https://doi.org/10.48417/technolang.2026.01.06
Research article

Technological Development Models in the Context of Speech Corpora Imbalance

Khumai Bairamova, Anton Gavrilov, Anastassia Kharitonova (✉), and Vladimir Nikolaev
ITMO University , Kronverksky Pr. 49, bldg. A, 197101, St. Petersburg, Russia
aekharitonova@itmo.ru

Abstract

The development of speech and language technologies in the era of artificial intelligence critically depends on the availability of large-scale, high-quality linguistic data. While low-resource languages have been widely studied, less attention has been paid to data imbalances among languages that are considered digitally well-supported. This paper examines the uneven distribution of open speech corpora across languages with established infrastructure of speech technologies and available datasets, showing that this disparity creates structural bottlenecks for sovereign AI development. We conduct a comparative analysis of open and non-commercial speech datasets, accounting for demographic factors, licensing conditions, and models of technological development. To quantify resource inequality, we propose the Digital Resource Saturation Index (DRSI), which relates the availability of speech data to the potential for content generation and consumption within language communities. Our findings reveal a strong dominance of English for open speech resources, while many non-Western languages – including Russian – remain systematically underrepresented. While interpreting these results through the lens of Western and non-Western technological modernization models, we suggest that language inequality in AI is not merely a technical or demographic issue, but a self-reinforcing structurally reproduced outcome of data governance, institutional coordination, and political choices regarding openness and digital sovereignty. The study further provides practical recommendations for mitigating these imbalances and fostering a more equitable technological landscape.

Keywords: Digital language divide, Speech corpora imbalance, Language inequality, Technological development models, Resource disparity analysis, Digital resource saturation index, DRSI.

Citation: Bairamova, K., Gavrilov A., Kharitonova A., & Nikolaev V. (2025). Technological Development Models in the Context of Speech Corpora Imbalance. *Technology and Language, 7*(1), 80-102. https://doi.org/10.48417/technolang.2026.01.06

УДК 316.422:004.522
https://doi.org/10.48417/technolang.2026.01.06
Научная статья

Модели технологического развития в контексте дисбаланса речевых корпусов

Хумай Бахруз кызы Байрамова, Антон Валерьевич Гаврилов,
Анастасия Евгеньевна Харитонова (✉) и Владимир Вячеславович Николаев
Университет ИТМО, Кронверкский пр., д. 49, стр. А, 197101, Санкт-Петербург, Россия
aekharitonova@itmo.ru

Аннотация
Развитие речевых и языковых технологий в эпоху искусственного интеллекта (ИИ) в решающей степени зависит от наличия крупномасштабных высококачественных лингвистических данных. В то время как языки с ограниченными ресурсами изучены достаточно широко, сравнительно мало внимания уделялось дисбалансу данных для языков, имеющих достаточную цифровую поддержку. В данной статье исследуется неравномерное распределение открытых речевых корпусов для языков, обладающих стабильной инфраструктурой речевых технологий и доступными корпусами, и утверждается, что эта асимметрия создает структурные ограничения для суверенного развития ИИ. Проводится сравнительный анализ открытых и некоммерческих речевых корпусов с учетом демографических факторов, условий лицензирования и моделей технологического развития. Для количественной оценки ресурсного неравенства предлагается индекс концентрации цифровых ресурсов (Digital Resource Saturation Index – DRSI), который соотносит объем доступных речевых данных с потенциалом генерации и потребления контента в пределах языковых сообществ. Полученные результаты выявляют явное доминирование английского языка в области открытых речевых ресурсов, в то время как многие незападные языки – включая русский – остаются недостаточно представленными. Интерпретация этих данных через призму западных и незападных моделей технологической модернизации показывает, что языковое неравенство в сфере ИИ является не просто технической или демографической проблемой, а структурно воспроизводимым результатом политики управления данными, институциональной координации и политического выбора в области открытости и цифрового суверенитета. Исследование также содержит практические рекомендации по снижению дисбаланса и формированию более справедливой технологической среды.

Ключевые слова: Цифровой языковой разрыв, Дисбаланс речевых корпусов, Языковое неравенство, Модели технологического развития, Анализ неравномерности ресурсов, Индекс концентрации цифровых ресурсов, DRSI

Для цитирования: Bairamova, K., Gavrilov A., Kharitonova A., Nikolaev V. Technological Development Models in the Context of Speech Corpora Imbalance // Technology and Language. 2026. № 7(1). P. 80-102. https://doi.org/10.48417/technolang.2026.01.06

INTRODUCTION

In the contemporary digital landscape, access to data has become a critical factor in technological development of digital sovereignty. The development of competitive artificial intelligence (AI) technologies requires large-scale training data with rich linguistic content. However, the availability of such data is highly asymmetric – a phenomenon commonly referred to as the *Digital Language Divide* (Gábor et al., 2023). While a substantial body of research has focused on languages with minimal digital resources, significant imbalance persists even among technologically advanced language ecosystems with tens or hundreds of millions of speakers. Analyzing this imbalance enables the identification of structural constraints that must be addressed to ensure the sustainable development of AI systems. Preliminary evidence indicates that the volume of open speech corpora available for English vastly exceeds that of other languages with comparable numbers of speakers (Joshi et al., 2020; Kreutzer et al., 2022). This imbalance reflects substantial differences in technological development trajectories and data governance practices. Western models of AI development typically rely on a combination of corporate and open data resources, whereas alternative strategies prioritize the construction of sovereign data infrastructures. The Russian language represents a particularly illustrative case: despite its large digital audience, it exhibits a shortage of open and well-structured speech data typical for many non-Western languages, placing it in a relatively vulnerable position within the global AI ecosystem, and making it dependent on external resources.

Modern large language models (LLMs), which are increasingly used across a wide range of applications, require massive textual corpora for training, with model quality and functional capabilities directly correlated with data scale and diversity. Speech processing systems, which underpin voice assistants and other speech-based technologies, are no exception: they similarly depend on large-scale, high-quality annotated speech corpora. The availability of extensive, structured, and diverse datasets improves predictive accuracy, enhances generalization capabilities, and accelerates both fundamental and applied research.

The Digital Language Divide constitutes a specific manifestation of broader digital inequality and is reflected in the weak correlation between the number of language speakers and the volume of digital resources available to support that language technologically (Gábor et al., 2023). Most publicly accessible linguistic corpora demonstrate a pronounced skew toward a small number of languages – primarily English – which complicates the training, evaluation, and reproducibility of contemporary AI models (Henning et al., 2023). This issue has often been discussed in the studies primarily as a consequence of technical and resource constraints (Bender et al., 2021). The dominant position of English is based on its historical role as the lingua franca of scientific communication and the emergence of a robust open-data ecosystem within the Western model of scientific and technological development (Joshi et al., 2020).

In contrast, the scale and structure of linguistic resources for Chinese are shaped within a distinct institutional context characterized by active state involvement in data regulation and the pursuit of digital sovereignty strategies (Roberts et al., 2021). The Russian language, despite its significant digital presence, remains constrained by limited

access to open and structured speech resources, which increases the risk of technological dependence within the global AI ecosystem (Joshi et al., 2020; Lau et al., 2025).

This asymmetry in the quantity and diversity of linguistically dependent data leads to degraded AI model performance for resource-constrained languages (Markl & McNulty, 2022, p. 6328), reduces model universality, and reinforces technological dependence on dominant Western platforms.

The primary objective of this study is to conduct a comparative analysis of open speech data availability for languages with a high level of digital support assessing the impact of different models of technological modernization on the formation of digital language inequality.

To achieve this objective, the study is structured around the following tasks:

1. The systematic assessment of open speech data availability for digitally well-supported languages.
2. The development of a relevant metric to quantify the gap between actual resource availability and demographic potential.
3. A comparative analysis of resource distribution across languages and interpretation of the observed disparities in relation to predominant technological development models.
4. A discussion of practical considerations for mitigating resource deficits in languages of intermediate development models, using Russian as an example.

This study advances the understanding of language inequality in AI not merely as a technical gap, but as a phenomenon deeply embedded in divergent models of technological modernization (Zapf, 2004). While existing research often describes the digital language divide at an aggregate level, we offer a comparative analysis of data imbalance among languages that are generally considered technologically well-supported.

To make this disparity visible and quantifiable, we propose the Digital Resource Saturation Index (DRSI) which assigns an individual score to each language. Rather than presenting DRSI as a definitive normative benchmark, we position it as a diagnostic tool to reveal the structural conditions under which non-Western technological projects must operate.

By interpreting our quantitative findings through the lens of different modernization trajectories, we aim to show how the observed asymmetries in open speech data are not accidental but reflect distinct political choices regarding data governance, openness, and digital sovereignty.

LANGUAGE INEQUALITY AND MODELS OF TECHNOLOGICAL DEVELOPMENT

This section situates language data inequality within broader models of technological modernization and examines the structural mechanisms through which disparities in linguistic resources are reproduced in AI systems.

Western, Chinese, and Intermediate Models of Technological Modernization

The development of language technologies in the AI era is closely tied to the ways in which states and corporations organize their technological development strategies. The most widespread approach is the Western model, which prioritizes open data and corporate leadership, reinforcing the global dominance of a few languages – a pattern that some scholars interpret as a new form of digital hegemony (Artyukhin et al., 2025). The dominance of major technology companies such as Apple, Alphabet, Microsoft and Amazon significantly shapes both the direction of AI development and the distribution of linguistic resources. Commercial investments are primarily directed toward languages that offer the highest economic gains and market scalability, most notably English. This dynamic is supported by an ecosystem of open science and corporate data sharing (Vincent et al., 2019) and contributes to the emergence of systemic Anglocentrism in language technologies, whereby models are trained on disproportionately large English-language corpora, while support for other languages remains secondary (Joshi et al., 2020).

In contrast, the Chinese model is grounded in the concept of digital sovereignty. Within this framework, data is treated as a strategic national resource, and the state plays an active role in regulating data access and fostering the development of a domestic language ecosystem (Christophe et al., 2023). This approach encourages the creation of large-scale local corpora and reduces reliance on Western platforms, resulting in a largely self-sufficient infrastructure for Mandarin and other regional languages of China (Roberts et al., 2021).

Between these two poles lie *intermediate models* of technological development, exemplified by countries such as Russia, India, and Brazil. In these contexts, the development of language resources remains fragmented. Despite the presence of large speaker populations, national languages often receive limited digital support due to constrained investment, the absence of long-term programs for building open corpora, and weak coordination among academic institutions, industry, and the state (Baishya et al., 2025; Adebara et al., 2025).

These different strategies are aligned with broader theoretical debates on modernization. Classical modernization theory viewed Western development as a universal template, but later scientists emphasized "multiple modernities" and the possibility of non-Western paths (Zapf, 2004). In the digital realm, this translates into competing notions of technological sovereignty (Neznamov et al., 2025). Taken together, these differences in scientific and technological development strategies give rise to persistent trajectories of language inequality, whereby some languages are able to participate fully in AI advancement, while others remain dependent on external centers of technological production.

Structural Drivers of the Global Digital Language Divide

Beyond national development strategies, a set of systemic factors consistently reproduces language inequality within the technological landscape.

Socio-digital asymmetry.

The majority of open data available on the internet is produced in a limited number of languages, primarily English, which directly shapes the composition of training corpora used by contemporary AI models (Akindotuni, 2025; Ranathunga & de Silva, 2022).

Market incentives.

From an economic perspective, it is more profitable for corporations to develop AI technologies for large global markets than for languages associated with smaller or less commercially attractive audiences. Consequently, technological development tends to reproduce economic priorities rather than cultural or linguistic diversity (Akindotuni, 2025; Joshi et al., 2020).

Infrastructure deficits.

In many countries, there is a lack of stable institutional frameworks for collection, licensing, storage, and open distribution of corpora. As a result, even languages with large speaker populations remain insufficiently represented in AI systems (Joshi et al., 2020; Akindotuni, 2025).

Cultural regulation of algorithms.

Language models trained on Western data impose standardized notions of "normative language", often disregarding dialectal variation and local linguistic practices. This produces a form of digital *epistemic inequality*, in which certain language varieties and communicative norms are systematically marginalized (Helm et al., 2024; Bird, 2020).

In addition, language inequality in AI is intensified by a self-reinforcing mechanism of data accumulation. Contemporary training corpora are largely derived from openly available internet data, which are themselves characterized by pronounced linguistic imbalance. As a result, AI models inherit this asymmetry and demonstrate superior performance for languages with extensive digital representation, primarily English. Improved model performance further promotes the use of these languages in digital platforms and services, leading to the generation of even larger volumes of data and the subsequent expansion of training corpora. This process constitutes a self-reinforcing loop – also described as a Matthew effect – in which languages with abundant digital resources continue to consolidate their advantage, while resource-constrained languages are systematically excluded from technological progress (Akindotuni, 2025).

Taken together, these factors demonstrate that the roots of the digital language divide lie not only in technical limitations, but also in data governance policies, economic structures, institutional arrangements, and cultural ideologies embedded in contemporary AI systems.

The Role of Alternative Models in Language Resource Development

The divide between Western and non-Western models of language resource development is neither natural nor inevitable. Rather, it is actively reproduced through institutional, economic, and technological configurations that define the contemporary data ecosystem. Recognizing this dependence has led to a growing demand in the research studies for alternative development trajectories that prioritize not only market efficiency, but also long-term scientific and societal goals.

Studies addressing language inequality increasingly emphasize several potential strategies for mitigating the divide, including expanding access to open corpus resources, supporting academic and civic initiatives for data collection, developing sustainable infrastructures for data storage and licensing, and fostering localized research ecosystems. These approaches are not presented as universal solutions, but rather as components of a broader strategy aimed at reducing dependence on dominant linguistic centers and promoting more inclusive and resilient language technologies. Such strategies include discussions on fostering technological sovereignty and developing balanced regulatory frameworks for national AI ecosystems (Neznamov et al., 2025).

DATA AND METHODOLOGY

Language Selection and Classification

For the purposes of this analysis, we selected 32 languages classified as having a *Thriving* level according to the Digital Language Support (DLS) scale (Simons et al., 2022; Derivation, 2025). This classification indicates that each selected language is supported by at least nine speech-processing tools (e.g., speech-to-text or text-to-speech systems) and at least one voice assistant. This criterion ensures the availability of multiple speech datasets for each language and allows for meaningful cross-linguistic comparison.

The selected languages were further categorized according to the following criteria.

General technological development model.

This criterion reflects differences in digital development strategies shaped by sociocultural, institutional, geopolitical, and economic characteristics of the countries in which the selected languages hold official or national status. Based on this perspective, we apply two macro-categories: *Western* and *Non-Western* in accordance with the theoretical framework outlined in Section 2. This distinction represents an analytical simplification and does not reflect linguistic properties of the languages themselves. A language is classified as *Western* if it is an official or state language in European or North American countries that are members or key partners of organizations such as the European Union or NATO, often conventionally referred to as the "West." Conversely, the *Non-Western* category includes languages of countries whose technological development follows alternative, non-Western trajectories.

EGIDS language status.

The Expanded Graded Intergenerational Disruption Scale (EGIDS) (Eberhard et al., 2025) is used to assess how a language's status in global communication affects the volume and accessibility of open data compared to languages with primarily national domains of use. According to data from Ethnologue and Derivation, six languages in our sample are classified as *International* (Arabic, English, French, Spanish, Russian, and Chinese), one language (Latin) has *Dormant* status, and the remaining languages are classified as *National*.

To analyze data availability, it was necessary to obtain demographic information on the number of native speakers (L1) and the total number of speakers (L1+L2). These data were sourced from Ethnologue, the most frequently cited repository of linguistic

demographic statistics. Several methodological adjustments were required for languages lacking conventional L1 speaker populations.

Latin (*la*) was excluded from further analysis due to its Dormant EGIDS status which is significantly lower than the National/International statuses of other languages, the lack of the native speakers, and no reliable estimates of L2 speakers.

Arabic presents a special case, as most datasets use the code *ar* that corresponds to the Arabic macrolanguage. Standard Arabic (*arb*), classified as *Thriving*, functions as a literary and formal standard widely used in education, media, and written communication, yet it does not have native speakers in a strict sense. In this study, the number of active users of Standard Arabic (334.8 million) is used as a proxy for L1, reflecting content generation potential, while the number of speakers of Arabic dialects (410.5 million) is used as L1+L2, accounting for functional proficiency in the standard variety. For comparability with existing datasets, the adapted category is denoted using the code *ar*.

Malay is similarly represented in datasets by the macrolanguage code *ms* (or *msa*), encompassing several closely related languages and dialects. Within this macrolanguage, two distinct varieties have *Thriving* status: Standard Malay (*zsm*), used as an official language in Malaysia, Brunei, and Singapore, and Local Malay (*zlm*), which comprises dialects spoken as native languages by Malay populations. For this study, the aggregated code *ms* was used to ensure compatibility with public datasets. Since Standard Malay has no native speakers, the number of Local Malay (*zlm*) speakers – 26.5 million – was adopted as the L1 value, reflecting content generation potential, while the total number of active users of Standard Malay (*zsm*) – 34.6 million – was used as the L1+L2 value. Unlike the Arabic case, this aggregation does not fully collapse the macrolanguage, as Indonesian (*id*), which also belongs to *ms*, was analyzed separately due to its independent *Thriving* status and distinct resource base.

The final language selection and categorization are summarized in Table 1 (presented later in Section 3.2), which additionally reports ISO-639 codes, linguistic genealogy (family and branch), DLS values, EGIDS status, and demographic indicators (L1 and L1+L2, in millions).

Selection of Speech Data Sources

To estimate the total volume of speech data available for the selected languages, a systematic dataset identification and filtering process was conducted. The primary objective was to obtain a representative sample of large-scale open datasets suitable for assessing inequalities in resource distribution.

The initial dataset list was compiled using a multi-source approach. Primary sources included are the following.

- **OpenSLR**, a public repository hosting a wide range of speech and language resources, from which datasets categorized under *Speech* were selected.
- **Systematic research reviews**, including survey papers providing curated overviews of publicly available speech datasets (Agnew et al., 2024; Longpre et al., 2024). These sources supplied lists of notable corpora along with metadata on duration, language coverage, and license types.

Other catalogues, such as the ELRA Language Resources Association Catalogue and the Linguistic Data Consortium Catalog, were considered but excluded as primary sources, as they predominantly list proprietary or paid datasets, which fall outside the scope of this study's focus on open-access resources.

Datasets identified across all sources were merged into a unified list and subsequently divided into monolingual and multilingual subsets. Both subsets were filtered using the following selection criteria: Both subsets were filtered using the following sequential criteria:

1. **Language relevance:** datasets not covering any of the 32 selected languages were excluded.
2. **Academic documentation:** datasets lacking an associated academic publication or detailed technical report were removed to ensure transparency and reproducibility.
3. **Metadata availability:** datasets without verifiable documentation of audio duration (in hours or convertible units) were excluded.
4. **License compatibility:** datasets with clearly specified licenses permitting the creation of derivative works were retained.

For licensing analysis, datasets were further divided into two categories: *Open*, which includes licenses permitting both commercial and non-commercial use (e.g., Apache 2.0, CC-BY, CC-BY-SA, MIT, Public Domain), and *NC*, which includes licenses restricting use to non-commercial purposes or requiring additional permissions (e.g., CC-BY-NC and similar licenses).

For multilingual datasets, the number of audio hours corresponding to each of the 32 target languages was extracted and tabulated to enable cross-linguistic comparison.

As a result of this process, 125 monolingual corpora and 22 multilingual datasets were selected, including Common Voice (Ardila et al., 2020), LibriVox, MediaSpeech (Kolobov et al., 2021), VoxPopuli (Wang et al., 2021), Multilingual LibriSpeech (Pratap et al., 2020), M-AILABS (Imdat, 2019), FLEURS (Conneau et al., 2022), CoVoST 2 (Wang et al., 2020), CML-TTS (Oliveira et al., 2023), MOSEL (Gaido et al., 2024), Emilia (He et al., 2025), and Yodas (Li et al., 2023).

Following this multi-stage filtering, summary statistics were aggregated into a Table 1 presenting the quantitative distribution of open and total speech data across the selected languages. This table forms the empirical basis for the subsequent inequality analysis.

Table 1. Language sample and speech corpora statistics

Language	ISO-639-1	Cat	Family	Branch	DLS	Status	L1, millions	L1+L2, millions	open, h	total, h
English	en	W	Indo-European	Germanic	1	Int	390.3	1527.9	1139368	1145394
French	fr	W	Indo-European	Romance	1	Int	74.2	311.9	89296	89296
German	de	W	Indo-European	Germanic	1	Nat	76.1	134	85687	85687
Chinese	zh	NW	Sino-Tibetan	Sinitic	1	Int	989.9	1183.8	78588	79588
Spanish	es	W	Indo-European	Romance	1	Int	484	558.5	70351	70351

Finnish	fi	W	Uralic	Finnic	0.89	Nat	4.96	5.62	63171	63171
Russian	ru	NW	Indo-European	East Slavic	0.96	Int	145.2	253.4	41772	61772
Italian	it	W	Indo-European	Romance	1	Nat	62.9	66.2	60851	60851
Dutch	nl	W	Indo-European	Germanic	1	Nat	23.74	25.39	49716	49716
Japanese	ja	NW	Japonic	-	1	Nat	123.6	125.6	45706	45706
Polish	pl	W	Indo-European	West Slavic	0.96	Nat	43.2	45.3	45223	45223
Portuguese	pt	W	Indo-European	Romance	1	Nat	249.5	266.6	44367	44367
Czech	cs	W	Indo-European	West Slavic	0.89	Nat	9.76	12.47	39405	39405
Romanian	ro	W	Indo-European	Romance	0.89	Nat	23.71	23.73	36911	36911
Hungarian	hu	W	Uralic	Ugric	0.89	Nat	12.14	12.15	36449	36449
Korean	ko	NW	Koreanic	-	1	Nat	81.2	81.6	22847	22849
Croatian	hr	W	Indo-European	South Slavic	0.89	Nat	5.15	6.45	22612	22612
Swedish	sv	W	Indo-European	Germanic	0.96	Nat	10.1	13.3	22413	22413
Hebrew	he	NW	Afro-Asiatic	Semitic	0.93	Nat	6.12	10.48	21022	21052
Danish	da	W	Indo-European	Germanic	0.96	Nat	5.81	5.82	16077	16077
Vietnamese	vi	NW	Austroasiatic	Mon-Khmer	0.93	Nat	86.1	97	10047	10547
Indonesian	id	NW	Austronesian	Malayo-Polynesian	0.93	Nat	75.2	252.4	9978	9978
Turkish	tr	NW	Turkic	Oghuz	0.96	Nat	85.2	91.3	4816	4816
Hindi	hi	NW	Indo-European	Indo-Aryan	0.93	Nat	345.4	609.1	3067	4328
Arabic	ar	NW	Afro-Asiatic	Semitic	1	Nat	334.8	410.5	1560	3601
Tamil	ta	NW	Dravidian	South	0.89	Nat	78.6	86.3	2140	2140
Ukrainian	uk	NW	Indo-European	East Slavic	0.89	Nat	32.1	39	1078	1078
Thai	th	NW	Kra–Dai	Kam–Tai	0.96	Nat	27.2	71.4	939	939
Norwegian	no	W	Indo-European	Germanic	0.96	Nat	5.49	5.5	834	834
Malay	ms	NW	Austronesian	Malayo-Polynesian	0.96	Nat	26.5	34.6	44	44
Serbian	sr	NW	Indo-European	South Slavic	0.89	Nat	8.23	8.26	41	41

Criteria for Resource Comparison and Disparity Estimation

To validate the presence of digital inequality among languages classified as Thriving on the DLS scale, we computed the Gini coefficient using the number of language speakers and the total volume of open speech data as parameters (Figure 1).

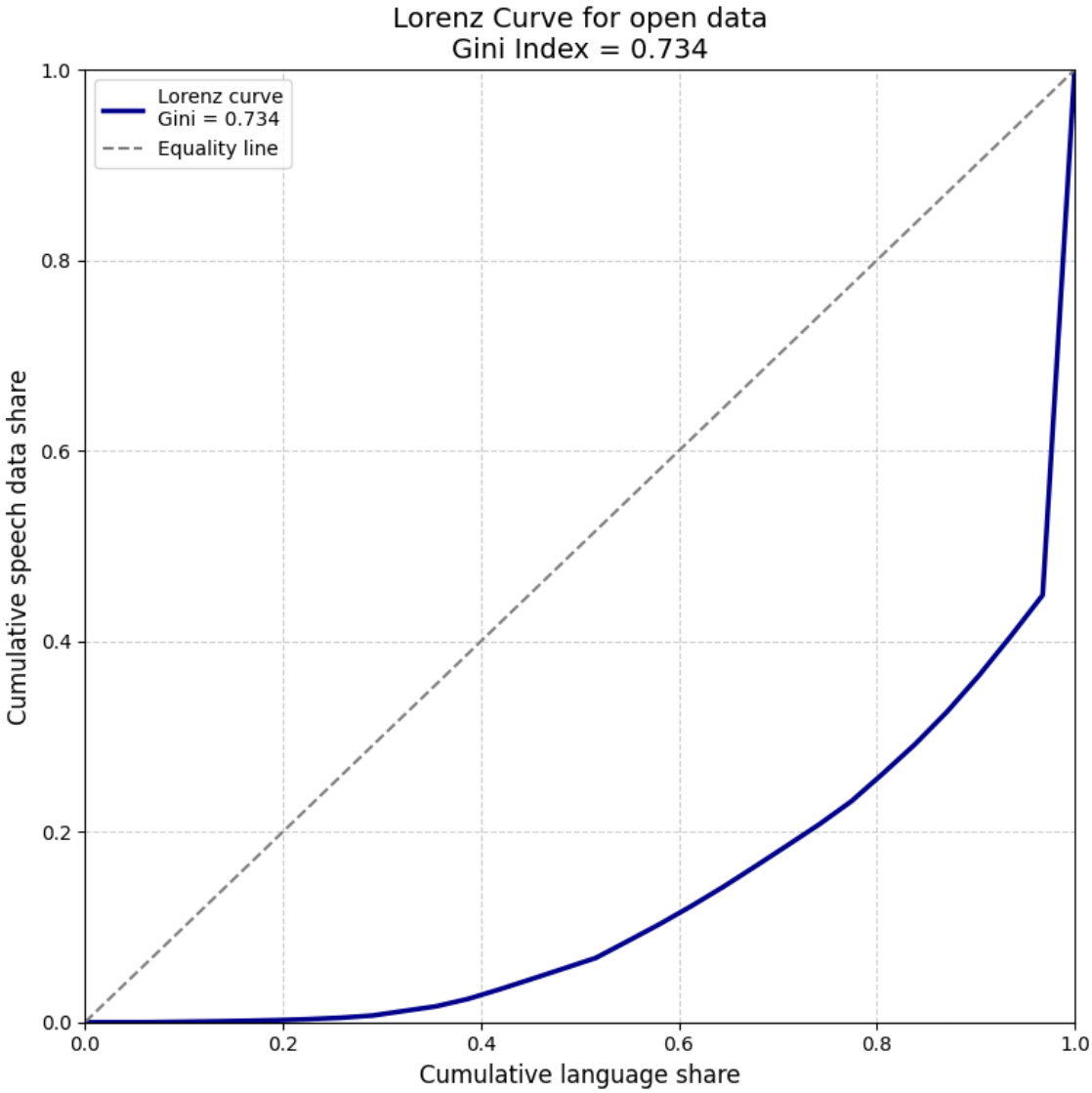

Figure 1. Lorenz curve of speech resource inequality

The corresponding Lorenz curve illustrates the cumulative distribution of open speech resources relative to the cumulative share of language populations, ordered from the least to the most resource-rich languages.

Unlike economic resources, digital data are non-rivalrous and do not imply direct redistribution among languages. Accordingly, in this study, the Gini coefficient is not interpreted as a measure of fairness, but rather as an indicator of structural concentration in the distribution of open speech resources. The obtained value of 0.734 indicates a pronounced dominance of a small number of languages – primarily English – within the AI data infrastructure.

Existing approaches to assessing technological language support include qualitative analyses documenting the existence of the digital divide (Bender et al., 2021), categorical evaluations (Simons et al., 2022), and quantitative methods based on aggregate resource assessments (Joshi et al., 2020) or downstream model performance (Blasi et al., 2022). For European languages, composite indices combining weighted technological and contextual factors have been proposed (Gaspari et al., 2022; Grützner-Zahn & Rehm, 2022; Rehm & Way, 2023).

To provide an assessment that captures resource availability at the level of individual languages, we introduce the Digital Resource Saturation Index (DRSI). This metric captures the relationship between the volume of available speech resources for a given language and its demographic resource potential, accounting for both content generation and consumption capacities under assumptions of proportionality and equality.

Demographic normalization

For each language i, we define demographics indicators based on the number of native speakers (L1) and the total number of speakers (L1+L2). They are normalized to the maximum value across the sample using parity-to-maximum scaling:

$$p_i^{L1} = \frac{L1_i}{L1_{max}}, \; p_i^{L1+L2} = \frac{(L1+L2)_i}{(L1+L2)_{max}}.$$

These normalized values represent the relative demographic scale of each language within the selected sample.

Normalization of speech resources

Analogously, we normalize the observed volumes of speech data:

$$pR_i^{open} = \frac{R_i^{open}}{R_{max}^{open}}, \; pR_i^{total} = \frac{R_i^{total}}{R_{max}^{total}},$$

where R_i^{open} denotes the total duration of speech data available under licenses permitting commercial use, and R_i^{total} includes both open and non-commercial datasets.

Expected resource levels

To estimate the expected level of speech resources for each language we introduce a parameter $\alpha \in [0,1]$, that corresponds to a social justice model and interpolates between a purely demographic expectation and an egalitarian baseline. When $\alpha = 0$, the expected resource level is fully proportional to the number of speakers; when $\alpha = 1$, all languages are assumed to have equal expected support.

From the perspective of data consumption, the expected resource level is defined as:

$$E^{cons}_i = (1 - \alpha)p^{L1+L2}_i + \alpha$$

From the perspective of data creation, the expected resource level depends on the number of native speakers and an additional generation coefficient $\gamma \in [0,1]$, which reflects the assumed institutional maturity and the capacity of a language community to produce original speech content:

$$E^{gen}_i = (1 - \alpha)\gamma p^{L1}_i + \alpha$$

Utilization and generation components

Using these expectations, we define two complementary components.

The *utilization component* measures how the availability of open speech data compares to the expected level of data consumption:

$$Util_i = \frac{\beta \cdot pR_i^{open}}{(1-\alpha) \cdot p^{L1+L2} + \alpha}$$

where $\beta \in [0,1]$ controls the relative emphasis on open-access data, reflecting the degree of digital data sovereignty governance.

The *generation component* captures how the total volume of speech data compares to the expected level of data creation:

$$Gen_i = \frac{(1-\beta) \cdot pR_i^{total}}{\gamma \cdot (1-\alpha) \cdot p^{L1} + \alpha}$$

Digital Resource Saturation Index

Finally, the Digital Resource Saturation Index for language i is defined as the sum of the utilization and generation components:

$$DRSI_i = Util_i + Gen_i$$

Values of DRSI close to zero indicate a substantial deficit of speech resources relative to the expected level, while higher values reflect stronger saturation of the language with available speech data. The index enables a comparative ranking of languages and highlights cases where demographic scale and digital resource availability are strongly misaligned.

RESULTS

The computed DRSI values for the selected languages, using mid-range parameter settings *($\alpha=\beta=\gamma=0.5$)*, are presented in Table 2.

The DRSI values in Table 2 reveal patterns that go beyond mere data availability; they reflect the underlying technological modernization regimes and civilizational clusters described and discussed in the comparative literature (Artyukhin et al., 2025; Zapf, 2004).

The results indicate a pronounced concentration of open speech resources in English, whose DRSI value substantially exceeds those of all other languages. English benefits from a large number of high-quality speech datasets, many of which are distributed under permissive open licenses. This reflects sustained research activity, a high level of digital infrastructure development, and the central role of English within Western research ecosystems. This is not simply a function of its speaker population, but a direct outcome of the Western model: a combination of corporate investment, a robust open-science ecosystem, and the path-dependent accumulation of datasets over decades (Zapf, 2004). English serves as the "default" language of AI research, and its resource base is both a cause and a consequence of its hegemonic position in global science and

commerce. Consequently, English-language data form the backbone of training pipelines for a wide range of contemporary machine learning models.

Table 2. DRSI by language

Language	Category	DRSI	Language	Category	DRSI	Language	Category	DRSI
English	Western	1.3353	Czech	Western	0.06854	Turkish	Non Western	0.008020
German	Western	0.1412	Portuguese	Western	0.06756	Hindi	Non Western	0.005142
French	Western	0.1402	Romanian	Western	0.06375	Arabic	Non Western	0.003768
Finnish	Western	0.1103	Hungarian	Western	0.06337	Tamil	Non Western	0.003575
Italian	Western	0.1027	Croatian	Western	0.03945	Ukrainian	Non Western	0.001849
Spanish	Western	0.0946	Swedish	Western	0.03897	Thai	Non Western	0.001596
Dutch	Western	0.0858	Korean	Non Western	0.03820	Norwegian	Western	0.001455
Chinese	Non Western	0.0852	Hebrew	Non Western	0.03665	Malay	Non Western	0.000076
Russian	Non Western	0.0817	Danish	Western	0.02805	Serbian	Non Western	0.000071
Polish	Western	0.0772	Vietnamese	Non Western	0.01712			
Japanese	Non Western	0.0746	Indonesian	Non Western	0.01591			

Western European languages (German, French, Italian, Spanish) form a second tier. The group also includes Finnish partly due to its relatively small speaker population. Their values, while lower than English, are still relatively high due to sustained national and EU-level funding for language technologies, as well as active participation in community-driven projects. This reflects the institutionalized support for linguistic diversity within the Western framework, although still subordinate to English. These languages correspond roughly to the “Protestant Europe” and “Catholic Europe” sociocultural clusters (Inglehart and Welzel, 2005), where secular-rational values and strong institutional frameworks support open scientific exchange.

Spanish and Portuguese present a particularly interesting case. Despite their classification as Western languages and their large global speaker bases (including across Latin America), their relatively lower DRSI values may reflect the “ambiguous dynamics” of modernization trajectories in the Latin American civilizational cluster. In this cluster, economic development has often outpaced the development of civic institutions and open science infrastructures (Inglehart and Welzel, 2005). The Catholic Church's historical role as a mediator between state and society, combined with persistent clientelism and corruption, has created a modernization trajectory where the production of openly accessible public goods – such as linguistic data – remains underdeveloped relative to demographic scale (Artyukhin et al., 2025). The DRSI thus captures not just a data gap, but a deeper structural feature of Latin American modernization: a pattern of “modernization through traditional structures” that does not automatically generate the kind of open, participatory data ecosystems characteristic of Northern Europe

Among non-Western languages, Chinese exhibits the highest level of resource availability. Nevertheless, the volume of openly accessible Chinese speech data remains limited relative to the size of the language community, which is consistent with findings

reported in prior studies (Tang et al., 2021; Zhou et al., 2025). Despite high user activity and a large domestic market, substantial portions of Chinese speech data are not released under open licenses, reflecting state-level data governance policies and the closed nature of many commercial platforms. At the same time, several large Chinese technology companies have made portions of their datasets publicly available, enabling continued progress in selected research directions. This apparent discrepancy is a hallmark of the Chinese model: the state's emphasis on digital sovereignty and centralized control leads to the creation of large domestic corpora, but these are often not released under open licenses (Roberts et al., 2021). The low DRSI thus does not indicate a lack of data per se, but a political choice to restrict openness – a form of "sovereignty through enclosure" that contrasts sharply with the Western paradigm. This aligns with the "Confucian civilizational cluster" (Inglehart and Welzel, 2005), where modernization proceeds through strong state guidance and corporate-group mentalities rather than through individualistic, open civic participation.

The Russian language occupies a mid-range position in the DRSI ranking, placed 12th overall, behind Chinese and slightly ahead of Polish. Its value is comparable to that of non-dominant European languages, yet it is significantly below its demographic potential. This is symptomatic of what might be termed a dependent modernization trajectory. Russia exhibits a post-Soviet political consciousness and a secular-rational value system (Artyukhin et al., 2025), but the institutions supporting open science and data sharing remain fragmented. Although the availability of high-quality open speech datasets for Russian has gradually increased (Andrusenko et al., 2020; Karpov et al., 2021), a significant share of relevant resources remains proprietary and inaccessible to the research community. The overall volume of openly available data remains insufficient for a number of speech-processing tasks, and data quality often requires additional preprocessing prior to use. Despite a large digital audience and significant domestic research activity, the absence of a coordinated open-data infrastructure and the historical reliance on Western platforms have created a structural deficit. This structural limitation constrains the development and evaluation of speech technologies based exclusively on domestic open resources. The Russian case illustrates how intermediate models, caught between openness and sovereignty, can result in a persistent resource gap – a form of "incomplete modernization" where demographic potential fails to translate into digital infrastructure.

India presents a unique and paradoxical case. Hindi, the most widely spoken language in India, has a very low DRSI, placing it among the most severely resource-constrained languages in our sample – despite having over 600 million speakers. Yet English, which is also an official language of India and widely used in administration, education, and technology, enjoys the highest possible resource saturation. This bifurcation reflects the enduring legacy of colonial linguistic hierarchies and "socio-cultural mosaicity" in the South Asian civilizational cluster. While India has developed robust policies for linguistic diversity and peaceful conflict resolution, the infrastructure for open data – particularly for non-English languages – remains profoundly underdeveloped (Artyukhin et al., 2025). The result is a form of internal digital stratification: English serves as the language of technological participation, while Hindi

and other Indic languages are systematically excluded from the foundational datasets of AI, despite their demographic weight. This challenges the assumption that democracy and demographic scale alone guarantee technological inclusion.

The lower end of the ranking is mostly occupied by languages of the Arab world and South-East Asia (Arabic, Tamil, Vietnamese, Indonesian, Thai, Malay). They exhibit extremely low DRSI values. These figures challenge the assumption that demographic scale automatically translates into technological influence in a multipolar world. Instead, they demonstrate how older global hierarchies are reproduced in digital form: despite hundreds of millions of speakers, these language communities remain largely invisible in the open-data landscape. This is not merely a technical deficit but a form of epistemic marginalization, where the data needed to build culturally and linguistically appropriate AI systems is systematically absent. In the “African-Islamic” and “South Asian” clusters, religious norms, clientelist political cultures, and post-colonial institutional weaknesses have created environments where open, inclusive data infrastructures struggle to emerge (Inglehart and Welzel, 2005). Importantly, the analysis includes only languages classified as *Thriving*, which implies the presence of at least a minimal level of speech-resource availability. Languages outside the scope of this study are therefore likely to face even more severe resource constraints.

In sum, the DRSI not only quantifies resource imbalance but also serves as an expression of the political and institutional regimes governing data. Low scores are not just a call for more data; they are an indicator of how different modernization paths – whether Western openness, Chinese sovereign enclosure, Latin American "modernization through tradition," South Asian linguistic stratification, or intermediate dependency – shape a language's capacity to participate in the AI revolution on its own terms. The index reveals that the digital language divide is not a simple function of demography, but a complex product of civilizational legacies, governance choices, and the structural inequalities embedded in the global technological order.

CONCLUSION AND LIMITATIONS

This study has several limitations that should be taken into account when interpreting the results.

Firstly, the analysis focuses exclusively on open speech datasets documented or widely cited as established reference resources in research papers. Commercial, proprietary, and most of the community-curated datasets were excluded, which may partially distort the overall picture of resource availability, particularly for languages developed under models with limited data openness, such as the Chinese model. In addition, no manual validation was conducted to assess the feasibility of obtaining missing data from excluded sources.

Secondly, the selection was restricted to languages with a relatively advanced digital ecosystem. While this ensures that all included languages have at least some level of speech-resource availability, it also excludes languages with weaker digital support, for which the degree of inequality may be even more pronounced.

Thirdly, the analysis represents a static snapshot of the data landscape at the time of collection and does not capture its dynamic evolution. New datasets – particularly multilingual ones – are continuously released and often expand both language coverage and data volume beyond previously available resources.

Fourthly, the proposed DRSI metric has inherent limitations and therefore is necessarily constrained by its reliance on readily measurable quantities – primarily audio hours and speaker counts. While this approach enables systematic cross-linguistic comparison, it does not incorporate qualitative characteristics of datasets, such as linguistic diversity, representativeness, annotation quality, or metadata completeness and other dimensions that are equally crucial for understanding technological capacity: the cultural representativeness of datasets, the diversity of voices and accents included, the quality of annotations, and the institutional mechanisms that sustain data production over time. These factors, though difficult to quantify, are essential for assessing whether a language community truly possesses the capacity to develop sovereign AI systems, or merely the raw material. Moreover, the sensitivity of the metric to variations in the parameters α, β, and γ requires further validation. The DRSI should therefore be understood as a diagnostic indicator of structural imbalance, not a comprehensive measure of technological readiness.

Finally, while we observe parallels between DRSI patterns and civilizational clusters described in comparative sociology (Inglehart & Welzel, 2005), this study does not establish correlation, causation, or explanatory linkage. These observations are intended as heuristic illustrations for future research, not as validated theoretical claims.

Despite these limitations, the results confirm the existence of a pronounced digital language divide in open speech resources even among languages classified as having high levels of digital support (DLS "Thriving"). Languages with comparable demographic scales and substantial digital audiences exhibit significant disparities in the availability of open speech corpora. English occupies a clearly privileged position: its resource base not only exceeds that of other languages in quantitative terms, but also demonstrates higher institutional stability and reproducibility due to the combination of an open scientific ecosystem and large-scale corporate contributions. At the same time, systematic asymmetries are observed even among major "developed" languages, with most languages exhibiting deficits in open resources relative to levels expected under demographic and fairness-based assumptions.

The application of the DRSI metric enables a shift from merely identifying inequality to providing a measurable description of resource imbalance and a ranking of languages by degree of disparity. Importantly, near-zero index values are observed not only for traditionally low-resource languages, but also for several languages with international status. In particular, Russian – despite a large speaker base and the presence of several large-scale initiatives – remains in a zone of moderate deficit in open speech resources and shows levels of imbalance comparable to those of European languages that do not occupy dominant positions in the global AI ecosystem. Chinese, while leading among non-Western languages, also exhibits a disproportionately low index value relative to its demographic scale, indicating a mismatch between the volume of open data

and the size of the language community and highlighting the role of political and institutional constraints on data openness.

Interpreting these findings through the lens of technological modernization models demonstrates that language inequality is shaped not only by technical factors, but also by institutional mechanisms. These include corporate centralization in the Western model, digital sovereignty strategies and restricted data openness in the Chinese model, and fragmented coordination and insufficient support for open infrastructures in intermediate development models. Accordingly, the digital language divide should be understood as a reproducible structure of digital inequality that reinforces technological dependence and constrains the development potential of national AI ecosystems.

Looking beyond the immediate findings, this study underscores a fundamental shift in how we understand technological development in the AI era. The classical modernization theory assumption that development follows a linear path toward a Western endpoint has given way to a recognition of "multiple modernities." Our DRSI analysis demonstrates that this plurality extends to the very raw materials of AI: speech data. The stark asymmetries we observe are not temporary gaps that will be closed by market forces or demographic growth alone. They are structural features of a global system where data governance regimes – the choices about openness, sovereignty, and institutional support – actively shape which languages can participate in the AI revolution and on what terms.

For languages in intermediate positions, such as Russian, the path forward requires more than simply increasing dataset hours. It demands strategic choices about data governance: whether to prioritize openness and global integration, risking continued dependence on Western platforms, or to pursue sovereign data infrastructures, potentially at the cost of interoperability and community contribution. For languages in the South Asian, Latin American, and African-Islamic clusters, the challenge is even more fundamental: building the institutional frameworks and civic cultures that can sustain open data production over the long term, often in the face of post-colonial legacies and persistent socio-economic constraints.

Ultimately, the digital language divide is not a technical problem awaiting a technical solution. It is a political and civilizational question about who gets to shape the future of human-machine interaction, whose voices are heard by AI systems, and what forms of linguistic and cultural diversity will survive the transition to a globally integrated technological order. The DRSI, for all its limitations, provides a lens through which these deeper questions can be seen – and, perhaps, addressed.

REFERENCES

Adebara, I., Toyin, H. O., Ghebremichael, N. T., Elmadany, A. A., & Abdul-Mageed, M. (2025). Where Are We? Evaluating LLM Performance on African Languages. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 32704–32731). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.1572

Agnew, W., Barnett, J., Chu, A., Hong, R., Feffer, M., Netzorg, R., Jiang, H. H., Awumey, E., & Das, S. (2024, October). *Sound Check: Auditing Audio Datasets*. arXiv. https://doi.org/10.48550/arXiv.2410.13114

Akindotuni, D. (2025). Resource Asymmetry in Multilingual NLP: A Comprehensive Review and Critique. *Journal of Computer and Communications*, *13*(7), 14–47. https://doi.org/10.4236/jcc.2025.137002

Andrusenko, A., Laptev, A., & Medennikov, I. (2020). Exploration of End-to-End ASR for OpenSTT – Russian Open Speech-to-Text Dataset. In A. Karpov & R. Potapova (Eds.), *Speech and Computer* (Vol. 12335, pp. 35–44). Springer, Cham. https://doi.org/10.1007/978-3-030-60276-5_4

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2020, March). Common Voice: A Massively-Multilingual Speech Corpus. *arXiv*. https://doi.org/10.48550/arXiv.1912.06670

Artyukhin O.A., Kritskaya A.A., & Samgurov A.S. (2025). Political modernization of non-Western countries: Socio-cultural and civilizational features. *State and Municipal Management. Scholar Notes, 2*, 163–170. https://elibrary.ru/UHSDXE

Baishya, Dr. D., Baruah, Dr. R., Bora, Dr. M., & Sarma, B. (2025). Processing Low-Resource Languages: A Review of Challenges and Strategies for Inclusive NLP And Sustainable Environment. *International Journal of Environmental Sciences*, *11*, 7730–7739. https://doi.org/10.64252/w55rwj24

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM. https://doi.org/10.1145/3442188.3445922

Bird, S. (2020). Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 3504–3519). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.313

Blasi, D., Anastasopoulos, A., & Neubig, G. (2022). Systematic Inequalities in Language Technology Performance across the World's Languages. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5486–5505). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.376

Christophe, B., Giron, A., & Verin, G. (2023). A comparative analysis with machine learning of public data governance and AI policies in the European Union, United States, and China. *Journal of Intelligence Studies in Business*, *13*(2), 61–74. https://doi.org/10.37380/jisib.v13i2.1084

Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., & Bapna, A. (2022, May). *FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech*. arXiv. https://doi.org/10.48550/arXiv.2205.12446

Derivation – Faster, easier, smarter multilingual business. (2025). https://derivation.co/

Eberhard D.M., Simons G.F., & Fennig C.D. (Eds.). (2025). *Ethnologue: Languages of the World. Twenty-eighth edition.* https://www.ethnologue.com/

Gábor, B., Helm, P., Koch, G., & Giunchiglia, F. (2023, July). *Towards Bridging the Digital Language Divide*. arXiv. https://doi.org/10.48550/arXiv.2307.13405

Gaido, M., Papi, S., Bentivogli, L., Brutti, A., Cettolo, M., Gretter, R., Matassoni, M., Nabih, M., & Negri, M. (2024). MOSEL: 950,000 Hours of Speech Data for Open-Source Speech Foundation Model Training on EU Languages. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 13934–13947). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.771

Gaspari, F., Gallagher, O., Rehm, G., Giagkou, M., Piperidis, S., Dunne, J., & Way, A. (2022). Introducing the Digital Language Equality Metric: Technological Factors. In I. Aldabe, B. Altuna, A. Farwell, & G. Rigau (Eds.), *Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference* (pp. 1–12). European Language Resources Association. https://aclanthology.org/2022.tdle-1.1/

Grützner-Zahn, A., & Rehm, G. (2022). Introducing the Digital Language Equality Metric: Contextual Factors. In I. Aldabe, B. Altuna, A. Farwell, & G. Rigau (Eds.), *Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference* (pp. 13–26). European Language Resources Association. https://aclanthology.org/2022.tdle-1.2/

He, H., Shang, Z., Wang, C., Li, X., Gu, Y., Hua, H., Liu, L., Yang, C., Li, J., Shi, P., Wang, Y., Chen, K., Zhang, P., & Wu, Z. (2025). Emilia: A Large-Scale, Extensive, Multilingual, and Diverse Dataset for Speech Generation. *IEEE Transactions on Audio, Speech and Language Processing*, *33*, 4044–4054. https://doi.org/10.1109/TASLPRO.2025.3612835

Helm, P., Bella, G., Koch, G., & Giunchiglia, F. (2024). Diversity and language technology: How language modeling bias causes epistemic injustice. *Ethics and Information Technology*, *26*(1), 8. https://doi.org/10.1007/s10676-023-09742-6

Henning, S., Beluch, W., Fraser, A., & Friedrich, A. (2023). A Survey of Methods for Addressing Class Imbalance in Deep-Learning Based Natural Language Processing. In A. Vlachos & I. Augenstein (Eds), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 523–540). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.38

Imdat, C. (2019, March). *M-AILABS Dataset*. Github. https://github.com/i-celeste-aurora/m-ailabs-dataset

Inglehart, R., & Welzel, C. (2005). Modernization, Cultural Change, and Democracy: The Human Development Sequence. *Cambridge: Cambridge University Press.* https://doi.org/10.1017/CBO9780511790881

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293).

Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

Karpov, N., Denisenko, A., & Minkin, F. (2021). Golos: Russian Dataset for Speech Research. *Interspeech*, 1419–1423. https://doi.org/10.21437/Interspeech.2021-462

Kolobov, R., Okhapkina, O., Platunov, A., Omelchishina, O., Bedyakin, R., Moshkin, V., Menshikov, D., & Mikhaylovskiy, N. (2021). *MediaSpeech: Multilanguage ASR Benchmark and Dataset*. OpenSLR. https://www.openslr.org/108/

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., Van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Suarez, P. O., … Adeyemi, M. (2022). Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, *10*, 50–72. https://doi.org/10.1162/tacl_a_00447

Lau, M., Chen, Q., Fang, Y., Xu, T., Chen, T., & Golik, P. (2025). *Data Quality Issues in Multilingual Speech Datasets: The Need for Sociolinguistic Awareness and Proactive Language Planning*. arXiv. https://doi.org/10.48550/ARXIV.2506.17525

Li, X., Takamichi, S., Saeki, T., Chen, W., Shiota, S., & Watanabe, S. (2023). Yodas: Youtube-Oriented Dataset for Audio and Speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 1–8). IEEE https://doi.org/10.1109/ASRU57964.2023.10389689

Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., Wu, X., Shippole, E., Bollacker, K., Wu, T., Villa, L., Pentland, S., & Hooker, S. (2024). A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, *6*(8), 975–987. https://doi.org/10.1038/s42256-024-00878-8

Markl, N., & McNulty, S. J. (2022). Language technology practitioners as language managers: Arbitrating data bias and predictive bias in ASR. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 6328–6339). European Language Resources Association. https://aclanthology.org/2022.lrec-1.680/

Neznamov, A., Chache, E., & Churilova, D. (2025). What is "Regulatory Path" for Russia? *Legal Issues in the Digital Age*, *6*(3), 4–22. https://doi.org/10.17323/2713-2749.2025.3.4.22

Oliveira, F. S., Casanova, E., Júnior, A. C., Soares, A. S., & Filho, A. R. G. (2023, June). *CML-TTS A Multilingual Dataset for Speech Synthesis in Low-Resource Languages*. arXiv. https://doi.org/10.48550/arXiv.2306.10097

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., & Collobert, R. (2020). MLS: A Large-Scale Multilingual Dataset for Speech Research. *Interspeech 2020*, 2757–2761. https://doi.org/10.21437/Interspeech.2020-2826

Ranathunga, S., & Silva, N. de. (2022). *Some Languages are More Equal than Others: Probing Deeper into the Linguistic Disparity in the NLP World*. arXiv. https://doi.org/10.48550/arXiv.2210.08523

Rehm, G., & Way, A. (2023). European Language Equality: Introduction. In G. Rehm & A. Way (Eds.), *European Language Equality: A Strategic Agenda for Digital Language Equality* (pp. 1–10). Springer International Publishing. https://doi.org/10.1007/978-3-031-28819-7_1

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. *AI & Society*, *36*(1), 59–77. https://doi.org/10.1007/s00146-020-00992-2

Simons, G. F., Thomas, A. L., & White, C. K. (2022, September). *Assessing Digital Language Support on a Global Scale*. arXiv. https://doi.org/10.48550/arXiv.2209.13515

Tang, Z., Wang, D., Xu, Y., Sun, J., Lei, X., Zhao, S., Wen, C., Tan, X., Xie, C., Zhou, S., Yan, R., Lv, C., Han, Y., Zou, W., & Li, X. (2021, August). KeSpeech: An Open Source Speech Dataset of Mandarin and Its Eight Subdialects. *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. https://openreview.net/forum?id=b3Zoeq2sCLq

Vincent, N., Hecht, B., & Sen, S. (2019). "Data Strikes": Evaluating the Effectiveness of a New Form of Collective Action Against Technology Companies. *The World Wide Web Conference*, 1931–1943. https://doi.org/10.1145/3308558.3313742

Wang, C., Rivière, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., & Dupoux, E. (2021, July). *VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation*. arXiv. https://doi.org/10.48550/arXiv.2101.00390

Wang, C., Wu, A., & Pino, J. (2020, October). *CoVoST 2 and Massively Multilingual Speech-to-Text Translation*. arXiv. https://doi.org/10.48550/arXiv.2007.10310

Zapf, W. (2004). Modernization theory – And the non-western world. *WeltTrends – Zeitschrift für Internationale Politik, Wissenschaftszentrum Berlin für Sozialforschung, 3*(44), 100–107. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-110362

Zhou, J., Wang, S., Zhao, S., He, J., Sun, H., Wang, H., Liu, C., Kong, A., Guo, Y., Yang, X., Wang, Y., Lin, Y., & Qin, Y. (2025). ChildMandarin: A Comprehensive Mandarin Speech Dataset for Young Children Aged 3-5. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 12524–12537). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.614

СВЕДЕНИЯ ОБ АВТОРАХ / THE AUTHORS

Байрамова Хумай, khumay.bayramova@gmail.com, ORCID 0009-0002-3281-057X

Гаврилов Антон, avgavriliov@itmo.ru, ORCID 0000-0002-9917-6609

Николаев Владимир, vladimir.nikolaev@cs.ifmo.ru, ORCID 0000-0002-3224-3934

Харитонова Анастасия, aekharitonova@itmo.ru, ORCID 0000-0001-6493-3801

Khumai Bairamova khumay.bayramova@gmail.com, ORCID 0009-0002-3281-057X

Anton Gavrilov, avgavriliov@itmo.ru, ORCID 0000-0002-9917-6609

Vladimir Nikolaev, vladimir.nikolaev@cs.ifmo.ru, ORCID 0000-0002-3224-3934

Anastassia Kharitonova, aekharitonova@itmo.ru, ORCID 0000-0001-6493-3801

Статья поступила 30 декабря 2025
одобрена после рецензирования 27 февраля 2026
принята к публикации 18 марта 2026

Received: 30 December 2025
Revised: 27 February 2026
Accepted: 18 March 2026