Global Language Justice*, edited by L. H. Liu and A. Rao. Columbia University Press, 2023, pp. 244–73. DOI: https://doi.org/10.7312/liu-21038-011

Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Profile Books, 2019.

TO CITE THIS ARTICLE:
Chonka, Peter, Stephanie Diepeveen and Yidnekachew Haile 2026 "No Language Left Behind?" Predictive Text, Generative AI and Dilemmas of Digital Inclusion for Marginalized Languages. *Modern Languages Open*, 2026(1): 6 pp. 1–16. DOI: https://doi.org/10.3828/mlo.v0i0.465

Published: 11 March 2026

COPYRIGHT:
© 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

Modern Languages Open is a peer-reviewed open access journal published by Liverpool University Press.