dossier

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

Ona de Gibert Bonet
Universitat d'Hèlsinki
ORCID: 0000-0002-7163-4807
ona.degibert@helsinki.fi

Ona de Gibert és estudiant de doctorat en Traducció Automàtica a la Universitat d'Hèlsinki, on contribueix al desenvolupament de tecnologies del llenguatge per a llengües amb pocs recursos. Graduada en Llengües i Literatures Modernes per la Universitat de Barcelona (2016), va obtenir un màster en Anàlisi i Processament del Llenguatge per la Universitat del País Basc - Euskal Herriko Unibertsitatea (UPV-EHU) (2018). El seu principal camp de treball gira entorn de la traducció automàtica, des de la seva implementació pràctica a l'empresa fins a la recerca de nous mètodes en la seva posició actual. També es dedica a la divulgació i l'activisme científics per a la preservació de les llengües minoritzades en l'era digital. Ha treballat en el projecte AINA, en l'àmbit estatal, i en el projecte BigScience, en l'àmbit internacional.

Resum

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 resultat 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*

Abstract

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

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*

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

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

Ona de Gibert Bonet

El ràpid avenç de la intel·ligència artificial (IA) en el segle XXI coincideix amb un desenvolupament imprevist de la tecnologia lingüística. El desenvolupament de grans models del llenguatge (LLM, *large language models*), sistemes neuronals de traducció automàtica (NMT, *neural machine translation*), aplicacions de reconeixement automàtic de la parla i d'assistents virtuals comercials estan obrint un nou espai de diàleg per al gran públic com no havia passat mai abans. Sembla que la tecnologia s'ha tornat més accessible a les masses i que és capaç de sortir-se'n de tasques difícils per a totes les llengües. Però, és realment així? El costat fosc del desenvolupament accelerat de la tecnologia lingüística comporta intrínsecament un biaix cap a una perspectiva anglocèntrica i eurocèntrica, cosa que dona com a resultat una representació i un reconeixement limitats de les llengües minoritàries.

Segons Bali *et al.* (2019), només entre deu i quinze llengües es veuen directament afavorides pels grans canvis en el paradigma de la IA; concretament, les llengües que tenen molts més recursos, com l'anglès, l'àrab i el castellà (vegeu la figura 1). A l'era digital en què vivim actualment, està demostrat que la presència d'una llengua a la Xarxa n'assegura la supervivència. En realitat, la distribució de llengües a Internet segueix una distribució de Zipf: hi ha molt poques llengües que hi apareixen molt (les llengües amb molts recursos) i, a l'inrevés, n'hi ha moltes que no hi són gaire presents. Més del 90 % de les llengües del món gairebé no tenen representació en línia (Choudhury, 2008). El 2012, META-NET va publicar un estudi en què afirmava que vint-i-una llengües europees, incloent-hi la catalana, estaven en perill d'extinció (Uszkoreit i Rehm, 2012). Woodbury (2019) considera que l'any 2100 el 90 % de les més de set mil llengües parlades al món poden haver desaparegut.

En part, això és degut al fet que el nucli de la tecnologia d'IA actual es troba en poder d'unes quantes empreses privades que segueixen interessos comercials i s'esforcen poc per desenvolupar un producte per a una llengua minoritzada. Si bé és cert que algunes grans corporacions poden incorporar llengües minoritzades a les seves tecnologies, moltes ho fan amb finalitats simbòliques; practiquen així el *language diversity washing*. Aquesta estratègia no aborda genuïnament els problemes subjacents de la preservació, la representació i l'equitat de les llengües i, per tant, els esforços en aquest sentit no sembla que siguin la solució al problema.

Una de les raons per les quals les corporacions no inverteixen en el desenvolupament de productes per a llengües amb menys recursos és la manca de dades, que és el primer pas per poder construir qualsevol sistema d'IA. En el cas de la tecnologia del llenguatge, amb *dades* ens referim a les grans quantitats de text o d'enregistraments de veu que fan falta per desenvolupar un model de NMT o un assistent de veu. En el context de la preservació d'una llengua en línia, les dades esdevenen un bé comú i és aquí on les organitzacions institucionals entren en joc. Els governs poden opinar sobre aquest assumpte i poden desenvolupar programes d'IA específics per a tecnologies lingüístiques que assegurin que hi ha la infraestructura lingüística necessària en termes de dades i models. Així ha estat en el cas de l'irlandès (Ní Chasaide *et al.*, 2019) i el gallès (Prys *et al.*, 2019), i més recentment en el del basc (GAITU), el gallec (Proxecto NÓS) i el català (Projecte AINA). Malauradament, aquest enfocament depèn del poder institucional i no és factible per a totes les

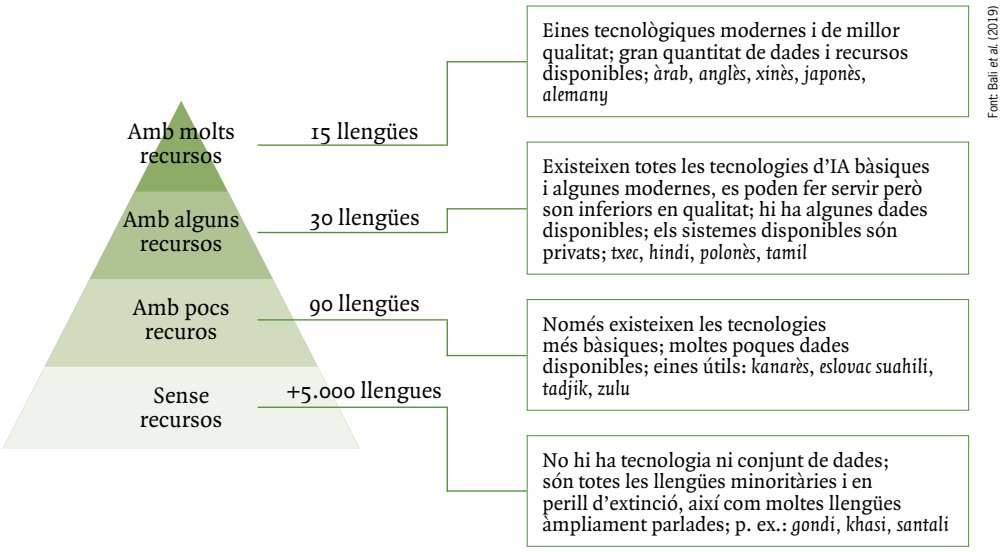

FIGURA 1. Classificació de les llengües segons la disponibilitat de tecnologia, d'eines i de recursos lingüístics

llengües. Tot i així, l'impuls creixent de la IA i la tecnologia del llenguatge pot encoratjar altres governs a desenvolupar plans similars.

En aquest panorama més aviat ombrívol és on trobem una escletxa d'esperança per a la digitalització de la llengua, que obriria noves oportunitats per a la conservació de les llengües i la superació de les barreres lingüístiques. Per primera vegada, el destí d'una llengua està lligat a la seva comunitat, independentment del nombre de parlants. En els darrers anys han sorgit múltiples iniciatives col·laboratives autogestionades, que han suposat una transformació radical dels seus llenguatges respectius. Amb aquesta nova perspectiva, la comunitat s'implica no només com a consumidora de tecnologia, sinó també com a productora de nous materials i recursos (Prys *et al.*, 2019).

Amb més de 10.048.969 parlants (Plataforma per la Llengua, 2018), el català ocupa el lloc 127 en la classificació de llengües d'Ethnologue segons el nombre de parlants (Eberhard *et al.*, 2023). Així mateix, hi ha més de 110.000 dominis «.cat» (Fundació.cat, 2022) i el català és la desena llengua europea més activa a Twitter i la dinovena a escala mundial (Plataforma per la Llengua, 2020). És clar, doncs, que la seva representació en el món en línia és molt més gran que la que tindria per nombre de parlants i estatus sociopolític (Irigoyen *et al.*, 2020; Melero, 2021); per tant, en comparació amb altres llengües minoritàries, el català té un alt potencial de mobilització de la comunitat (Külebi, 2021).

Una de les principals forces de promoció del català a la Xarxa és Softcatalà, una associació sense ànim de lucre fundada el 1997 amb l'objectiu de promoure l'ús del català en l'àmbit de les tecnologies de la informació (TI). Són els desenvolupadors d'una àmplia gamma d'eines de codi obert, d'entre els quals, els més coneguts són un corrector ortogràfic gramatical i un traductor. També van ser els desenvolupadors del primer assistent de veu en català, anomenat Ona, que es va basar en Catrotron, el primer sintetitzador de veu català construït per Col·lectivaT, una altra associació sense ànim de lucre que desenvolupa tecnologies lingüístiques per al català. Tots són projectes de codi obert i, per tant, representen un exemple clar de com la comunitat es beneficia de la comunitat.

Softcatalà també s'ha implicat en la promoció del projecte Common Voice, un projecte col·laboratiu iniciat per Mozilla el 2017 que té com a objectiu recopilar grans quantitats de dades de veu oberta per crear sistemes de reconeixement automàtic. Actualment, el projecte AINA també hi ha participat i ha portat el català a ocupar la posició número dos pel que fa a hores de gravació recollides en el projecte Common Voice, només cent hores per darrere de l'anglès. Altres exemples reeixits del projecte Common Voice per a llengües amb menys recursos han tingut lloc a Islàndia (Mollberg *et al.*, 2020) i Ruanda (Muhire, 2020).

Un altre projecte internacional i imprescindible en què el català té una gran participació és la Viquipèdia; l'enciclopèdia en línia, oberta i gratuïta. La comunitat catalana a la Viquipèdia és molt activa i ha situat la versió en català en la posició vint en nombre d'articles a escala mundial, de les 333 Viquipèdies que existeixen, amb més de 725.000 articles. En termes de qualitat, la Viquipèdia en català també lidera el rànquing de qualitat dels 1.000 articles que tota Viquipèdia hauria de tenir (Hinojo, 2020). La importància de la Viquipèdia rau en el fet que permet la reutilització dels seus continguts i això la converteix en un recurs molt valuós per al desenvolupament de la tecnologia del llenguatge.

Aquests són només alguns exemples, però ja podem veure que totes les iniciatives no només es caracteritzen pel voluntariat i la col·laboració de la societat civil, sinó també pel fet que tots els recursos i sistemes generats són de codi obert i amb llicències permissives que en permetin la reutilització.

Una altra iniciativa que segueix la ciència oberta va néixer el 2022 des de l'acadèmia anomenada BigScience. Com a resposta al ràpid creixement dels LLM de propietat privada, més de nou-cents investigadors de tot el món van unir forces per desenvolupar el primer model massiu generatiu i en obert, BLOOM (Scao *et al.*, 2020). Curiosament, les llengües d'aquest model inclouen el català, el basc i el nigerocongolès, mentre que hi falten algunes de les llengües «grans» (vegeu la figura 2). Això va ser perquè qualsevol persona interessada podia participar-hi i incloure-hi el seu idioma. Les comunitats actives darrere dels idiomes esmentats anteriorment es van assegurar que els seus hi fossin presents i en van contribuir així a la preservació, tot aprofitant el poder del col·lectiu.

En conclusió, hem vist per què la indústria de la intel·ligència artificial es concentra en gran manera en un grapat de llengües franques i com una comunitat activa pot ser el factor clau per a la preservació de les llengües en els temps actuals. La generació de nous recursos, la implicació constant de la comunitat, així com la creació de les infraestructures públiques necessàries, poden alterar el destí d'una llengua en perill d'extinció digital i convertir-la en una llengua en expansió digital.

És clar que «sense comunitat no hi ha dades» (Muhire, 2020) i, per tant, no hi ha IA. No obstant això, en el cas concret del català, podem capgirar la citació i afirmar que, si hi ha una comunitat, hi ha dades i, amb dades, el desenvolupament de la tecnologia lingüística del català acaba de començar. De fet, d'acord amb l'últim informe del European Language Equality Project (Melero *et al.*, 2022), el català és la llengua d'àmbit regional que té el suport més gran en tecnologia lingüística (vegeu la figura 3), seguida del basc, el gallec i el gal·lès; totes són llengües que reben suport institucional.

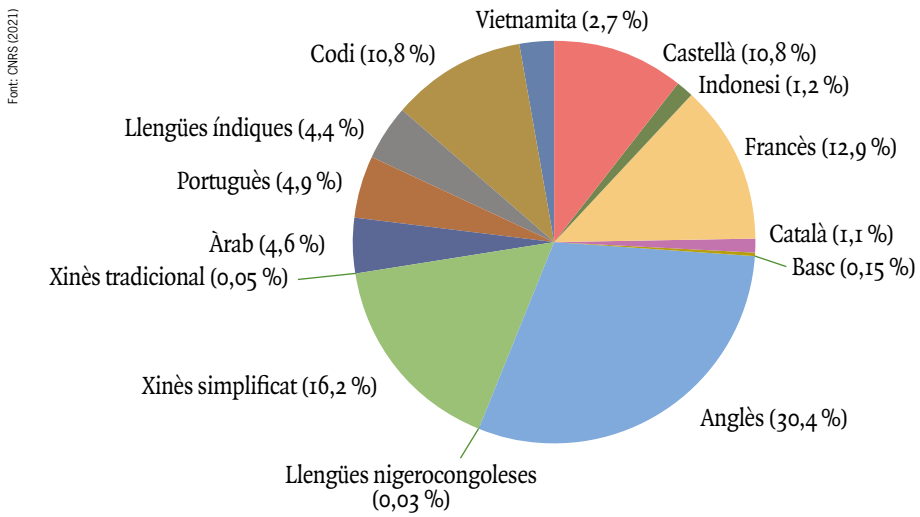

FIGURA 2. Llengües utilitzades per entrenar BLOOM

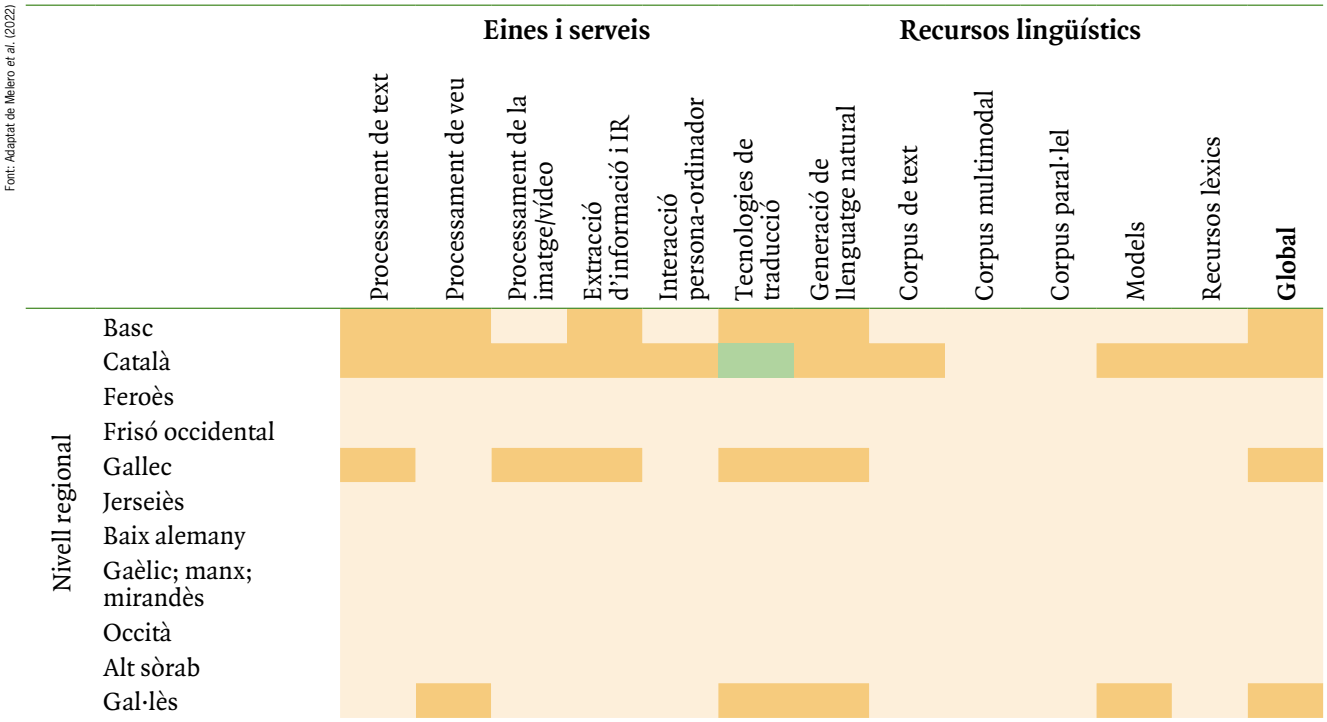

FIGURA 3. Estat de suport tecnològic, el 2022, de les llengües europees regionals, pel que fa a les eines i serveis i als recursos lingüístics (groc clar: suport feble o gens de suport; groc: suport fragmentari; verd clar: suport moderat)

La distribució entre llengües amb molts recursos i amb menys recursos és un espectre que canvia contínuament. Fa uns quants anys, el català era una llengua de recursos escassos, però gràcies a l'activa comunitat en línia que ha establert les bases de la tecnologia lingüística, ja no és així. Per tant, si vols contribuir a preservar la teva llengua, «només» necessites escriure, parlar i fer servir les xarxes socials en català. D'aquesta manera, seràs un activista de la llengua, digitalment resilient.

Bibliografia

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* [en línia]. Comunicat de premsa. <https://www.cnrs.fr/en/release-largest-trained-open-science-multilingual-language-model-ever> [Consulta: 29 de març de 2023].

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

Fundació.cat (2022). *Estat del català a Internet i les TIC* [en línia]. <https://observatori.fundacio.cat/#evolucio> [Consulta: 29 de març de 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* [en línia], 73, 133-145. <https://doi.org/10.2436/ rld.i73.2020.3424> [Consulta: 29 de març de 2023].

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

Melero, Maite (2021). «La tecnologia obre portes a la diversitat lingüística digital». *Pensem* [en línia]. <https://www.pensem.cat/noticia/233/maite-melero--tecnologia-obre-portes-diversitat-linguistica-digital> [Consulta: 29 de març de 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* [en línia]. Mozilla Foundation. <https://foundation.mozilla.org/en/blog/how-rwanda-making-voice-tech-more-open/> [Consulta: 29 de març de 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* [en línia], p. 177-181. <https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.45.pdf> [Consulta: 29 de març de 2023].

Plataforma per la Llengua (2018). *InformeCAT 2018: 50 dades sobre la llengua catalana* [en línia]. <https://www.plataforma-llengua.cat/media/upload/pdf/informecat2018_1528713023.pdf> [Consulta: 29 de març de 2023].

— (2020). *InformeCAT 2020: 50 dades sobre la llengua catalana* [en línia]. <https://www.plataforma-llengua.cat/media/upload/pdf/informecat-2020_267_11_2406.pdf> [Consulta: 29 de març de 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* [en línia], 73, p. 146-153. <https://doi.org/10.2436/rld.i73.2020.3396>. [Consulta: 29 de març de 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» [en línia]. <https://arxiv.org/abs/2211.05100. [Consulta: 29 de març de 2023].

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

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