

ChatGPT-like models boom, but small languages remain in shadows

by *Naiara Bellio*

A lack of source material, investment, and commercial prioritization are all holding back the development of generative models and automated moderation for languages spoken in smaller countries and regions.

STORY 5 JUNE 2023 #CHATGPT #NLP #SPAIN

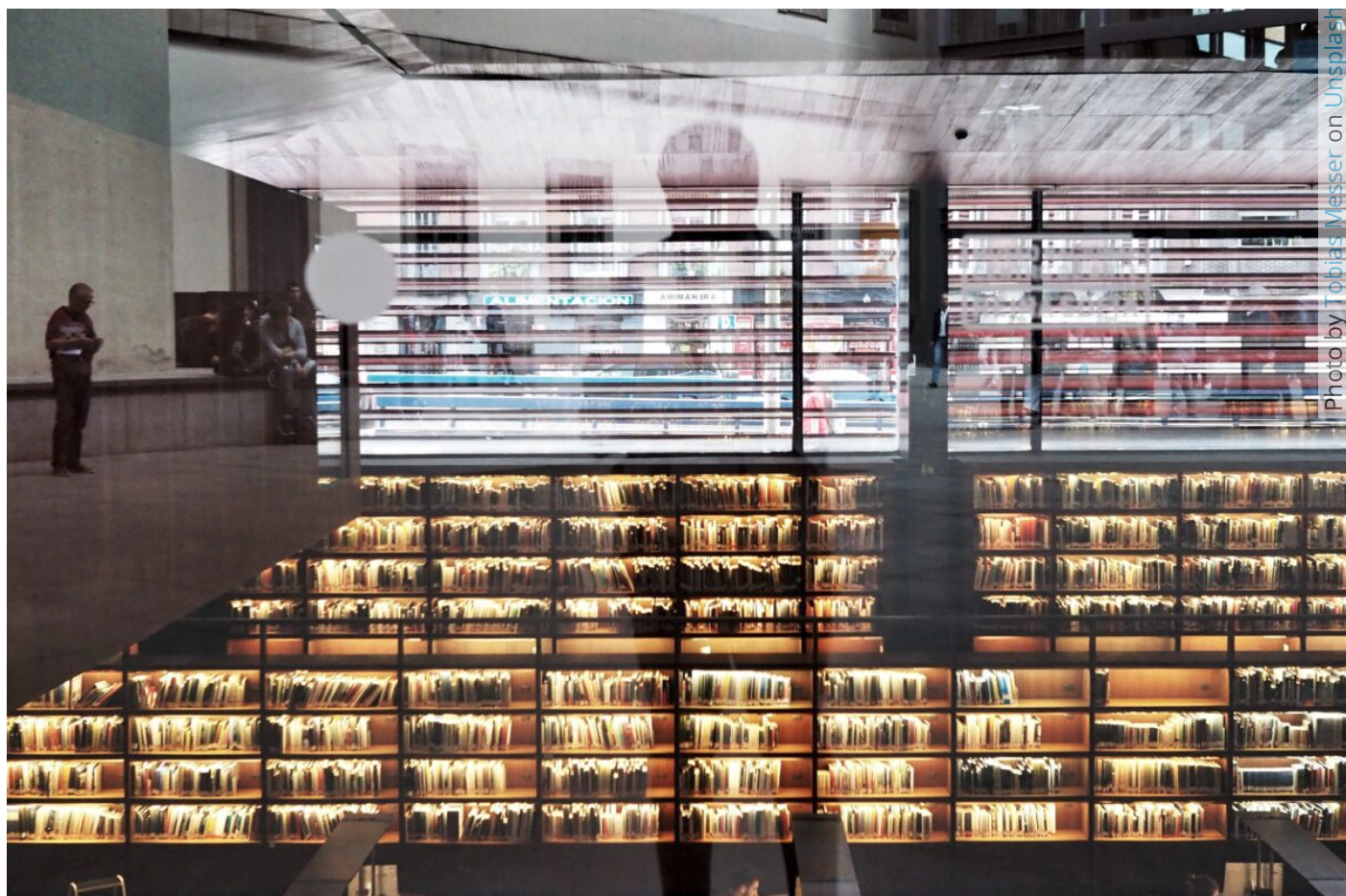


Photo by Tobias Messer on Unsplash

“Do you want to know how to say ‘I love you’ in Italian? Ask Alexa and start learning a new language today.”

This is a popular ad that Echo home assistant devices broadcast in Spain between music tracks and other programs. However, it is misleading. If you ask Alexa how to say ‘good morning’ in Galician, it will admit that it doesn’t speak this language yet. Or if you want to know how to ask for the time in Basque, it will sometimes say that “possibly” it hasn’t understood you well and then advertises its language services to you: “If you want to translate

something, you can ask me how to wish a good day in French," the automated voice says in that case.

It is not a secret that English is the ruling language in technology and that plenty of less-spoken languages are left behind when it comes to automation, but it seems funny that popular systems like Alexa play these kinds of ads in regions where they ignore the official language, like the Basque Country. Local companies have seen this before and are now under pressure to not fall behind with the new 'boom' concerning language and technology: ChatGPT. The problem is the money and the efforts are still gathered in one place (yes, it's Silicon Valley and yes, it's big tech).

Repressed languages trying to survive technology

Although not being able to ask Alexa for the road conditions in Basque is not a hindrance in all of Spain, it is a real handicap for a large part of the population who don't actually communicate in Spanish day to day. One in five people in the country [speak other languages](#), specifically Basque, Galician, and Catalan (including Valencian) as their mother tongue. This is true to the point that if you travel to fishing towns in the Basque region of Biscay, like Bermeo or Lekeitio, you'll encounter people who do not speak Spanish — but may be on Facebook, or even own an Echo device if, for example, they live in multi-generational households.

Some digital services have been sold and marketed with [the premise of democratizing](#) access to technologies and giving voice to the people, but hardly any fulfil the needs of those who don't interact in widely-spoken languages. Although this is a problem that affects communities of all kinds across the globe, Spain's situation is particularly sensitive due to the political restrictions that were imposed during Francisco Franco's dictatorship, which ended in 1975: people were punished for speaking in Catalan or Basque, the languages were suppressed from schools and even names were rewritten after people were born.

Still, Catalan, Basque and Galician are considered co-official languages in the country, and official in their respective autonomous regions. This leads to companies based in these regions to demand that their products and services, such as chatbots, be deployed with the same degree of development that would be accomplished in Spanish or English. In most

cases, this is not possible due to the lack of interest and investment in adapting existing technologies, which come mostly from the United States.

ChatGPT is making the problem more acute

“There has been a wake-up call since ChatGPT entered the scene,” says Carlos Rodríguez, member of the Language Technology team at the [Barcelona Supercomputing Centre \(BSC\)](#), who remembers how companies were also keen to be present in the metaverse when it was announced. “They know they need to use it, but they do not know how it fits in their business model. Many start-ups are now in dire straits because they based their model on things that now are easily done by anyone with GPT-4,” he explains.

Rodríguez is one of the people at the BSC overseeing the introduction of Catalan in language technologies. They are part of the small community of researchers working in Spain to introduce regional languages into the technological landscape, as most resources and efforts in the country are focused on improving how these technologies work in Spanish, which is itself underrepresented. While the BSC works in Catalonia, [IXA Group](#) and [Proyecto Nòs](#) are the academic groups working in the Basque Country and Galicia, respectively.

“A medical start-up came to us because they saw that, using ChatGPT, they could obtain interesting results, but they did not want to pay an American company to use the tool without the capacity to train it themselves,” adds Pablo Gamallo, a computational linguistics expert at Proyecto Nòs.

Even though [firms and experts long](#) for a Galician or a Catalan ChatGPT to thrive, it is such a complex aspiration that it becomes almost impossible. But why? The reasons are varied, but boil down to the lack of linguistic resources to train such a large model and the extreme amount of hardware and energy that is required to do so – two obstacles that large English-speaking companies step over easily.

“They have a brutal amount of linguistic resources, in particular Google, with its handling of the web, and tremendous computational capacity due to its revenues,” points out Gamallo. He tries to pin it down further: “For our supercomputer, we have 128 GPUs [Graphic Processing Units, the

processors used to train large language models], [compared with the 10,000 GPUs](#) that were used for the GPT model alone. If we were to pool all the GPUs in Europe, it would still not amount to all the GPUs that OpenAI has.”

Raiders of the Lost Data

Even if it was possible to match the money of market-leading companies, the researchers believe that it would be very complicated to train a model that would work as well as ChatGPT with the scarce resources available. The Spanish National Library is one of the main sources for large language models developed in the country, but does not always have enough data available.

“Many people think we use content from a newspaper, a book or a website, when there is actually a lot of other content on the Internet that needs to go on the models too for them to work properly, especially conversation dynamics that cannot be retrieved from copyright-protected content in Catalan,” explains Rodriguez. While [larger models built in Spanish like MarIA](#) use a corpus of up to 135 billion words, in Catalan researchers, developers and academics have to settle for around 1.7 billion words.

Gamallo from Proyecto Nòs affirms that their “gold mine” relies in adapting resources from Portuguese thanks to its similarities with Galician: “Half of the data [for training models] comes from Portuguese adapted to Galician. Even so, we have to work hard on it, it is not just a matter of taking Portuguese and adding it, but at least we do have that vein that Basque, for example, doesn't have.”

Basque does not share any roots with other European languages. Finding enough material to train the models is a problem. The large amounts of text you may find for English or Spanish simply do not exist for Basque, at least for some of the tasks required to develop what customers ask for.

Researcher Rodrigo Agerri, from Basque Center for Language Technology ([HiTZ Zentroa](#)), confirms that “such a large amount of texts do not even exist”, online or offline.

There are several applications that could improve social interaction with technology however, like systems to help with writing, correcting, scanning large databases, assisting older people or helping with medical reviews,

Agerri affirms: "In Basque, the problem is that the text production is rather low, so we are punching above our weight to develop state-of-the-art AI-based language technology with the available textual resources we have, as we did 30 years ago with automatic translation."

In fact, automatic translation is the closest researchers are getting to match big tech's services in their proper languages. It is also one of the largest advances in the natural language processing (NLP) field and probably what has been addressed first. [Leon Derczynski](#), a seasoned NLP and Machine Learning specialist who has led major research on the same issue in Denmark, stated: "The thing with minority or marginalized languages is that they don't transfer well. People represent this kind of thing in a linguistic and socially idiomatic way, it's going to be specific to that culture, so you need data in that language for things to work."

Who is moderating the results?

Derczynski points out translation technologies as one of the main areas where it is critical for NLP to prosper along with content moderation, which is somehow related to it. "When we started looking at Danish we saw that hate and toxicity look different depending on what country you're looking at, the way they construct negative arguments differs a lot. So even if you get data in every language, the phenomena that you're trying to detect might be different.

When AlgorithmWatch tried to dive deeper into the state of content moderation in Basque, Galician, and Catalan, it came as a surprise to find out that it was not being dealt with yet. The lack of investment and the prioritization of other applications mean that it is left to digital platforms to figure out how this content is moderated. Normally, the answer relies on the use of multilingual models, which are systems that apply the same analysis techniques from one language to another (for example, English to Danish). Facebook-owner Meta is one of the largest: Its model is trained to [detect harmful content in over 100 languages](#).

Asking large platforms like Meta, Google, or TikTok how they would moderate an ad in Basque, Galician, or Catalan or if they had specialized staff that spoke these languages did not bear fruit. TikTok did not reply while Google and Meta gave assurances that their teams and technologies had the

means to review content in more than 50 languages.

Meta specifically named Catalan, Basque, and Galician, but [a simple test conducted by AlgorithmWatch](#) in order to check whether their automated systems would catch non-compliant text in any of these languages showed that they do not always spot them. An ad that purposely discriminated against Black people and women was approved both by Meta and TikTok in Basque, Galician, and Catalan... and even in English, proving an even bigger problem to address.

“We have asked Google, Meta, and Amazon several times to collaborate. They don't seem to be interested in outsourcing their work,” Rodríguez from the BSC says. This is a general trend for these firms and not surprising in itself. But their assurances that they have things under control when they clearly do not is more troubling.


Edited on 7 June to better reflect Rodrigo Agerri's quotes.

Did you like this story?

Every two weeks, our newsletter [Automated Society](#) delves into the unreported ways automated systems affect society and the world around you. **Subscribe now to receive the next issue in your inbox!**

 **Naiara Bellio** (she/her)

Reporter

 Spanish, English

 bellio@algorithmwatch.org

 [@naiablm](https://twitter.com/naiablm)

 [9 Articles by Naiara Bellio](#)

Naiara Bellio covers the topics privacy, automated decision-making systems, and digital rights. Before she joined AlgorithmWatch, she coordinated the technology section of the [Maldita.es](#) foundation, addressing disinformation related to people's digital lives and leading international research on surveillance and data protection. She also worked for Agencia EFE in Madrid and Argentina and for [elDiario.es](#). She collaborated with organizations such as Fair Trials and AlgoRace in researching the use of algorithmic systems by administrations.