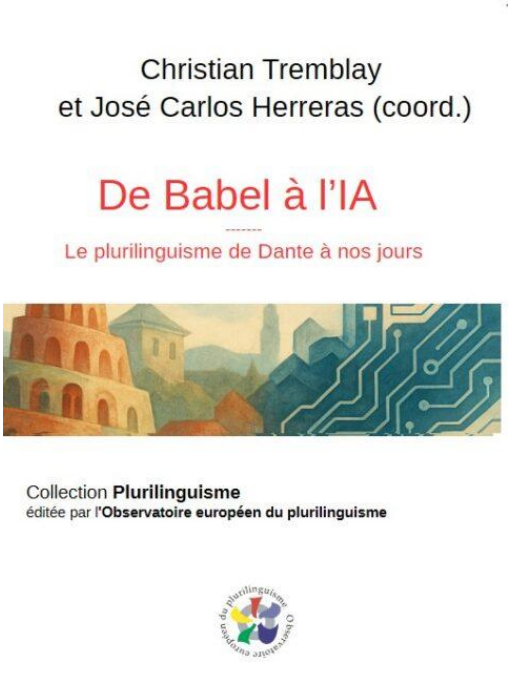
Linguistic diversity on the Internet

- Daniel Pimienta -



This is the translation made by the author from French of the corresponding chapter in "*From Babel to AI – Plurilingualism from Dante to nowadays*."

Chapitre du livre « *De Babel à l'intelligence artificielle - Le plurilinguisme de Dante à nos jours* », Christian Tremblay, José Carlos Herreras (coord.). Collection plurilinguisme, édité par l'Observatoire européen du plurilinguisme. ISBN : 9791042488130 – 1/2026 - deuxième édition. – pp 498-510
Description : https://www.observatoireplurilinguisme.eu/les-actions/collection-plurilinguisme/18704-de-babel-à-l-intelligence-artificielle-le-plurilinguisme-de-dante-à-nos-jours-coord-christian-tremblay-et-josé-carlos-herreras-deuxième-édition
Achat : https://livres.bookelis.com/documents/74996-De-Babel-a-l-IA.html

The public's perception of linguistic diversity on the Internet—and often, unfortunately, that of the academic and research communities as well—rests on a profound misunderstanding, itself fueled by chronic misinformation. This misinformation is perpetuated by sources whose structural biases stem primarily from a denial of what constitutes the very essence of the Internet: plurilingualism. This is all encapsulated in the pithy statement, " *English is the lingua franca of the Internet* ," repeated like a mantra, in direct contradiction to the facts.

The misunderstanding stems from the belief that what is true for the world of research and business – an undeniable dominance of the English language – should necessarily also be true in the digital world.

The essential difference between these three areas, which must be integrated to clear up this misunderstanding, lies in a tool that is still poorly understood or little known: demolinguistics, that is to say, the demography of languages in the world.

A global survey of researchers would almost certainly demonstrate that English is a second language for the vast majority of those who do not have it as their native tongue. Not that fluency in English is a determining factor for becoming a researcher, but more prosaically, because the intense pressure to publish scientific work in this language acts as an unavoidable incentive. Indeed, the quantity of publications, but even more so the publication venue and the number of citations generated by this work, are the essential indicators that determine researchers' careers. In the procedures for evaluating these indicators, the obligatory venue for obtaining points remains English-language scientific journals. The rise of open access indirectly improves the situation for plurilingualism in scientific publications, but this evolution remains slow and its effects are still barely perceptible.

In the business world, particularly in international trade, a similar line of reasoning prevails: there is a strong incentive to conduct business in English and, therefore, to master the language, even if this constraint is less pronounced than in the world of research.

It is legitimate to assert today that "English is the lingua franca of research" and, albeit to a lesser extent, that it is also the lingua franca of business. This smaller proportion is explained in particular by the fact that, in e-commerce—which represented more than 20% of global trade in 2020 and continues to grow rapidly—consumers overwhelmingly prefer platforms that address them in their native language and are reluctant to buy in a foreign language they do not fully understand. This factor also explains the already significant, and growing, plurilingualism of e-commerce websites.

A demolinguistic misunderstanding perpetuated by chronic bias

Why should it be any different for the Internet as a whole, and how is this a demolinguistic misunderstanding? The research and business communities represent very specific segments of the world's population, concentrating a tiny percentage of humanity—closer to 0.1% than 1% for researchers, and an order of magnitude larger for those involved in international trade, though still marginal. Analysis shows that the demolinguistic characteristics of these segments are far removed from global averages. The high percentage of English speakers within these communities is in no way comparable to world percentage, which is several times lower.

Today, the Internet is accessible to more than 60% of the world's population; in many countries, the rate of Internet access exceeds 90%. Considering the global Internet user population, the proportions are therefore approaching those of the literate population, estimated at around 85% of the world's population. It is thus necessary to apply global demographic and linguistic data to understand the linguistic reality of the Internet.

According to Ethnologue (2024), the native English-speaking population is just over 380 million people, representing slightly more than 5% of the world's population. Including first and second language speakers, the proportion remains below 15%. Even using the most generous estimates—up to two billion English speakers—less than 20% of humanity understands English. A *lingua franca* understood by less than 20% of the target population, " *francamente* ," is simply not viable.

This misunderstanding about the demolinguistic reality of the Internet, structurally incompatible with the dominance of English, has been perpetuated by biased sources. Its reach has been amplified by the "loudspeakers" of search engines, Wikipedia, and even some companies specializing in the commercialization of statistics. For example, Statista.com was still claiming in 2022, with a strong marketing veneer, that " *English is the universal language of the Internet* "[1]. Likely lacking any solid argument to counter a contradictory claim, this company carefully avoided responding to the letter[2] sent by the Observatory of Linguistic and Cultural Diversity on the Internet (OBDILCI[3]).

---

1 https://web.archive.org/web/20220221144826/https://www.statista.com/chart/26884/languages-on-the-internet/ The content of this page was changed in 2024 and for this reason the link provided is to the valuable website for preserving the memory of the Web, archive.org.

2 https://obdilci.org/wp-content/uploads/2024/04/Statista-W3Techs.pdf

3 https://obdilci.org

At the same time, OBDILCI published the results of its model of language indicators on the Internet and concluded that " *the transition of the Internet, from the dominance of European languages, English in the lead, towards Asian languages and Arabic, Chinese in the lead, is well advanced. The real winner of this transition is multilingualism, even if African languages are still slow to take their place*."

This situation of biased data, produced by marketing companies using unpublished methods, has persisted for years. The significant influence of this pseudo-data, fabricated without the required scientific rigor and massively disseminated by the media—and all too often by researchers—is also explained by the near absence of academics on this front[4]. OBDILCI, with the support of the International Organisation of the Francophonie (OIF) and the General Delegation for the French Language and the Languages of France (DGLFLF), has waged a long-term, albeit isolated, battle against this misinformation. But in the media landscape, science, especially when produced by civil society, carries little weight compared to the power of marketing (Pimienta, 2022-2).

From the originally English-speaking Internet to plurilingualism

It is true that at the origins of digital language, in 1992, the year the Web was born, the context was quite different and, in some ways, even more distinctly Anglophone than that of the research world. The Internet was born from the encounter between the world of computing, heavily influenced by English, and that of scientific research. The creation resulting from such a cross-fertilization could only be overwhelmingly Anglophone, at least in its early stages.

This offspring, however, paid a heavy price, with a birth defect that was eventually healed, but whose scars remain visible: an initial non-inclusive encoding. English, unlike most languages using the Latin alphabet, does not have diacritics. This peculiarity allowed for a complete encoding—lowercase, uppercase, punctuation marks, and special characters—contained within 128 characters, or 7 bits of information: the ASCII code (*American Standard Code for Information Interchange*). This reduced encoding penalized, for several years, the many

[4] The few university initiatives, such as the " The Language Observatory Project " in Japan, which brought together a network of universities and worked in collaboration with the now-defunct global network for linguistic diversity, MAAYA (https://web.archive.org/web/20170606000521/http://www.maaya.org/), unfortunately did not stand the test of time.

languages whose alphabetic inventory exceeded this threshold, causing legitimate frustration.

As an aside, a summer course organized by the Complutense University of Madrid in 2002 was titled " *Interñet en español* ". The presence of the tilde (~), impossible to represent correctly in the digital space at the time, phonetically rendered it as "Inter *niet* en español," an ironic testament to the anger provoked by this exclusion of a symbol that touched upon Spanish sovereignty. During this initial phase, English was more than a *lingua franca* : it was a veritable *lingua absoluta* of the Internet. However, this was only a transitional period, relatively brief in the grand scheme of things.

The world of research, international by nature, and that of computer science, creative and flexible by nature, have gradually addressed this original wound. After some a stopgap bandage[5], a universal encoding was finally implemented with Unicode[6], removing the fundamental obstacle to plurilingualism. European languages quickly found their place, followed by Asian languages, then Arabic, and finally plurilingualism, the natural flow of the "network of networks," appears in full force. It is also worth recalling that domain names were able to break free from the exclusive dominance of the Latin alphabet and become eligible to be defined in various codified alphabets, thanks to the creation of "internationalized domain names" in 2010, after a long development process that began in 1998. At the same time, the use of diacritics was permitted in domain names using the Latin alphabet. Thus, domains such as . みんな, for Japanese, or españa.com, for the revenge of the tilde, or even déjàvu.com have become possible…

The Internet is now the most multilingual space ever created by humankind, even if this plurilingualism still only concerns a minority of existing languages – less than 10%. Persistent misinformation continues to obscure this structural and undeniable multilingual reality.

The evolution of the percentage of English-language content clearly illustrates this transformation. Starting from nearly 100% in 1992, the proportion fell to around 50% in the late 2000s, before reaching an asymptotic level slightly above 20% today, now sharing first place with Chinese. Spanish occupies third position, at around 7% of content, while French is in fourth place, at around 3.5%, tied with Hindi, Arabic, Russian, and Portuguese.

---

[5] MIME, a protocol designed to extend the number of ASCII character combinations, see https://en.wikipedia.org/wiki/MIME.
[6] https://unicode.org

Medium and long-term projections favor Hindi, whose demographic growth could allow it to overtake Spanish. The outlook for French, on the other hand, largely depends on Africa's population, expected to double by 2050, as well as on a remarkably high presence of French language on the African Web, relative to the number of French speakers in the countries concerned[7]. In the long term, these factors could allow French to reach fifth place.

Selective plurilingualism

The structural difficulty many languages face in fully existing in the digital world is clearly apparent in the following data, which roughly reflects the current situation while also highlighting the distance still to be covered:

- Of the nearly 7,500 languages in existence worldwide, only about 10% have a minimal digital existence, that is, a codification allowing their representation in computer systems.
- Of the approximately 750 languages thus codified, only a third benefit from a sufficient level of technological support – for example, the possibility of being processed by machine translation programs.
- Of these 750 codified languages, about half have a sufficiently large and diverse volume of content on the Web.
- Of the fewer than 400 languages for which content actually exists, less than a quarter have reasonably "discoverable[8]" content, that is, content highlighted by dominant search tools.
- Finally, among the hundred or so languages that can actually be discovered, less than half have linguistic corpora of sufficient size to be integrated into large language models.
- Fewer than twenty languages currently benefit from fully functional major language models.

---

[7] The State of Web Multilingualism, Technical Report #7: *Propensity to Use French in the Web Ecosystem of Francophone African ccTLDs* . OBDILCI - 10/2025 – (in French) https://obdilci.org/wp-content/uploads/2025/10/WebMulti7.pdf

[8] "Discoverability", a concept introduced in Quebec a few years ago for cultural content, refers to the ability of major platforms to recommend such content (songs, films, etc.), without which it would be difficult for it to gain significant traction. The concept can be extended to all online content that is "discoverable" insofar as search engines rank it highly in search results based on relevant keywords. It's worth noting that discoverability increasingly relies on visibility within AI systems, which complicates the issue.

How can we explain such a bottleneck in the path of languages towards a full presence in the digital world?

There is a structural paradox linking the number of languages to the number of their speakers, which sheds light on the selective economic equation inherent in linguistic diversity in the digital world. Languages with more than one million speakers number 336 (before grouping into macro-languages), representing less than 5% of all existing languages. Nevertheless, these languages are spoken by more than 95% of the world's population.

At the other end of the curve, approximately 95% of the world's languages are spoken by less than 5% of the human population. Given the strong correlation between a language's digital existence and having more than a million speakers, this situation translates into a striking equation: nearly 95% of humans could, in theory, access the Internet in their own language, while approximately 95% of languages are excluded from the Internet.

If we seek to estimate the cost of the path to providing a language with the conditions for a full digital existence – from the codification of its writing system to its integration into artificial intelligence tools, via robust technological support, content covering a wide range of themes and real discoverability – and if this cost is related to the number of speakers, the equation appears as an almost insurmountable wall for the thousands of languages with fewer than 10,000 speakers, which in effect represents or more than half of the world's languages.

The transition to digital therefore requires a difficult but necessary alliance within language families, in order to pool efforts and reduce the cost per speaker. Given that approximately 10% of languages have no codified writing system and that nearly half lack a universally accepted orthography, the first challenge is to create, or adopt, common writing systems within these language families. Experience shows, however, that this objective is easier said than done: linguists, often jealous guardians of their language, are hesitant to relinquish certain particularities to adapt to the constraints of the digital age.

Towards genuine plurilingualism: translation, AI and paradigm shift

Despite the economic constraints that severely impact minority languages—and especially Indigenous languages—and keep many of them isolated from the digital world, it is undeniable that the Internet is now one of the most multilingual spaces ever created by humankind. Moreover, bridges between languages are beginning to be built at increasingly affordable costs.

In fact, translation has become the true *lingua franca* of the Internet. Thanks to artificial intelligence, it is entering a historic moment that can be described as an accelerated paradigm shift. This is manifested by a rapid proliferation of linguistic tools: first and foremost, those that aid multilingual inter-comprehension[9], but also, increasingly, those that assist with translation itself, or even automatic translation and interpretation – at least for the major languages.

The Internet has abolished geographical borders; the true borders of cyberspace are now linguistic. However, the digital traveler encounters a growing number of easily accessible bridges, allowing them to cross these borders and discover new horizons. Conceiving of a cyber-geography whose only borders are languages helps us understand why plurilingualism is inherent to the very nature of the Internet, and why AI is the historical tool capable of fully revealing its multilingual dimension.

Today, a person with a minimal mastery of digital tools can, with no investment other than their time – and saving a significant amount of it thanks to these tools – accomplish a set of tasks that were once cumbersome and costly:

> – maintain a multilingual website by integrating translation tools into editing software, significantly reducing the marginal cost of managing plurilingualism;
>
> – to allow visitors, by means of a few simple instructions, to access all the pages of a site "dynamically translated" into more than 250 languages at the time of consultation – an impressive ease, but one which should be described as an aid to inter-understanding rather than translation, given the still mediocre quality of the results for many languages;
>
> – watching foreign language videos on platforms like YouTube by requesting automatic subtitles in one's own language; this aid to intercomprehension now covers 250 languages, although the results can sometimes border on the grotesque when minority languages are involved;
>
> – organize remote conferences with automatic interpretation;

[9]In fact, the tools are not defined as such except as translation tools; however, in terms of use, an imperfect translation tool, but with a sufficient level of quality, remains very useful in the function of intercomprehension between the languages involved.

– to obtain translations of documents in foreign languages, admittedly imperfect, but which nevertheless constitute a powerful tool for helping linguistic inter-understanding;

– significantly reduce human translation time through translation assistance tools that preserve all the formats of the original document.

Significant progress is expected in the coming years in terms of the quality of results, and the prospect of meetings in which everyone can listen and speak in their native language is no longer science -fiction.

High-level translation and interpreting professionals are not necessarily threatened by these developments. However, they will need to integrate the impact of these tools and make their expertise heard in the ethical debates they generate. The Italian pun *Traduttore , traditore (translator, traitor)* is destined to take on a new meaning : while the "betrayal" of the human translator can be understood as a symptom of the fact that literary translation is also an act of co-creation—the translator bringing their own sensibility to the work—the betrayal of AI systems is of a different, far more worrying nature. It stems from biases embedded in the corpora that have fueled their deep learning, biases whose effects remain, for the time being, largely opaque.

It is therefore essential that this profession fully engages in the action networks emerging within civil society in order to ensure the ethics of ongoing AI developments, and that it actively participates in the transparency requirements concerning algorithms and data sources.

The nature of Internet plurilingualism: initial indicators

The nature of plurilingualism on the Internet remains largely unknown, mainly due to the lack of data on the subject until recently. This veil is only now beginning to lift thanks to exploratory studies conducted in 2025 by OBDILCI [10]. This work was made possible by access to a database describing, according to several parameters, including languages, more than 80% of the approximately 200 million existing websites.

The initial results reveal a dynamic and rapidly expanding reality, but one marked by considerable disparities depending on the criteria used. On average, the proportion of multilingual websites is between 11 and 12%. However, this figure varies dramatically from country to country: it exceeds 50% for sites hosted in

[10] Consult seven studies on multilingualism on the Web: https://www.obdilci.org/projects/other/mlreports/

Monaco, Moldova, Kuwait, Ukraine, Mauritania, or Luxembourg, while it falls below 4% for those located in China, South Africa, or South Korea.

Analysis by language reveals equally striking contrasts. More than half of the websites written in Basque, Ukrainian, Latvian, Catalan, or Estonian are multilingual, while for websites in Chinese, Korean, or Japanese, less than 4% offer access in another language. These results highlight the crucial role of sociolinguistic and geopolitical contexts in web language strategies.

Another revealing indicator concerns the average number of languages per multilingual site, which is clearly increasing: it has risen from around five languages in 2023 to seven languages in 2026. This trend confirms the acceleration of plurilingualism on the Web.

A key indicator, particularly revealing when compared to that of humanity as a whole, is the rate of plurilingualism. For human populations, this rate is typically calculated as the ratio of the number of L1+L2 speakers to the number of native L1 speakers. According to Ethnologue, it reached 1.43 in 2024. For the Web, a similar calculation can be proposed, by dividing the total number of language versions of websites by the number of existing websites. The resulting value is around 1.8, and it is also growing rapidly.

Is it any wonder that the Web appears more multilingual than humanity itself? It is undeniably easier for a website to "learn a new language" than for a human being, and the trend toward adding new language versions can only intensify as translation tools become integrated into website editing software. We should therefore expect rapid growth in these indicators in the coming years.

Some fear that the widespread adoption of AI-powered translation tools will discourage people from learning foreign languages. But what if the opposite were true? What if linguistic appetite were actually fueled by more easily crossing linguistic boundaries? These questions open up a vast field of research and represent key indicators to observe closely in the coming years, at the heart of the digital language revolution.

Cyber-geography of plurilingualism and the place of the Francophonie

Initial analyses reveal a highly contrasting geography of plurilingualism on the Internet. Websites in Arab and European countries are among the most multilingual, while the lowest levels are found in the web of major Asian countries and in English-speaking countries as a whole. This distribution underscores that web

plurilingualism does not automatically reflect a country's demographic or economic weight, but rather results from specific political, cultural, and economic choices.

Several specific cases strikingly illustrate this complexity. Luxembourg, for example, ranks very highly among countries for multilingual websites. Conversely, Luxembourgish-language websites are among the least multilingual in the language ranking. This discrepancy highlights the potential disconnect between national language policies and the actual language strategies employed for digital content.

North America generally performs poorly in terms of plurilingualism, although this trend cannot be attributed to Canada, whose results are above average. Among European languages, Portuguese also scores relatively low, despite Portugal's strong position. This apparent contradiction is explained by the dominance of the Brazilian web, which concentrates over 90% of Portuguese-language content and remains largely unmultilingual.

What about the Francophonie in light of these findings? Is its digital practice consistent with the promotion of plurilingualism advocated by the International Organisation of the Francophonie? French-language websites rank highly, with approximately 21% being multilingual, ahead of those in Spanish, Polish, or German, but behind those in Finnish, Dutch, or Italian. However, a country-by-country analysis reveals a more nuanced picture: France is among the lower-ranked European countries, though still above the world average, suggesting significant room for improvement. With regard to one of the most multilingual applications on the Internet, Wikipedia, with encyclopedic articles in 343 languages, and also, under the umbrella of the Wikimedia association, other highly multilingual open sharing applications (Wiktionary, WikiSource, Wikibooks , etc.), the French language stands out in third position in a ranking based on the average percentages of presence in all these applications.

The most encouraging results, however, concern the strong correlation observed between plurilingualism and the economic impact of websites, particularly in the e-commerce sector. Platforms with high economic added value are also those that invest most decisively in plurilingualism. The "linguistic compass" of the Web thus clearly points towards a continued acceleration of plurilingualism.

Internet governance and conclusion: the era of digital plurilingualism

The institutional Francophonie has been deeply involved in defending linguistic diversity on the Internet. This commitment has been expressed, in particular, through UN multilateral diplomacy processes linked to the World Summit on the Information Society, initiated in the early 2000s and continued, in various forms, within the

broader framework of Internet governance. Despite this sustained mobilization, linguistic diversity has never attained the strategic priority it deserves, overshadowed by the urgency of the digital divide and by a dominant vision—often championed by actors from the technology sector —excessively focused on connectivity, to the detriment of a more holistic approach to digital technology.

Priorities are, however, shifting. Linguistic and cultural diversity, as well as digital literacy—and more specifically its informational dimension—are beginning to receive the attention and priority they deserve. Several factors explain the beginning of this shift: the recognition of the damage caused by insufficient information literacy in the face of disinformation and its detrimental effects on democratic processes; the progressive saturation of connectivity levels in many countries, which allows attention to be shifted to other priorities; and, finally, the rapid emergence of artificial intelligence in the digital ecosystem, which is disrupting established certainties, reconfiguring central architectures—such as those of search engines and the underlying issue of discoverability—and raising major new ethical challenges.

As early as 2006, British linguist David Graddol warned young monolingual English speakers that their professional future would be jeopardized in a Europe where they would find themselves competing with peers who, in addition to their native language, were fluent in English and often another language[11]. Twenty years later, a similar warning is in order for those responsible for monolingual websites: their digital impact will be permanently compromised if they do not commit to developing multiple language versions of their content. The supposed absolute dominance of English on the web now belongs to a bygone era.

The digital age is all about plurilingualism. French, with its rich history, its geographical reach, its significant proportion of second-language speakers, and its unwavering commitment to linguistic diversity, possesses real advantages to capitalize on this dynamic and become an influential player. However, -these advantages must be fully leveraged in digital strategies, both institutional and economic. From Babel to artificial intelligence, the linguistic history of the digital age reminds us that linguistic plurality is not an obstacle to overcome, but a structural asset to be recognized, organized, and amplified.

[11] "English Next", https://wteachingenglish.org.uk/sites/teacheng/files/pub_english_next.pdf

References

OIF, 2023, *La présence de la langue française dans le cyberespace*, dans "La langue française dans le monde 2019–2022", 3/2022 – Gallimard/OIF– ISBN : 9782072976865
Chapter available online: https://observatoire.francophonie.org/wp–content/uploads/2022/08/OIF2_Extrait_p313–330.pdf

Pimienta D., 2022, *Indicators of the presence of languages on the Internet* , in Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under–Resourced Languages, pages 83–91, Marseille, France. European Language Resources Association. https://aclanthology.org/2022.sigul–1.11/

Pimienta D., 2022, *Une histoire très brève de l'observation des langues dans l'Internet* dans Culture et Recherche, N° 143, AUTOMNE–HIVER 2022, La recherche culturelle à l'international, page 128–131.
https://culture.gouv.fr/fr/content/download/319703/4810462

Pimienta D., 2024, *Is it true that more than half of web content is in English? If the multilingualism of the web were seriously taken into account, then no!* in *Forum for Linguistic Studies* , *6* (5), 201–212. https://doi.org/10.30564/fls.v6i5.7144