Digital Resilience: Harnessing the Power of the Collective for Language Preservation, A Success Story for Catalan

Ona de Gibert Bonet
University of Helsinki
ORCID: 0000-0002-7163-4807
ona.degibert@helsinki.fi

Ona de Gibert Bonet is a PhD student in Machine Translation at the University of Helsinki, where she contributes to the development of language technologies for low-resource languages. She has a degree in Modern Languages and Literature from the University of Barcelona (2016), and a Master's degree in Language Analysis and Processing from UPV-EHU (2018). Her main field of work revolves around machine translation, from its practical implementation in companies to the research of new methods in her current position. She has also worked on dissemination and scientific activism for the preservation of minority languages in the digital age, participating in the AINA project at the national level and the BigScience project at the international level.

Abstract

The widespread growth of Artificial Intelligence (AI) has made language technology more accessible than ever before, bringing language technology into our daily lives. Nevertheless, the rapid development of language technology intrinsically carries a bias towards Anglo-centric and Euro-centric perspectives, leading to limited representation and recognition for minoritized languages. It is a fact that the online presence of a language ensures its survival. This article explores the success story of Catalan, a minoritized language with an active online community, which has laid the foundations for language technology development. The Catalan-speaking community's success story demonstrates how communities can play a significant role in the survival and evolution of minoritized languages in the digital age.

KEYWORDS: *artificial intelligence; Catalan; language preservation; open-source*

Resum

Resiliència digital: aprofitar el poder del col·lectiu per preservar la llengua. Una història d'èxit per al català

El creixement generalitzat de la intel·ligència artificial (IA) ha fet que les tecnologies de la llengua siguin més accessibles que mai i ha portat aquesta tecnologia a la nostra vida diària. Tanmateix, el desenvolupament accelerat de la tecnologia lingüística comporta intrínsecament un biaix cap a una perspectiva anglocèntrica i eurocèntrica, que té com a una representació i un reconeixement limitats *de les llengües minoritzades. És ben sabut que la presència d'una llengua en línia en garanteix la supervivència. En aquest article explorem la història d'èxit del català, una llengua minoritzada amb una comunitat activa en línia, que ha establert les bases per al desenvolupament de les tecnologies de la llengua. La història d'èxit de la comunitat catalanoparlant demostra que les comunitats poden tenir un paper important en la supervivència i en l'evolució de les llengües minoritzades en l'era digital.*

PARAULES CLAU: *intel·ligència artificial; català; preservació de la llengua; codi obert*

TERMINÀLIA 27 (2023): 85-89 · DOI: 10.2436/20.2503.01.191
ISSN: 2013-6692 (impresa); 2013-6706 (electrònica) · *http://terminalia.iec.cat*

The rapid progress of Artificial Intelligence (AI) in the 21st century goes hand-in-hand with the unplanned development of language technology. Large language models (LLM), neural machine translation (NMT) systems, automatic speech recognition applications and commercial chatbots are opening up a new space for public dialogue that has never before taken place. Technology seems to have become more accessible to the masses and is capable of excelling at difficult tasks for all languages. But is this really the case? The dark side to the rapid development of language technology is host to an intrinsic bias towards Anglo-centric and Euro-centric perspectives, leading to limited representation and recognition for minoritized languages.

According to Bali et al. (2019), only 10-15 languages are directly impacted by the great changes in the AI paradigm—high-resourced languages such as English, Arabic and Spanish (see Figure 1). In the digital era we currently live, it is a fact that the online presence of a language ensures its survival. Currently, language distribution on the Internet follows a Zipfian distribution: only a few languages appear a lot (high-resource languages) and, in turn, many languages have a scant presence. Indeed, over 90% of the world's languages have almost no online representation (Choudhury, 2008). In 2012, META-NET published a study highlighting 21 European languages, including Catalan, in danger of extinction (Uszkoreit & Rehm, 2012). Woodbury (2019) considers that by 2100, 90% of the more than 7,000 languages spoken in the world may have disappeared.

In part, this is due to the fact that the core power of current AI technology lies with just a few private companies pursuing commercial interests that put little effort into developing products for a minoritized languages. While it is true that certain large corporations incorporate minoritized languages into their technologies, they do so for tokenistic purposes, practising "language diversity washing". This strategy fails to genuinely address the underlying issues of language preservation, representation and equity. Therefore, these efforts do not seem to offer a solution to the problem.

One of the reasons why corporations will not invest in developing products for lesser-resourced languages is the lack of data, which is the first step when building any AI system. Data for language technology refers to large amounts of text or voice samples needed to develop a model or a voice assistant. In the context of online language preservation, data becomes a common good and institutional organizations come into play. Governments can have a say in this matter by developing specific AI programs for language technologies to ensure that the necessary linguistic infrastructure in terms of data and models exists. This has been the case for Irish (Ní Chasaide et al., 2019) and Welsh (Prys et al., 2019), and more recently Basque (GAITU), Galician (Proxecto NÓS), and Catalan (Projecte AINA). Unfortunately, this approach depends on institutional power and is not feasible for all languages. However, the growing momentum in AI and language technology may encourage other governments to develop similar plans.

In this rather gloomy scenario is where we find a silver lining to the digitalization of language. It opens up new opportunities for language preservation and allows us to overcome linguistic barriers. For the first time, the fate of a language is bound to its community, independently of the number of speakers. Multi-

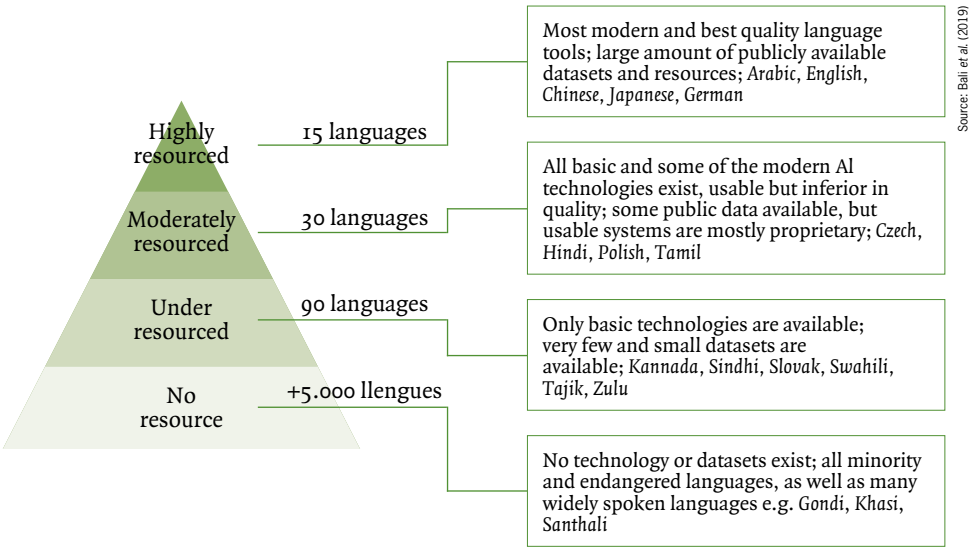

FIGURE 1. Classification of languages according to the availability of language technology, tools and resources

ple collaborative initiatives, self-organised in nature, have emerged in recent years, resulting in a radical transformation for their respective languages. With this new perspective, the community is involved not only as a consumer of technology but as a producer of new materials and resources (Prys et al., 2019).

With over 10,048,969 speakers (Plataforma per la Llengua, 2018), Catalan sits at 127th in the Ethnologue listing of languages according to number of speakers (Eberhard et al, 2023). Even so, there are over 110,000 .CAT domains (Fundació.cat, 2022) and Catalan is the 10th most active European language on Twitter and 19th worldwide (Plataforma per la Llengua, 2020). It is clear that Catalan's online footprint is much larger than its number of speakers and socio-political status would suggest (Irigoyen et al., 2020; Melero, 2021). In comparison with other minoritized languages, Catalan has a high potential to mobilize the community (Külebi, 2021).

Softcatalà is one of the driving forces in promoting Catalan online. This non-profit association was founded in 1997 with a view to promoting Catalan use in Information and Communication Technology (ICT). It has developed a wide range of open-source tools, including a grammar and spellcheck and translator tool. It also developed "Ona", the first voice assistant in Catalan, based on Catrotron—the first Catalan voice synthesizer built by Col·lectivaT, another non-profit association that develops language technologies for Catalan. Since all these tools are open-source projects, they are a clear example of how community begets community.

Softcatalà has also been involved in promoting Common Voice, a collaborative project started by Mozilla in 2017 with a view to collecting large amounts of open voice data to create automatic recognition systems. Currently, Projecte AINA also takes part in promoting the project; this has led to Catalan sitting in second place in terms of the collected number of hours in the Common Voice project, just 100 hours behind English. Other successful examples from the Common Voice project for lesser-resourced languages include Icelandic (Mollberg et al., 2020) and Rwanda (Muhire, 2020) .

The online, open-source and free encyclopaedia, Wikipedia, is a further international and equally essential project where Catalan has enjoyed mass participation. The Catalan community on Wikipedia is highly active and placed the Catalan version in 20th position for articles worldwide (725,000), out of the current 333 Wikipedia versions. In terms of quality, the Catalan version of Wikipedia also leads the quality ranking of the top 1,000 articles that every Wikipedia version should have (Hinojo, 2020) . Moreover, Wikipedia is important as it allows its content to be reused, making it a highly valuable resource for language technology development.

Although these are just a few examples, we can already see how they are all characterized not only by the involvement of volunteers and collaboration with civil society, but also by the fact that all generated resources and systems are open-source with licences enabling reuse.

Another open science initiative emerged from academia in 2022: BigScience. As a response to the rapid growth of privately-owned LLM, over 900 researchers worldwide joined forces to develop BLOOM—the first massive open generative model (Scao & al, 2020). Interestingly, the model includes Catalan, Basque and Niger-Congo languages, while some of the "big" languages are missing (see Figure 2). This was down to the fact that anyone interested could participate and include their own language. The active communities behind the aforementioned languages ensured their presence, contributing to their preservation by harnessing the power of the collective.

In conclusion, we have seen why the AI industry is largely focused on a handful of linguae francae, and how an active community can be the key factor for language preservation today. The generation of new resources, constant community involvement and the creation of necessary public infrastructure can alter the fate of a language that is in danger of digital extinction, and turn it into one enjoying online expansion.

It is clear that without community, there are no data (Muhire, 2020), and thus no AI. However, in the specific case of Catalan, we could pivot this affirmation and state with a community come data, and with data comes the development of language technology which, for Catalan, has only just got started. In fact, according to the latest European Language Equality report (Melero et al, 2022), at regional level, Catalan has the strongest support in language technology (see Figure 3), followed by Basque, Galician and Welsh. All of these languages enjoy ongoing institutional support.

The distribution between high- and low-resource languages is a continuously shifting spectrum. Just a few years ago, Catalan would have been a low-resource language; however, thanks to an active online community providing the foundations for language technology, it is not deemed so anymore. Therefore, at a personal level, if you wish to contribute to preserving your language, you "merely" need to write, speak and use social media in said language. In this way, you will become a digitally resilient language activist.

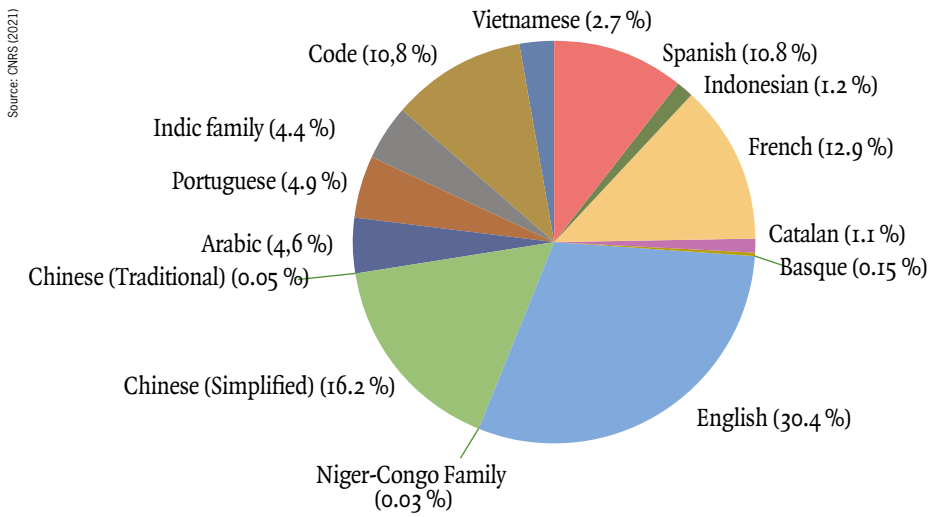

FIGURE 2. Languages used to train BLOOM

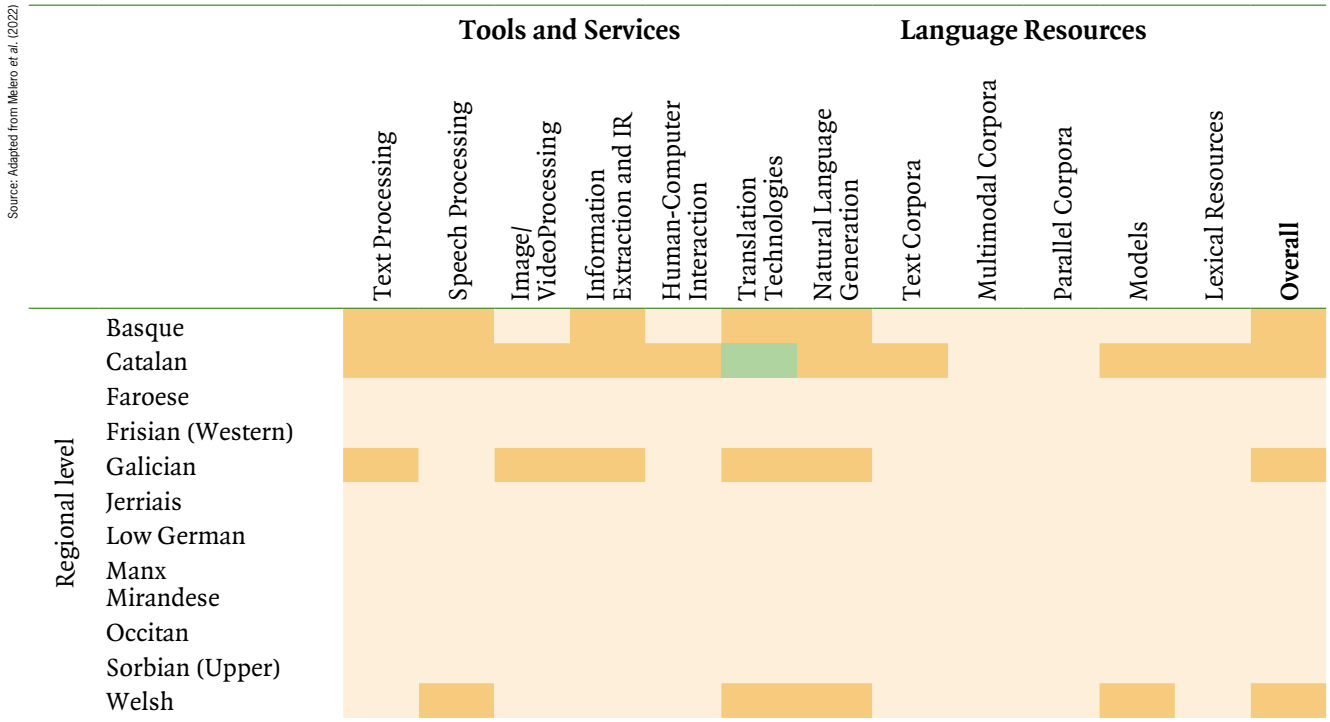

FIGURE 3. State of technology support in 2022 for selected European languages with regard to core Language Technology areas and data types, as well as overall level of support (light yellow: weak/no support; yellow: uneven support; light green: moderate support)

Bibliography

Bali, Kalika; Choudhury, Monojit; Sitaram, Sunaya; Seshadri, Vivek (2019). «ELLORA: Enabling low resource languages with technology». *Proceedings of the 1st International Conference on Language Technologies for All*, p. 160-163.

Choudhury, Monojit (2008). «Breaking the Zipfian Barrier of NLP». *Proceedings of the IJCNLP- 08 Workshop on NLP for Less Privileged Languages*.

CNRS (2021). *Largest trained open-science multilingual language model ever* [online]. Comunicat de premsa. <https://www.cnrs.fr/en/release-largest-trained-open-science-multilingual-language-model-ever> [Accessed: March 29, 2023].

Eberhard, David M.; Simons, Gary F.; Fennig, Charles D. (ed.) (2023). *Ethnologue: Languages of the world* [online]. 26a ed. Dallas, Texas: SIL International. <http://www.ethnologue.com> [Accessed: March 29, 2023].

Fundació.cat (2022). *Estat del català a Internet i les TIC* [online]. <https://observatori.fundacio.cat/#evolucio> [Accessed: March 29, 2023].

Hinojo, Àlex (2020). «Somien els viquipedistes en enciclopèdies elèctriques? Present i futur de la Viquipèdia i el rol de la comunitat catalanoparlant». *Revista de Llengua i Dret / Journal of Language and Law* [online], 73, 133-145. <https://doi.org/10.2436/ rld.i73.2020.3424> [Accessed: March 29, 2023].

Külebi, Baybars (2021). «El fet diferencial del català: la comunitat de programari lliure i obert». *Pensem* [online] <https://www.pensem.cat/noticia/226/fet-diferencial-catala-la-comunitat-programari-lliure-obert> [Accessed: March 29, 2023].

Melero, Maite (2021). «La tecnologia obre portes a la diversitat lingüística digital». *Pensem* [online]. <https://www.pensem.cat/noticia/233/maite-melero--tecnologia-obre-portes-diversitat-linguistica-digital> [Accessed: March 29, 2023].

Melero, Maite; Figueras, Blanca C.; Rodríguez, Mar; Villegas, Marta (2022). «Report on the Catalan language». *Language Technology Support of Europe's Languages in 2020/2021 - European Language Equality Project*.

Mollberg, David Erik; Jónsson, Ólafur. Helgi; Porsteinsdóttir, Sunneva; Steingrímsson, Steinpór; Magnúsdóttir, Eydis Huld; Guðnason, Jón (2020). «Samrómur: Crowd-sourcing data collection for Icelandic speech recognition». *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings* (maig), p. 3463-3467.

Muhire, Remy (2020). *How Rwanda is making voice tech more open* [online]. Mozilla Foundation. <https://foundation.mozilla.org/en/blog/how-rwanda-making-voice-tech-more-open/> [Accessed: March 29, 2023].

Ní Chasaide, Ailbhe; Ní Chiaráin, Neasa; Berthelsen, Harald; Wendler, Chrisyoph; Murphy, Andrew; Barnes, Emily; Gobl, Christer (2019). «Can we defuse the digital timebomb? Linguistics, speech technology and the Irish language community». *Proceedings of the 1st International Conference on Language Technologies for All* [online], p. 177-181. <https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.45.pdf> [Accessed: March 29, 2023].

Plataforma per la Llengua (2018). *InformeCAT 2018: 50 dades sobre la llengua catalana* [online]. <https://www.plataforma-llengua.cat/media/upload/pdf/informecat2018_1528713023.pdf> [Accessed: March 29, 2023].

— (2020). *InformeCAT 2020: 50 dades sobre la llengua catalana* [online]. <https://www.plataforma-llengua.cat/media/upload/pdf/informecat-2020_267_11_2406.pdf> [Accessed: March 29, 2023].

Prys, Delith; Jones, Dewi B.; Prys, Gruffud (2019). «Planning for language technology development and language revitalization in Wales». *Proceedings of the 1st International Conference on Language Technologies for All*, p. 367-370.

Riera Irigoyen, Marc; Ivers Ribes, Xavier; Orga Esteve, Pere; Montané Camacho, Joan, Mas Hernández, Jordi; Vicedo Cremades, Artur (2020). «Softcatalà: nous reptes per garantir la vitalitat del català a les tecnologies». *Revista de Llengua i Dret / Journal of Language and Law* [online], 73, p. 146-153. <https://doi.org/10.2436/rld.i73.2020.3396>. [Accessed: March 29, 2023].

Scao, Teven Le; Fan, Angela; Akiki, Christopher; Pavlick, Ellie; Ilić, Suzana; Hesslow, Daniel; [...];Wolf, Thomas. (2022). «Bloom: A 176b-parameter open-access multilingual language model» [online]. <https://arxiv.org/abs/2211.05100. [Accessed: March 29, 2023].

Uszkoreit, Hans; Rehm, Georg (2012). *META-NET White Paper Series: Press Release* [online]. <http://www.meta-net.eu/whitepapers/press-release. [Accessed: March 29, 2023].

Woodbury, Anthony C. (2019). *What is an endangered language?* Washington: Linguistic Society of America.