Reliably exploring the presence of languages on the Internet

The Internet is a precious resource for linguists as it offers an easily accessible and broad space where they can observe the evolution of languages over time. Daniel Pimienta, Head of the Observatory of Linguistic and Cultural Diversity in the Internet, has developed a method to measure the presence of languages on the Web, which was greatly enhanced in recent years. This method informed the development of a comprehensive database that could support linguistic research, language-related public policy, and e-commerce strategies.

The advent of the Internet and its widespread use opened up new avenues for the study of languages. For instance, the presence of different languages online offers valuable hints about their future use and development. Reliable estimates of the use of languages on the Internet could ultimately guide the development of public policies aimed at influencing their presence in the cyberspace. The Observatory of Linguistic and Cultural Diversity in the Internet, a research institute founded in 1996, specialises in the development of effective methods to measure the presence and evolution of languages online.

While language recognition algorithms – computational tools that can identify written languages – might seem like ideal tools for determining the prevalence of languages online, the Web has now become so large that applying these tools to all online content is highly impractical. Some studies thus used these algorithms to analyse subsets of online content, yet this experimental approach has been found to be ineffective, leading to biased and often unreliable results.

Up until recently, the most consulted source of statistics related to the use of languages online (W3Techs) relied on algorithms to analyse websites classified as the most visited. While these statistics offer some interesting insight, they might not accurately reflect the presence of languages online due to the lack of consideration of the often very multilingual nature of websites which trigger important biases.

In 2017, the Observatory of Linguistic and Cultural Diversity in the Internet devised a new approach that could help to better follow the progress and prevalence of languages online. Using this approach, Daniel Pimienta and his colleagues were able to identify meaningful indicators outlining the presence of 343 languages on the Internet.

**INDICATORS OF ONLINE LANGUAGE PRESENCE**

Back in 1998, research at the observatory had introduced one predominant approach to study the presence of seven languages online, which relied on data collected by search engines such as AltaVista and Google. In 2007, however, they started realising that search engine reports were becoming unreliable, so looked for alternative methods.

The new approach introduced in 2017 addresses the strong biases associated with previous efforts in this field of research. Initially, the researchers applied this approach to 138 languages, namely those with over 5 million native speakers, but they were recently able to extend it to 343 languages, those natively spoken by over 1 million people.

Using their proposed method, Pimienta and his colleagues compiled a set of indicators of the presence of these 343 languages online. These indicators were divided into three broad categories, namely intermediate, macro, and advanced indicators.

Intermediary indicators, all of which are expressed as percentages, include the number of internauts (i.e., speakers of a given language connected to the Internet), the usage of specific Internet services or applications, and the traffic reported on websites or applications. They also include an approximation of the digital language level of support and so-called indexes, which are country ratings based on Information Society parameters weighted into language ratings.

The second set of indicators, referred to as macro-indicators or model outputs, is comprised of connected speakers (i.e., the percentage of first- and second-language speakers globally who are connected to the Internet), the percentage of Web content in each language, content productivity (the ratio between Web content and internauts), and virtual presence (the ratio between Web content and speakers).

Finally, the more advanced indicators identified by the researchers include the cyber-geography of languages or, in other words, the division of languages on the Web into geographical groups (i.e., European, Asian, Arabic, American, and African) and the so-called Cyber-Globalisation indicator. This value, calculated using some of the other indicators, summarily sums up the ‘strategic advantages’ of a given language on the Internet.

**A NEW METHOD TO STUDY THE PRESENCE OF LANGUAGES**

The observatory’s new approach consists of indirectly approximating the relative amount of Web content per language. In doing so, it also considers crucial factors that are often ignored when describing a language’s Internet presence, but should be considered to prevent errors or biases.

Firstly, the team considers the likely existence of an ‘economie law’ relating to online communication, which links the offer (i.e., Web content available in a language) with the demand (i.e., number of speakers of that language who are connected to the Internet). Past findings suggest that the more speakers of a given language are connected to the Internet, the more webpages in that language tend to exist.

In addition, past research suggests that Internet users often prefer to communicate in their mother tongue when content is available in that language, yet they are happy to use their second language or languages in the absence of this content. In some cases, Internet users might also create content in their second language for economic reasons and could use translation services to do so.

A language’s presence online is also linked to the amount of Internet traffic in different places, the number of subscriptions to social networks, and the progress of different countries in terms of Internet-related services. The indicators of Internet presence created by the researchers collectively consider all these factors, thus painting a more detailed picture of how much and in what ways different languages exist online.

**A READILY ACCESSIBLE DATABASE**

Using their approach, Pimienta and colleagues set out to calculate indicators of online presence for languages natively spoken by over 1 million people globally. This allowed them to compile a comprehensive database summarising the presence of these languages online, which the observatory plans to update every year.

The values derived are very interesting, as they often do not match with values attained by other linguistic research efforts. For instance, between 2011 and 2022, the website W3Techs presented statistics suggesting that English has a steady presence online, representing over 50% of online content, yet the observatory’s analysis indicates that this is not the case.

Between 2011 and 2022, W3Techs presented statistics suggesting that English has a steady presence online, representing over 50% of online content, yet the observatory’s analysis indicates that this is not the case.
Behind the Research
Dr Daniel Pimienta

Bio
Dr Daniel Pimienta reads Applied Mathematics and Computer Sciences in Nice (France). First a Telecommunication System Architect at IBM, he later launched FUNREDES, a pioneer research-action NGO in ICT4D and managed the World Network for Linguistic Diversity (MAAYA). He is currently responsible for CREDDI, producing indicators on the presence of languages in the Internet.

Funding
Organisation Internationale de la Francophonie, Brazilian government, Instituto Internacional de Lingua portuguesa, Direction du français et des langues de France du ministère de la Culture de France

Collaborators
UNESCO Chair On Language Policies For Multilingualism

Japanese-speaking persons appear to be the most virtually present, generating proportionally more content than connected speakers of other languages. Over 1 million people in Africa is greater than the number of languages spoken in each of the other geographical regions. African language speakers are less connected to the Internet than people speaking languages from other regions, although the recent trend is for growth.

Finally, the Cyber-Globalisation indicator emphasised the strategic advantage of speaking English and French. In other words, speaking these two languages appeared to open new future opportunities for Internet users.

When it was able to revive again standing in a new and promising approach allowing to extend the scope of languages processed.

Both nationals and expatriates who are the most connected to the Internet, with a staggering 98.8% of connected speakers, followed by Danish (96.7%), Swiss German (94.1%), Catalan (94.1%), and Finnish (92.8%) speakers.

In addition, Japanese-speaking persons appear to be the most virtually present, generating proportionally to their connected number, more content than connected speakers of other languages.

Indicators related to the cyber-geography of languages also provided valuable insights. For instance, they showed that the amount of English progressively declined from 80% back in 1998 to 20% today. Notably, this does not mean that the amount of English content has diminished over time, but rather that online content in many other languages increased, in turn reducing the percentage of English content.

Collectively, the remaining 215 languages were found to account for 18–26% of Web content, followed by Spanish (7–9%), Arabic, Hindi, Russian, French, and Portuguese (3–4%), Japanese, German, and Malay (2–3%), Bengali (1.5–2.3%), and Turkish, Vietnamese, Italian, Korean and Persian (0.8–1.2%).

Overall, the team’s results suggest that content online has become increasingly multilingual, with the prevalence of English progressively diminishing over time. The observatory’s database is freely available to linguists exploring the presence of different languages online.