OPEN ACCESS | ORIGINAL ARTICLE
DOI: 10.47743/ lincu-2025-si-0435

WORLD LANGUAGES IN COMPETITION: THE DIGITAL FACTOR

Christian MAIR
University of Freiburg

Abstract

This paper examines the evolving dynamics of global multilingualism in the digital age, focusing on the hierarchical structure of the World Language System (WLS) and the intensifying competition among languages online. Two major trends shaping linguistic diversity in the 21st century are identified: the ongoing decline of offline linguistic diversity and the selective, economically driven nature of online multilingualism. Drawing on recent data, the analysis demonstrates that while fears of an English-language internet monoculture have not materialized, digitalization has further entrenched English as the global lingua franca, creating sharper hierarchies and new digital divides. The paper explores the implications for smaller national languages, such as Romanian, which risk marginalization despite robust offline vitality. The author argues that the future of linguistic diversity depends on active digital language planning and intelligent language activism, advocating for policies that support both the global spread of English and the technological empowerment of other languages.

Keywords: *World Language System (WLS); digital divide; English as Lingua Franca; global multilingualism; language hierarchies; language activism*

1. INTRODUCTION: A HIERARCHICAL WORLD LANGUAGE SYSTEM

Two decisive developments will shape global multilingualism in the 21st century.

Received: 1 September 2025 | Accepted: 01 October 2025 | Published: 15 December 2025
LINGUACULTURE VOL. 16, SPECIAL ISSUE, 2025 – *Language(s) and (re)translation in a changing world*, edited by Sorina Ciobanu
ISSN (print): 2067-9696 | ISSN (online): 2285-9403 | www.journal.linguaculture.ro

(1) The rapid decline in the world's linguistic diversity 'offline' will continue. Of the approximately 7,000 languages currently spoken (*Ethnologue*), half could disappear by the end of the century. The trend is irreversible, but its extent can be influenced by intelligent language activism.
(2) As for language use online, multilingualism will prevail, though in a form that will be more selective and more economically driven than multiligualism offline (Dor). Fears that the internet and the World Wide Web might develop into an English-language monoculture have been proved wrong (Danet and Herring), but competition and power differentials among languages are likely to be even sharper online than offline.

Late 20th century global mulilingualism has been best described by Abram de Swaan in his model of the World Language System (WLS), a four-tier hierarchy in which English occupies the top, as the undisputed current global lingua franca (de Swaan *Words of the World*; "Language Systems"; "The Unequal Exchange"). On the level below, there is a small number (less than 20) of world languages, i.e. languages with well above-average numbers of native speakers, typically used in more than one nation-state, whether as second or official languages in postcolonial contexts or as widely taught foreign languages and/or lingua francas. Below this 'oligopoly', there is a number of fully standardised and codified languages (totalling less than 200), which enjoy prestige, official status and institutional support in their national home-bases, but are not widely taught as foreign languages or used in transnational lingua-franca interaction. As can be seen from this distribution, the WLS presents itself as a four-layer pyramid with a very broad base, comprising the vast majority of the world's ca. 7,000 languages. Demographically, this period is inverted, with the languages on the top two tiers accounting for 80 per cent of the world's population, currently estimated at around 8 billion people, and thousands of languages spoken by small and frequently economically precarious communities forming the WLS's periphery.

To make concrete sense of what this hierarchy means in practice, it is instructive to look at the five most widely used languages of the world (Figure 1):

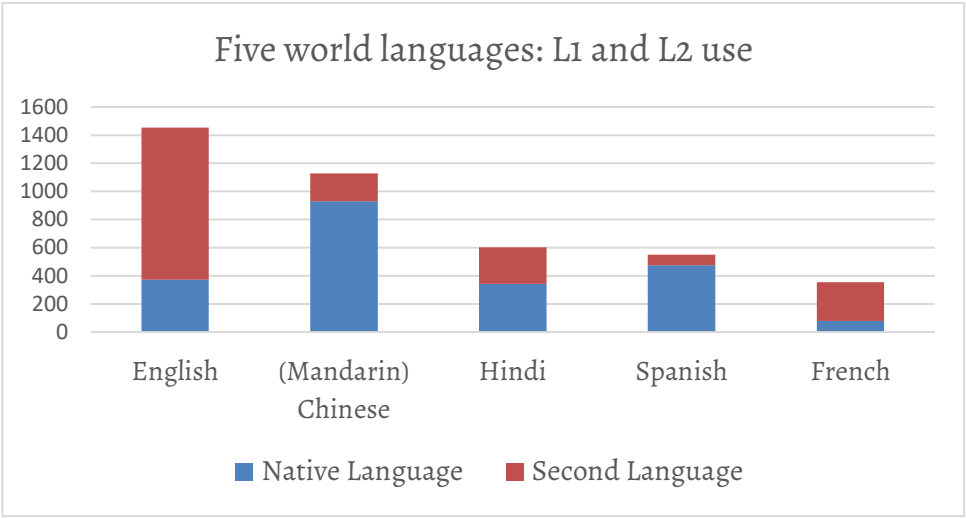


Figure 1. Five world languages, L1 and L2 use (*Ethnologue*)

The diagram clearly shows that the number of native speakers is not the decisive criterion for defining the global position of a language. (Mandarin) Chinese and Spanish have more native speakers than English, but do not rival it as international languages. Based on native-speaker numbers, French would not remotely have made it into the top five; it owes its position to the sizable proportion of second-language speakers.

Language hierarchies of this kind have been in place for most of recorded human history, ever since people started settling in cities and forming extended states, kingdoms and empires. What is unique to the present, though, is the pre-eminence of one single language, English, around which the whole WLS is integrated. This, of course, does not mean that everybody in the world understands or uses English, nor is such a global linguistic monoculture likely to emerge. The Ethnologue estimate of around 1.4 billion English speakers comprises around 400 million native speakers and a further billion of second-language speakers, which in Ethnologue's categorisation broadly means speakers in postcolonial official-English contexts such as India or Nigeria. Adding to this, fluent users of English as a Foreign Languae (EFL) or lingua franca (ELF) in continental Europe and other parts of the world, from Latin America to China, we arrive at a generous estimate of 2 billion, which is still just one fourth of the world population. Three quarters of the world's population do

not have a command of the global language, which often makes it a medium of indirect rather than direct communication. In spite of these restrictions, the lingua-franca use of English has become an indispensable asset, as airline pilots starting and landing in faraway airports, Romanian tourists asking for information in Finland, or Malaysian bankers doing business in Brazil will appreciate.

The move from an oligopoly of competing world languages (note the plural!) to the monopoly of a single global lingua franca (note the singular!) was so smooth and effortless that, although it took place within living memory of my older readers, we rarely take proper note of this world-historical first. As a reminder, it is enough to consider international academia. While for the greater part of the 20th century, two major European colonial languages, French and English, dominated international politics and diplomacy, German played an important role as a language of publication in the natural sciences until World War II. Today, English reigns supreme in all these domains.

2. EVEN SHARPER HIERARCHIES ONLINE

What is the relationship between linguistic diversity online and offline? The null hypothesis, namely that online competition among languages merely replicates offline constellations, can be confidently rejected. As hinted above, digitalisation has further strengthened the position of the language that needs such additional support the least, namely English. Not unexpectedly, most of the languages on the second tier of the WLS (conventionally referred to as *world languages* or *international languages*) have benefitted in one way or another. But especially where smaller national languages, such as Icelandic, Slovene or Estonian (all of them on tier three), are concerned, a new Digital Divide has opened. As fully standardised and codified languages, the three are ethnolinguistically vital in daily use and enjoy prestige in their home-bases, as carriers of national culture and identity, and ethnolinguistically vital in daily use. Yet all three are losing out to other languages (placed above them in the WLS) when their speakers go online. This is shown by Figure 2, an extract of a map of the world that shows Wikipedia users' language preferences by nation:

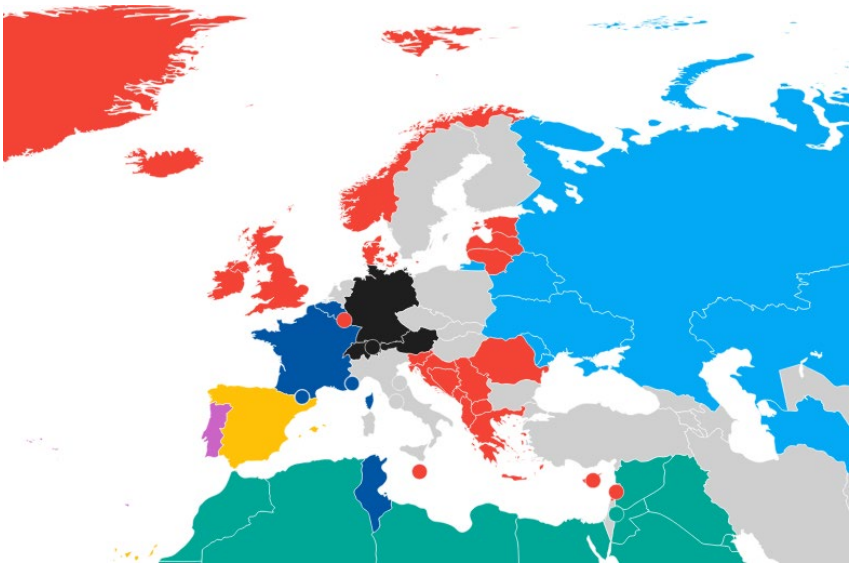


Figure 2. Most popular edition of Wikipedia by country[1]

The six languages that are colour-coded are Arabic (green), English (red), French (dark blue), German (black), Portuguese (violet), and Russian (light blue). In countries shaded grey, such as Italy and Romania, the dominant local languages, in this case Italian and Romanian, are the preferred choices when consulting Wikipedia. In comparison to the World Wide Web as a whole, Wikipedia is a domain that is friendly to linguistic and cultural diversity, because it is supported by authors that are not usually driven by commercial interests or click numbers.

Yet even in this friendly niche the pull of more powerful languages makes itself felt. The locally dominant offline language cede to English in Scandinavia (Denmark, Iceland, Norway, though not Finland and Sweden), in the Baltic region (Estonia, Latvia, Lithuania), in Luxemburg, and in large parts of southeastern Europe (Slovenia, Croatia, Bosnia-Herzegovina, Serbia, Montenegro, Northern Macedonia, Albania, Greece, Romania, Malta and countries of the eastern Mediterranean (Cyprus, Israel). The Russian Wikipedia

[1] Source: https://commons.wikimedia.org/wiki/File:Most_popular_edition_of_Wikipedia_by_country.svg.

is the most popular edition in Moldova and Ukraine. French makes modest inroads on Arabic in the North African Maghreb (Tunisia).

On a purely descriptive-diagnostic level, the map may reveal where in Europe Wikipedia editions in the local languages lag behind in terms of quantity and quality of information in comparison to competitors – and of course also where the local population reads English, French and Russian well enough to avail themselves of these alternatives. To exemplify the situation with the case of Romania: The current size of *Wikipedia în limba română* is 517,101 entries, with 2.139 active users, as compared to 7,047,293 entries and 112,428 active users for English. This is a yawning gap that probably cannot be closed by traditional language activism, for example campaigns to recruit new active contributors to the Romanian Wikipedia. An indirect route to success might be AI-assisted machine translation from English and other languages into Romanian, which would bring this information to the attention of monolingual Romanians. However, as humanities scholars, we are, of course, aware that technology is not a panacea in such situations. First of all, it is necessary to ensure that AI-based machine translation from, say, English to Romanian works well enough for practical purposes, to avoid content being "lost in translation." Secondly, and more importantly, we also need to be clear about the fact that foreign Wikipedias, while containing numerous useful entries on geographically or culturally neutral topics, may have gaps in areas that would matter to Romanian users. Whereas the battle between *Wikipedia în limba română* and English Wikipedia appears as one between David and Goliath, the race is much closer between Ukrainian (1,393,844 articles; 6,363 active users) and Russian (2,087,237; 7,342).

Note that, with around 19 and 36 million speakers (*Ethnologue*) neither Romanian nor Ukrainian would be considered small languages. Both are firmly entrenched among the roughly 200 to be found on tier three of the WLS. And yet they risk ending up on the wrong side of a new digital divide that is opening up between a very small number of technologically fully resourced languages and the vast majority of the rest of the world's tongues. Surveying the situation, Joshi *et al.* propose a six-tier hierarchy of digital inclusion/exclusion, which I list from top (maximal participation) to bottom (total exclusion). The illustrative examples are those given by the authors, but I would encourage Romanian readers to think about where they would place their multilingual nation's several

major and minor languages. The technical term *labelled data* refers to professionally curated digital language data such as linguistic corpora; *unlabelled data* denotes born-digital material mostly scraped from the web, such as text messages, Wikipedia entries and YouTube videos:

- *The Winners*: Languages in this group have massive online presence and substantial industrial and governmental support, making them the leaders in terms of resources. They benefit fully from ongoing technological advancements in NLP. Examples: English, Spanish, German, Japanese, French.
- *The Runners-up*:[2] These languages have active research communities, large amounts of unlabelled data, but fewer labelled datasets, which makes it very likely for them to eventually succeed in digital space. Examples: Russian, Hungarian, Vietnamese, Dutch, Korean.
- *The Rising Stars*: These languages are actively used online and have benefited from unsupervised pre-training, but are handicapped by a lack of labelled data. Examples: Indonesian, Ukrainian, Cebuano, Afrikaans, Hebrew.
- *The Hopefuls*: These languages have small amounts of labelled data and play some role in current NLP research. Examples: Zulu, Konkani, Lao, Maltese, Irish.
- *The Scraping-Bys*: These languages have some unlabelled, but no labelled data. A digital future for them is unlikely without organised language activism on their behalf. Examples: Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo.
- *The Left-Behinds*: These languages have not been introduced to digital space yet and are highly likely to remain invisible and ignored by NLP developers. Examples: Dahalo, Warlpiri, Popoloca, Wallisian, Bora.

The number of languages increases dramatically as one moves down the scale. The vast majority of the world's languages are left behind or scraping by. As for Romanian, it broadly meets the criteria for the 'Rising Star' class. It has a

[2] I take the liberty of replacing Joshi *et al.*'s term *underdogs* with *runners-up*, because of the former's potentially negative connotations.

solid corpus-linguistic working environment with labelled datasets[3] and a significant presence in social media and the internet.

3. HOW TO COMPETE WITH ENGLISH IN THE CYBERSPACE

Any analysis of digital language use begins with the question of which languages occur on the internet/World Wide Web, in what proportions, and in what hierarchical relationships to each other. Given that there is no consensus on what precisely constitutes the 'zone of interest' (consider internet-based social-media communication not indexed by search engines), none of these questions admits a straightforward answer.

Early approaches relied on extrapolation from offline speaker numbers and rates of access to the internet for the relevant populations. The good news suggested by such statistics is that, after an initial phase dominated by English, the internet has become much more multilingual over the decades, with large numbers of people from Brazil or Arab countries getting connected to the Web. There is certainly some truth in this, but the method has a major blind spot, because it ignores how many people use languages other than those they use offline when they go online. Countries in which language preferences offline closely correspond to those online are rare cases, an example being Japan. Most of the 126 million speakers of Japanese live in the country, have access to the internet almost from the beginning and were also fortunate enough to find content in their language in sufficient quantity and quality. The pressure to switch to sources in other languages, such as English, has been low. Conversely, few people outside Japan are able to benefit from Japanese web content.

The situation is completely different for most other languages, whether they rank above or below Japanese in the WLS. World languages tend to come out even stronger online than offline. In its full version, the Wikipedia map, whose European part is displayed in Figure 1, shows that English, and not

[3] Compare the list of resources available through SketchEngine (https://www.sketchengine.eu/corpora-and-languages/romanian-text-corpora/) and the Europarl Corpus project (https://www.clarin.eu/resource-families/parliamentary-corpora). This could certainly be expanded for broader and more balanced coverage of text types and genres, particularly in spoken language.

Hindi/Urdu, is the preferred language on the internet in India and Pakistan. French dominates in all former French colonies in West and Central Africa.

Whether productive or receptive, mulilingualism and multilingual repertoires are likely to be as common online as offline. For Romanian internet users, monolingual 'Romanian only' and bilingual 'Romanian and English' will be the most widespread constellations. Romanians who speak Hungarian may well rely on an expanded trilingual 'Romanian/Hungarian and English' repertoire. Similarly, 'Estonian only', 'Russian only', 'Estonian and English', 'Estonian, English and Russian' are all plausible scenarios in Estonia. It follows from the global dominance of English offline and online that all over the world such bi- and multilingual repertoires are highly likely to include English as one element. It should also be added that while web searches are usually formulated in either one language or the other, users' contributions to social media and even blogs may often be in the form of code-switching and other forms of language mixing.

The best current indicators of the 'cyber-power' of languages are content-based statistics, usually focussing on language-use in popular websites. Again, findings need to be taken with a grain of salt, because some websites may be multilingual and are thus difficult to assign to one language. Such minor problems notwithstanding, there is general agreement that more than half of the internet content measured in this way is in English.[4] At this point I would like to refer readers to an infographics site that condenses data from *Ethnologue* and recent research on online language use in a graphically appealing manner (*Visual Capitalist*[5]). I draw on this site for the following two charts (Figures 3 and 4).

[4] See, for example: https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/; https://w3techs.com/technologies/overview/content_language.

[5] https://www.visualcapitalist.com/the-most-used-languages-on-the-internet/.

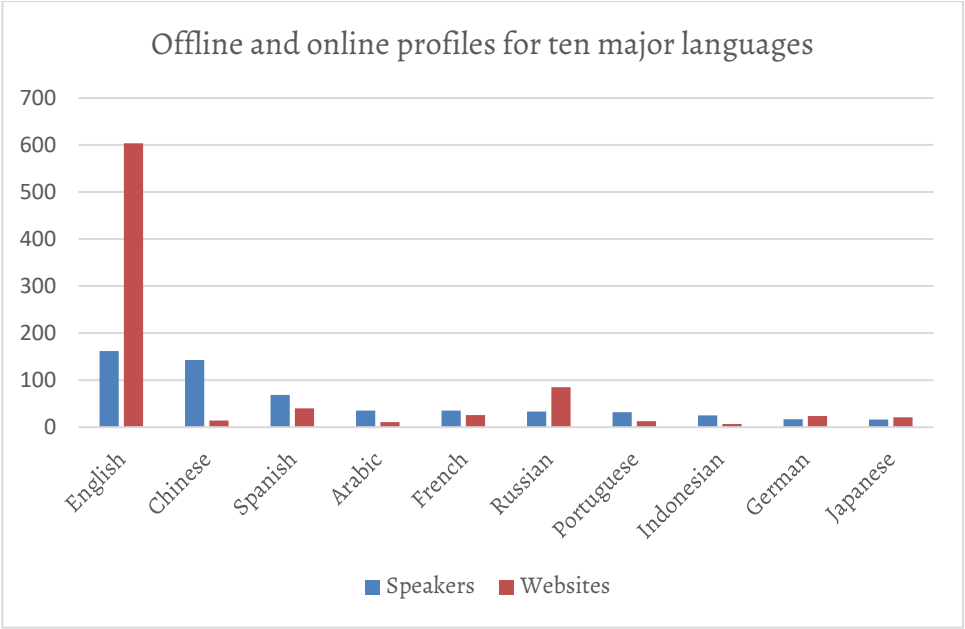


Figure 3. Online and offline presence of ten prominent internet languages[6]

From a wider pool of languages, I have selected ten that are prominent as internet languages. The measures are parts per thousand, which easily translates into per cent. Speaker figures are the same as used in Figure 1, based on *Ethnologue* and including L1 and L2 speakers. The estimates for web-content are based on the ten million most popular sites a the time of collection. For English, read the results as follows: L1 and L2 speakers account for 14 per cent of the global population, and English-language websites make up 60.4 per cent of the content measured – powerful evidence for the dominance of English online. In fact, the dominance of English is so overwhelming that it is necessary to remove it from the chart to get an idea of what the situation is like in the other languages:

[6] Source: https://www.visualcapitalist.com/the-most-used-languages-on-the-internet/.

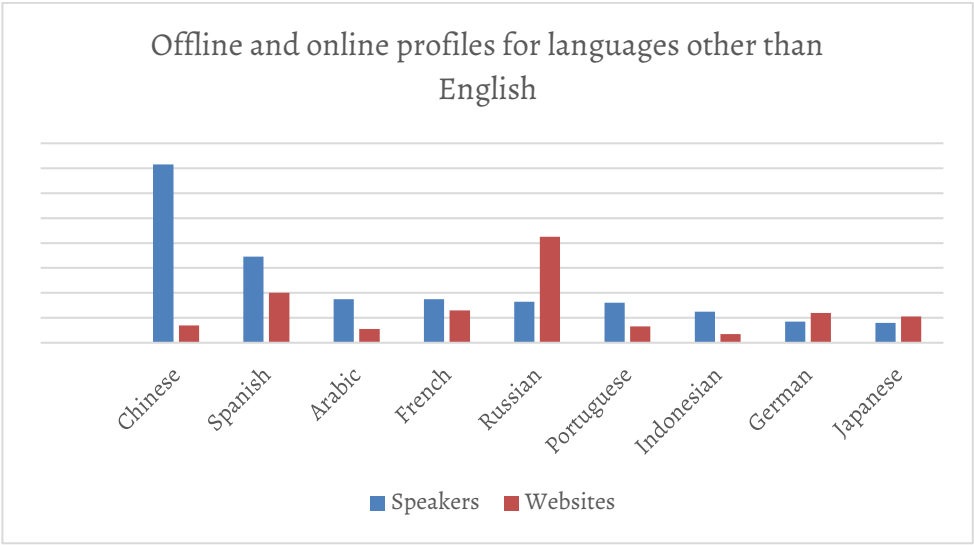


Figure 4. Online and offline presence of languages other than English[7]

Russian is the only case comparable to English in that its presence is significantly greater online than offline, although on a much smaller scale in terms of absolute numbers. Chinese, Arabic, Portuguese and Indonesian are severely under-represented online in relation to their speaker numbers offline. The picture is balanced for the remaining languages.

4. DIGITAL LANGUAGE COMPETITION AND THE FUTURE OF ENGLISH STUDIES

The digital revolution has made English even more dominant than it was before. But, as I will argue, it has also changed the language itself, and it has massively intensified contact between English and a large number of other languages in the world. As non-native speaking scholars in English Studies, whether in linguistics or in literary and cultural studies, whether based in Romania or in Germany, we should take up these new challenges.

As regards the English language itself, there is a standardisation paradox. The digital revolution hit English when it had already become a

[7] Source: https://www.visualcapitalist.com/the-most-used-languages-on-the-internet/.

pluricentric language. By the end of the First World War, American English had established itself alongside British English as the second globally relevant reference variety. From the 1960s onwards, further standards with more limited reach emerged, first in the area of native varieties (e.g. Australian and Canadian English), and later also in second language varieties (e.g. Indian and Nigerian English). The internet, and even more so, recent developments in AI-assisted language technologies will profoundly change the nature of this pluricentric constellation.

From the outset, U.S. usage dominated the internet, and AI will reinforce and entrench this domination massively. On the Web, if not on the ground, British Standard English—still the second most widely used and recognised norm offline—is being increasingly marginalised. It cannot be reliably generated by ChatGPT and other Large Language Models (LLMs) (Mair, "More Uniformity"). LLMs are able to generate largely correct lists of the essential features of British English in comparison to American English. They will even 'promise' to implement British conventions when prompted to do so, but will hardly ever consistently live up to this promise. Universal use of the new technologies will thus produce more homogenisation, though not always on the lines of the traditional standard of the textbooks and usage guides.

Contrary to this homogenising trend are developments towards de-standardisation, which are caused by the considerable presence of selected nonstandard varieties in the training data used for the LLMs. Language that is restricted regionally or stigmatised socially thus becomes accessible and 'generatable' througout the world. AI tools that prove incapable of consistently generating British English produce acceptable texts in Jamaican Creole English, Nigerian Pidgin, or the Spanish-English mixed repertoire informally known as Spanglish, which has become established in parts of the Latino community in the United States (Mair, "Migration, Media").

Scholars in English Studies have to recognise that their object of inquiry has outgrown the boundaries of their discipline. When used as a native language, English is a natural language like any other and should be described as such, and literary works written in English can be interpreted within national traditions and contexts. When used as lingua franca and a global language, however, many more aspects come into play, and we can and should learn from scholars in economics, political science and law:

> English is no longer solely a language for ethnic or national identification as languages are conventionally considered. For better or for worse, it is an economic skill, a marketable commodity and a form of cultural capital. English is the most marketable language in today's globalized economy. It is more sought after than any conventional commodity in the market, pervading the entire range of social and business relations in which it is used and discussed. […] It is the language of global communication, both driving the knowledge economy and gaining from it. (Salomone 8-9)

Last but not least, the continuing advance of English has also made it more necessary than ever to study the language together with other languages – as the world's most important contact language and as (admittedly usually dominant) part of multilingual constellations. English functions as a global language in a world that, fortunately, will remain multilingual. Research on English will be carried out in a world that has to navigate its path between two extreme scenarios.

The dystopian scenario foresees a mostly economically and technologically driven form of multilingualism in which English continues to spread rapidly as the global lingua franca. In this world, of the approximately 7,000 languages currently in existence, no more than 100 – essentially the languages of the many and the rich – will be able to hold their own in the digital space alongside English. In many cases, non-native speakers' English language skills will not keep pace with the demands made on them. Bad English produced by people combined with biassed and dubious or even hallucinatory machine-generated content is a toxic combination for communication across political, linguistic and cultural boundaries. And it will be the end authentic sociocultural diversity. This is what may await us if we continue with policies of *laissez faire*.

The utopian scenario requires active and intelligent digital language planning. All those who want or need to use the global lingua franca should be given the opportunity to learn it at the required level. Conversely, native speakers of English must recognise that monolingualism is also a social and cultural handicap for them. Instead of accepting that only about 100 languages will survive in the digital space, as is likely in the *laissez-faire* scenario, smart cultural policies and language activism will ensure that up to 1,000 languages are standardised and technologised to such an extent that they remain present in the digital text universe. Digital technologies could additionally be employed to

maintain endangered languages that will not be technologically resourced fully. For all those languages that disappear despite these efforts, digital technologies can at least help to document them comprehensively as intangible world heritage. In this bright scenario, the World Wide Web of the future will provide space for more domains such as Wikipedia, where cooperation, linguistic diversity and democratic discourse can flourish across national, ethnic, religious and cultural boundaries. Non-native speaking scholars of English should embrace their role as mediators who can help their communities to 'domesticate' the world's most powerful language in ways that maximises the benefits and minimises negative potential negative impact in the local community.

Works Cited

Blommaert, Jan. *The Sociolinguistics of Globalization*. Cambridge University Press, 2010, DOI: https://doi.org/10.1017/CBO9780511845307.

Crawford, Kate. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, 2021, DOI: https://doi.org/10.12987/9780300252392.

Crystal, David. *English as a Global Language*. 2nd ed., Cambridge University Press, 2003, https://doi.org/10.1017/CBO9780511486999.

Danet, Brenda, and Susan C. Herring, editors. *The Multilingual Internet: Language, Culture, and Communication Online*. Oxford University Press, 2007.

de Swaan, Abram. *Words of the World*. Polity, 2001.

---. "Language Systems." *The Handbook of Language and Globalization*, edited by Nikolas Coupland, Blackwell, 2010, pp. 56–76, DOI: https://doi.org/10.1002/9781444324068.ch2.

---. "The Unequal Exchange of Texts in the World Language System." *Bridging Linguistics and Economics*, edited by Cécile Vigouroux and Salikoko Mufwene, Cambridge University Press, 2020, pp. 203–23, DOI: https://doi.org/10.1017/9781108783101.009.

Dor, Daniel. "From Englishization to Imposed Multilingualism: Globalization, the Internet, and the Political Economy of the Linguistic Code." *Public Culture*, vol. 16, 2004, pp. 97–118, DOI: https://doi.org/10.1215/08992363-16-1-97.

Ethnologue: Languages of the World. 2025.

Joshi, Pratik, et al. "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, arXiv:2004.09095v1 [cs.CL], DOI: https://doi.org/10.18653/v1/2020.acl-main.560.

Mair, Christian. "Migration, Media and the Emergence of Pidgin and Creole-Based Informal Epicentres." *World Englishes*, vol. 41, 2022, pp. 414–28.

---. "More Uniformity and Diversity at the Same Time? Artificial Intelligence and the Standardisation Paradox in Twenty-First Century English." *World-wide Perspectives on English Usage: Into the Third Millennium*, edited by Peter Collins and Adam Smith, Cambridge University Press, forthcoming 2026.

Salomone, Rosemary. The Rise of English: Global Politics and the Power of Language. Oxford University Press, 2022.

Seargeant, Philip. The Future of Language: How Technology, Politics and Utopianism Are Transforming the Way We Communicate. Bloomsbury, 2024.

BIONOTE

Christian MAIR obtained his PhD and 'Habilitation' at the University of Innsbruck, Austria, and was appointed to a Chair in English Linguistics at the University of Freiburg in Germany in 1990. His research has focussed on the corpus-based description of Modern English grammar and on variability and change in Standard Englishes world-wide. Mair's current research centres on the role of global English in a multilingual world, on multilingual and nonstandard language practices in computer-mediated communication, and on the sociolinguistics of diaspora and migration, where he is advocating the use of corpora and digital tools to complement existing approaches. He has published several research monographs (e.g. Twentieth-Century English: History, Variation, Standardization, Cambridge University Press, 2006), widely used textbooks and more than 200 chapters and articles. In 2005 he was awarded the degree of Doctor honoris causa by Alexandru Ioan Cuza University.
E-mail: christian.mair@anglistik.uni-freiburg.de