Versteasch du mi? Computational and Socio-Linguistic Perspectives on GenAI, LLMs, and Non-Standard Language

Verena Platzgummer **John McCrae** **Sina Ahmadi**
University of Galway University of Zurich
{verena.platzgummer,john.mccrae}@universityofgalway.ie sina.ahmadi@uzh.ch

Abstract

The design of Large Language Models (LLMs) and generative artificial intelligence (GenAI) has been shown to be "unfair" to less-spoken languages (Petrov et al., 2023) and to deepen the digital language divide (Bella et al., 2023). Critical sociolinguistic work has also argued that these technologies are not only made possible by prior socio-historical processes of linguistic standardisation, often grounded in European nationalist and colonial projects (Migge and Schneider, 2025), but also exacerbate epistemologies of language as "monolithic, monolingual, syntactically standardized systems of meaning" (Schneider, 2024, p. 5). In our paper, we draw on earlier work on the intersections of technology and language policy (Kelly-Holmes, 2019) and bring our respective expertise in critical sociolinguistics and computational linguistics to bear on an interrogation of these arguments. We take two different complexes of non-standard linguistic varieties in our respective repertoires–**South Tyrolean** dialects, which are widely used in informal communication in South Tyrol, Italy (Alber et al., 2024), as well as varieties of **Kurdish**–as starting points to an interdisciplinary exploration of the intersections between GenAI and linguistic variation and standardisation. We discuss both how LLMs can be made to deal with non-standard language from a technical perspective, and whether, when or how this can contribute to "democratic and decolonial digital and machine learning strategies" (Migge and Schneider, 2025, p. 12), which has direct policy implications.

1 Introduction

The rapid advancement of Large Language Models (LLMs) and Generative Artificial Intelligence (GenAI) has transformed digital communication, yet these technologies systematically privilege standardized and, in computational linguistic terms, high-resource languages while marginalizing millions of speakers who communicate primarily through non-standard varieties and dialects. Recent scholarship demonstrates that LLM design is fundamentally "unfair" to less-spoken languages (Petrov et al., 2023) and deepens the digital language divide (Bella et al., 2023). Critical sociolinguistic work argues that these technologies not only emerge from historical processes of linguistic standardization rooted in colonial and nationalist projects (Migge and Schneider, 2025), but also reinforce epistemologies of language as "monolithic, monolingual, syntactically standardized systems of meaning" (Schneider, 2024, p. 5). When AI systems fail to process non-standard varieties, they exclude speakers from full participation in digital citizenship. Further, technical decisions made in the architecture of models impose a "tokenization tax" that increases costs and degrades performance for languages less represented in the input data to LLMs.

This paper brings together critical sociolinguistics and computational linguistics to examine these issues through two complementary case studies: South Tyrolean German and Kurdish varieties. South Tyrolean dialect comprises non-standard forms of German widely used in informal communication in South Tyrol, Italy (Alber et al., 2024), yet remains largely absent from LLM training data despite being the primary medium of everyday interaction. Kurdish represents a dialect continuum spoken by over 40 million people across multiple nation-states, characterized by orthographic diversity, systematic political suppression, and acute digital underrepresentation. While Central Kurdish (Sorani) and Northern Kurdish (Kurmanji) have achieved modest computational resources, varieties such as Southern Kurdish, Hawrami, and Zazaki remain almost entirely invisible to language technol-

ogy. Through these cases, we analyze the technical barriers preventing LLMs from processing non-standard varieties, examine how evaluation frameworks reproduce standardization biases, and explore policy implications for diverse actors—from Big Tech and states to civil society and academia.

Building on earlier work examining technology and language policy intersections (Kelly-Holmes, 2019), we argue that addressing LLM marginalization of non-standard varieties requires more than technical fixes; it demands a reorientation toward "democratic and decolonial digital and machine learning strategies" (Migge and Schneider, 2025) that center linguistic agency and treat variation as constitutive of human communication rather than noise to eliminate. This has direct policy implications, from requiring "dialect gap" reporting to ensuring community data sovereignty. Ultimately, whether LLMs should "handle" non-standard varieties is not primarily a technical question but a fundamentally political one, with implications for digital sovereignty, cultural preservation, and social justice in an age where the choice between enforcing standardization or embracing variation will shape the vitality of the world's linguistic diversity in digital spaces. In the following sections, we will first address sociolinguistic (section 2) and computational linguistic perspectives on standard and non-standard language and language technologies (section 3), before giving an overview of the policy landscape in relation to linguistic variation and LLMs (section 4). We will proceed by examining the cases of South Tyrolean dialects (section 5) and varieties of Kurdish (section 6) and end with overarching conclusions and policy implications.

2 Sociolinguistic Perspectives on (Non-)Standard Language and LLMs

The concept of *standard language*, as well as processes of linguistic standardisation, have been objects of sociolinguistic investigation since the inception of the field itself (Haugen, 1959; Ferguson, 1962). The impetus for this type of work largely came from concurrent processes of standardization in newly independent states, often postcolonial realities. As such, it is closely linked to the beginnings of language planning and policy as a subfield of applied linguistics. As characteristics of standard languages, Auer (2005) lists their validity across regional borders, their being considered as the 'best' language within their realm of validity, and their being codified in norms. As Deumert (2004, 2010) notes, it is their institutional enforcement that distinguishes standard languages from other types of linguistic norms, both from "always emergent, variable, and never 'fixed' conventions" (Deumert, 2010, p. 244) and from language standards that go beyond such conventions in that they become morally imperative, but are not quite standard languages as they are not institutionally enforced.

Processes of standardisation usually involve some degree of reduction in linguistic variability (Deumert, 2004; Milroy and Milroy, 2012). As such, these processes can also be understood as linguistic hierarchisations (Kristiansen and Coupland, 2011; Costa et al., 2017; Gal, 2017), whereby specific sets of linguistic features, as well as entire varieties conceived as bundles of such features, are imbued with legitimacy for public use, while others are considered acceptable only in private (Costa et al., 2017) . As Gal (2017) argues, this legitimacy and authority of linguistic forms stem the two ideological complexes of anonymity and authenticity that go back to rationalist and romantic philosophies of the 17th and 18th centuries.

It is commonly recognised that "[s]tandardization is [...] best approached as an ideological phenomenon" (Gal, 2017, p. 222) . Processes of linguistic standardisation have tended to be part and parcel of processes of nation-building (Deumert, 2010; Costa et al., 2017; Erdocia et al., 2025) and the concomitant creation of an apparently neutral public sphere, as well as of colonialism, and are quintessentially modernist projects (Costa et al., 2017). Standard language and non-standard language – be it referred to as dialect, patois, etc. – are constructed in opposition to one another, whereby only standard language indexes modernist values like progress and development, and is associated with the future (Gal, 2017). It is easy to see how, within such standardisation regimes, the majority of the world's languages that do not have standardised forms (Romaine, 2008) can become constructed as 'lacking' and 'backwards' (Gal, 2017), and even more so the myriad of primarily oral language practices (Bird, 2020). In fact, the epistemologies of language underlying standard language ideologies are those of language "as an autonomous and unitary system whose main function is the effective and precise transmission of information" (Deumert, 2010, p. 245), disregarding a view of language as

situated and embodied social practice (Costa et al., 2017; Schneider, 2024).

It has been argued that the development of LLMs has been based both on the epistemological notions underlying standardising regimes and on the effects of these regimes on language practices: According to Schneider (2024), LLMs "build on a foundation of prior technological and sociopolitical conditions including phonetic spelling, standardized print literacy, and linguistic nationalism" (Schneider, 2024). It is thus the diffusion of notions of standard language along with universal education and mass literacy that have produced a quantity of text homogeneous enough to be probabilistically modellable – and these models, in turn, provide the basis for the generation of statistically likely sequences of tokens by generative AI (Schneider, 2024).

While linguistic standardisation is usually examined within the bounds of a single nation, a different perspective is necessary for investigating the development and effects of language technologies like LLMs and generative AI. As Schneider (2022, p. 381) notes, in contrast with the modernist projects of nation-building and linguistic standardisation, these technologies do not aim "to homogenize language in order to create a linguistically homogenous national population" but instead produce a global digital public and are motivated, in most cases, by commercial interests. Previous work on the intersections of technology and language policy in relation to the development of the internet has shown how "technological advances and breakthroughs occur in particular ideological and cultural spaces, and the shape of those technological advances bears the imprint of those cultural and ideological norms" (Kelly-Holmes, 2019, p. 27). Dividing the de facto language policy of the internet into four distinct periods, Kelly-Holmes (2019) argued that we are currently witnessing the period of idiolingualism, characterised by mass linguistic customization. It is to be expected that this type of customization or personalization will impact on the linguistic direction that LLMs will take, and that user needs – as well as their power of consumption – will contribute to steering their development (Kelly-Holmes, 2025; Schneider, 2022)

In this manner, language technologies may reinforce and potentially also reconfigure linguistic hierarchisations (Schneider, 2022; Leblebici and Rostom, 2025). As we will also show in our next section, the hierarchies originating from the interplay between linguistic standardisation and colonialism are already being exacerbated, with LLMs further contributing to the dominance of English and of other European-derived standard languages (Schneider, 2022), and of Mandarin (Arora, 2019). At the same time, however, the jury is still out on the effects that this technology will have on variation within what is commonly constructed as one language. For instance, the fact that the acceptance of standard languages as 'best' language tends to be more wide-spread than their use (Kristiansen and Coupland, 2011; Gal, 2017) might mean that it might not be those forms of language that LLMs will reinforce, if other forms end up being used more frequently – analogous to destandardisation tendencies that have been identified in different European contexts for some time now (Auer, 2005). It thus becomes a highly relevant question how LLMs and generative AI will impact on hierarchisations of language practices and their associated types of speakers.

3 Computational Linguistic Perspectives on LLMs & (Non-)Standard Language

Underlying those generative AI systems that generate text, Large Language Models (LLMs) have demonstrated remarkable capabilities in what computational linguistics refers to as high-resource linguistic environments (Phan et al., 2025), especially English. However, their deployment across the global linguistic landscape remains deeply asymmetrical. In computational linguistic terms, the "digital divide" between languages has translated into lesser-used languages in the digital sphere becoming classed as *Under-Resourced Languages* (URLs), lacking representative training data and thus *language resources*. At the same time, these languages are impaired by Anglo-centric architectural biases of LLMs which negatively affect performance (Xuan et al., 2025). While this challenge is substantial for national or regional minority languages, like Irish or Basque, non-standardized dialects and varieties such as South Tyrolean or Kurdish varieties have been barely considered when evaluating factors such as LLM performance, even though LLMs can work to some degree in such languages (Faisal et al., 2025). Applying LLMs to these languages is often complex due to issues such as the lack of formal orthographies, diglossic tension with standard versions (like Standard German) (Lin et al., 2025) and they are even fre-

quently deliberately excluded from the massive web-scrapes that form LLM training sets (Gao et al., 2021). Consequently, these languages face a double marginalization where both the data scarcity and the structural assumptions of modern NLP fail to capture their unique phonetic, syntactic, and cultural nuances. We see three main areas where current research on LLMs can be improved for non-standard languages: The architecture of LLMs, especially around the tokenization of the input text, the handling of morphological complexity and orthographic variation, and the creation of relevant benchmarks.

The technical infrastructure of language technologies significantly impacts dialects and determines whether a variety is even visible to digital tools. One issue in this regard is the assignment of ISO-639 codes, which acts as a primary gatekeeper; for instance, German dialects like Upper Saxon or Bavarian have assigned ISO codes, which allows them to be catalogued in major Natural Language Processing (NLP) resources like OPUS and the Virtual Language Observatory. Without such codes, whose colonial history Migge and Schneider (2025) have elaborated on, a variety is effectively invisible to language technology, making it nearly impossible to track its representation or performance in AI models. Moreover, the transcription of some dialects may make use of diacritics or other writing methods not supported by the Unicode standard, and as such cannot be processed. The graphemic representation of any language must therefore either remain within the existing coding, or the respective signs must be added to the Unicode standard.

To see why LLMs struggle with non-standard languages like South Tyrolean or Kurdish, we must also look at the representation system that stands between human text and the machine, namely the **tokenizer**. Modern AI models do not *read* text as humans do, but instead convert the input text into a sequence of numbers that can be processed by a neural network. The process of segmenting the text is called tokenization and most current LLMs use methods based on Byte Pair Encoding (Sennrich et al., 2016, BPE). Instead of breaking text into whole words (which would create a vocabulary too large for the computer to manage) or individual characters (which are too small to carry much meaning), BPE identifies "subword" units based on how frequently they appear in a training dataset. As such models break words based on their statistical frequency, they will have more tokens assigned to languages that occur more frequently in the corpus. In this way, words for well-resourced languages will often be broken into fewer tokens and in ways that are more meaningful and related to the morphology of the language. For example, the English word 'tokenization' is divided into two tokens by the widely used BERT model[1] (Devlin et al., 2019), namely 'token+ization', separating the root and the suffix in a morphologically sound way. In contrast, the Irish word 'ionchomharthú' (meaning 'tokenization') is divided into 5 tokens, 'ion+cho+m+hart+hú', and this tokenization not only does not bare any resemblance to the word's morphological components (ion+chomh+arth+ú), but even breaks the word across digraphs such as 'mh' and 'th'[2].

When a tokenizer encounters an under-resourced language or non-standard language, its training hasn't resulted in any specialized subword entries for that language. Instead, it produces a sequence of tokens formed from the subwords that it already has in its vocabulary. This is computationally inefficient, as a single dialectal word might be shattered into 5 or 6 tiny fragments (e.g., individual characters or meaningless byte-strings), whereas the English equivalent would be a single token. This computational inefficiency results in extra costs, e.g. for using more electricity to process a query in these languages. Providers often pass on this cost to users, charging them per token, which causes what is referred to as a "tokenization tax"; so a speaker of Kurdish or South Tyrolean literally pays more when they prompt generative AI in these varieties[3] to process the same amount of information given in English. Performance, in terms of the model's ability, will also naturally be degraded as the model is working with small tokens without specific meaning, which carry less input than large subwords or whole words. Finally, most AI models have a context window (Beltagy et al., 2020), which is a hard limit on how many tokens it can consider at one time. Because dialects require more tokens to

[1] bert-based-uncased

[2] Similar results hold for tokenizer in more recent LLMs. OpenAI provide a platform for this at https://platform.openai.com/tokenizer

[3] Petrov et al. (2023) reports that most languages pay $2-3\times$ the price of English, with languages that use non-Latin scripts such as Arabic paying even higher costs up to $5\times$ and severely under-resourced languages such as Odia paying over $15\times$ the cost

express the same idea, they fill up the model's memory faster, leading to poorer reasoning and shorter possible conversations. In this light, the "tokenization tax" is not merely a technical detail (Ahia et al., 2023); it is a form of algorithmic discrimination (Benjamin, 2019) that systematically increases the cost and decreases the quality of AI services for marginalized linguistic communities.

A second major issue that further compounds the challenge of tokenization is that many under-resourced languages have high morphological complexity and non-standard languages have substantial orthographic variation. English and Chinese, languages spoken by many LLM developers, are relatively "morphologically poor" (such as measured by Čech and Kubát (2018)), having fewer morphemes and simpler morphotactics. Tokenization methods are inherently biased towards morphologically poor languages, as languages with a lower Type-To-Token (TTR) ratio (Kettunen, 2014) can be represented with a smaller vocabulary of subwords. On the other hand, Kurdish for example is a fusional language, where a single word can be built by combining morphemes for functions such as tense, person or negation. As such, a whole sentence in English may be compressed into a single complex Kurdish word. This challenge is intensified by the presence of allomorphs: Because Kurdish grammar is highly sensitive to the sounds surrounding a prefix or suffix, a single grammatical marker (like a plural or a tense indicator) might change its spelling or sound depending on the verb it attaches to.

For languages without a unified orthography, multiple spellings may exist based on local subdialects. Moreover, much of what LLMs see of such languages is extracted from informal contexts such as social media, which often contain errors due to carelessness (i.e., typos) as well as linguistic variation. Further, for a language such as South Tyrolean, there is a strong linguistic pull towards the associated standard language (i.e., Modern High German) not only for the speakers but also for the model, which will have substantial training on the standard language. If AI tools are used for regional governance or service delivery, their inability to handle non-standard spelling can lead to exclusionary bias. Citizens who write in their native dialect may find themselves misunderstood or ignored by automated systems that were optimized for a standardized "prestige" language.

The final barrier to linguistic equity is how we measure LLM success. In AI development, "what gets measured gets built"[4] (Sewak, 2025); however, the current tools for evaluating model performance are fundamentally ill-suited for non-standard and under-resourced languages. Measurement of an LLM's general knowledge is achieved through benchmarks such as MMLU (Hendrycks et al., 2021), which contains questions from standardised textbooks and similar sources. The MMLU benchmark is highly US-centric, including topics such as US History, US Law and US Accounting, with approximately 28% of the questions requiring specific knowledge of Western cultures and a staggering 84.9% of geographic questions focus exclusively on North America or Europe (Faisal et al., 2025). To address this, Global MMLU was introduced, which covers 42 languages, including highly under-resourced languages such as Nyanja or Telugu and explicitly marks questions as culturally sensitive. This avoids a distorted ranking, where a model might appear highly capable in a target language simply because it has memorised Western facts while failing to grasp the specific cultural, legal, or social nuances relevant to actual speakers of that language.

This bias is largely due to the fact that most benchmarks are created by translation of English benchmarks into other languages, and while this has been shown to correlate with human judgements (Thellmann et al., 2024), benchmarks are even worse for lower-resourced languages. This creates a translation pivot trap, where models are optimized to perform well on translated English concepts rather than achieving true native-level reasoning. To correct this bias, high-quality, culturally grounded benchmarks need to be developed, which is an immense financial and logistical undertaking. For example, the development of **MMLU-ProX** (Xuan et al., 2025, covering 29 languages) involved a rigorous expert-review process to ensure cultural relevance, with development costs approaching $80,000 at market rates. Furthermore, the quality of benchmarks created by translation varies significantly with STEM-related tasks exhibiting strong correlation with human judgements (0.70-0.85), while other tasks, such as question answering, have very poor correlation (0.11-0.30) (Wu et al., 2025).

[4] Paraphrasing the famous quote by "what gets measured gets managed" by Peter Drucker

The path forward requires moving away from translated benchmarks toward "culturally and linguistically tailored benchmarks" (Wu et al., 2025). A recent example is **IRLBench** (Tran et al., 2025), a benchmark for the Irish language derived from the Irish Leaving Certificate exams. Recently, DialectBench (Faisal et al., 2024) has introduced a benchmark covering 281 dialects across 10 different tasks, providing the first benchmark that covers a wide variety of non-standard languages. However, South Tyrolean is absent from this benchmark and the support for Kurdish dialects is only partial. Further, there are specific aspects of relevance to non-standard languages that are excluded from benchmarks derived from English that are of relevance to speakers of the community. Firstly, being able to distinguish between dialects and measure whether a text is truly in the dialect or is just outputting standard language with a few dialectal words thrown in. This has led researchers to propose a Dialect Fidelity Score (Park et al., 2025, DFS) to measure this. Secondly, more culturally relevant questions, for example, explaining the moral of a specific proverb, would be relevant, where success requires an understanding of the cultural metaphors that do not exist in the model's English-centric training data. Finally, as non-standard languages are primarily used orally or in informal digital spaces, tasks in these benchmarks should reflect this bias. For example, VoxLect (Feng et al., 2025) uses speech foundation models to evaluate how well AI understands regional accents and phonetic variations that are never captured in written text. Similarly, using "noisy" data from social media, such as in the NorDial benchmark for Norwegian dialects (Barnes et al., 2021) helps determining whether models can handle inconsistent orthography and non-standard spelling without crashing or defaulting to English or standard languages.

Considering the caveats of most current benchmarks, it is all the more alarming that even the best-performing models still show a persistent gap (Ahuja et al., 2023) and produce significantly worse results in languages other than English. For non-standard varieties like South Tyrolean or Kurdish, where formal academic benchmarks do not exist or are fragmentary, the "performance cliff" is likely even steeper. Without a policy-driven investment in local, non-translated evaluation data, speakers of non-standard varieties will not be able to make use of generative AI in these varieties.

4 Linguistic Variation, Language Policy, and LLMs

In this section, we turn to intersections between linguistic variation, language policy and LLMs and show how the way in which LLMs work is based on earlier language policy, and which types of policy actors respond in which kinds of ways to this technology and its development.

4.1 LLM Functioning as Outcomes of Language Policy

Non-standard languages, in computational linguistic terms, are often referred to as "severely" or "acutely" under-resourced" (Krauwer, 2003; Jimerson and Prud'hommeaux, 2018). Traditionally, language documentation has been framed as an intervention intended to "save" endangered languages by capturing their lexico-grammatical structures as data before they are lost (Bird, 2020). This model typically involves linguists creating resources, such as orthographies, lexicons, and grammars, to support the production of pedagogical materials for formal language programs, but also increasingly to develop artificial intelligence systems (Himmelmann, 2006). This is often founded on a deep-seated ideological bias of 'scriptism' (Harris, 1980), which treats written language as the primary, superior, or only "proper" form of language, often reducing speech to a mere derivative of text. In the context of language technology and documentation, scriptism manifests as the insistence on standardizing orthographies and prioritizing textual data as a prerequisite for technological support. Scriptism has led LLMs to ignore non-standardised languages by prioritising the creation of massive text-based datasets, which effectively excludes primarily oral or non-standard speech from the "resource horizon" of modern AI (Faisal et al., 2024). By treating standardized writing as the only suitable data and by developing systems that disregard linguistic variation, LLM technologies overlook the linguistic practices used by many in the world to communicate.

Beyond the bias of scriptism, the evolution of large language models is fundamentally linked to the hegemony of English as the global lingua franca of digital infrastructure and academic research. AI development has seen English as the default language (Bender, 2019) and as such, computational benchmarks treat English as a universal means of expression and the primary pivot language for mul-

tilingual capabilities. The reliance on English, thus, not only affects the amount of data available but also imposes Anglo-centric semantic structures and pragmatic norms onto other languages. Further, educational policies which are rooted in nationalistic and colonial projects (Migge and Schneider, 2025) have further exacerbated this by ensuring that the majority of high-quality digitized texts, such as textbooks, academic papers and documentation, are produced in standardized language. This creates a feedback loop, where LLMs trained on these corpora understand these languages and registers as the means for intellectual discussions and thus generate descriptions in these registers when asked to explain complex and challenging topics. This leads to LLMs considering content that falls outside of these languages and registers to be of lesser intellectual value and less prestigious (Bui et al., 2025).

4.2 Policy Actors and Responses in the LLM Era

The rapid proliferation of LLMs and Generative AI has led to questions about how language policy can support the development of these technologies in a manner that supports language equality. By analyzing the motivations and limitations of diverse policy actors from Big Tech and state entities to civil society and academic institutions, we argue for a shift towards democratic machine learning strategies that recognize linguistic variation not as statistical noise, but as a vital component of human communication and digital citizenship.

The motivation for big technology companies like Meta, Google or OpenAI to move beyond standardized English is rarely purely altruistic but is governed by a tension between *economic scalability* and *socio-technical responsibility*. Big Tech actors occupy a contradictory space as both the primary enforcers of linguistic standardization and the only entities with the compute power to technically reduce the dialect gap. The dominant logic of the large scale favours standardisation to minimise computational costs; however, increasingly, performance on English cannot easily be improved and, as such, under-resourced and non-standard languages have become strategic for improving overall model performance. Similarly, high-resource language markets, such as English, Spanish or Mandarin, are saturated, while the "next billion web users" (Arora, 2019) speak a diverse range of languages. Development in this logic is likely to remain mostly extractive, unless governed by policies that support linguistic equality. A key technical measure could be the measurement and reporting of a *dialect gap*, which measures the reduction in performance in dialects versus English and by requiring this to be explicitly reported or even supported by means of Corporate Social Responsibility (CSR) credits. In this manner, the paradigm could be shifted towards more explicit support for non-standardized languages.

State actors have emerged as critical counterweights to the English-centricity of commercial GenAI, reframing linguistic diversity as a pillar of *digital sovereignty*. Initiatives such as **OpenEuroLLM**[5] support development in particular in EU official languages and the democratic participation of all its citizens. By leveraging public supercomputing infrastructure, such as the *EuroHPC* network[6], states can subsidize the high computational cost of training models on acutely under-resourced dialect data. Furthermore, the implementation of the **EU AI Act** in 2026 provides a regulatory framework that requires high-risk AI systems to be transparent and non-discriminatory. This provides a legal hook: if an AI used in Italian public administration fails to understand a South Tyrolean citizen, it may be deemed a violation of fundamental rights to non-discrimination. While most modern states now frame linguistic diversity as a public good, historically, states have implemented nationalistic policies that marginalize non-standard varieties. This legacy has both led to the data voids that create issues with current LLMs, as well as reduced the trust among speakers of public initiatives. As such, state-led AI initiatives should avoid replicating the discriminatory practices of the past, e.g. by adopting a policy framework of *linguistic agency*, centered on the principle of "nothing about us without us." A democratic strategy (Migge and Schneider, 2025) requires that speakers of non-standard varieties retain *data sovereignty* over their linguistic repertoires.

Community and civil society organisations (CSOs) serve as the essential connective tissue between the technical requirements of LLM development and the lived reality of speech communities. For non-standard languages, like South Tyrolean dialect or varieties of Kurdish, CSOs can

[5] https://openeurollm.eu/
[6] https://eurohpc.eu/

transition from being passive subjects of study to active data stewards and algorithmic auditors. Unlike Big Tech's extractive scraping, CSOs can run *citizen science* initiatives (Hilton, 2021) (e.g., using platforms like Mozilla Common Voice (Ardila et al., 2020)) to collect authentic speech and text with explicit community consent. Furthermore, CSOs can serve as *algorithmic auditors*, performing socio-pragmatic *red-teaming* to identify where models fail to respect local norms or inadvertently enforce standardization. Similarly, academic institutions and national language institutes serve as the primary bridge between technical innovation and socio-historical depth. While Big Tech prioritizes computational scale, universities provide the sociolinguistic granularity necessary to prevent the erasure of non-standard varieties. Furthermore, by fostering interdisciplinary collaboration between Natural Language Processing engineers and sociolinguists, academics ensure that the models do not merely replicate standardized norms, but reflect a linguistic reality grounded in actual usage.

5 Case Study: LLMs and South Tyrolean dialect(s)

South Tyrol is the northernmost Italian province and is known for its contested history since its annexation to Italy in 1919, its autonomy provisions and its institutional multilingualism. German is one of the official languages in the Province – legally on par with Italian – but predominantly it is not Standard German that is spoken in social life in the province, but an ensemble of non-standard forms of German (Platzgummer, 2021). Research has stressed the social relevance of these local dialects (Risse, 2016), which are also increasingly being used in informal writing, particularly in digital contexts such as on social networks or on WhatsApp (see e.g. (Glaznieks et al., 2018; Alber et al., 2024)). Standard German, in contrast, has been shown to mostly be used as a spoken language in a school context (Risse, 2010), and while South Tyrolean German speakers seem to orient to a standard from Germany as the norm with the highest prestige, they have been shown to consider this standard as somewhat 'foreign' to them (Ciccolone, 2010). At the same time, however, most formal writing still takes place in standard German.

When evaluating the resourcefulness, in terms of available language resources and technologies, of South Tyrolean dialects, one first runs up against the problem that South Tyrolean does not have a specific ISO-639 code, which, as previously discussed, means it is not catalogued in NLP resources. As Table 1 shows, there are only ISO-codes for broader groups of German dialects, such as Upper Saxon, Allemanic, Swabian or Bavarian, which makes only language resources with these codes trackable. South Tyrolean dialects are thereby included within the broader group of Bavarian dialects, spoken through most of Austria (with the exception of Vorarlberg in the west) and most of Bavaria. While the Table shows that there are some language resources for Bavarian - especially in comparison to many other German dialect groups represented - it is impossible to see at first glance whether this actually also represents South Tyrolean varieties.

The same then holds for evaluating the availability of pre-trained language models or benchmarks. While, as Faisal et al. (2025) note, LLMs might work to some degree in dialects - and in fact several Generative AI models responded affirmative when we asked *"Versteasch du mi?"* [Do you understand me? Standard German *Verstehst du mich?*] - dialects without a specific ISO code are so far not officially supported, and LLMs performance is not evaluated against them. However, the same is also true for many non-standard varieties with an ISO code. Thus, neither South Tyrolean nor Bavarian appear in any of the relevant multilingual evaluation benchmarks, such as translation-oriented benchmarks like FLORES-200 and NTREX-128, in the cross-lingual topic classification dataset SIB-200 (Adelani et al., 2024), or in GlobalPIQA (Chang et al., 2025) or Global MMLU. The only benchmark that Bavarian is included in is the recent *DialectBench* (Faisal et al., 2024) mentioned in Section 2.

Nevertheless, there has been some demand and language technology development in relation to South Tyrolean varities. More specifically, models are currently being fine-tuned specifically for automated transcription and subtitling of audio(visual) material (Ducceschi and Franzini, 2025). The use case being addressed is that of transposing non-standard audio into standard German writing, implying that the aim is to allow broader access to non-standard audio(visuals) to speakers of Standard German and/or to provide a basis for machine translating South Tyrolean German into other languages. This, in turn, can be expected to have

Name	ISO Code	Speakers	OPUS Words	VLO Resources
Upper Saxon	sxu	2,000,000	0	0
Kölsch	ksh	250,000	6,940	8
Palatine	pfl	400,000	122	0
East Franconian	vmf	4,900,000	0	0
Allemanic	gsw	7,162,000	3,030	1,282
Swabian	swg	820,000	203	8
Walser	wae	22,780	0	0
Bavarian	bar	15,000,000	156,237	124

Table 1: Representation of German dialects spoken in Germany, Austria or Switzerland with an assigned ISO code, in major resources for natural language processing. Speaker counts are based on the best information in Wikipedia. OPUS words is the size of the largest resource in OPUS. VLO refers to the number of listed resources in the CLARIN Virtual Language Obvservatory.

ramifications for speakers of South Tyrolean German. For instance, this might mean that they might be able to use non-standard forms in digitally mediated interaction with non-speakers of South Tyrolean German, in recorded interactions intended for a broader audience, such as TV broadcasts, or to voice-controlled digital assistants. However, as Schneider (2022) has shown, whether such capabilities actually correspond to user needs and will be taken up in user practices is far from clear, as language technologies are ideologically associated with standard language varieties.

6 Case Study: LLMs and Kurdish Varieties

Kurdish is an Indo-European language belonging to the Northwestern Iranian branch, spoken by over 40 million people across Western Asia, primarily in Iraq, Turkey, Iran, Syria, and Armenia, as well as among substantial diaspora communities worldwide (Haig and Matras, 2002). Rather than constituting a single unified language, Kurdish represents a dialect continuum comprising several distinct varieties with varying degrees of mutual intelligibility. The principal varieties include Northern Kurdish (Kurmanji), spoken by an estimated 15–20 million speakers predominantly in Turkey, Syria, northern Iraq, and northwestern Iran; Central Kurdish (Sorani), with approximately 10 million speakers concentrated in Iraqi Kurdistan and Iranian Kurdistan; and Southern Kurdish, spoken in Kermanshah, Ilam, and Lorestan (Matras, 2019). Additionally, the Zaza-Gorani languages, including Zazaki and Hawrami, are spoken by communities who identify as ethnic Kurds, though their linguistic classification remains debated; some scholars group them within the broader Kurdish language family while others consider them closely related but distinct Northwestern Iranian languages (Arslan, 2019).

6.1 Language Suppression and Standardization Challenges

Divided across multiple nation-states, Kurdish has historically faced systematic suppression and assimilation campaigns, including outright bans on public use in Turkey (1923–1991) (Ergil, 2000) and Persianization and Arabization policies targeting its use in formal contexts (Romano et al., 2025, p. 268). (Weisi, 2021) documents how such policies have led Kurdish-speaking parents in certain regions to use the dominant state language with their children, contributing to intergenerational language shift. Nevertheless, the communicative spaces for Kurdish in media and education have expanded over the past decades, mainly thanks to the establishment of the Kurdistan Regional Government in Iraq, where Kurdish now enjoys official status and institutional support (Öpengin, 2012). Unlike Northern Kurdish and Central Kurdish, other varieties such as Southern Kurdish, Hawrami, and Zazaki remain severely under-represented, with Hawrami classified as "definitely endangered" by UNESCO (Moseley, 2010).

The orthographic landscape of Kurdish reflects its political fragmentation. Northern Kurdish is written using a Latin-based alphabet, while Central Kurdish employs an Arabic-based script. Soviet-era Kurdish communities used a Cyrillic-based system. This orthographic diversity, while enabling written expression within political boundaries, creates significant barriers to cross-dialectal communication, pan-Kurdish linguistic unity (Hassanpour,

Variety	Supported Domains	Score (/4)
Central Kurdish (Sorani, ckb)	Education, Media, Press, Official	4
Northern Kurdish (Kurmanji, kmr)	Education, Media, Press, Official	4
Southern Kurdish (sdh)	Press, Publishing	2
Zazaki (zza)	Media, Publishing	2
Hawrami (Gorani, hac)	Publishing	1
Laki (lki)	None	0

Table 2: Distribution of institutional support for Kurdish varieties across key sociolinguistic domains (education, media, publishing, and official use). Northern and Central Kurdish shows the widest domain coverage, while other varieties remain restricted to non-institutional or community-based domains.

2012) and efficient language technology (Ahmadi and Anastasopoulos, 2023).

Table 2 summarizes the distribution of institutional support across Kurdish varieties in four key sociolinguistic domains: education, media, publishing, and official use. Central Kurdish and Northern Kurdish, the two most widely spoken varieties that are also standardised to some degree (Matras, 2019), exhibit full domain coverage, benefiting from formal recognition in educational curricula, broadcast media, print press, and governmental functions. In contrast, Southern Kurdish and Zazaki maintain a more limited presence, primarily confined to press and publishing activities, reflecting their exclusion from official and educational institutions. Hawrami receives support only in publishing despite ongoing efforts for its official recognition along with Central Kurdish and Northern Kurdish (Sheyholislami, 2017; Gialdini, 2023). As the least supported variety, Laki lacks institutional backing entirely. This uneven distribution underscores the hierarchical nature of language vitality within the Kurdish continuum, where political recognition and demographic weight correlate strongly with institutional investment, a disparity that poses significant challenges for language preservation efforts and the development of inclusive Kurdish language technologies.

6.2 Implications for Computational Linguistics and NLP

The sociolinguistic situation of Kurdish has direct implications for its computational processing and representation in language and speech technologies. The historical suppression of the language resulted in limited written corpora, leading to Kurdish being consistently classified as a low-resource language in NLP research. A survey of the NLP literature carried out by AUTHOR reveals that among over 100 papers published in this field, only a handful address varieties other than Central and Northern Kurdish. To further illustrate this disparity, we assess resourcefulness from the three essential pillars of modern language technology: (1) data (text and audio) needed for training models, (2) pretrained models, including embeddings essential for representation, and (3) evaluation benchmarks. Additionally, given its importance as an NLP application, we consider machine translation support as an indicator of technological investment.

Table 3 presents an overview of available resources across Kurdish varieties in terms of parallel text corpora, audio data, pretrained language models, and machine translation services. Central Kurdish followed by Northern Kurdish emerges as the most resourced variety, with over 300 million tokens of textual data and approximately 200 hours of transcribed audio, alongside support from multilingual models such as MADLAD-400 (Kudugunta et al., 2023) and TranslateGemma (Finkelstein et al., 2026), as well as commercial translation services from Google and Microsoft. Critically, Southern Kurdish, Laki, Zazaki, and Hawrami remain entirely unsupported across all categories, reflecting a near-total absence from the modern NLP ecosystem.

Table 4 provides a complementary perspective by surveying the inclusion of Kurdish varieties in prominent multilingual evaluation benchmarks. Benchmarks are essential for assessing LLM performance in an era of rapid technological advancement. Central and Northern Kurdish appear in translation-oriented benchmarks such as FLORES-200 and NTREX-128, as well as in the cross-lingual topic classification dataset SIB-200 (Adelani et al., 2024). Central Kurdish is additionally represented in GlobalPIQA (Chang et al., 2025) for commonsense reasoning. However, neither variety

Variety	Data		Models				Machine Translation	
	Bitext	Audio	XLM-R	BERT	MADLAD-400	TranslateGemma	Google	Microsoft
Central Kurdish	<300M	200h	×	KuBERT	✓	✓	✓	✓
Northern Kurdish	<300M	200h	✓	×	✓	×	✓	✓
Southern Kurdish	<10M	10h	×	×	×	×	×	×
Laki	<1M	2h	×	×	×	×	×	×
Zazaki	<10M	2h	×	×	×	×	×	×
Hawrami	<10M	20h	×	×	×	×	×	×

Table 3: Resourcefulness from the perspective of datasets, pretrained models and machine translation

Variety	FLORES-200	NTREX-128	SIB-200	GlobalPIQA	Global MMLU
Central Kurdish	✓	✓	✓	✓	×
Northern Kurdish	✓	✓	✓	×	×
Southern Kurdish	×	×	×	×	×
Laki	×	×	×	×	×
Zazaki	×	×	×	×	×
Hawrami	×	×	×	×	×

Table 4: Resourcefulness from the perspective of benchmarks

is included in Global MMLU, and the remaining four varieties—Southern Kurdish, Laki, Zazaki, and Hawrami—are absent from all surveyed benchmarks entirely.

It should be noted that the models and benchmarks presented here are not intended to be exhaustive; rather, they are selected to highlight the stark discrepancy in computational support among varieties of Kurdish. This technological marginalization mirrors and reinforces the sociolinguistic hierarchies discussed earlier, posing substantial barriers to the development of inclusive, pan-Kurdish language technologies.

6.3 The Vicious Circle of Underinvestment and Closed Resources

The persistent underrepresentation of Kurdish varieties in NLP can be attributed, in part, to a structural lack of investment from governments and policymakers. Unlike languages that benefit from state-sponsored digitization initiatives, corpus development programs, or dedicated research funding, Kurdish, particularly its lesser-resourced varieties, has received negligible institutional support for computational linguistics research. This absence of top-down investment stands in stark contrast to the resources allocated to dominant state languages in the regions where Kurdish is spoken (Ahmadi et al., 2025).

In the absence of such funding, the development of modern LLMs has come to rely predominantly on publicly available web data, operationalizing the concept of the "Web as corpus" (Kilgarriff and Grefenstette, 2003). However, this approach inherently disadvantages languages with limited digital presence, creating a feedback loop in which low online visibility leads to exclusion from training corpora, which in turn perpetuates technological marginalization. For Kurdish, the historical suppression of the language in public and institutional domains has directly constrained its digital footprint, leaving vast gaps in web-crawled datasets.

Moreover, where data collection efforts do exist, whether by individual researchers, diaspora communities, or private enterprises, the resulting resources frequently remain proprietary or unpublished. This reluctance or inability to release data under open-source licenses compounds the scarcity problem, as subsequent research cannot build upon prior work. The result is a vicious circle: the lack of open resources discourages new investment, while the absence of investment limits the creation of shareable datasets. Until this cycle is disrupted through coordinated open-source initiatives, sustained funding, or policy intervention, the majority of Kurdish varieties will remain stranded at the margins of language technology development.

7 Conclusion

The development of LLMs for non-standard languages like South Tyrolean German represents more than a technical challenge; it is a battleground

for linguistic and digital sovereignty. As we have argued, the digital language divide is maintained by a complex interplay of market forces, historical state policies, and the inherent biases of standardized data pipelines. The cases of South Tyrolean and Kurdish varieties illustrate how linguistic hierarchies, rooted in standardization, colonialism, and scriptism, are reproduced and intensified through tokenization practices, morphological biases, and evaluation frameworks that privilege standardised, high-resource languages. Ultimately, closing the dialect gap requires a multi-faceted approach. We advocate for:

Big Tech Accountability The implementation of dialect gap reporting and Corporate Social Responsibility (CSR) credits to incentivize support for under-resourced varieties.

Linguistic Agency Adopting the principle of "nothing about us without us," ensuring that communities retain data sovereignty and lead the creation of modern technological terminology.

Interdisciplinary Synthesis Leveraging the expertise of academics and language institutes to provide the sociolinguistic granularity that prevents "hallucinated standardization."

By moving away from extractive data practices and toward participatory stewardship, we can ensure that AI serves to revitalize rather than erase non-standard linguistic repertoires.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 226–245. Association for Computational Linguistics.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Sina Ahmadi and Antonios Anastasopoulos. 2023. Script normalization for unconventional writing of under-resourced languages in bilingual communities. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14466–14487. Association for Computational Linguistics.

Sina Ahmadi, Rico Sennrich, Erfan Karami, Ako Marani, Parviz Fekrazad, Gholamreza Akbarzadeh Baghban, Hanah Hadi, Semko Heidari, Mahîr Dogan, Pedram Asadi, Dashne Bashir, Mohammad Amin Ghodrati, Kourosh Amini, Zeynab Ashourinezhad, Mana Baladi, Farshid Ezzati, Alireza Ghasemifar, Daryoush Hosseinpour, Behrooz Abbaszadeh, and 14 others. 2025. PARME: parallel corpora for low-resourced middle eastern languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 30032–30053. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Birgit Alber, Jennifer-Carmen Frey, Aivars Glaznieks, Alexander Glück, and Joachim Henri Kokkelmans. 2024. Verschriftungsprinzipien im geschriebenen dialekt: Whatsapp-nachrichten aus südtirol. *Linguistik online*, 127(3):25–49.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and*

Evaluation Conference, pages 4218–4222, Marseille, France. European Language Resources Association.

Payal Arora. 2019. *The Next Billion Web Users: Digital Life Beyond the West*. Harvard University Press.

Sevda Arslan. 2019. Language, religion, and emplacement of zazaki speakers. *Journal of Ethnic and Cultural Studies*, 6(2):11–22.

Peter Auer. 2005. Europe's sociolinguistic unity, or: A typology of european dialect/standard constellations. *Perspectives on variation: Sociolinguistic, historical, comparative*, 7:7–42.

Kozhin Muhealddin Awlla, Hadi Veisi, and Abdulhady Abas Abdullah. 2025. Sentiment analysis in low-resource contexts: BERT's impact on Central Kurdish. *Language Resources and Evaluation*, pages 1–31.

Jeremy Barnes, Petter Mæhlum, and Samia Touileb. 2021. NorDial: A preliminary corpus of written Norwegian dialect use. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 445–451, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Gábor Bella, Paula Helm, Gertraud Koch, and Fausto Giunchiglia. 2023. Towards bridging the digital language divide. *arXiv preprint arXiv:2307.13405*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Emily Bender. 2019. The #benderrule: On naming the languages we study and why it matters. *The Gradient*.

Ruha Benjamin. 2019. *Race after technology: Abolitionist tools for the New Jim Code*. Social Forces, Medford, MA.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Minh Duc Bui, Carolin Holtermann, Valentin Hofmann, Anne Lauscher, and Katharina von der Wense. 2025. Large language models discriminate against speakers of German dialects. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8212–8240, Suzhou, China. Association for Computational Linguistics.

Radek Čech and Miroslav Kubát. 2018. *Morphological Richness of Text*, pages 63–77. Springer International Publishing, Cham.

Tyler A. Chang, Catherine Arnett, Abdelrahman Eldesokey, Abdelrahman Sadallah, Abeer Kashar, Abolade Daud, Abosede Grace Olanihun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarinayum Meerajita Sharma, Aditi Gupta, Afitab Iyigun, Afonso Simplício, Ahmed Essouaied, Aicha Chorana, Akhil Eppa, Akintunde Oladipo, Akshay Ramesh, Aleksei Dorkin, and 319 others. 2025. Global PIQA: Evaluating physical commonsense reasoning across 100+ languages and cultures. *Preprint*, arXiv:2510.24081.

Simone Ciccolone. 2010. Tutela delle lingue minoritarie come tutela del repertorio: riflessioni dal caso Sudtirolo. In *Les droits linguistiques: droit à la reconnaissance droit à la formation. Actes des deuxièmes Journées des droits linguistiques. Teramo, 20-22 mai 2008*, pages 159–168. Aracne Editrice.

James Costa, Haley De Korne, and Pia Lane. 2017. Standardising minority languages: Reinventing peripheral languages in the 21st century. In *Standardizing minority languages*, pages 1–23. Routledge.

Ana Deumert. 2004. *Language standardization and language change*. John Benjamins Publishing Company.

Ana Deumert. 2010. Imbodela zamakhumsha–reflections on standardization and destandardization. *Multilingua-Journal of Cross-Cultural and Interlanguage Communication*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Luca Ducceschi and Greta H. Franzini. 2025. Speech transcription from South Tyrolean dialect to Standard German with Whisper. In *26th Annual Conference of the International Speech Communication Association, Interspeech 2025, Rotterdam, The Netherlands, 17-21 August 2025*. ISCA.

Iker Erdocia, Britta Schneider, and Bettina Migge. 2025. Language in the age of AI technology: From human to non-human authenticity, from public governance to privatised assemblages. *Language in Society*, pages 1–21.

Dogu Ergil. 2000. The Kurdish question in Turkey. *Journal of Democracy*, 11(3):122–135.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: an NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 14412–14454. Association for Computational Linguistics.

Fahim Faisal, Md Mushfiqur Rahman, and Antonios Anastasopoulos. 2025. Dialectal toxicity detection: Evaluating LLM-as-a-judge consistency across language varieties. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Miami, Florida, USA. Association for Computational Linguistics.

Tiantian Feng, Kevin Huang, Anfeng Xu, Xuan Shi, Thanathai Lertpetchpun, Jihwan Lee, Yoonjeong Lee, Dani Byrd, and Shrikanth Narayanan. 2025. Voxlect: A speech foundation model benchmark for modeling dialects and regional languages around the globe. *CoRR*, abs/2508.01691.

Charles A Ferguson. 1962. The language factor in national development. *Anthropological linguistics*, pages 23–27.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. 2026. TranslateGemma technical report. *Preprint*, arXiv:2601.09012.

Susan Gal. 2017. Visions and revisions of minority languages: Standardization and its dilemmas. In *Standardizing minority languages*, pages 222–242. Routledge.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The Pile: An 800GB dataset of diverse text for language modeling. *CoRR*, abs/2101.00027.

Cecilia Gialdini. 2023. One minority, one language? evaluating linguistic justice for the Kurdish minority in Iran and Iraq. *Diversity and Governance Papers (Eurac Research)*.

Aivars Glaznieks, Jennifer-Carmen Frey, and 1 others. 2018. Dialekt als norm? zum sprachgebrauch südtiroler jugendlicher auf facebook. In *Jugendsprachen. Aktuelle Perspektiven internationaler Forschung*, pages 859–890. de Gruyter.

Geoffrey Haig and Yaron Matras. 2002. Kurdish linguistics: a brief overview. Otto-Friedrich-Universität.

Roy Harris. 1980. *The Language-Makers*. Cornell University Press, Ithaca, NY.

Amir Hassanpour. 2012. The indivisibility of the nation and its linguistic divisions. *International journal of the sociology of language*, 2012(217).

Einar Haugen. 1959. Planning for a standard language in modern norway. *Anthropological linguistics*, pages 8–21.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Nanna Haug Hilton. 2021. Stimmen: A citizen science approach to minority language sociolinguistics. *Linguistics Vanguard*, 7(s1):20190017.

Nikolaus Himmelmann. 2006. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel, editors, *Essentials of Language Documentation*, pages 1–30. Mouton de Gruyter, Berlin.

Robert Jimerson and Emily Prud'hommeaux. 2018. ASR for documenting acutely under-resourced indigenous languages. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, pages 4161–4166.

Helen Kelly-Holmes. 2019. Multilingualism and technology: A review of developments in digital communication from monolingualism to idiolingualism. *Annual Review of Applied Linguistics*, 39:24–39.

Helen Kelly-Holmes. 2025. Artificial intelligence and the future of our sociolinguistic work. *Journal of Sociolinguistics*.

Kimmo Kettunen. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the Web as corpus. *Computational linguistics*, 29(3):333–347.

Steven Krauwer. 2003. The Basic Language Resource Kit (BLARK) as the first milestone for the Language Resources Roadmap. In *Proceedings of the International Workshop on Speech and Computer (SPECOM)*, pages 8–15.

Tore Kristiansen and Nikolas Coupland. 2011. *Standard languages and language standards in a changing Europe*. Novus Press Oslo.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. 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*.

Didem Leblebici and May Rostom. 2025. "alexa learned arabic": A translanguaging and multimodal perspective on language and media ideologies. *Discourse, Context & Media*, 66:100909.

Pin-Jie Lin, Daniel Zeman, Guy Davidson, and Udo Kruschwitz. 2025. Construction-based reduction of translationese for low-resource languages: A pilot study on Bavarian. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025)*, pages 114–121, Vienna, Austria. Association for Computational Linguistics.

Yaron Matras. 2019. Revisiting Kurdish dialect geography: Findings from the Manchester database. *Current issues in Kurdish linguistics*, 1:225.

Bettina Migge and Britta Schneider. 2025. The material making of language as practice of global domination and control: continuations from european colonialism to ai. *AI & SOCIETY*, pages 1–13.

James Milroy and Lesley Milroy. 2012. *Authority in Language: Investigating Standard English*. Routledge.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. UNESCO.

Ergin Öpengin. 2012. Sociolinguistic situation of Kurdish in Turkey: Sociopolitical factors and language use patterns. *International Journal of the Sociology of Language*, 2012(217):151–180.

Keunhyeung Park, Seunguk Yu, and Youngbin Kim. 2025. Steering LLMs toward Korean local speech: Iterative refinement framework for faithful dialect translation. *CoRR*, abs/2511.06680.

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

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Verena Platzgummer. 2021. *Positioning the self: A subject-centred perspective on adolescents' linguistic repertoires and language ideologies in South Tyrol*. Ph.D. thesis, University of Vienna.

Ildikò Erika Stephanie Risse. 2016. Societal multilingualism in South Tyrol: Language contact, diglossia and the dominance of the German-speaking minority. In *Alpen-Kaukasus, Natur- und Kulturraum im Vergleich: Ergebnisse der internationalen Sommerschule Innsbruck 2015*, pages 97–104. Innsbruck University Press.

Stephanie Risse. 2010. Zugehörigkeitsstiftendes und zugehörigkeitsdemonstrierendes sprachliches handeln–versuch einer katalogisierung jenseits des konzepts von" identität "und" alterität ". *Geschichte und Region/Storia e regione*, 19(2):120–135.

Suzanne Romaine. 2008. *Linguistic diversity and language standardization*, volume 9. De Gruyter.

David Romano, Hedi Rashid Hamad Amin, and Farhang Faraydoon Namdar. 2025. Arabization, Turkification, and Persianization of the Kurds and the question of genocide. In *The Palgrave Handbook of Kurdish Genocides*, pages 257–281. Springer.

Britta Schneider. 2022. Multilingualism and ai: The regimentation of language in the age of digital capitalism. *Signs and Society*, 10(3):362–387.

Britta Schneider. 2024. A sociolinguist's look at the "language" in large language models. *Critical AI*, 2(1).

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

Mohit Sewak. 2025. AGI: What gets measured gets built. Medium article. `https://medium.com/aiguys/measuring-agi-706f7862c918`, (accessed March 2026).

Jaffer Sheyholislami. 2017. Language status and party politics in Kurdistan-Iraq: The case of Badini and Hawrami varieties. *Zazaki–yesterday, today and tomorrow. Survival and standardization of a threatened language. Dieter Halwachs: Grazer Plurlingualismus Studien (GPS 04). Graz: GLM*.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and Mehdi Ali. 2024. Towards multilingual LLM evaluation for European languages. *Preprint*, arXiv:2410.08928.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang D. Nguyen. 2025. IRLBench: A multi-modal, culturally grounded, parallel Irish-English benchmark for open-ended LLM reasoning evaluation. *Preprint*, arXiv:2505.13498.

Hiwa Weisi. 2021. Language dominance and shift among Kalhuri Kurdish speakers in the multilingual context of Iran: Linguistic suicide or linguicide? *Language Problems and Language Planning*, 45(1):56–79.

Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. The bitter lesson learned from 2,000+ multilingual benchmarks. *Preprint*, arXiv:2504.15521.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.