

Computer and Decision Making

An International Journal

www.comdem.org
eISSN: 3008-1416



Globalization and Preservation of National Languages: Challenges of the Digital Age

Nadiia Senchylo-Tatlilioglu^{1,*}, Tetiana Nedashkivska², Lesia Kushmar³, Lilia Pryimak⁴, Anzhelika Cherniuk⁵

¹ Faculty of Literature, Ataturk University, Erzurum, Turkey

² Department of Slavic and Germanic Philology and Translation, Educational and Scientific Institute of Philology and Journalism, Zhytomyr Ivan Franko State University, Zhytomyr, Ukraine

³ Department of Foreign Philology and Translation, State University of Trade and Economics, Kyiv, Ukraine

⁴ Department of Foreign Languages and Country Studies, Vasyl Stefanyk Carpathian National University, Ivano-Frankivsk, Ukraine

⁵ Department of Foreign Philology and Translation, Interregional Academy of Personnel Management, Kyiv, Ukraine

ARTICLE INFO

Article history:

Received 1 July 2026

Received in revised form 25 July 2026

Accepted 30 July 2026

Available online 31 July 2026

Keywords:

Globalization; linguistic diversity; national languages; digitalization; artificial intelligence; Internet; language policy; low-resource languages; linguistic inequality; web content.

ABSTRACT

The article explores how globalization and digitalization affect the preservation of national or minority languages in the modern world. The aim of the study was to identify and interpret the relationships between global linguistic diversity, the level of digitalization of countries and their socio-economic development, as well as to assess the potential risks and opportunities provided by the digital environment for linguistic heritage. The theoretical basis of the study includes modern sociolinguistic approaches, digital studies and language policy. The empirical base of the study is taken from international data from open sources (Our World In Data, World Bank, W3Techs, Ethnologue and UNESCO). The following statistical methods were used to process statistical data: descriptive statistics, Spearman and Kendall rank correlations, Kraskel-Wallis test for nonparametric statistical testing, and comparative analysis of temporal trends. In the course of the study, the authors found that linguistic diversity is unevenly distributed throughout the world, a comparative analysis was carried out by continents and countries of the world. A statistically significant negative correlation was found between the amount of living languages and the extent of internet use in the countries of the world, which suggest an asymmetry between language representation and internet inclusion. Statistically significant differences between groups of countries in terms of income level were also found: the average number of living languages in high-income countries is less than in lower-income countries. An analysis of the linguistic composition of web materials in the world from 2015 to 2026 showed that the English language retains its dominance, but the shares of Spanish and Portuguese are growing. The scientific novelty of the study lies in the complex combination of sociolinguistic and statistical analysis of global language processes in the context of the digital economy. The practical significance of the study lies in the formation of recommendations for national and international language policy, in particular the development of digital data corpora, the localization of technological platforms, the support of low-resource languages in artificial intelligence systems, and the expansion of language communities' access to digital infrastructure.

* Corresponding author.

E-mail address: matst@ukr.net

1. Introduction

Globalization is the term to describe the unprecedented integration of economies, cultures and technologies that characterizes the modern world. It has undoubtedly transformed our world for the better: transition of knowledge across the borders becoming faster, fresh ideas and innovations spreading more quickly and stimulation for economic development. But it has also posed serious threats to the cultural and linguistic diversity of our world. Languages, which for centuries had been the conveyors of people's worldviews, knowledge of nature, cultures and identities, have become subject to enormous pressure.

In today digital era, not only these processes persist but they are also radically speeded up. The Internet, social media, search engines and cutting-edge AI models are establishing a new world in which high-resource lingos receive disproportionately heavy investment while the vast majority of national languages are pushed aside in an ever-more distinct digital world. Algorithms that decide what information appears on our screen, which services are easily available to us and even our way of studying are mainly "tuning" in the English language or a few other popular lingos. Everything else is ignored by the giants of technologies [1].

The figures behind these actions are striking. There are approximately 7170 living languages in the world today according to Ethnologue [2]. Only approximately 20 of these major languages account for more than 88% of the people in the entire world. The majority of these languages are spoken by small assemblies. As per UNESCO [3], at least 40% of all the languages are endangered, with a foreign language becoming extinct every two weeks on an average.

Digitalization does only establish new pressure on this fact. As in 2026, 49.4% [4] of www content will be in English. It will be much more difficult for billions of nonEnglish speakers to have access to information, education, business and culture exchange via the digital space. Native languages, even large like Ukrainian, will have less than 1% of share and be invisible for modern technologies.

It is particularly frightening that the loss of a language is not merely the extinction of a specific tool of expressiveness comprising a set of words and a system of rules governing its use. The loss of a language is also the loss of a particular conception of the world which humans have developed over many centuries and handed down from generation to generation in the practice of medicine, the study of ecology, the intimate interaction with nature, and the experience of human relations.

The topic is of particular relevance today. The United Nations has declared the International Decade of Indigenous Languages 2022-2032 and has officially expressed their awareness of their need for committed action [5]. In this context, scientific research that combines the analysis of globalization processes, digital technologies and specific strategies for the preservation of languages becomes not only theoretical, but also practically important.

The purpose of this article is to comprehensively analyze the challenges faced by national and minority languages in the context of globalization and digital transformation, to show real statistical trends with the help of verified data and graphs, as well as to outline possible ways and tools for their preservation in the modern world.

2. Literature Review

Modern scientific literature on the interaction of globalization with digital technologies, as well as with the preservation of national and minority languages, gives a fairly clear answer: globalization

is realized both as a destructive and as a saving force. Economists, as they did a few years ago, support the consensus that economic integration, mass migration [6], and the dominance of two or three world languages lead to linguistic conformism, while to the emergence of an increasing number of digital tools for documenting and revitalizing languages that are threatened.

The first large block of works is devoted precisely to how the changes brought by globalization affect the diversity of languages. Huanova [7] and Sheraliyeva [8] clearly illustrate how globalization is transforming English into a modern lingua franca that marginalizes regional and minority languages. The authors emphasize that economic motives, media and international trade are forcing the younger generation to switch to dominant languages, which leads to the gradual disappearance of cultural identity. Kearney [9] analyzes the impact of globalization on linguistic culture and notes that socioeconomic factors can contribute to changes in the use of languages, in particular through the growing role of more "globally appropriate" languages in communication. Lalombo [10] provides evidence that the process of globalization is contributing to a transition towards Indonesian and English, which is explained by their functional role in education, on social media and in socio-economic opportunities. Zhang [11] critically analyses the process and gives the examples that globalization could have both negative effects on small languages as well as new opportunities for protection and spread through digital technologies and new media. The author highlights that the impact of these processes is uneven. It largely depends on the state language policy, socio-economic and technological conditions.

The second significant part of the research, then, shall deal with the digital aspects of the same problem called digital divide with respect to low-resource languages. Hutson *et al.* [1] offer a sophisticated framework for the preservation of linguistic as well as cultural traditions relying on machine learning and big data. The authors argue that digital technologies provide new opportunities for documenting languages, which allows for faster data collection and analysis as compared to manual procedures [12]. Mor [13] developed the notion of a digital sidelining, illustrating structural disparities pervading the training of LLMs using Anglocentric corpora. As a result, low-resource languages become underrepresented in computing technologies, which, in turn, results in their invisibility in AI systems. Yet, even more, undeniable proof of this phenomenon is provided by the recent data from Stanford News [14]: lower-resource languages are inted in translation and text generation with disastrously low accuracy, thereby deepening worldwide disparities in accessing "education, health care, and knowledge".

Regional studies are also encouraging. Muawanah *et al.* [15], in their case study of the Indonesian context, explore the potential for digital literacy, artificial intelligence and e-learning to support efforts to safeguard regional languages. They highlight that the effectiveness of a tool is largely contingent on the presence of both digital and language data of sufficient quality. Mullah and Zainon [16], on the theme of the linguistic digital divide, stress that "out of more than 7000 living languages in the world, most remain 'digitally invisible' and this not only presents a risk to languages, but to the like-ing knowledge systems that they encapsulate". Wang [17] also observed that the needs of small languages (and linguistic minorities) are often overlooked by national AI policy in multilingual countries, which could further marginalize their digital presence and include them less frequently in technology.

The third sections of the sources then address conservation methods and "savior" technology. Bird [18] warned a few years ago: technology can be used to accelerate the disappearance of languages, but it can also be a powerful tool for documenting and reviving languages, however only with the ethical participation of language communities themselves. Kumar *et al.* [19] and Nasrullah [20] analyze the potential of modern natural language processing technologies (in particular,

machine translation and AI-related techniques) for a low-resource region. The authors emphasize that these approaches can be used to support documentation efforts, as well as to adapt and improve language resources, but their success is highly dependent on the availability of data and the technological limitations specific to each language environment. Chang *et al.* [21] demonstrate using the example of the Hakka language how a combination of augmented search generation (RAG) and large models can improve translation quality from 12% to 31% on the BLEU scale, but only from curated data. The results of a study by the University of Hawai'i [22] confirm these conclusions with Formosan material on the languages of Taiwan: without the active participation of communities and technologies for creating open datasets, they remain ineffective.

Special attention in modern scientific research is paid to socio-cultural and psychological challenges that accompany globalization processes in the educational space. Researchers emphasize the need to implement strategies that take into account intercultural aspects of learning to overcome barriers in a multinational environment [23]. An important component of integration is the adaptation of persons who have received temporary shelter to new educational conditions, which requires comprehensive psychological and pedagogical support [24]. Along with this, the development of the digital presence of languages is closely related to the investment approaches of universities to the creation and improvement of educational platforms that contribute to the preservation of knowledge [25]. At the same time, the formation of self-efficacy mechanisms and comparison in specific types of activities helps to better understand the structure of adaptation and development of the individual in the face of modern challenges [26].

Overall, the literature shows that globalization and digital transformation simultaneously threaten linguistic diversity and offer new ways to preserve it. The most important problem noted by the researchers is the exhaustiveness of ethical, community-driven language information and the lack of coordination of language policy at the international and national levels. This creates a research gap that this study fills by studying current digital trends and the role of artificial intelligence in protecting national languages.

3. Materials and Methods

The study combines several areas of knowledge: sociolinguistics, digital studies, comparative economics and applied statistics. Such a combination is justified by the fact that modern trends in the preservation or disappearance of languages are formed by mutually cultural, economic and technical factors. The task of the empirical part was to identify the links between the linguistic diversity of countries, the level of digitalization and the socio-economic characteristics of states.

The work is quantitative and comparative in nature and is based on the analysis of secondary international data. The main analysis was carried out at the cross-sectional level of the countries of the world, while the temporal analysis for 2015–2026 was used to assess changes in the language presence on the Internet. This made it possible to combine the spatial and dynamic dimensions of the study.

The source base was made up of open international resources. Sources on the number of living languages in the country are taken from Our World in Data [27], based on Ethnologue. The indicator of Internet penetration among the population and the classification of countries by income are taken from the World Bank [28]. To analyze languages in global web content, W3Techs [4] statistics were used. UNESCO and Ethnologue materials were used to interpret the results.

The key variables were: the number of living languages in the country, the level of Internet penetration, the state's income group, continental affiliation, and the share of languages in web

content. Before the analysis, the data was cleaned, brought to a single format and combined by country.

The total sample included 193 countries for which data on the number of living languages as of 2025 were available from Our World in Data (based on Ethnologue). Internet penetration (% of population) is taken from the World Bank for 2023 (the most recent year with the most complete coverage); for countries where 2023 was missing, 2022 data was used. The country income classification is based on the World Bank version for fiscal year 2025. Information on language shares in web content covers the period from January 2015 to April 2026 according to W3Techs monthly reports. Missing values were excluded by pairwise deletion for correlation analysis; only countries for which both language count and income group were available ($n = 189$) were used for the Kruskal–Wallis test. The reconciliation of country names between different sources was performed according to the ISO 3166-1 alpha-3 standard. The use of non-parametric methods (Spearman and Kendall coefficients, Kruskal–Wallis test) is due to the fact that the distributions of the number of living languages and the level of Internet penetration significantly deviated from normal (according to the Shapiro–Wilk test $p < 0.001$), as well as the presence of outliers. Rank methods do not require assumptions about linearity and normality, which is consistent with the structure of the data.

Taking into account the uneven distribution of indicators and the presence of extreme values, nonparametric methods were used. To identify the relationship between the number of languages and the level of Internet use, the Spearman and Kendall correlation coefficients were used. To check the differences between groups of countries by income, the Kruskal–Wallis criterion was used, followed by the Bonferroni correction. The dynamics of languages in the web space is studied according to the descriptive-trend approach.

The results are presented in the form of diagrams, cartographic visualizations, boxplot graphs and time time series. The use of open international data ensures the reproducibility of the study, although the number of living languages and the share of languages in web content do not fully reflect their real viability or everyday use. Therefore, the conclusions should be interpreted as a macro-level analysis of global trends.

4. Results

To better understand how linguistic diversity is distributed in the world today, it is advisable to analyze the number of living languages by continent and country of the world based on data from Our World in Data [27]. A living language is understood as a language that at least one person speaks as a native language. Such an indicator makes it possible to assess the current state of the linguistic wealth of mankind and identify regions where the issue of language preservation is especially relevant.

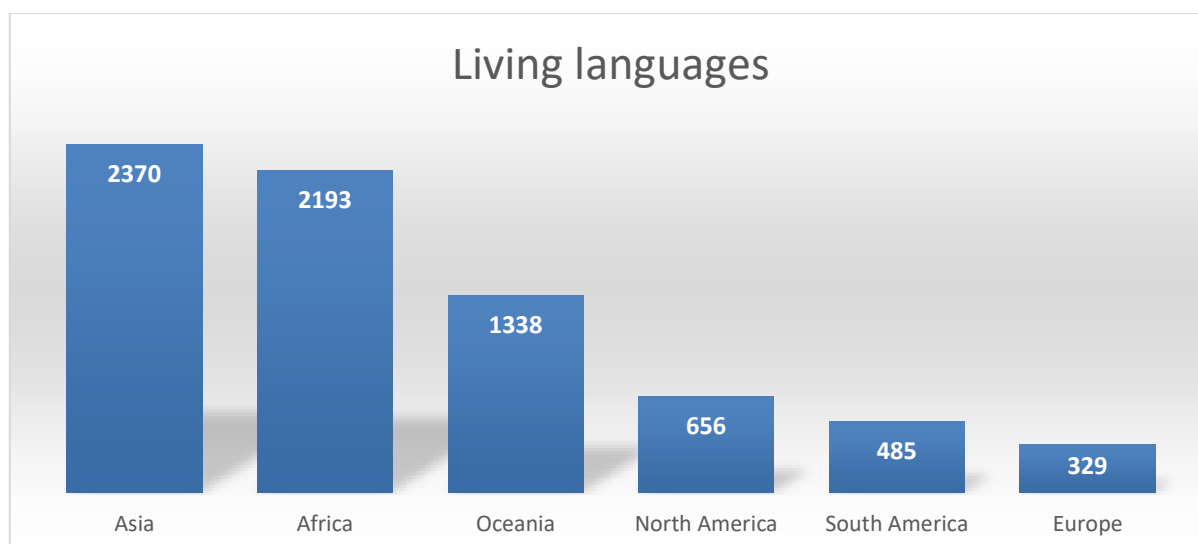


Fig. 1. Number of living languages in the world by continent as of 2025
Source: Built by the authors based on data Our World in Data [27]

The data collected show a very uneven distribution of linguistic diversity across continents (Figure 1). Most living languages are concentrated in Asia (2370). Almost the same amount falls on Africa (2193). It is these two regions that are the main centers of the linguistic wealth of the planet. This situation is explained by the large population, vast territories, the centuries-old history of the development of various civilizations and the existence of many ethnic and indigenous communities that have preserved their linguistic traditions.

The third position is occupied by Oceania with 1338 living languages. This is impressive, because the region is not too large in terms of population. Therefore, the number of languages is not always closely related to the number of people. The high linguistic diversity of Oceania is mainly due to Papua New Guinea and the numerous islands of the Pacific Ocean, where for centuries isolation geographical has contributed to the emergence of many individual languages.

There are 656 living languages in North America and 485 in South America. For both continents, the current linguistic situation is largely shaped by the history of colonization, the spread of English, Spanish, Portuguese and French, and the gradual displacement of many indigenous languages. At the same time, they remain important centers of linguistic heritage, especially where autochthonous communities live.

Europe has the fewest living languages (329). Such a low figure can be explained by the long processes of centralization of states, the formation of national standards, school standardization, migration to cities, and the predominance of a limited number of official languages in the public sphere.

In general, the data suggest that the world's linguistic wealth is mainly concentrated in Asia, Africa and Oceania. It is these regions that require special attention in language policy and digital development, because many languages here are simultaneously numerous in number, but vulnerable to globalization, migration processes and poor representation in modern technologies.

At the same time, continental indicators do not reflect the internal heterogeneity of the linguistic space, since linguistic diversity is very unevenly distributed within each region. Within the continents, there are leading countries that concentrate a significant part of all living languages in the region, while other states have much more modest indicators (Figure 2).

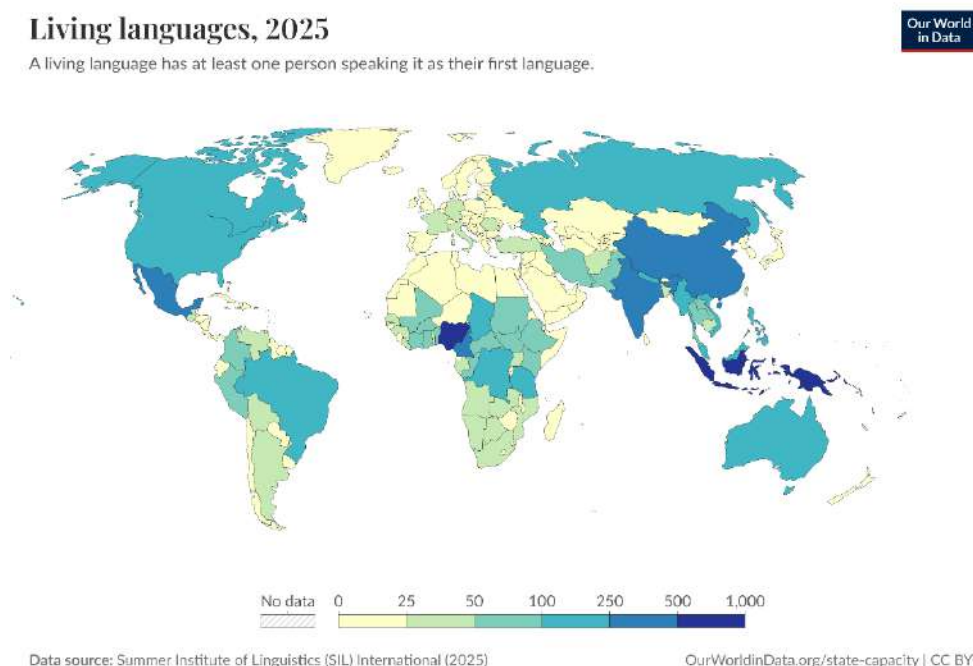


Fig. 2. Geographical distribution of the number of living languages in the countries of the world as of 2025
Source: Our World in Data [27]

In Asia, where there are a total of 2370 living languages, linguistic diversity is concentrated mainly in a few countries. Most languages are used in Indonesia (709), which is explained by the country's extensive archipelagic structure, the long isolation of the islands and the rich ethno-cultural mosaic. A large chain of linguistic diversity is also visible in India (454 languages) and China (309), where the richness of languages has been formed thanks to the long history of civilizations and significant regional heterogeneity. At the same time, in many countries of Western Asia and the Middle East, the number of living languages is much lower.

Africa, which ranks second in the world in terms of linguistic diversity (2193), has a high internal concentration. Nigeria has the most languages in use (503), making it one of the key language centers on the planet. It is interesting to emphasize the countries: Cameroon(280), DRC(211) and Ethiopia(92). As far as the North-African countries are concerned, a more homogeneous landscape in terms of languages because the presence of Arabic is very strong, and also, a process of standardization, which is very long.

There is quite an important difference in Oceania. Although the population there is not very large, the region has 1338 living languages, and almost all of them are concentrated in one country - Papua New Guinea, where 843 languages are recorded. This is a record for an individual state, a world leader with such an indicator. This situation is explained by the complex relief of mountains, the isolation of communities and the long-term autonomous development of local communities. For comparison, in many other states of Oceania, there are much fewer languages.

In North America, 656 languages have been recorded, but even there, concentration is observed in several countries. The United States (239), Mexico (296), and Canada (101) have the most languages. Much of the diversity is related to indigenous languages and immigrant language communities. On the other hand, in the smaller states of the Caribbean, the number of languages is much lower.

In South America, which has 485 languages, the following countries stand out most clearly: Brazil (222), Peru (95), Colombia (89) and Bolivia (44). It is here that the numerous indigenous languages of the Amazon and the Andean region are preserved. On the other hand, in more urbanized countries, where there is a stronger dominance of Spanish or Portuguese, linguistic diversity is significantly lower.

Europe, which has the fewest living languages (329), is also not homogeneous. The highest diversity is shown by Russia (119), Spain (21), France (36) and Italy (36) due to their multinational population structure, regional languages and historical minorities. At the same time, a significant part of European states have only a few dominant languages, and the language space as a whole is much more standardized compared to other continents.

This shows how the linguistic diversity is not concentrated between continents but also within themselves. In each region, there are separate countries that act as global or regional centers of linguistic wealth. They need special attention in the field of language policy, digital representation and programs for the preservation of language heritage.

Additionally, we will investigate the relationship between the number of living languages in countries and the level of Internet use (% of population) based on World Bank data [28]. Spearman and Kendall rank correlations were used to assess the relationship, since the variables have different scales of measurement and do not imply a normal distribution. The results showed a statistically significant negative relationship between the variables: Spearman's $\rho = -0.299$ ($p < 0.01$) and Kendall's $\tau = -0.206$ ($p < 0.01$), suggesting that countries with greater linguistic diversity have, on average, lower levels of internet use. From the point of view of interpretation, this may indicate that limited access to the Internet contributes to the preservation of linguistic diversity, since less digitalization reduces the dominance of global languages and slows down the processes of linguistic unification. At the same time, the relationship is weak–moderate, so it reflects a general trend rather than a deterministic causality.

It was also checked whether the number of living languages differs between countries of different income groups (IncomeGroup) using the non-parametric Kruskal–Wallis test. The classification of countries by income groups in 2025 is given in the World Bank [29]. The results showed statistically significant differences between the groups ($H = 17.090$; $df = 3$; $p = 0.001$), which allows us to reject the null hypothesis of the same distribution of the number of languages in all income groups.

Pairwise comparisons with the Bonferroni adjustment (Table 1) showed that the most significant differences were observed between high-income countries and lower-middle-income countries ($p = 0.006$) and low-income countries ($p = 0.012$). In these comparisons, high-income countries have a statistically significantly lower number of living languages. At the same time, the differences between adjacent groups (high vs upper-middle, upper-middle vs lower-middle, lower-middle vs low) do not reach the level of statistical significance after correction.

Table 1
Pairwise Comparisons of IncomeGroup

Sample 1 - Sample 2	Test Statistic	Standard Error	Standardized Test Statistic Value	Significance	Adjusted Significance ^a
High income-Upper middle income	-16.891	9.727	-1.736	0.082	0.825

High income-Lower middle income	-33.835	9.851	-3.435	0.001	0.006
High income-Low income	-42.864	13.230	-3.240	0.001	0.012
Upper middle income-Lower middle income	16.944	10.561	1.604	0.109	1.000
Upper middle income-Low income	25.974	13.766	1.887	0.059	0.592
Lower middle income-Low income	9.029	13.854	0.652	0.515	1.000

Note: Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is 0.05.

a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

The results obtained indicate the presence of a structural relationship between the level of economic development of a country and linguistic diversity: wealthier countries are characterized by fewer living languages, while lower-income countries demonstrate higher linguistic diversity (Figure 3).

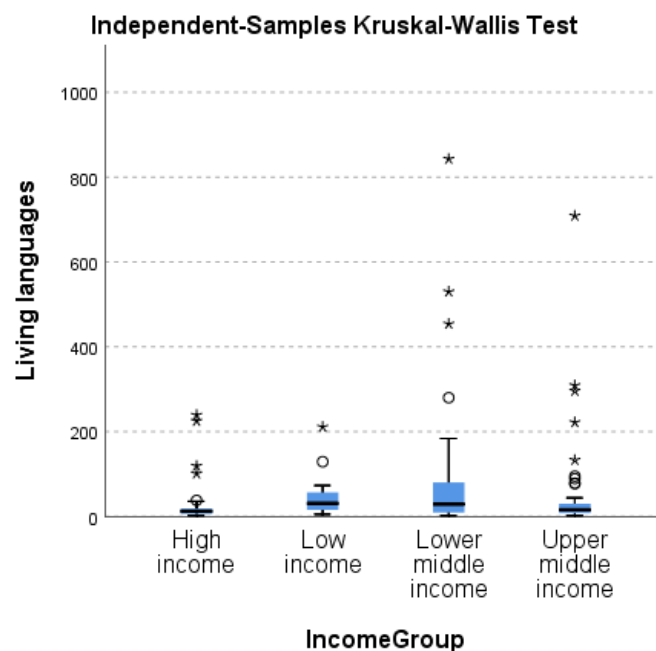


Fig. 3. Distribution of the number of living languages by groups of countries by income level (Kruskal–Wallis test)

Further, as part of the analysis of the impact of globalization and digital transformation on linguistic diversity, it is important to understand how the real presence of languages in the global Internet space is changing. To do this, we looked at historical annual trends in the use of languages in web content from 2015 to 2026 according to W3Techs [4]. Figure 4 shows the dynamics of the

share of major languages that exceeded 3% of the total number of websites at least in certain years (English is shown on an additional scale for better perception).

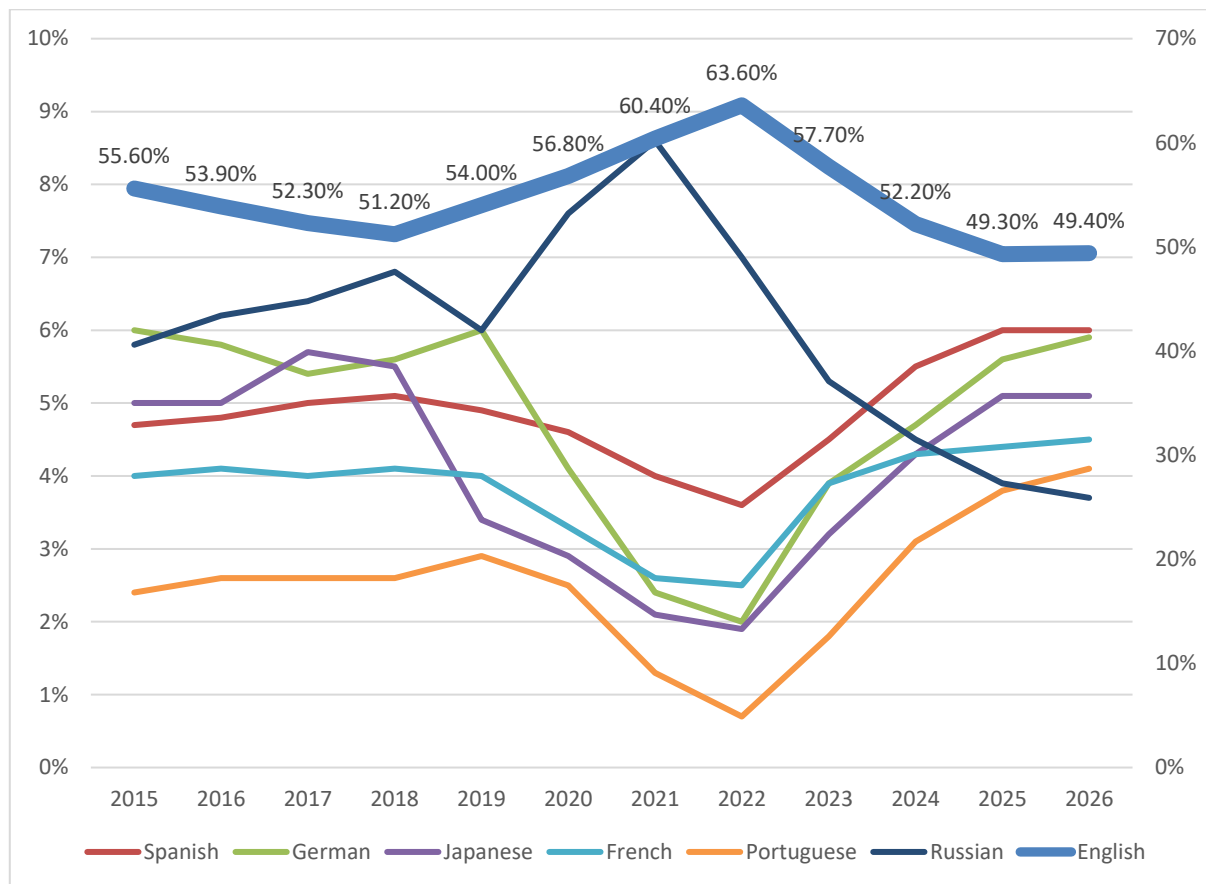


Fig. 4. Historical annual trends in the use of languages in website web content (2015–2026) for major languages with a share of more than 3%
Source: W3Techs [4]

In the last twelve years, English was dominating, but shows a clear descending slope, if in 2015 was 55.6, in 2022 got a maximum of 63.6, from where starts a decreasing process to end with 49.4 in 2026, losing during this period 6.2 percentage points. Absolute English monopoly in the digital sphere is not so absolute any more.

Iberian languages have been the one with the greatest improvement. Spanish increased from 4.7% in 2015 to 6.0% in 2026 (+1.3%), and Portuguese – from 2.4% to 4.1% (+1.7%). Portuguese developed in particular dynamically, which with a temporary pushdown in 2021–2022 showed the greatest positive dynamics among all analyzed languages. In this case this is a reflection of the active development of digital content in Latin America and Portugal.

European languages (German and French) are relatively stable. The German one was nearly flat 0.1% at about 5-6%, the French one grew a little (+0.5%), finishing at 4.5% in 2026. The value of Japanese dollar was also almost at the same level, increased just by 0.1%.

In the context of the general tendency towards diversification, the position of the Russian language is especially bright. The share of its content in the world web in the pre-invasion period [before the full scale invasion of Ukraine by Russia in 2022] was gradually increasing and reached maximum 8.6% in 2021. After 2022 this tendency was radically changed to decrease – in 2023 its

value was 5.3% and in 2026 – 3.7%. That means, that during only 4 years the Russian language lost more than 4.9 percentage points of its share being the most dynamically decreasing among the big languages.

This trend is explained by a set of factors related to the war: a massive blocking of Russian and pro-Russian sites in a number of countries, increased sanctions, a drop in the international visibility of Russian-language content, the ongoing processes of de-racification and transition to the state languages of the republics of the former USSR, especially Ukraine, the Baltic states and the Central Asian countries. The falling of Russian's popularity in the digital realm clearly shows how geopolitical developments may have a great influence on language presence on the global net, accelerating the redistribution of the digital space towards other languages.

In general, all the above-mentioned information demonstrates the following crucial process: English remains, obviously, the most popular language on the Web, but it is losing its absolute relative superiority and the number of other languages used on the Internet tends to grow accordingly. Thus, it generates a series of new challenges and perspectives for the national language preservation.

5. Discussion

The results obtained substantiate one of the key hypotheses of modern sociolinguistics – in the 21st century linguistic diversity is more and more defined not only by historical or demographic prerequisites but by access to digital infrastructure, economic inequality and even by the design of the world information system. If colonization, urbanization or assimilation state politics was a reason of possible death for small languages before, now a new reason appears – digital language choice when the role of algorithms, platforms and large language models is no less significant.

The scientific novelty of this study lies not only in the confirmation of known trends, but also in the comprehensive combination of four disparate global databases (Ethnologue, World Bank, W3Techs, UNESCO) with the use of non-parametric statistical methods to simultaneously analyze the spatial distribution of linguistic diversity, the level of digitalization and the economic status of countries. In contrast to previous works, which mostly focused on individual regions or were limited to descriptive statistics, this is the first time that a statistically significant negative relationship between the number of living languages in a country and the level of Internet penetration has been quantitatively demonstrated at the global level, and it has also been shown that high-income countries have systematically lower linguistic diversity. The methodological contribution is the integration of spatial-cartographic, comparative-temporal (web content for 2015–2026) and multidimensional statistical analysis in a single analytical framework, which made it possible to reveal not only a static picture, but also the dynamics of the displacement of some languages by others in the digital space under the influence of geopolitical events. The theoretical result is the conceptualization of the “digital language asymmetry” as a multidimensional phenomenon that is not reduced exclusively to economic or demographic explanations.

Special mention should be made of the finding of a statistically significant negative link between the extent of language use in the country and the use of the Internet. While the correlation in this case is relatively modest, this begins to take on a conceptually relevant dimension: the most linguistically-disaggregated countries tend to be the least connected to the digital economy. It is worth emphasizing that the identified correlations should not be interpreted as a direct causal relationship. The cross-sectional design of the study allows us to record statistical interdependencies, but does not allow us to draw conclusions about whether digitalization causes a decrease in linguistic diversity, or vice versa. The relationship between the number of languages and the level of Internet

use can be simultaneously influenced by numerous factors that are left out of the model: the level of education, urbanization, ethnic fractionalization, the historical experience of colonialism, state language policy and the volume of GDP per capita. To establish causal mechanisms, it is necessary to use multivariate regression methods or quasi-experimental approaches that take into account potential confounders. Therefore, the results obtained should be considered as preliminary evidence in favor of the hypothesis of an asymmetric digital presence of languages, and not as evidence of the unequivocal impact of Internetization on language preservation. That is, the world's most linguistically-disaggregated regions – Africa, Oceania and much of Asia – are also the most vulnerable to exclusion in the digital sphere. According to Mullah and Zainon [16], these figures appear to support the proposition that most of the world's languages are still “digitally invisible” – that is, nowhere in existence in any large technological system.

At this point, caution is in order concerning the understanding of the above relation. A lower level of Internet penetration is not the “shield” for languages in itself, but just a temporary suspension of the assimilation processes. When the digital access grows and there is no special language policy, there will be a fast switching of users to the lingua francas (English, Spanish, French or whatever else). The challenge is not the Internet itself, but the characteristics of the integration of individual language communities in this space.

The results of the income groups analysis are equally insightful. The fact that high-income countries indeed have statistically fewer living languages also, partly, echoes the long run processes of modernization: standardization of education, centralization of states, urbanization and, more generally, the historical weakened of local linguistic identities. Conversely, in the lower-middle and low-income countries we find much more linguistically rich countries, but where such materials are not critically and institutionally supported. This is the paradox of modernity: language wealth is found where we have the least means to maintain it. Such remark is of utmost importance for the international language policy: it raises the question of a global, and not merely national responsibility.

The evolution of web linguistic profiles can also be seen as the sign of a structural evolution of the digital space. The decline in proportions of English, while keeping its preponderance, is not a sign of every language losing ground, but a slow process of Dutchification of the Internet. The We say in French, the increase of spanish, portugues and partly french, is a response to the demographic dimensions of digital markets and a sign of the regionalization of web content. It is a crucial message: the digital space is not necessarily a unilingual space. Under the proper balances, national languages may even gain ground in the global network.

Particular emphasis should be given to the great downfall of share in the use of Russian language after 2022. The collected data clearly demonstrate that linguistic dominance on the Internet does not only depend on the numerical strength of population but also on the geopolitical legitimacy, sanctions system, country reputation and transnational flow of contents. In other words, the English language could gain its digital hegemony rather quickly under the influence of political events. This nuanced approach will contribute enormously to the classical models of language competition which largely only considered the demographic and economic factors.

6. Conclusions

The empirical results, directly confirmed by statistical analysis, are as follows:

- there is a statistically significant negative correlation between the number of living languages in a country and the level of Internet use (Spearman's $\rho = -0.299$, $p < 0.01$);

- high-income countries have, on average, a statistically smaller number of living languages compared to lower-income countries;
- the share of English in web content decreased from 55.6% to 49.4% between 2015 and 2026, while Spanish and Portuguese increased; Russian lost more than 4.9 percentage points after 2022.

Based on these facts, as well as taking into account the wider scientific literature, the following policy implications can be formulated: supporting national languages in the digital age requires systemic investments in the creation of text corpora, dictionaries, speech recognition systems, keyboard layouts and interfaces; Artificial intelligence technologies should be developed with the involvement of language communities on ethical grounds; state language policy should cover not only the educational sphere, but also digital platforms (search engines, voice assistants, social networks). For Ukraine, the growth of Ukrainian-language data corpora, machine translation and voice models remains relevant, which will contribute to the growth of the presence of the Ukrainian language in the global digital space.

Further research should be directed towards:

- multidimensional modeling of the relationship between linguistic diversity and digitalization, taking into account confounders;
- micro-level case studies of individual language communities that have successfully integrated into the digital environment;
- assessing the effectiveness of specific AI tools for underserved languages;
- monitoring the dynamics of language representation in web content under the influence of geopolitical changes.

This combination of macro- and micro-analysis will allow us to move from identifying global trends to practical mechanisms for preserving linguistic heritage.

Author Contributions

N.S.-T. and T.N.; methodology, N.S.-T. and L.P.; software, A.C.; validation, N.S.-T., T.N., and L.K.; formal analysis, T.N. and L.P.; investigation, N.S.-T., L.K., and A.C.; resources, L.K.; data curation, A.C. and L.P.; writing—original draft preparation, N.S.-T. and T.N.; writing—review and editing, N.S.-T., T.N., L.K., L.P., and A.C.; visualization, A.C. and L.P.; supervision, N.S.-T.; project administration, N.S.-T. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

The data used in this study were obtained from publicly available sources, including Our World in Data (based on Ethnologue), the World Bank, W3Techs, and UNESCO. All datasets are openly accessible through the respective official websites.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was not funded by any grant.

References

- [1] Hutson, J., Ellsworth, P., & Ellsworth, M. (2024). Preserving Linguistic Diversity in the Digital Age: A Scalable Model for Cultural Heritage Continuity. *Journal of Contemporary Language Research*, 3(1), 10–19. <https://doi.org/10.58803/jclr.v3i1.96>
- [2] Ethnologue. (2026). *Population*. <https://www.ethnologue.com/#population>
- [3] UNESCO. (2021, June 11). *The International Year of Indigenous Languages*. <https://www.unesco.org/en/articles/international-year-indigenous-languages>
- [4] W3Techs. (2026, April). Historical trends in the usage statistics of content languages for websites. https://w3techs.com/technologies/history_overview/content_language/ms/y
- [5] UNESCO. (n.d.). *Indigenous Languages Decade (2022–2032)*. <https://www.unesco.org/en/decades/indigenous-languages?hub=67103>
- [6] Pavlovskiy, O., Blikhar, M., Akimova, L., Kotsur, V., Akimov, O., & Karpa, M. (2024). International migration in the context of financial and economic security: The role of public administration in the development of national economy, education, and human capital. *Edelweiss Applied Science and Technology*, 8(6), 1492–1503.
- [7] Husanova, Z. (2025). The impact of globalization on language diversity: Challenges and opportunities. *Global Science Review*, 7(1), 92–95. <https://global-science-review.com/ojs/index.php/gsr/article/view/2452>
- [8] Sheraliyeva, M. (2025). The impact of globalization on language change. *International conference on multidisciplinary studies and education*, 2(5), 245–247. <https://eoconf.com/index.php/icmse/article/view/578>
- [9] Kearney, A. (2016). *Exploring the impact of globalization on linguistic culture* (Honors thesis, Western Michigan University). https://scholarworks.wmich.edu/honors_theses/2725
- [10] Lalombo, A. S. (2024). The Influence of Globalisation on the Shift in Local Language and Cultural Identity. *Journal Corner of Education, Linguistics, and Literature*, 4(001), 731–738. <https://doi.org/10.54012/jcell.v4i001.492>
- [11] Zhang, F. (2024). Globalisation and minority languages: A critical analysis of the impact of globalisation on minority languages. *International Journal of Frontiers in Sociology*, 6(4), 21–26. <https://doi.org/10.25236/IJFS.2024.060404>
- [12] Khrapatyi, S., Tokarieva, K., Hlushchenko, O., Paramonova, O., & Lvova, I. (2024). Research on performance evaluation of higher vocational education informatization based on data envelopment analysis. *STEM Education*, 4(1), 51–70.
- [13] Mor, N. (2025). It's a global village (if you speak the right language): On language models, digital sidelining, and participation. *Wisconsin International Law Journal*, 42(3), 329–388. <https://doi.org/10.59015/wilj.SCTW4297>
- [14] Stanford News. (2025, May 19). *How AI is leaving non-English speakers behind*. <https://news.stanford.edu/stories/2025/05/digital-divide-ai-llms-exclusion-non-english-speakers-research>
- [15] Muawanah, U., Marini, A., & Sarifah, I. (2024). The interconnection between digital literacy, artificial intelligence, and the use of E-learning applications in enhancing the sustainability of Regional Languages: Evidence from Indonesia. *Social Sciences & Humanities Open*, 10, 101169. <https://doi.org/10.1016/j.ssaho.2024.101169>
- [16] Mullah, N. S. & Zainon, W. M. (2026). AI and Low-Resource Languages: Bridging the Linguistic Divide. In L. Yaqin, Y. Alas, & A. Habibi (Eds.), *Reshaping Language and Cognition in Education Through AI* (pp. 77–128). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3373-4397-6.ch003>
- [17] Wang, D. (2025). Mind the gap in AI integration: a comparative study of language teachers' responses in a national survey. *Language, Culture and Curriculum*, 1–18. <https://doi.org/10.1080/07908318.2025.2574633>
- [18] Bird, S. (2022). Local languages, third spaces, and other high-resource scenarios. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, 7817–7829. <https://aclanthology.org/2022.acl-long.539/>
- [19] Kumar, S., Anastasopoulos, A., Wintner, S., & Tsvetkov, Y. (2021). Machine translation into low-resource language varieties. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 110–121). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-short.16>
- [20] Nasrullah, R. (2025). Artificial intelligence in language revitalization: a systematic review of global challenges and opportunities. *Prosiding Konferensi Linguistik Tahunan Atma Jaya (KOLITA)*, 23(23). <https://doi.org/10.25170/kolita.v23i23.7172>
- [21] Chang, C.-C., Li, C.-F., Lee, C.-H., & Lee, H.-S. (2025). Enhancing low-resource minority language translation with LLMs and retrieval-augmented generation for cultural nuances. *arXiv*. <https://arxiv.org/abs/2505.10829>

- [22] University of Hawai'i. (2025, September 5). *Endangered languages AI tools developed by UH researchers*. <https://www.hawaii.edu/news/2025/09/05/endangered-languages-ai-tools/>
- [23] Bakhov, I., Terebushko, Y., Osaulchyk, O., Ryhina, O., & Vedenieiev, V. (2024). Cultural challenges in education: strategies for consideration of various intercultural aspects in the educational process. *Multidisciplinary Science Journal*, 6, 55.
- [24] Ishchenko, Y., Rusnak, A., Artemov, V., Syniavskiy, P., & Soroka, I. (2024). Psychological and pedagogical aspects of adaptation of students who received temporary shelter to the educational environment of another country. *Journal of Higher Education Theory and Practice*, 24(1).
- [25] Bobro, N., Ivanova, D., Pyvovarov, K., Shatskaya, Z., & Kucheriavyi, V. (2025). Investment approach of higher education institutions to the development of educational platforms. *Salud, Ciencia y Tecnología - Serie de Conferencias*, 4, 1392. <https://doi.org/10.56294/sctconf20251392>
- [26] Raievska, Y., Savchuk, O., Huias, I., Khokhlov, A., Hoian, I., Syniakova, V., ... & Lukashov, O. (2024). Mechanism of comparison in the structure of self-efficacy in junior athletes' sporting activities. *Journal of Physical Education and Sport*, 24(6), 1349-1359.
- [27] Our World in Data. (2025). *Living languages*. <https://ourworldindata.org/grapher/living-languages>
- [28] World Bank. (2025a). *Individuals using the Internet (% of population)*. World Development Indicators. <https://data.worldbank.org/indicator/IT.NET.USER.ZS>
- [29] World Bank. (2025b). *World Bank country and lending groups*. World Bank Data Help Desk. <https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups>