Prompt...

LA IA EN EL FUTUR DE LES LLENGÜES EUROPEES NO HEGEMÒNIQUES

OPORTUNITATS, DESAFIAMENTS I ESTRATÈGIES DE PRESERVACIÓ

ALBERT CUESTA

FUNDACIÓ IRLA

Coppieters foundation

Accent obert

Prompt...

LA IA EN EL FUTUR DE LES LLENGÜES EUROPEES NO HEGEMÒNIQUES

OPORTUNITATS, DESAFIAMENTS I ESTRATÈGIES DE PRESERVACIÓ

FEBRER DE 2026

This publication is financed with the support of the European Parliament (EP). The EP is not responsible for any use made of the content of this publication. The editor of the publication is the sole person liable.

Dipòsit legal: B 2377-2026
ISBN: 978-84-09-81831-0

LA IA EN EL FUTUR DE LES LLENGÜES EUROPEES NO HEGEMÒNIQUES

OPORTUNITATS, DESAFIAMENTS I ESTRATÈGIES DE PRESERVACIÓ

ALBERT CUESTA

SUMARI

RESUM EXECUTIU

La irrupció de la Intel·ligència Artificial (IA) representa una de les transformacions tecnològiques més profundes de l'actualitat, amb implicacions de gran abast per a tots els sectors de la societat. Aquest estudi acadèmic detalla l'impacte d'aquesta tecnologia, amb un èmfasi particular en la variant generativa i els models de llenguatge grans (LLM), sobre el futur de les llengües europees no majoritàries. L'anàlisi explora tant els efectes positius, que ofereixen noves vies per a la preservació i revitalització lingüística, com els desafiaments inherents, que podrien exacerbar les desigualtats digitals existents.

Es defineix la IA com un camp dedicat a la creació de sistemes que emulen la intel·ligència humana, i la IA generativa com la capacitat d'aquests sistemes per produir contingut nou i realista. Els LLM, un subconjunt de la IA generativa, es presenten com el motor d'aquesta revolució lingüística, entrenats en vastes quantitats de dades per comprendre i generar llenguatge humà. No obstant això, la dependència d'aquests models de grans corpus de dades revela una vulnerabilitat crítica per a les llengües amb pocs recursos.

L'informe destaca les oportunitats que la IA ofereix per a la documentació, l'aprenentatge i l'accessibilitat de les llengües no majoritàries, així com el foment de la creativitat i el desenvolupament econòmic local. No obstant això, també s'aborden els riscos significatius, com l'ampliació de l'escletxa digital, la propagació de biaixos algorítmics, la manca de transparència dels models i el perill d'homogeneïtzació cultural.

A través de casos d'estudi com el Projecte AINA per al català, l'experiència d'Islàndia amb OpenAI i iniciatives reeixides en basc i kalaallisut, es demostra que la proactivitat, la inversió en dades, la col·laboració publicoprivada i l'enfocament en aplicacions pràctiques són fonamentals per a la supervivència digital. L'anàlisi dels *benchmarks* [índexs de referència][1] comparatius subratlla la necessitat de mètriques més justes i culturalment rellevants per avaluar el rendiment dels LLM en llengües amb pocs recursos.

Finalment, l'estudi examina la Llei d'IA de la UE, que, tot i no abordar explícitament les llengües, estableix un marc ètic i de seguretat que pot influir indirectament en la diversitat lingüística. Es conclou que el futur de les llengües europees no majoritàries en l'era de la IA depèn d'un enfocament col·laboratiu, estratègic i centrat en l'ésser humà, que transformi els desafiaments en oportunitats per a la inclusió i la preservació cultural.

1 Indicadors objectius de qualitat i rendiment.

1

INTRODUCCIÓ

1.1. Context i importància de les llengües europees no hegemòniques

Europa és un continent que es distingeix per la seva extraordinària riquesa lingüística, un mosaic de centenars de llengües que reflecteixen una història i una diversitat cultural profundes. Dins d'aquest panorama, conviuen llengües amb milions de parlants amb nombroses llengües regionals, minoritzades o minoritàries, algunes de les quals es troben en una situació de vulnerabilitat o fins i tot en perill d'extinció.[2] Aquesta diversitat lingüística no és merament un fet estadístic, sinó que representa un component fonamental i insubstituïble del patrimoni cultural europeu.[3] Cada llengua encarna una cosmovisió única, un sistema de coneixement, tradicions orals i una identitat col·lectiva que enriqueix la humanitat en el seu conjunt.

Més enllà de les llengües majoritàries, que gaudeixen d'un ús generalitzat i un reconeixement consolidat, coexisteixen nombroses llengües no hegemòniques que constitueixen una part intrínseca i valuosa del patrimoni cultural europeu. Aquestes llengües, parlades per comunitats sovint més reduïdes dins dels estats on romanen, contribueixen de manera fonamental a la diversitat cultural i

2 McMONAGLE, «Autochtone Minderheitensprache».

3 HULATT, «Langue minoritaire».

al mosaic lingüístic del continent. La vitalitat d'aquestes llengües és un reflex directe de la riquesa cultural d'Europa, i la seva preservació és un objectiu clau per a la humanitat en el seu conjunt, ja que la pèrdua de qualsevol llengua implica la desaparició de coneixement ancestral i de formes úniques de percebre i comprendre el món.

1.2. Definició i marc conceptual de les llengües europees no hegemòniques

Per tal de comprendre de manera rigorosa l'abast i la naturalesa de l'impacte de la IA en les llengües europees no majoritàries, és imprescindible establir primer una definició clara i un context adequat per a aquest terme dins l'àmbit europeu. La conceptualització d'aquestes llengües no és unívoca i depèn de diversos marcs de referència, principalment la Carta Europea de les Llengües Regionals o Minoritàries (ECRML)[4] i els criteris de la UNESCO sobre el perill d'extinció lingüística.

Definicions segons la Carta Europea de les Llengües Regionals o Minoritàries (ECRML) i la UNESCO

La Carta Europea de les Llengües Regionals o Minoritàries (ECRML), un tractat europeu adoptat el 1992 sota els auspicis del Consell d'Europa, proporciona una de les definicions més influents i utilitzades per a aquestes llengües. Segons la Carta, les «llengües regionals o minoritàries» són aquelles que es fan servir tradicionalment dins d'un territori determinat d'un estat per persones membres d'aquest estat que formen un grup numèricament inferior a la llengua hegemònica de la resta de població. És crucial destacar que aquestes llengües han de ser diferents de la llengua o llengües oficials d'aquest estat i la definició no inclou ni els dialectes de la llengua o llengües oficials ni les llengües dels migrants. La Carta també preveu les llengües no territorials, que són utilitzades per nacionals de l'estat però que no es poden identificar amb una àrea geogràfica particular. La determinació de quines llengües s'inclouen sota aquesta definició recau en cada estat signatari, que ha de considerar aspectes psicològics, sociològics i polítics, sense especificar un nombre o percentatge mínim de parlants.

En contrast, la UNESCO ofereix una perspectiva diferent, centrada en el grau de perillositat o risc de desaparició d'una llengua. Segons la UNESCO,[5] una llengua es considera en perill quan es troba en un camí cap a l'extinció. Aquesta situació es manifesta quan els seus parlants deixen d'utilitzar-la progressivament, la

4 COUNCIL OF EUROPE, «About the European Charter for Regional or Minority Languages».

5 MOSELEY, «Atlas of the world's languages in danger».

fan servir en un nombre cada vegada menor de dominis comunicatius i, de manera crítica, quan no la transmeten d'una generació a la següent, quan no hi ha nous parlants, ja siguin adults o nens. La transmissió intergeneracional és, per tant, un factor clau per determinar el nivell de perill d'una llengua. La UNESCO opera amb quatre nivells de perill d'extinció: vulnerable –no parlada pels nens fora de la llar–, en perill definitiu –els infants no la parlen–, seriosament en perill –només parlada per la gent gran– i críticament en perill –només parlada per la gent molt gran, de manera parcial i infreqüent.

Classificació i exemples de llengües no hegemòniques a Europa

Més enllà del terme «llengües regionals o minoritàries», s'utilitzen altres denominacions per referir-se a aquestes llengües, com ara «llengües minoritàries», «llengües minoritzades», «llengües regionals», «llengües menys utilitzades», «llengües comunitàries», «llengües locals» o «llengües patrimonials». La diversitat tipològica d'aquestes llengües és considerable. Es poden classificar en diverses categories, incloent-hi:

- Llengües d'una comunitat en un sol país on no són la majoria lingüística: exemples d'aquesta categoria són el sòrab a Alemanya i el gal·lès al Regne Unit.
- Llengües d'una comunitat en dos o més països on tampoc no són la majoria: el basc o el sami són exemples de llengües que es parlen en diversos estats però que no són majoritàries en cap d'ells.
- Llengües d'una comunitat que és minoritària en un país però majoritària en un altre: el danès a Alemanya il·lustra aquesta categoria, en què una llengua és minoritària en un context però oficial i majoritària en un altre.
- Llengües sense cap territori fix: el romaní o el ídix/jiddisch són exemples de llengües parlades per comunitats disperses sense un territori geogràfic delimitat.

En el context europeu, la magnitud d'aquesta diversitat és notable. Es calcula que hi ha més de 250 llengües indígenes a Europa. D'aquestes, entre 40 i 50 milions de persones parlen una de les aproximadament 60 llengües regionals, minoritàries i minoritzades dins de la UE. Malauradament, un nombre significatiu d'aquestes llengües es troben en risc de desaparició.

Les llengües de «pocs recursos»

Un aspecte crucial per entendre l'impacte de la IA és la distinció entre la definició sociolingüística de «llengua no hegemònica» i la definició tècnica de *low-resource language* (LRL) [llengua de pocs recursos].[6]

6 LAUMANN, «Low-resource language: what does it mean?».

Mentre que moltes llengües no hegemòniques són, de fet, llengües de baixos recursos en l'àmbit digital, els termes no són sinònims, i aquesta distinció té implicacions significatives per al desenvolupament de la IA.

En el cas de les llengües, el concepte «de pocs recursos» fa referència principalment a la manca de dades digitals suficients i de qualitat, així com a la limitació de recursos computacionals i d'investigació necessaris per entrenar models d'IA eficaços. Aquesta manca de dades pot incloure corpus de text i veu, diccionaris i altres recursos lingüístics estructurats. Per exemple, menys del 5% de les aproximadament 7.000 llengües parlades al món tenen una representació significativa en línia. Les dades disponibles per a algunes d'aquestes llengües es limiten sovint a textos religiosos, documents legals o articles de Wikipedia, que poden estar fins i tot traduïts automàticament i no ser representatius de l'ús quotidià de la llengua.

Aquesta definició de «baixos recursos» és fonamentalment una definició operativa i tècnica, centrada en la disponibilitat de dades per a l'entrenament de models d'IA, més que no pas en el nombre de parlants o el seu estatus sociolingüístic. Per exemple, el català, amb uns deu milions de parlants, és una llengua no majoritària a l'Estat espanyol, però gràcies a iniciatives com el Projecte AINA –definit en l'apartat 7.1 d'aquest estudi–, ha generat una quantitat significativa de recursos digitals que ha reduït la seva qualificació de «baixos recursos» en l'àmbit de la IA. En canvi, llengües amb grans poblacions com el birmà o el suahili són considerades de baixos recursos en el context de la IA a causa de la manca de dades digitals.

La implicació d'aquesta distinció és crucial: els esforços per al desenvolupament de la IA per a les llengües han de centrar-se específicament en la digitalització i la creació de recursos de dades d'alta qualitat, independentment de l'estatus demogràfic o polític de la llengua. La superació dels «baixos recursos» digitals és el principal obstacle per a la integració efectiva d'aquestes llengües en l'ecosistema de la IA. Això requereix inversions estratègiques en R+D, una major inclusió global en la investigació d'IA i una propietat de dades més equitativa.

La irrupció de la IA en les primeres dècades del segle XXI, i especialment en els últims anys, ha marcat una transformació tecnològica sense precedents. Aquesta tecnologia, amb el seu potencial disruptiu enorme, ja és present en moltes de les eines que utilitzem diàriament, des dels navegadors web fins a les aplicacions d'internet, i està impulsant avenços significatius en sectors tan diversos com la sanitat o la defensa.[7]

7 LÓPEZ, «Els robots i els sistemes intel·ligents revolucionen la medicina».

La capacitat de la IA per processar i generar informació a una escala i velocitat inaudites té implicacions profundes per a la vitalitat i el futur de totes les llengües. No obstant això, aquest impacte és particularment crític per a aquelles llengües amb menys recursos digitals, que s'enfronten a desafiaments únics en aquesta nova era.

És fonamental comprendre que la IA no opera en un buit sociolingüístic; la seva implementació i el seu desenvolupament estan intrínsecament lligats a les dinàmiques de poder i a la disponibilitat de recursos existents en el món real. Per tant, l'impacte de la IA en les llengües no hegemòniques no es pot considerar exclusivament una qüestió tècnica, sinó que esdevé una extensió i, potencialment, una amplificació de les desigualtats sociolingüístiques preexistents. Aquesta comprensió és crucial per abordar els reptes i aprofitar les oportunitats que la IA presenta.

La raó d'aquesta dinàmica rau en el fet que els sistemes d'IA, especialment els models de llenguatge avançats, s'alimenten de grans quantitats de dades per aprendre i funcionar.[8] Les llengües no hegemòniques, per la seva pròpia definició i per factors històrics i socioeconòmics, ja pateixen una manca estructural de dades digitals exhaustives i de qualitat.[9] Si el desenvolupament de la IA es concentra principalment en dades de llengües hegemòniques, les llengües minoritzades i minoritàries es trobaran inherentment en una posició de desavantatge. Aquesta situació no és el resultat d'un error tecnològic, sinó una conseqüència directa de l'«escletxa digital» preexistent i de la «minorització lingüística» que ja les afecta. En aquest sentit, l'impacte de la IA sobre aquestes llengües no és merament una qüestió tècnica de rendiment del model, sinó que reprodueix i pot amplificar les asimetries de poder lingüístic que ja caracteritzen el panorama global.[10] Per tant, l'anàlisi d'aquesta interacció requereix una perspectiva que transcendeixi la tecnologia per abraçar les seves implicacions socials i culturals.

8 «Generative Artificial Intelligence».

9 PAVA, «Mind the (Language) Gap: Mapping the challenges of LLM development in low-resource language contexts».

10 BASTARDES, «Les polítiques de la llengua i la identitat a l'era 'glocal'».

1.3. Objectius de l'estudi

Aquesta recerca es proposa els objectius fonamentals següents per aprofundir en la comprensió de la interacció entre la IA i les llengües europees no hegemòniques:

- Definir la IA i explicar la importància de la seva variant generativa, així com els models de llenguatge grans (LLM). Es proporcionarà una base conceptual sòlida per entendre les tecnologies subjacents a la discussió.
- Analitzar els efectes positius i negatius de la IA en les llengües europees no hegemòniques. Es desglossaran les oportunitats que pot oferir per a la preservació i revitalització, així com els riscos i desafiaments que planteja per a la seva supervivència digital.
- Presentar casos d'estudi concrets, com el català i l'islandès, per il·lustrar les oportunitats i els desafiaments. L'examen d'exemples reals permetrà contextualitzar la teoria i extreure'n lliçons pràctiques.
- Explorar l'abordatge de les llengües per part de les grans empreses tecnològiques i la rellevància dels *benchmarks* [indicadors de rendiment] comparatius. S'analitzarà com els actors dominants del sector tecnològic tracten la diversitat lingüística i la importància de les eines d'avaluació per a les llengües amb pocs recursos.
- Extreure'n lliçons per a les llengües europees mitjanes, com l'alemany, el francès i l'italià. Es buscaran sinergies i estratègies que les llengües no hegemòniques poden oferir a altres llengües minoritàries i minoritzades per afrontar els reptes de la IA.
- Examinar les referències a les llengües en l'AI Act de la UE. S'analitzarà el marc regulador europeu per identificar com aborda o pot influir en la diversitat lingüística.
- Formular conclusions i recomanacions per a la preservació lingüística en l'era de la IA. Es proposaran línies d'acció per garantir que la IA esdevingui una eina per a la inclusió i el foment de la diversitat lingüística, en lloc d'un factor d'erosió.

1.4. Estructura de l'informe

Aquest informe s'estructura en les següents seccions per abordar de manera sistemàtica els objectius plantejats:

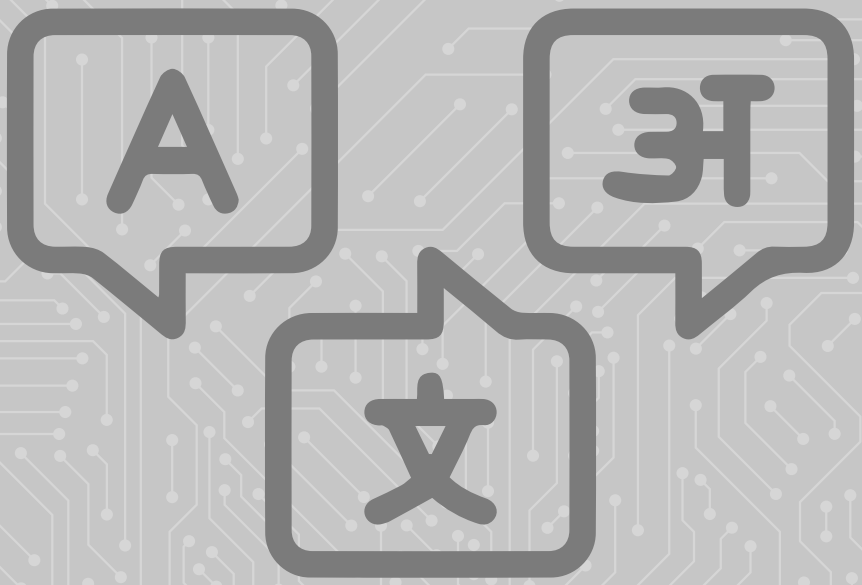
La introducció estableix el context de la diversitat lingüística europea, la irrupció de la IA, la importància de les llengües no hegemòniques i els objectius de l'estudi. A continuació, es defineixen la IA, la IA generativa i els models de llenguatge grans (LLM), i se n'explica el funcionament i la rellevància.

El panorama lingüístic europeu analitza les llengües hegemòniques, minoritàries i minoritzades, classificant i presentant la distribució de les llengües a Europa, incloent-hi una taula amb el nombre de parlants, i analitzant la vulnerabilitat de les llengües amb pocs recursos. Seguidament, s'examina com la IA serveix les llengües no hegemòniques, detallant les tecnologies actuals de processament del llenguatge natural i les seves aplicacions per a les llengües minoritàries i minoritzades.

L'estudi també aborda els efectes positius de la IA en les llengües europees no hegemòniques i en detalla les oportunitats que ofereix per a la preservació, revitalització, comunicació, creativitat i desenvolupament econòmic. Paral·lelament, s'exploren els riscos i desafiaments associats a l'esborrament de la diversitat lingüística, l'escletxa digital, els biaixos algorítmics, la manca de transparència, la desinformació, l'impacte laboral i l'erosió cultural.

Es presenten casos d'estudi i iniciatives d'èxit en llengües no majoritàries i s'analitzen la IA conversacional i extenses amb OpenAI i altres casos d'èxit a Europa. S'observa també l'abordatge de les llengües a les grans empreses tecnològiques i els *benchmarks* comparatius, analitzant com les grans tecnològiques gestionen la diversitat lingüística i la importància dels *benchmarks* per a llengües amb pocs recursos.

La Llei d'Intel·ligència Artificial de la UE i les llengües s'examina, així com les disposicions de l'AI Act, amb relació a la diversitat lingüística i els drets fonamentals. Finalment, s'ofereixen recomanacions basades en l'experiència de les llengües minoritàries per a les llengües com l'alemany, el francès i l'italià, se'n sintetitzen els resultats i se'n proposen estratègies per a la supervivència i el desenvolupament lingüístic en l'era de la IA.



2

FONAMENTS DE LA IA I ELS MODELS DE LLENGUATGE

2.1. Què és la IA?

La IA es pot definir com un camp interdisciplinari de la informàtica dedicat al desenvolupament de sistemes i màquines que són capaços de realitzar tasques que, tradicionalment, requereixen intel·ligència humana. Aquestes tasques abasten un ampli espectre de capacitats cognitives, incloent-hi l'aprenentatge a partir de dades, el raonament lògic, la resolució de problemes complexos, la percepció de l'entorn —visual i auditiva— i la comprensió i generació del llenguatge humà. L'objectiu fonamental de la IA és dotar les màquines de la capacitat d'imitar i, en certs casos, superar les habilitats cognitives humanes en dominis específics.

Els sistemes d'IA no són programats explícitament amb totes les regles per resoldre una tasca, sinó que aprenen a partir de l'experiència. La seva construcció es basa principalment en l'aplicació de tècniques d'aprenentatge automàtic —*machine learning*—, un subcamp de la IA que permet as sistemes millorar el seu rendiment amb l'exposició a més dades. Dins de l'aprenentatge automàtic, les arquitectures de xarxes neuronals profundes —*deep neural networks*— han esdevingut la pedra angular de la IA moderna. Aquestes xarxes, inspirades en l'estructura del cervell humà, consisteixen en múltiples capes de «neurones» interconnexes que

processen la informació de manera jeràrquica i permeten la detecció de patrons complexos en grans conjunts de dades. La capacitat d'aquestes xarxes per aprendre representacions sofisticades de les dades ha estat clau per als avenços recents en la IA.

2.2. La importància de la IA generativa

Dins del vast camp de la IA, la Intel·ligència Artificial generativa –IA generativa, GenAI o GAI– ha emergit com una de les àrees més innovadores i de ràpida expansió. Es tracta d'un subcamp de la IA que se centra en la creació de models capaços de produir dades noves i realistes, que no són meres còpies de les dades d'entrenament, sinó que reflecteixen les seves característiques subjacents. Aquests models generatius aprenen els patrons i les estructures inherents als seus conjunts de dades d'entrenament i posteriorment utilitzen aquest coneixement per generar contingut original basat en una entrada determinada, sovint en forma de peticions en llenguatge natural, conegudes com a indicacions o *prompts*. La IA generativa pot produir una àmplia varietat de modalitats de dades, incloent-hi text, imatges, vídeos, àudio o codi de programari.

L'auge de la IA generativa, especialment a partir de la dècada de 2020, ha estat impulsat per avenços significatius en les xarxes neuronals profundes basades en arquitectures de transformadors, una innovació que ha estat fonamental per al desenvolupament dels models de llenguatge grans (LLM). Aquesta tecnologia ha permès la creació d'eines que han revolucionat nombrosos sectors. Per exemple, en la creació i augment de contingut, els sistemes d'IA generativa poden produir esborranys de text en un estil i longitud desitjats, generar dades sintètiques per a l'entrenament de models o automatitzar la creació de contingut per a màrqueting i publicitat, augmentant-ne dràsticament l'eficiència.[11] En el camp de la salut, la IA generativa és crucial per accelerar el descobriment de fàrmacs, creant estructures moleculars amb característiques desitjades, i per generar imatges radiològiques que serveixen per entrenar models de diagnòstic, fet que facilita així una presa de decisions mèdiques més ràpida i econòmica. En el sector financer, la seva utilitat es manifesta en la generació de conjunts de dades per entrenar models, l'automatització de la generació d'informes amb capacitats de resum en llenguatge natural, i la personalització de les comunicacions amb els clients, que en millora l'eficiència i redueix costos operacionals.

11 AMAZON WEB SERVICES, «¿Qué es la IA generativa?».

L'educació també s'hi ha vist profundament impactada, ja que les eines d'IA generativa permeten personalitzar l'aprenentatge mitjançant la creació de proves, materials d'estudi i composició d'assajos, que ha beneficiat tant professors com alumnes. A més, la IA generativa augmenta la productivitat de diversos tipus de treballadors, ja que pot resumir, simplificar i classificar continguts, generar i verificar codi de programari i millorar el rendiment dels bots de conversa i agents virtuals.

La capacitat de la IA generativa per produir text, imatges i àudio a gran escala i de manera autònoma la converteix en una eina de doble tall per a les llengües no hegemòniques. Per una banda, pot ser un catalitzador potent per a la creació de contingut original i per a la revitalització lingüística, ja que ofereix recursos que abans eren inaccessibles. Però, per altra banda, representa una amenaça significativa si els models subjacents no estan entrenats amb dades representatives i de qualitat d'aquestes llengües, perquè perpetua així la seva invisibilitat digital i la seva marginació enfront de les llengües dominants.

La capacitat de la IA generativa per «produir text, imatges, vídeos o altres formes de dades» i per «crear nous continguts i idees»[12] implica que, si les llengües no hegemòniques no tenen una presència adequada i representativa en les dades d'entrenament d'aquests models, els sistemes generatius no en podran produir contingut de qualitat. El contingut generat podria ser inexacte, contenir errors gramaticals o reflectir biaixos culturals aliens, com s'ha observat en el cas de l'islandès amb GPT-4.[13] Aquesta situació significa que la «revolució» de la IA generativa, en lloc de ser universalment beneficiosa, podria ser-ho de manera desproporcionada en les llengües ja dominants, i augmentar l'escletxa digital existent i exercir una pressió addicional cap a l'homogeneïtzació lingüística i cultural. Així, el potencial transformador de la IA generativa es manifesta de manera asimètrica, i depèn críticament de la inclusió i representació lingüística en les seves bases de dades.

2.3. Models de llenguatge grans (LLM): definició i funcionament

Un model de llenguatge gran (LLM) constitueix un tipus avançat de programa d'IA que opera mitjançant l'aprenentatge profund —*deep learning.* Aquests models utilitzen una arquitectura de xarxa neuronal artificial per processar i comprendre la informació lingüística.[14] La seva característica distintiva és que estan entre-

12 AMAZON WEB SERVICES, «¿En qué consiste la IA generativa?».

13 KULP, «Studies explore challenges of AI for low-resource languages».

14 UNITED STATES MILITARY ACADEMY LIBRARY, «Large Language Models».

nats amb quantitats ingents de dades de text, sovint a una escala d'internet, la qual cosa els habilita per comprendre, generar i manipular el llenguatge humà amb una precisió i fluïdesa notables.

Els LLM són un subconjunt crucial de la IA generativa, ja que la seva funció principal és la creació de contingut textual. Es centren específicament en tasques relacionades amb el llenguatge, que els diferencia d'altres models generatius que poden produir imatges o àudio.[15] Mentre que inicialment molts LLM es limitaven a acceptar entrades de text, els avenços recents han portat al desenvolupament de models multimodals, com el GPT-4 d'OpenAI, que poden processar diverses modalitats d'entrada, incloent-hi text, imatges i àudio, que amplien significativament les seves capacitats.

El funcionament bàsic d'un LLM es basa en la predicció probabilística. Un cop entrenat, el model és capaç de predir la paraula següent en una seqüència de text a partir de les paraules anteriors i el seu context, seleccionant la paraula més probable segons les distribucions de probabilitat que ha après.[16] Aquesta capacitat, aparentment simple, esdevé sorprenentment competent en la resolució de tasques complexes gràcies a la immensa quantitat de dades d'entrenament que processa.

Els LLM s'apliquen en una gamma àmplia de tasques basades en text. Són utilitzats per a la traducció de llenguatges, la generació de contingut –des de resums fins a articles complets–, la personalització de comunicacions i per alimentar bots de conversa i assistents virtuals que interactuen amb els usuaris de manera conversacional. La seva versatilitat els ha convertit en una eina fonamental en nombrosos àmbits digitals.

La dependència dels LLM d'«enormes quantitats de dades de text» és la causa fonamental de la seva infrarepresentació i mal rendiment en llengües amb pocs recursos. Aquesta dependència crea un cicle de retroalimentació negativa, on la manca de dades porta a models deficients, que al seu torn desincentiven la creació de més dades en aquestes llengües, perpetuant així la seva marginació digital.

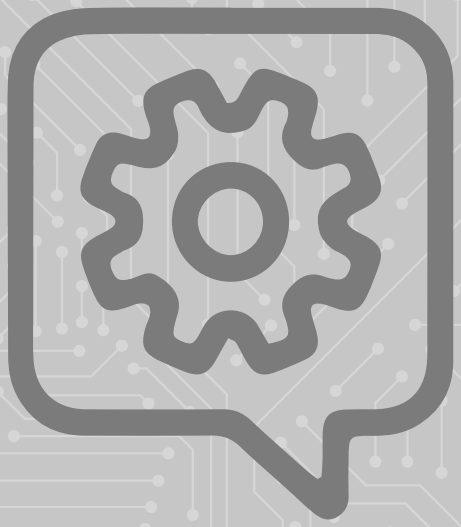
L'explicació d'aquesta dinàmica es pot traçar pas a pas: els LLM es defineixen per la seva necessitat intrínseca d'«enormes quantitats de dades de text» per assolir fluïdesa i precisió. No obstant això, les llengües

15 BELL, «Generative AI vs. Large Language Models (LLMs): What's the Difference?».

16 KLEIBER, «Was ist generative künstliche Intelligenz (KI)?».

amb pocs recursos es caracteritzen precisament per una manca de dades lingüístiques completes, ja sigui en forma de documentació escrita, eines digitals o investigació acadèmica. Aquesta escassetat de dades d'entrenament, tant en quantitat com en qualitat, condueix directament a un mal rendiment dels LLM quan s'apliquen a aquestes llengües.

La conseqüència d'aquest rendiment deficient és que les empreses i els usuaris troben que els models no són prou útils o fiables per a les seves necessitats en aquestes llengües, la qual cosa desincentiva la seva adopció i ús. La manca d'ús, al seu torn, significa que es generen menys dades noves en aquestes llengües a través de la interacció digital, i tanquen el cercle i perpetuen l'escletxa de recursos. Aquest mecanisme subratlla que el desavantatge de les llengües no hegemòniques en l'era dels LLM no és un problema aïllat, sinó un fenomen arrelat en la infraestructura de dades i en les dinàmiques d'adopció tecnològica.



3

EL PANORAMA LINGÜÍSTIC EUROPEU: LLENGÜES HEGEMÒNIQUES I LLENGÜES MINORITZADES I MINORITÀRIES

3.1. Classificació i distribució de les llengües a Europa

Europa és un continent que conté una diversitat lingüística excepcional, amb centenars de llengües que reflecteixen la seva complexa història i les seves múltiples cultures. La majoria d'aquestes llengües pertanyen a la família indoeuropea, que es subdivideix en diverses branques principals:[17]

- Llengües romàniques: inclouen el català, el francès, l'italià, el castellà, el romanès, el portuguès, el sicilià, el venecià, el gallec, el sard, l'occità, l'aragonès, l'asturià, el francoprovençal, el friülà, el ladí, el còrnic i el mirandès.
- Llengües germàniques: comprenen l'alemany, l'anglès, el neerlandès, el suec, el danès, el noruec, el frisó, el luxemburguès, l'islandès i el ídix/jiddisch, entre d'altres.
- Llengües eslaves: engloben el rus, l'ucraïnès, el polonès, el serbocroat, el txec, el búlgar, l'eslovac, el bielorús, l'eslovè i el macedoni.
- Llengües cèltiques: inclouen el gaèlic irlandès, el gaèlic escocès, el gal·lès, el bretó, el còrnic i el manx.

17 «Languages of Europe».

- Llengües bàltiques: representades pel lituà i el letó.
- Llengües hel·lèniques: principalment, el grec.
- Altres llengües indoeuropees: com l'albanès i l'armeni.

Més enllà de la família indoeuropea, Europa també acull llengües que pertanyen a altres troncs lingüístics o que són aïllades:

- Llengües no indoeuropees: inclouen el basc, una llengua aïllada sense relació coneguda amb altres famílies. També hi ha les llengües uralianes, com el finès, l'estonià i l'hongarès.[18] Altres famílies presents són les llengües turqueses –com el turc o el gagaús– i diverses llengües caucàsiques –com el txetxè o l'àvar.

Dins del marc de la UE, hi ha 24 llengües oficials reconegudes.[19] No obstant això, la riquesa lingüística real del continent va molt més enllà, amb més de 60 llengües regionals o minoritàries autòctones que són parlades per aproximadament 40 milions de persones. Aquestes llengües, tot i no gaudir sovint d'un estatus oficial a nivell estatal, constitueixen una part vital del patrimoni cultural i social d'Europa. La seva protecció i promoció són objectius clau de la Carta Europea de les Llengües Regionals o Minoritàries del Consell d'Europa.

3.2. Llista de llengües europees per nombre de parlants

La taula següent presenta una selecció de llengües europees, incloent-hi tant les majoritàries com una representació significativa de les no hegemòniques, ordenades pel nombre estimat de parlants nadius. Aquesta informació és crucial per visualitzar la disparitat en la mida de les comunitats lingüístiques, un factor directe que influeix en la disponibilitat de dades per a l'entrenament de sistemes d'IA. La quantitat de parlants nadius i totals és un indicador clau de la «riquesa de recursos» d'una llengua en el context digital. Les llengües amb menys parlants sovint tenen menys contingut digital disponible, cosa que les converteix en «llengües amb pocs recursos» per a l'entrenament d'IA. Aquesta taula, per tant, estableix la base empírica per a l'anàlisi de l'escletxa digital que es desenvoluparà en seccions posteriors de l'informe.

18 TATUTRAD TRADUCTORES, «¿Qué idiomas se hablan en Europa?».

19 En el moment de tancar l'estudi a l'octubre del 2025, la negociació per fer oficial el català a la UE encara no ha conclòs.

Taula 1. Llengües europees per nombre de parlants estimats a tot el món (nadius i totals)

Llengua	Parlants nadius	Parlants totals	Principals territoris a Europa (Estatal/Subestatal)
Anglès	380.000.000	1.457.000.000	Regne Unit, Irlanda, Malta
Castellà	519.000.000	636.000.000	Estat espanyol
Francès	74.000.000	312.000.000	França, Luxemburg, Mònaco, Suïssa / Valònia, Vall d'Aosta
Portuguès	250.000.000	267.000.000	Portugal
Rus	145.000.000	253.000.000	Rússia, Bielorússia, Ucraïna
Alemany	95.000.000	180.000.000	Alemanya, Àustria, Bèlgica, Liechtenstein, Luxemburg, Suïssa / Tirol del Sud
Turc	85.000.000	91.000.000	Turquia
Italià	65.000.000	85.000.000	Itàlia, San Marino, Suïssa, Vaticà / Ístria
Polonès	40.000.000	43.000.000	Polònia
Ucraïnès	32.000.000	39.000.000	Ucraïna
Neerlandès	25.000.000	30.000.000	Països Baixos / Flandes
Romanès	22.000.000	-	Romania, Moldàvia
Serbocroat	17.000.000	-	Sèrbia, Croàcia, Bòsnia i Hercegovina, Montenegro, Kosovo
Hongarès	14.000.000	-	Hongria
Grec	13.500.000	-	Grècia, Xipre
Suec	10.000.000	13.000.000	Suècia / Åland
Txec	9.800.000	12.000.000	Txèquia
Català	4.800.000	10.000.000	Andorra / Catalunya, País Valencià, Illes Balears, Aragó, Catalunya Nord, L'Alguer
Búlgar	6.800.000	7.900.000	Bulgaria
Albanès	7.500.000	-	Albània, Kosovo, Macedònia del Nord, Montenegro
Eslovac	5.000.000	7.100.000	Eslovàquia, Txèquia / Vojvodina
Bielorús	5.000.000	6.300.000	Bielorússia
Danès	6.000.000	-	Dinamarca / Schleswig-Holstein
Finès	5.000.000	-	Finlàndia

Llengua	Parlants nadius	Parlants totals	Principals territoris a Europa (Estatal/Subestatal)
Romaní	4.600.000	-	/ Diverses zones
Noruec	4.320.000	-	Noruega
Lituà	4.000.000	-	Lituània
Eslovè	2.500.000	-	Eslovènia
Gallec	2.400.000	-	/ Galícia
Irlandès	300.000	2.100.000	Irlanda, Irlanda del Nord
Macedoni	1.700.000	-	Macedònia del Nord
Occità	1.000.000	1.670.000	/ Occitània, Aran
Letó	1.500.000	-	Letònia
Estonià	1.200.000	-	Estònia
Basc	800.000	1.200.000	/ País Basc, Navarra, Iparralde
Sard	1.000.000	-	/ Sardenya
Gal·lès	610.000	724.000	/ País de Gal·les
Reto-romànic	500.000	660.000	/ Friül, Ladinia, Grisons
Maltès	570.000	-	Malta
Asturià	100.000	550.000	/ Astúries
Frisó	470.000	-	/ Frísia
Luxemburguès	400.000	-	Luxemburg
Islandès	390.000	-	Islàndia
Gaèlic	70.000	200.000	/ Escòcia
Arpità	157.000	-	
Cors	150.000	-	/ Còrsega
Bretó	120.000	-	/ Bretanya
Feroès	70.000	-	/ Fèroe
Groenlandès	57.000	-	/ Groenlàndia
Aragonès	25.000	50.000	/ Aragó
Sami	23.000	-	/ Lapònia
Sòrab	20.000	-	/ Lusàcia
Còrnic	563	-	/ Cornualla

Font: Elaboració pròpia.

3.3. La vulnerabilitat de les llengües amb pocs recursos en l'era digital

Les «llengües amb pocs recursos» (LRLs) representen una categoria crítica en el panorama lingüístic global, particularment en l'era digital. Aquestes llengües es caracteritzen per una manca significativa de recursos lingüístics complets, la qual cosa inclou una escassa documentació escrita, una absència o limitació d'eines digitals per al seu processament i una investigació acadèmica insuficient. A Europa, moltes de les llengües no hegemòniques presentades a la Taula 1 encaixen plenament en aquesta definició, enfrontant-se a desafiaments únics en la seva adaptació al món digital.

La manca de dades digitals és un problema estructural i crític per a la supervivència d'aquestes llengües en l'ecosistema de la IA. Les dades disponibles per a algunes d'aquestes llengües sovint es limiten a corpus molt específics, com ara textos religiosos –la Bíblia–, documents legals o articles de la Viquipèdia. Un problema addicional és que fins i tot aquests recursos limitats poden haver estat generats mitjançant traducció automàtica, la qual cosa significa que no són representatius de l'ús quotidià i natural de la llengua i poden contenir errors o biaixos. Aquesta situació es veu agreujada pel fet que, com hem dit anteriorment, menys del 5% de les aproximadament 7.000 llengües parlades al món tenen una representació significativa a internet.

La història de moltes llengües indígenes i no hegemòniques està marcada per la pressió i el perill d'extinció, sovint a causa de circumstàncies tràgiques com l'assimilació lingüística forçada. L'escletxa digital actual pot exacerbar dràsticament aquest perill. Les comunitats que parlen aquestes llengües queden excloses dels beneficis de la tecnologia moderna, ja que els models d'IA no funcionen eficaçment per a elles. Aquesta exclusió es manifesta en un «forat negre digital» que impedeix l'accés a serveis en línia essencials com l'educació, l'ocupació o la informació sanitària, i que crea una nova capa de desigualtat.

La manca de dades digitals de qualitat i en quantitat suficient per entrenar models d'IA amenaça d'ampliar les divisions globals existents excloent parts del món de la tecnologia transformadora. Aquesta situació no només es tradueix en una inconveniència, sinó en una «exclusió sistemàtica» que priva cultures i comunitats senceres dels avantatges econòmics i educatius que les llengües majoritàries obtenen de la tecnologia.

La vulnerabilitat de les llengües amb pocs recursos en l'era de la IA no és simplement una conseqüència tècnica de la manca de dades, sinó un reflex i una amplificació de desigualtats socioculturals i històriques

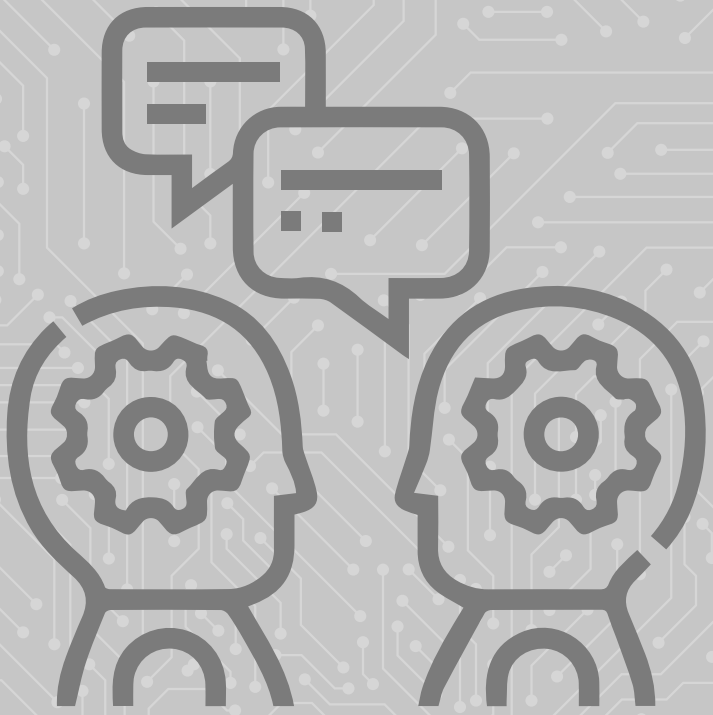
profundament arrelades. La minorització en l'àmbit digital es transforma en una forma de colonització digital, on els patrons i perspectives de les llengües dominants continuen imposant-se i posen en perill la diversitat de coneixements i visions del món inherents a cada llengua.

Aquesta dinàmica es pot entendre com una cadena de causalitat: la definició mateixa de «llengües amb pocs recursos» inclou la manca de documentació escrita, eines digitals o investigació acadèmica. Aquesta escassetat de recursos no és un fenomen aïllat, sinó que té arrels històriques profundes, com l'assimilació lingüística forçada que han patit, i pateixen, moltes comunitats.

El resultat directe d'aquesta mancança és que menys del 5% de les llengües tenen representació significativa en línia.[20] Aquesta infrarepresentació condueix a una exclusió sistemàtica de les comunitats que parlen aquestes llengües dels beneficis de la IA. Aquesta exclusió, al seu torn, es tradueix en una pèrdua de coneixement indígena i una erosió cultural, ja que les visions del món encapsulades en aquestes llengües esdevenen inaccessibles en l'àmbit digital.

Aquesta interconnexió de factors revela que la IA, sense una intervenció deliberada i conscient, no només reflecteix les desigualtats existents, sinó que les magnifica, produeix un desequilibri de poder que transcendeix la mera tecnologia i es converteix en una qüestió de justícia lingüística i cultural.

20 PAVA (et al.), «Mind the (Language) Gap: Mapping the Challenges of LLM Development in Low-Resource Language Contexts».



4

LA IA AL SERVEI DE LES LLENGÜES MINORITZADES I MINORITÀRIES: APLICACIONS ACTUALS DE PROCESSAMENT DEL LLENGUATGE NATURAL

La IA i, de manera més específica, el processament del llenguatge natural (PLN),[21] ofereixen un conjunt d'eines i tecnologies amb un potencial considerable per influir positivament en el futur de les llengües europees no hegemòniques. Aquestes aplicacions poden abordar molts dels reptes que afronten aquestes llengües en l'era digital, especialment la manca de recursos i la dificultat per generar contingut i interactuar en l'àmbit digital.

4.1. Traducció automàtica

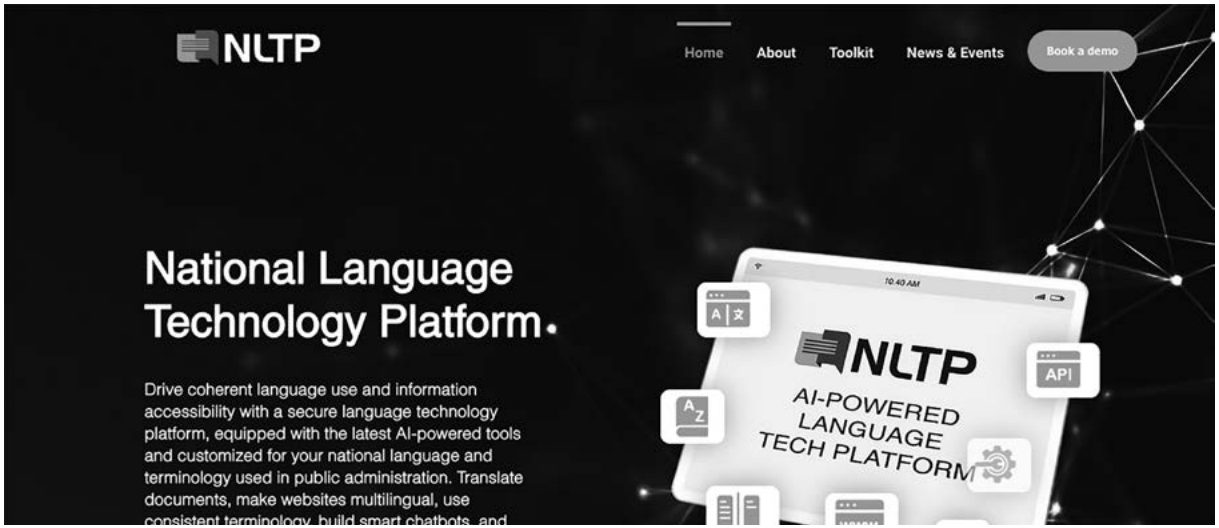
La traducció automàtica (TA) basada en la IA s'ha convertit en una de les aplicacions més rellevants per a les llengües no hegemòniques. Aquestes eines estan facilitant la documentació de llengües menys conegudes, cosa que permet als lingüistes i a les comunitats enregistrar i traduir formes orals i escrites, i crear així recursos digitals essencials que abans eren inassolibles. La TA impulsada per IA té la capacitat de millorar substancialment l'accessibilitat a la informació i als serveis per als parlants de llengües no hegemòniques, facilitant la comunicació

21 Disciplina informàtica que s'encarrega de tractar computacionalment els llenguatges humans.

amb aquells que parlen llengües majoritàries o dominants. Aquest fet pot tenir un impacte significatiu en la vida quotidiana i en la capacitat de participació pública de les persones que utilitzen aquestes llengües.

Diversos projectes i iniciatives il·lustren aquest potencial. La Plataforma Nacional de Tecnologia Lingüística (NLTP)[22], impulsada per la UE i ja activa a Croàcia, Estònia, Islàndia, Letònia, Malta, se centra específicament a proporcionar serveis de traducció automàtica a les administracions públiques, i posa un èmfasi particular en les llengües no hegemòniques i l'adaptació a les necessitats de l'administració pública local. La mateixa Comissió Europea treballa en el desenvolupament de sistemes de traducció automàtica per a les llengües de la UE, incloent-hi aquelles amb un menor nombre de parlants. A més, empreses tecnològiques com Google i Microsoft estan col·laborant activament amb universitats i centres de recerca per desenvolupar sistemes de traducció sofisticats dissenyats específicament per a llengües en risc d'extinció.[23]

Imatge 1. National Language Technological Platform [Plataforma Nacional de Tecnologia Lingüística]



Font: National Language Technological Platform.

22 NLTP, «National Language Technology Platform».

23 BAPNA, «Building machine translation systems for the next thousand languages».

Per a llengües no hegemòniques específiques, com el basc, ja hi ha motors de traducció automàtica disponibles. El projecte Itzuli,[24] implementat pel Govern basc, és un exemple concret d'aquesta aplicació, que utilitza la IA per permetre la traducció entre el basc, el castellà, el francès i l'anglès. Aquests sistemes, tot i que encara en desenvolupament, representen un pas endavant crucial per a la visibilitat i la funcionalitat de les llengües no hegemòniques en l'entorn digital.

4.2. Reconeixement de veu

El reconeixement de veu (RV) basat en IA també presenta un potencial revolucionari per a l'ús de les llengües no hegemòniques. Aquesta tecnologia permet la interacció amb dispositius i serveis mitjançant la parla, la qual cosa és especialment rellevant per a llengües amb una tradició oral forta o per a usuaris amb un baix nivell d'alfabetització. Projectes com el desenvolupament de Macsen,[25] un assistent de veu en llengua gallesa, il·lustren clarament aquest potencial.

Investigadors de diverses institucions estan treballant activament en la creació d'eines de reconeixement i transcripció de veu per a llengües com el sami. Una estratègia prometedora consisteix en l'adaptació de models de reconeixement de veu que han estat entrenats amb grans conjunts de dades de llengües majoritàries per millorar la precisió en el reconeixement de llengües no hegemòniques. No obstant això, per aconseguir sistemes de reconeixement de veu precisos per a aquestes llengües, la creació de corpus de veu específics és una tasca essencial. La disponibilitat de dades de veu de qualitat és un factor determinant per a l'èxit d'aquestes aplicacions. El reconeixement de veu basat en IA té la capacitat de fer que la tecnologia sigui molt més accessible per als parlants de llengües no hegemòniques, especialment per a aquells amb dificultats per escriure, obrint noves vies per a la interacció digital i per a la creació de contingut en aquestes llengües.

4.3. Generació de text

La generació de text mitjançant IA és una altra aplicació amb un gran potencial per a les llengües no hegemòniques. Eines com Cardamom Workbench[26] han estat dissenyades per ajudar els usuaris a generar text

24 EUSKADI.EUS, «Traductor neuronal».

25 WELSH NATIONAL LANGUAGE TECHNOLOGIES PORTAL, «Macsen».

26 IRISH RESEARCH COUNCIL, «Cardamom. Comparative deep models for minority and historical languages».

per a llengües que tenen pocs recursos digitals disponibles. Els grans models lingüístics (LLM) han demostrat avenços significatius en la generació de text, però el seu rendiment per a llengües no hegemòniques sovint es veu limitat per la manca de grans quantitats de dades d'entrenament.

Per abordar aquesta limitació, la generació de text per aquest tipus de llengües requereix l'ús d'estratègies específiques per a la recopilació de dades d'internet que siguin rellevants i suficients per a l'entrenament dels models. La IA té el potencial d'ajudar a generar contingut escrit en aquestes llengües, cosa que pot ser particularment útil per a l'educació, la creació de materials culturals i la preservació de la llengua. No obstant això, la qualitat i la rellevància d'aquest contingut estan directament relacionades amb la quantitat i la qualitat de les dades d'entrenament que s'utilitzen, fet que posa de manifest la importància de la curació de dades i la implicació de la comunitat.

4.4. Altres aplicacions de PLN basades en IA

Més enllà de la traducció, el reconeixement de veu i la generació de text, el PLN basat en IA ofereix una àmplia gamma d'altres possibilitats. Per exemple, els bots de conversa o assistents de veu es poden fer servir per facilitar l'aprenentatge d'idiomes i per proporcionar accés a serveis públics en llengües no hegemòniques, superant les barreres lingüístiques i millorant la comunicació. Aquests assistents conversacionals poden oferir pràctica de la llengua en un entorn interactiu i personalitzat.

La IA també pot ser una eina valuosa per estructurar dades de cara a activitats culturals i per generar subtítols automàtics per a contingut multimèdia en llengües no hegemòniques, i augmentar la seva accessibilitat i abast. Aquestes aplicacions tenen el potencial de fer que aquesta tipologia de llengües siguin més visibles, utilitzades i valorades en diversos contextos, i contribueixen significativament a la seva vitalitat i presència en el món digital.

Sistemes de conversió de text a veu (TTS) i de conversió de veu a text (STT)

Les tecnologies de conversió de text a veu –TTS, per les sigles en anglès– i de conversió de veu a text (STT) són particularment útils per a aquestes llengües, ja que ajuden els educadors a crear materials didàctics amb recursos mínims, generen àudios a partir de textos o transcriuen la parla dels aprenents. Aquestes eines poden facilitar enormement la creació de contingut educatiu i materials de formació, especialment en llengües amb pocs recursos escrits.

Doblatge automàtic de vídeos

En aquest àmbit podem incloure el sistema de doblatge automàtic de vídeos al català[27] creat per Softcatalà. Encara en proves, integra eines existents com són Whisper[28] –reconeixement de la parla–, Matxa[29] –síntesi de veu–, Pyannote[30] –identificació del parlant–, Demuc[31] –separació de la parla–, Audeering[32] –reconeixement del gènere del parlant, nmt-softcatala[33] –traducció anglès-català–, Apertium[34] –traducció castellà-català–, Open-dubbing»[35] –sistema de doblatge automàtic– i subdub-editor[36] per a l'edició.

Anàlisi de sentiments i processament semàntic

L'anàlisi de sentiments, que identifica i extreu informació subjectiva a partir de dades textuals, i el processament semàntic adaptat a llengües no hegemòniques pot ajudar a comprendre millor les opinions i actituds de les comunitats parlants, així com millorar la moderació de continguts en plataformes digitals. El processament del llenguatge natural (PNL) pot ajudar a traduir llengües no hegemòniques per millorar la moderació de continguts en plataformes digitals. Això és especialment rellevant per a la detecció de discursos d'odi o desinformació, fins i tot en llocs web ultraespecialitzats o amb poca visibilitat, on la supervisió humana és inviable.

Aplicacions educatives interactives

Les plataformes d'aprenentatge d'idiomes impulsades per IA poden oferir experiències d'aprenentatge interactives i personalitzades per a llengües no hegemòniques. Aquestes eines poden adaptar-se al nivell de desenvolupament i a les necessitats individuals dels estudiants i proporcionar suport adaptatiu a escala. Les aplicacions mòbils basades en IA poden ampliar l'accés a una educació lingüística de qualitat per a comunitats que es troben en zones remotes i aïllades, i superar així les limitacions geogràfiques i la manca de professors nadius.

---

27 Veure a: <www.softcatala.org/doblatge>.

28 Veure a: <www.openai/whisper>.

29 Veure a: <www.projecte-aina/matxa-tts-cat-multiaccent>.

30 Veure a: <www.github.com/pyannote/pyannote-audio>.

31 Veure a: <www.adefossez/demucs>.

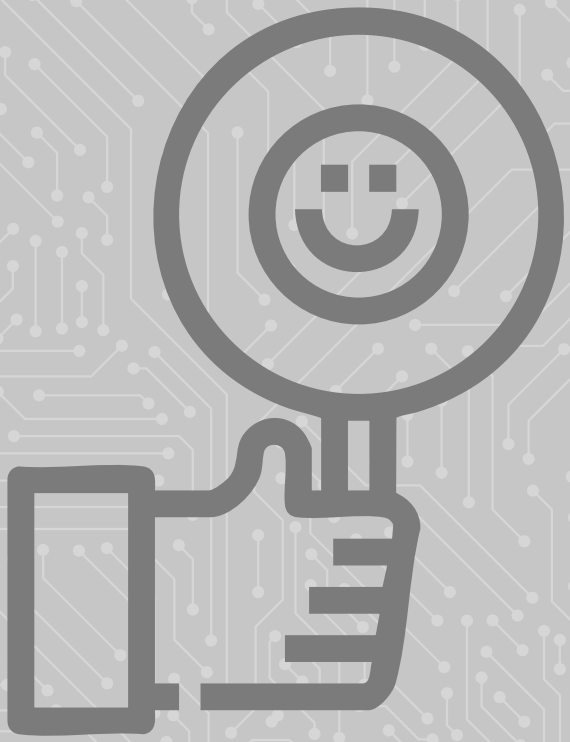
32 Veure a: GITHUB, «Model for Age and Gender Recognition based on Wav2vec 2.0.».

33 Veure a: <www.softcatala/nmt-softcatala>.

34 Veure a: APERTIUM, «Apertium. Plataforma lliure i de codi obert per a la traducció automàtica».

35 Veure a: <www.softcatala/open-dubbing>.

36 Veure a: <www.softcatala/subdub-editor>.



5

EFECTES POSITIUS DE LA IA EN LES LLENGÜES EUROPEES NO HEGEMÒNIQUES

La IA, malgrat els desafiaments que planteja, ofereix un ventall d'oportunitats significatives per a la preservació, revitalització i promoció de les llengües europees no hegemòniques. Aquestes oportunitats es manifesten en diverses àrees clau, i transformen la manera com les comunitats lingüístiques poden interactuar amb la seva llengua en l'àmbit digital.

5.1. Preservació i documentació lingüística

La IA pot ser una eina instrumental en l'arxivament i la preservació digital de llengües en perill d'extinció. Permet la creació de repositoris digitals exhaustius de textos escrits i parlats, que serveixen com a bases de dades inestimables per a lingüistes i aprenents d'idiomes.[37] Per a les llengües dominants orals, els models d'IA poden transcriure automàticament enregistraments d'àudio, converteixen la llengua parlada en forma escrita. Aquest procés automatitzat redueix dràsticament la càrrega de treball manual per a lingüistes i investigadors de camp, que abans requerien hores de feina intensiva. A més, el text generat per IA, un cop

37 SPECTOR, «AI in languagepreservation: Safeguarding low-resource and indigenous languages».

validat, pot ajudar els lingüistes a desenvolupar de manera més eficient diccionaris, guies gramaticals i sistemes de notació fonètica. Aquestes eines són fonamentals per a la codificació i l'estudi sistemàtic de la llengua, són components crítics per a la seva preservació a llarg termini.

La IA també pot facilitar la creació d'arxius digitals dinàmics que no només emmagatzemen informació lingüística, sinó que la fan accessible i consultable de maneres innovadores. Mitjançant tècniques de processament del llenguatge natural, aquests arxius poden permetre cerques sofisticades per patrons lingüístics, estructures gramaticals específiques o evolució semàntica, i proporcionen eines valuoses per a la investigació lingüística i la documentació cultural.

5.2. Revitalització i aprenentatge de llengües

Una de les aplicacions més prometedores de la IA és en el camp de la revitalització i l'aprenentatge de llengües. Les eines d'IA poden fer que l'aprenentatge de llengües sigui més interactiu, personalitzat i àmpliament disponible mitjançant el desenvolupament d'aplicacions mòbils i plataformes d'aprenentatge electrònic. Les tecnologies de conversió de text a veu (TTS) i de veu a text (STT) són particularment útils, ja que ajuden els educadors a crear materials didàctics amb recursos mínims, generen àudios a partir de textos o transcriuen la parla dels aprenents.

Les plataformes habilitades per IA poden facilitar la creació d'aules virtuals, per connectar parlants nadius i aprenents a través de barreres geogràfiques. Això és especialment valuós per a comunitats disperses o per a aquells que no tenen accés a professors presencials. La IA també ofereix avantatges clau en l'aprenentatge personalitzat, adaptant els continguts al ritme i a les necessitats individuals de cada alumne. La retroacció immediata, per exemple, en la pràctica de la pronunciació o la correcció gramatical, és un element que millora significativament l'adquisició del vocabulari i la fluïdesa. La ludificació, amb elements lúdics integrats per la IA, pot augmentar la motivació i el compromís dels aprenents.

Els bots de conversa o assistents de veu i tutors virtuals

Els bots de conversa o assistents de veu i els tutors virtuals, impulsats per IA, poden simular converses en la llengua no hegemòniques i proporcionar als estudiants una pràctica immersiva i la possibilitat de millorar les seves habilitats comunicatives en un entorn relaxat i atractiu. Aquests assistents d'IA poden oferir assistència, respondre preguntes i facilitar sessions de pràctica replicant converses quotidianes. La IA pot

identificar les emocions dels alumnes i desxifrar el llenguatge no literal, que pot ajudar els estudiants que tenen dificultats amb termes metafòrics.

Materials d'aprenentatge personalitzats

La IA pot ser utilitzada per generar materials d'aprenentatge personalitzats, com ara exercicis, qüestionaris i històries interactives adaptats a les necessitats individuals de l'alumnat. Aquesta capacitat de personalització permet crear experiències d'aprenentatge úniques que s'adapten al nivell, els interessos i l'estil d'aprenentatge de cada persona, que maximitzen l'eficàcia del procés educatiu.

5.3. Millora de la comunicació i l'accessibilitat

La IA té el potencial de millorar dràsticament la comunicació en llengües no hegemòniques, especialment a través de la traducció automàtica. Models avançats com DeepL[38] han demostrat la capacitat d'oferir traduccions més fluides i de major qualitat que els mètodes estadístics anteriors, gràcies a l'ús de xarxes neuronals.[39] Aquesta millora en la traducció automàtica pot trencar barreres lingüístiques, i facilitar la comprensió intercultural i l'accés a informació en diverses llengües.

La traducció automàtica en temps real pot facilitar la comunicació entre parlants de diferents llengües, incloses les no hegemòniques, superant les barreres lingüístiques en diversos entorns, com ara l'educació o la sanitat. Els assistents de veu basats en IA poden permetre als usuaris interactuar amb la tecnologia utilitzant la seva llengua minoritària, fent que els dispositius i les plataformes digitals siguin més accessibles i fàcils d'utilitzar.

En aquest sentit, els assistents de veu comercials, vinculats o no a dispositius específics de maquinari –Alexa d'Amazon, Siri d'Apple, Assistant de Google– incorporen fins ara una varietat molt limitada d'idiomes, però estan evolucionant cap a la incorporació de models LLM en la capa de diàleg amb l'usuari,[40] fet que hauria d'acabar facilitant l'ampliació del ventall de llengües que admeten.

38 DEEPL.COM, «DeepL Translator».

39 IBANEZ, «L'impact de l'intelligence artificielle sur l'avenir de la traduction».

40 PANAY, «Introducing Alexa+, the next generation of Alexa».

Accessibilitat per a persones amb discapacitat

La IA pot jugar un paper crucial en la millora de l'accés a la informació i a la tecnologia per a les persones amb discapacitat que parlen llengües no majoritàries. Per exemple, la conversió de text a veu pot beneficiar les persones amb discapacitat visual, perquè els permet consumir contingut escrit en la seva llengua materna. Aquestes tecnologies poden proporcionar descripcions d'àudio, de text a veu i plataformes interactives per a un aprenentatge inclusiu que maximitzi el potencial de l'estudiant.

Eliminació de barreres geogràfiques

La IA pot connectar parlants de llengües no hegemòniques que es troben dispersos geogràficament, crear comunitats virtuals que abans eren impossibles. Això és especialment valuós per a llengües amb comunitats de parlants reduïdes i distribuïdes en àrees geogràfiques àmplies, o en diferents estats, ja que permet l'intercanvi cultural i lingüístic, que reforça la identitat comunitària.

5.4. Foment de la creativitat i la producció de contingut

La IA generativa obre noves perspectives per al foment de la creativitat i la producció de contingut en llengües no hegemòniques. Té la capacitat de produir grans volums de materials lingüístics, tant escrits com parlats, de manera gairebé instantània. Aquests materials poden ser utilitzats en una àmplia gamma d'aplicacions, des d'esforços educatius i campanyes de revitalització cultural fins a la creació d'arxius digitals dinàmics.

Això inclou la generació de textos, àudios i vídeos diversos que es poden adaptar a diferents estils d'aprenentatge i interessos, que proporcionen una riquesa de material que abans era difícil d'obtenir.[41] La IA pot donar suport a tasques creatives generant múltiples prototips basats en entrades i restriccions específiques, i pot optimitzar dissenys existents a partir de la retroalimentació humana. Aquesta capacitat pot empoderar creadors de contingut, artistes i educadors en llengües minoritàries i minoritzades per produir material atractiu i rellevant a una escala sense precedents.

Creació de contingut multimèdia

Les eines de text a veu (TTS) basades en IA poden crear contingut d'àudio amb un so natural en llengües minoritàries i minoritzades, i fer producció d'audiollibres, podcasts i altres materials d'àudio, la qual cosa és

41 KERNTRAINING, «Chancen und Risiken von KI-generierten Lerninhalten im Fremdsprachenunterricht».

especialment útil per a la creació de materials didàctics amb recursos mínims. A més, la IA pot facilitar la creació de subtítols automàtics per a vídeos en aquestes llengües, de manera que el contingut multimèdia sigui més accessible per a un públic més ampli.

Un exemple interessant de l'ús de la IA en la creació de contingut en una llengua no hegemònica és el projecte *The Christmas Miracle*[42] de la radiotelevisió pública de Suècia, que va utilitzar IA generativa per crear una sèrie de televisió en sami, involucrant nens sami en el procés creatiu. Aquesta iniciativa demostra com la IA pot democratitzar la creació de contingut en llengües no hegemòniques, atès que permet que un nombre més gran de persones produeixin i comparteixin materials en la seva llengua materna.

Eines per a creadors de contingut

La IA esdevé una eina valuosa per ajudar a generar articles, blocs i publicacions a les xarxes socials en llengües no hegemòniques, perquè agilita el procés de producció de contingut i augmenta la presència d'aquestes llengües en línia. Això pot augmentar significativament la presència i la visibilitat d'aquestes llengües en l'esfera digital, tot i que la qualitat i la rellevància d'aquest contingut estan directament relacionades amb la quantitat i la qualitat de les dades d'entrenament que s'utilitzen.

5.5. Oportunitats econòmiques i desenvolupament tecnològic local

La IA té el potencial d'impulsar el desenvolupament de nous productes i serveis, crear nous canals d'ingressos i augmentar la productivitat de les persones treballadores en diversos sectors. Per a les llengües no hegemòniques, això es pot traduir en la creació de nous models de negoci al voltant de la tecnologia lingüística. Un exemple il·lustratiu és el cas de *Sermitsiaq*, el principal editor de premsa de Groenlàndia, que ha monetitzat les dades de la seva hemeroteca històrica per finançar el periodisme independent gràcies a una eina de traducció automàtica d'IA.[43]

El disseny de models de llenguatge grans (LLM) multilingües pot promoure el sector de la IA i el PLN a nivell regional, optimitzar processos empresarials, internacionalitzar productes i millorar els serveis oferts per

42 EUROPEAN BROADCASTING UNION, «The Advent of AI: The Sami kids whose story saved Christmas».

43 CHAUVET, «How using AI translation tools for minority languages can boost subscriptions».

les administracions públiques.[44] Això no només genera oportunitats econòmiques locals, sinó que també contribueix a la sobirania tecnològica lingüística.

Nous models de negoci

La creació d'eines i serveis especialitzats en llengües no hegemòniques pot obrir nous mercats i oportunitats empresarials. Empreses locals poden desenvolupar solucions tecnològiques específiques per a les seves comunitats lingüístiques i crear ecosistemes econòmics que reforcin l'ús i la vitalitat de la llengua.

Innovació en serveis públics

Les administracions públiques poden utilitzar la IA per crear sistemes més inclusius i accessibles per a tota la ciutadania i millorar l'accés als serveis en llengües no majoritàries. Això pot incloure bots de conversa per a consultes administratives, sistemes de traducció per a documents oficials o eines de reconeixement de veu per a la interacció amb serveis públics digitals.

5.6. Preservació del patrimoni cultural i transmissió intergeneracional

La IA pot facilitar la transmissió de coneixements culturals i lingüístics entre generacions, crear ponts entre parlants de diferents edats i diferents nivells de competència lingüística. Mitjançant la creació d'assistents virtuals que reprodueixin les característiques de parlants nadius tradicionals es pot facilitar l'aprenentatge de joves que potser no tenen accés directe a aquestes fonts de coneixement.

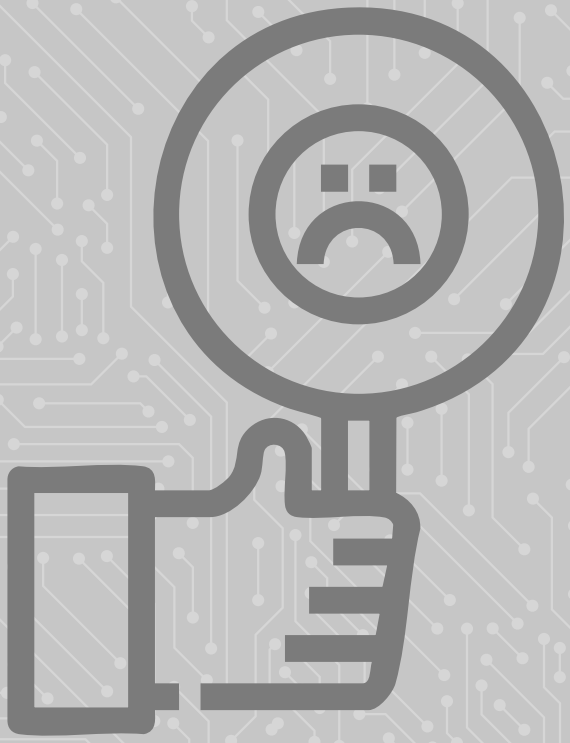
A més, la IA pot crear arxius digitals interactius que vagin més enllà de la simple documentació textual per ajudar a preservar i transmetre aspectes culturals específics associats amb la llengua, com ara expressions idiomàtiques, tradicions orals, cançons tradicionals i narracions culturals.

Els efectes positius de la IA per a les llengües no hegemòniques són principalment habilitadors, ofereixen eines per escalar esforços de preservació i revitalització que abans eren inabastables. No obstant això, la realització d'aquest potencial depèn críticament de la inversió proactiva en la creació de dades i la personalització dels models, ja que la IA no pot «crear» una llengua on no hi ha una base digital.

---

44 PROJECTE AINA, «Promoting the use of Catalan in the digital age».

La capacitat de la IA per amplificar i preservar dades lingüístiques i per produir grans volums de materials lingüístics és innegable. Aquestes capacitats, a més, complementen els esforços humans i redueixen la càrrega de treball per als lingüistes. No obstant això, totes aquestes funcionalitats es basen en la premissa fonamental de la disponibilitat de dades d'entrenament suficients. Si una llengua no posseeix una base de dades digitals sòlida i representativa, la IA no podrà generar contingut de qualitat ni oferir eines personalitzades que siguin realment útils. Per tant, la IA es presenta com una eina poderosa i transformadora si hi ha una inversió prèvia i estratègica en la digitalització i la creació de corpus lingüístics. Aquesta inversió transforma un repte preexistent –la manca de dades i recursos– en una oportunitat tangible –l'escalabilitat de la preservació i revitalització lingüística–, que subratlla la necessitat d'una acció proactiva per part de les comunitats i institucions.



6

EFECTES NEGATIUS I DESAFIAMENTS DE LA IA PER A LES LLENGÜES NO HEGEMÒNIQUES

Si bé la IA ofereix un gran potencial per a les llengües no hegemòniques, també planteja una sèrie de desafiaments significatius i riscos que podrien exacerbar les vulnerabilitats existents. Aquests efectes negatius es deriven principalment de la naturalesa de la IA actual, que depèn en gran mesura de la quantitat i qualitat de les dades i recursos d'entrenament i de les dinàmiques de poder inherents al desenvolupament tecnològic global.

6.1. L'escletxa digital i la desigualtat de dades

El principal obstacle per a la inclusió de les llengües no hegemòniques en l'era de la IA és la profunda escletxa digital existent. La majoria dels models de llenguatge grans (LLM) dominants al mercat tenen un rendiment notablement inferior per a les llengües no angleses i, de manera més acusada, per a les llengües amb pocs recursos. Aquesta deficiència es deu directament a l'escassetat i la mala qualitat de les dades digitals disponibles per a aquestes llengües.

Les dades disponibles per a algunes d'aquestes llengües són, com hem dit anteriorment, limitades a corpus molt específics, com textos religiosos, documents

legals o articles de baix nivell. Per exemple, els corpus de l'irlandès, el maltès, les llengües bàltiques, l'eslovè, l'eslovac, el croat i l'hongarès estan força esbiaixats cap al contingut jurídic, com a fruit del seu caràcter oficial dins de la UE que obliga a traduir les normatives comunitàries. A més, cal remarcar que alguns d'aquests recursos no tenen per què estar disponibles recollint l'ús quotidià i normal de la llengua, sinó que poden ser fruit de traduccions automàtiques o altres situacions.

Aquesta disparitat en la disponibilitat de dades i recursos crea una «monocultura tecnològica» que pot excloure comunitats senceres de l'accés a serveis en línia essencials com l'educació, l'ocupació o la informació sanitària. La manca de tecnologia lingüística funcional en la seva llengua materna priva aquestes comunitats dels beneficis que altres obtenen de la IA, cosa que amplia les desigualtats econòmiques i socials.

6.1.1. Homogeneïtzació lingüística

Un dels principals riscos associats al desenvolupament de la IA és la possible homogeneïtzació lingüística. La majoria dels sistemes d'IA actuals, especialment els grans models lingüístics (LLM), es basen en models que s'entrenen principalment amb conjunts massius de dades en anglès i, en menor mesura, en unes poques llengües dominants. Aquesta circumstància amenaça d'excloure els parlants de la resta de les més de 7.000 llengües que es parlen al món. La preponderància de l'anglès en les dades d'entrenament i en el disseny dels models d'IA és tan profunda que, fins i tot, alguns models d'IA generativa que estan dissenyats per respondre a indicacions en altres llengües sembla que «pensen» en anglès, la qual cosa subratlla la influència dominant d'aquesta llengua en el desenvolupament de la IA.

La dependència creixent de la IA en aplicacions quotidianes com la traducció automàtica, els assistents virtuals i la creació de contingut podria consolidar encara més la centralitat de l'anglès com a llengua dominant en l'esfera digital. Alguns experts adverteixen que la IA podria accelerar el procés d'homogeneïtzació lingüística que va començar amb l'educació colonial, i marginar i, eventualment, fer que les llengües no hegemòniques esdevinguin no només menys utilitzades sinó obsoletes.

6.2. Biaixos algorítmics i propagació d'estereotips

Els sistemes d'IA aprenen dels patrons presents en les dades amb les quals són entrenats. Conseqüentment, si aquestes dades històriques contenen biaixos socials o culturals, els models d'IA corren el risc de

perpetuar i fins i tot reforçar les desigualtats existents.[45] Els conjunts de dades no curats –desestructurats, sense supervisar–, especialment aquells extrets directament d'internet, sovint codifiquen i objectifiquen la visió dominant o hegemònica, i reflecteixen els biaixos de les poblacions que més contribueixen al contingut digital –per exemple, homes joves i acadèmics en el cas de la Viquipèdia.[46]

Això pot portar a la manifestació de biaixos de gènere, ètnics o religiosos. Un exemple documentat és l'observació que la paraula italiana «musulmana» –femení de musulmà– va ser valorada més negativament que «musulmano» en un model d'IA,[47] cosa que suggereix un biaix arrelat tant en la identitat de gènere com en l'ètnica o religiosa. Aquests biaixos tenen implicacions serioses quan els sistemes d'IA s'utilitzen en contextos sensibles o en llengües minoritzades i minoritàries, en què la discriminació pot ser encara més perjudicial.

6.2.1. Manifestacions específiques del biaix en llengües no hegemòniques

En el context de les llengües no hegemòniques, els biaixos algorítmics es poden manifestar de diverses formes:

- Rendiment inferior per a accents o dialectes no estàndard: els sistemes de reconeixement de veu poden prioritzar les varietats «estàndard» en l'entrenament dels models d'IA i tenir un rendiment inferior per a parlants amb accents regionals o dialectes no estàndard de la llengua no hegemònica. Això pot conduir a una discriminació envers els parlants d'altres varietats lingüístiques.
- Generació de text culturalment inadequada: la generació de text pot no reflectir adequadament les normes culturals o els usos lingüístics de la comunitat parlant, i perpetuar estereotips o associacions negatives. Per exemple, s'ha demostrat que els models lingüístics grans poden associar atributs negatius i treballs menys prestigiosos amb els parlants de dialectes minoritaris de l'anglès.
- Errors en la detecció i la interpretació: fins i tot els detectors d'IA utilitzats per a detectar el plagi acadèmic poden assenyalar erròniament l'escriptura de parlants no nadius d'anglès com a generada per IA, cosa que podria tenir conseqüències negatives per a aquests estudiants. A més, els sistemes d'anàlisi de sentiments poden marcar erròniament textos en variants lingüístiques minoritàries amb puntuacions sentimentals més negatives.

---

45 METTA-WINDISCHER, «Reframing minority rights amid global challenges: The role of AI and algorithmic fairness in promoting diversity and inclusion».

46 SCHNEIDER, «Multilingualism and AI: The regimentation of language in the age of digital capitalism».

47 EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHT, «Bias in algorithms. Artificial Intelligence and discrimination».

6.3. Manca de transparència: l'efecte «caixa negra» dels models

Un altre desafiament significatiu dels sistemes d'IA actuals és la seva manca de transparència. Molts models, especialment els més complexos com els LLM, funcionen com a «caixes negres». Això significa que la lògica interna darrere dels seus processos de presa de decisions o de generació de contingut és opaca i extremadament difícil d'escodrinyar o comprendre per part dels humans.

Aquesta manca de transparència té implicacions directes en la capacitat de detectar i corregir biaixos. Si no es pot entendre com un model arriba a una determinada conclusió o genera un text esbiaixat, resulta gairebé impossible identificar la font del problema i implementar mesures correctores efectives. Aquesta opacitat limita la rendició de comptes i la possibilitat de garantir que els sistemes d'IA siguin justos i equitatius, especialment en contextos multilingües on els biaixos poden ser més subtils i complexos.

6.3.1. El cas de Meta AI amb el català

Un cas recent de manca de transparència s'ha produït durant el mes de juny de 2025 amb Meta AI, la IA de Meta Platforms. Gràcies a la inclusió del seu bot de conversa en plataformes de l'empresa com Instagram i, sobretot, WhatsApp, nombrosos usuaris catalanoparlants han provat de dialogar-hi en català, i poques dècimes de segon després de respondre correctament redactada en català, el mateix bot ha rectificat substituint la resposta original per un altre missatge on diu que «encara no sé català, però t'avisaré quan l'aprengui». Altres llengües amb què s'ha observat el mateix fenomen són el basc, el gallec i el grec. En canvi, altres usuaris han pogut dialogar-hi prou bé en asturià i en aragonès.

Davant les nombroses reclamacions dels usuaris catalanoparlants, Meta Platforms ha comunicat a Accent Obert[48] que Meta AI només inclou tretze idiomes, el menor dels quals, el tailandès, té 70 milions de parlants, i que només responen pel comportament en aquests tretze idiomes. L'empresa també assegura que treballen en ampliar l'abast lingüístic de Meta AI, però sense comprometre's en cap termini.

Després de diverses proves, Accent Obert ha comprovat que el comportament de Meta AI es repeteix quan l'usuari intenta dialogar-hi en català des de Romania, Polònia i els Països Baixos, mentre que respon quan s'hi dialoga des dels EUA. Això permet inferir que en realitat Meta AI sí que ha après català i altres

48 CUESTA, «Bon dia. Tinc resposta de @Meta sobre el català. No us agradarà: Ara com ara #MetaAI només contempla 13 idiomes, el menor dels quals —tailandès— té 70 milions de parlants. Diu que treballen en ampliar però sense comprometre cap data».

llengües –sigui per entrenament específic o bé per aprenentatge creuat–, però considera que no les sap prou bé per donar-hi suport explícit. El fet que es comporti diferent als EUA i a la UE també permet deduir que la plataforma no es desplega de manera uniforme a tot el món, sinó que té variants diferents a cada continent,[49] probablement en compliment de la regulació europea, més estricta, i això impedeix fer servir plenament les seves capacitats lingüístiques, per imperfectes que siguin, en algunes regions i no en altres.

6.3.2. La confusió sobre les capacitats lingüístiques dels bots de conversa d'IA

La incidència de Meta AI amb el català ha portat a revisar[50] les capacitats lingüístiques d'altres bots de conversa. A finals de juny de 2025, els onze provats –ChatGPT (Open AI), Claude (Anthropic), Copilot (Microsoft), DeepSeek, Gemini (Google), Grok (X AI), LeChat (Mistral AI), Lumo (Proton), Perplexity, Meta AI i Qwen (Alibaba)– són capaços de dialogar en català amb l'usuari per escrit, si bé només tres[51] –ChatGPT, Copilot i Qwen– inclouen explícitament aquesta llengua entre les que suporten formalment i inclouen en la interfície d'usuari. Els altres vuit entenen i responen en català, tot i que alguns canvien espontàniament d'idioma. Això fa que molts usuaris renunciïn a interactuar-hi en català, precisament quan continuar fent-ho ajudaria a entrenar els models i demostraria una demanda de compatibilitat amb l'idioma. D'aquesta realitat se'n deriva la conveniència d'educar els usuaris perquè interpel·lin els bots de conversa en el seu idioma i valorin si les respostes són pertinents.

6.4. Risc de desinformació i de manipulació lingüística

La capacitat de la IA generativa per produir contingut de manera autònoma i a gran escala comporta un risc inherent de generar informació inexacta o enganyosa, un fenomen conegut com a «al·lucinacions».[52] Aquests errors no són meres imprecisions, sinó que poden ser presentats amb una fluïdesa i coherència que els fan semblar plausibles, de manera que enganyen els usuaris.

---

49 CUESTA, «La meva interpretació: si en algun moment heu dialogat en català (o gallec, basc, grec) ha sigut per un afortunat accident, causat per les diferències entre la MetaAI original (EUA) i la variant europea. Poseu-hi paciència o activeu una altra IA al vostre WhatsApp».

50 CUESTA, «La IA de WhatsApp es nega a respondre en català (tot i saber-ne)».

51 Amb aquest estudi ja tancat, Google ha anunciat a finals de novembre que el seu bot Gemini 3 Pro afegirà 30 idiomes europeus compatibles, entre ells el català, el gallec, el basc, el croat, el serbi, l'eslovac i el lituà, entre una majoria d'idiomes del subcontinent indi i del sud-est asiàtic.

52 «Hallucination (artificial intelligence)».

Aquesta capacitat pot ser explotada per a finalitats malicioses, com la ciberdelinqüència, la creació de notícies falses –*fake news*– o la producció de vídeos o àudios manipulats –*deepfakes*– que poden ser utilitzats per enganyar o manipular persones a gran escala. El poder retòric de la IA pot fer creure als usuaris que el sistema «entén» el llenguatge com els humans, tot i que en realitat estigui generant multitud de falsedats o informació sense fonament. Aquest risc és particularment preocupant en llengües no hegemòniques, en què la verificació de la informació pot ser més difícil a causa de la manca de recursos o de la menor visibilitat en línia.

Impacte en la salut i en la seguretat

Els errors de traducció o interpretació en llengües no hegemòniques poden tenir conseqüències greus en escenaris de «vida o mort», especialment en àmbits com la medicina. Errors de traducció en aquesta tipologia de llengües han causat malentesos[53] sobre dosis de medicaments, procediments mèdics o instruccions de seguretat, que arriben a posar en risc la vida de les persones.

6.5. Impacte en l'ocupació i en la dependència tecnològica

La introducció massiva de la IA en diversos sectors pot portar a la substitució de tasques que tradicionalment realitzaven humans, amb el potencial de provocar una substitució massiva de llocs de treball humans. En el context de les llengües minoritzades i minoritàries, això podria significar una major dependència de bots de conversa o d'interfícies en llengües hegemòniques en l'àmbit professional i econòmic. Per exemple, si una empresa islandesa no pot utilitzar GPT-4 en islandès per als seus bots de conversa, haurà de recórrer a l'anglès, i marginarà així la llengua pròpia en un context crucial de servei al client.

Aquesta situació podria ampliar la desigualtat econòmica existent, ja que els treballadors fluids en llengües hegemòniques estarien en una posició avantatjosa per progressar en un mercat laboral cada cop més dominat per la IA, mentre que d'altres s'enfrontarien a barreres tecnològiques i lingüístiques per a l'ocupació si les eines d'IA no són accessibles en la seva llengua.

53 MORNINGSIDE, «The Real Cost of Errors in Medical Translations».

6.6. Erosió cultural i homogeneïtzació lingüística

El domini aclaparador de les llengües hegemòniques, especialment l'anglès, en el desenvolupament de la IA condueix a una preocupant tendència cap a l'homogeneïtzació lingüística. Aquesta situació pot produir la pèrdua de coneixements, de visions del món i de marcs cognitius únics que estan encapsulats en les llengües maternes. Els processos de segmentació [*tokenització*],[54] les codificacions semàntiques i els models de llenguatge que impulsen els sistemes d'IA moderns afavoreixen intrínsecament les llengües amb grans corpus escrits, i en perpetuen així la dominació.

Sense una intervenció deliberada i estratègica, la IA pot crear una «monocultura tecnològica» que no només reflecteixi sinó que magnifiqui els desequilibris de poder lingüístic i cultural existents, en lloc de reduir-los. Això envia un missatge subtil però poderós que algunes llengües –i, per extensió, els seus parlants– importen menys en el futur col·lectiu, fet que amenaça el ric patrimoni lingüístic de la humanitat.

Pèrdua de diversitat cognitiva

La dominació de poques llengües en l'àmbit digital pot portar a la pèrdua de formes úniques de pensament i conceptualització del món que són inherents a cada llengua. Aquesta «racisme lingüístic» o «imperialisme cultural» pot alienar les comunitats no occidentals i limitar l'accessibilitat de les llengües pròpies, perpetuant desigualtats històriques en nous contexts tecnològics.

6.7. Manca de dades i qualitat insuficient

Un repte important per a la IA i les llengües minoritàries i minoritzades és la manca de dades suficients i de qualitat per entrenar models d'IA efectius. A diferència de les llengües majoritàries, que tenen grans corpus de text i veu disponibles, moltes llengües amb un menor nombre de parlants no disposen de les dades necessàries per entrenar models d'aprenentatge automàtic robustos. Aquesta escassetat de dades pot provocar imprecisions significatives en la documentació de la llengua i en el rendiment de les aplicacions d'IA dissenyades per a ella.

La manca de corpus paral·lels –textos originals i les seves traduccions– i de recursos lingüístics ben establerts dificulta especialment el desenvolupament de sistemes de traducció automàtica d'alta qualitat

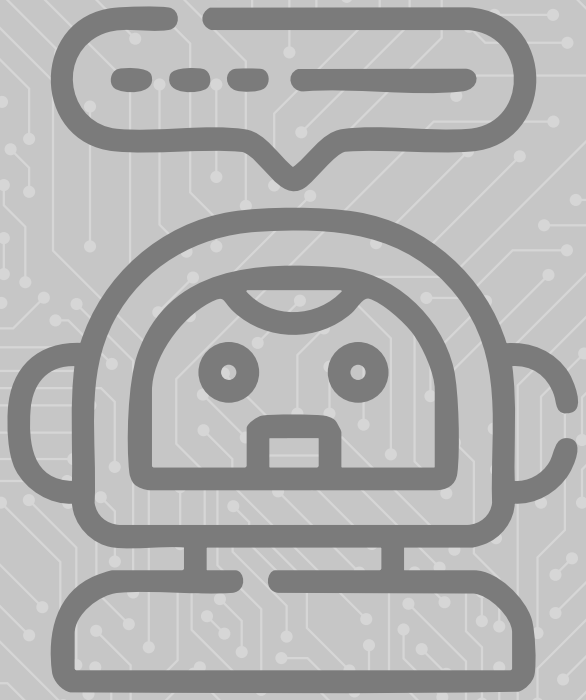
54 Procés de dividir text en unitats més petites –segments o *tokens*– que els models d'IA poden processar, com ara paraules, subparaules o caràcters.

per a llengües amb pocs recursos digitals. La recopilació de dades per a aquestes llengües pot ser un procés complex i costós, ja que sovint són parlades per un nombre reduït de persones o en comunitats geogràficament disperses. A més, la qualitat de les dades disponibles pot ser insuficient, ja que poden ser fragmentades, inconsistents o fins i tot contenir errors de traducció automàtica.

Sense dades adequades, els models d'IA no poden aprendre les subtileses, les complexitats gramaticals i els matisos culturals d'aquestes llengües, cosa que limita la seva utilitat i pot fins i tot generar respostes incorrectes o «al·lucinacions». Aquesta problemàtica no és només tècnica, sinó que també està arrelada en problemes socials com ara pràctiques de recerca d'IA no diverses, excloents i fins i tot explotadores.

Els efectes negatius de la IA per a les llengües no hegemòniques no són meres «limitacions» tècniques; són «riscos sistèmics» que amenacen la diversitat lingüística i cultural en la seva essència. La perpetuació de biaixos i la creació de desinformació en llengües no hegemòniques poden tenir conseqüències, fins i tot greus en escenaris d'emergència com la mèdica, i erosionar la confiança en la tecnologia, creant una nova forma de discriminació digital.

Aquesta interconnexió de problemes es pot desglossar de la manera següent: la manca de dades és la causa arrel que condueix a la manifestació de biaixos i la generació d'informació inexacta o enganyosa per part dels models d'IA. Aquesta inexactitud no és trivial. A més, la «manca de transparència» dels models, que operen com a «caixes negres», dificulta enormement la detecció i la correcció d'aquests problemes, i limita la rendició de comptes. Simultàniament, el domini de les llengües majoritàries en el desenvolupament de la IA pot provocar una «erosió cultural» i la «pèrdua de visions del món» úniques, ja que les llengües no hegemòniques queden marginades en l'àmbit digital. Aquesta concatenació de problemes revela que la IA no és simplement una eina neutral, sinó que actua com un «mirall infinit» que «s'entesta a mostrar-nos el que ja coneix», i reforça les desigualtats existents si no es dissenya i s'implementa amb una consciència ètica i cultural profunda.



7

CASOS I INICIATIVES D'ÈXIT EN LLENGÜES EUROPEES NO HEGEMÒNIQUES

L'anàlisi dels efectes de la IA en les llengües no hegemòniques no estaria completa sense examinar casos concrets on s'han implementat estratègies i projectes per abordar els reptes i aprofitar les oportunitats. Aquests exemples il·lustren la viabilitat de la preservació i revitalització lingüística en l'era digital.

7.1. El cas del català: Projecte AINA

El Projecte AINA representa un model integral i proactiu per a la supervivència digital del català, una llengua minoritzada amb una comunitat de parlants significativa però que encara es considera no majoritària en el context global de la IA.

Objectius i infraestructura

El Projecte AINA, impulsat per la Generalitat de Catalunya des del 2020 a través de la Secretaria de Polítiques Digitals a partir d'una iniciativa de Softcatalà i desenvolupat pel Barcelona Supercomputing Center (BSC-CNS), té com a objectiu fonamental generar una infraestructura digital robusta per al català. Aquesta infraestructura es basa en la potència de supercomputació del MareNostrum

5, que permet el processament de dades massives i l'execució de models lingüístics avançats i pioners en el sector. L'objectiu és entrenar models que no només en millorin el rendiment, sinó que també avancin en termes de seguretat i eficiència de resposta. A més, AINA treballa per implementar i incloure mòduls i biblioteques catalanes en entorns i plataformes de referència, garantint així una cobertura adequada de la llengua en l'ecosistema digital global. El nom *Aina* és un homenatge a la filòloga Aina Moll Marquès –1930-2019–, primera directora general de Política Lingüística de la Generalitat de Catalunya restaurada, subratllant la importància històrica i cultural del projecte.

Recursos i col·laboracions

El projecte AINA opera mitjançant la recollida i curació de dades de text i veu, que rep gràcies a la col·laboració activa d'usuaris i l'aprofitament d'altres recursos disponibles. Col·labora estretament amb entitats de la comunitat lingüística i altres sectors per tal de recopilar i processar grans volums de dades.

Imatge 2. Projecte Aina, el corpus de Common Voice: optimisme i reptes pendents

Font: Projecte Aina.

Un èxit notable del projecte és la consolidació del català com la llengua amb més hores enregistrades i validades[55] a la plataforma Common Voice[56] de Mozilla. Aquesta iniciativa de dades obertes és particularment rellevant, ja que aquests repositoris han inspirat models lingüístics de grans empreses com el PaLM2 de Google.[57] Aquesta estratègia de dades obertes és crucial per superar la «manca de dades» que afecta moltes llengües no hegemòniques.

Desenvolupament de models i recursos

El projecte ha desenvolupat diversos conjunts de dades per a l'afinació, la instrucció i l'avaluació de models de text, incloent-hi corpus paral·lels amb gallec, italià, francès, portuguès i xinès. AINA ha desenvolupat models de generació de text –com FLOR-6.3B, FLOR-1.3B, Aguila-7B–, models de reconeixement automàtic de la parla (ASR) i models de text a veu (TTS) per al català. El primer recurs generat va ser el corpus català més gran creat fins ara, amb 1.770 milions de metadades associades a paraules.[58]

Iniciatives principals

El Projecte AINA ha llançat diverses iniciatives per fomentar el desenvolupament de la IA en català:

- L'«Aina Challenge» és un concurs dotat amb un milió d'euros que finançarà fins a 22 projectes d'IA i tecnologies del llenguatge en català. Aquesta iniciativa busca promoure l'adopció de solucions innovadores i impulsar la competitivitat del teixit productiu català, especialment empreses emergents i pimes.[59] Els reptes proposats inclouen el desenvolupament de serveis i aplicacions d'IA/TL en català, la creació de sistemes de control i monitorització d'aquests models, i el desenvolupament de recursos oberts per enriquir l'ecosistema d'AINA.[60]
- AINA participa activament en esdeveniments científics de rellevància internacional i nacional, com el Deep Learning Barcelona Symposium i la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN), en què les seves tecnologies han estat reconegudes i presentades.[61]

---

55 PROJECTE AINA, «El corpus de Common Voice; optimisme i reptes pendents».

56 COMMON VOICE DE MOZILLA, «Tecnologia que parla la vostra».

57 GHAHRAMANI, «Presentamos PaLM 2».

58 BSC, «Sobre Aina».

59 PROJECTE AINA, «The Aina Challenge, the competition endowed by the Catalan Government with €1M to finance up to 22 AI and language technology projects in Catalan, has started».

60 PROJECTE AINA, «Success in the Aina Challenge call to accelerate AI projects in Catalan».

61 PROJECTE AINA, «Aina participates in the Deep Learning Barcelona Symposium 2024».

- El projecte també impulsa hackatons, com l'Aina Hack, per resoldre reptes concrets de l'Administració catalana utilitzant els recursos d'IA generats pel Projecte AINA.[62] Aquestes competicions fomenten la innovació i l'aplicació pràctica de les tecnologies desenvolupades.
- AINA Lab és una iniciativa per dinamitzar una comunitat de desenvolupadors i crear una xarxa activa de professionals que treballin per a la innovació tecnològica en català.
- Col·laboració amb altres projectes i iniciatives: AINA col·labora amb altres projectes com Common Voice de Mozilla per a la recopilació de dades de veu en català. També s'ha establert una col·laboració amb OpenAI per millorar el suport del català en les seves tecnologies d'IA.

Imatge 3. Concurs del Projecte Aina per promocionar iniciatives de foment de la llengua

Font: Projecte Aina.

Casos d'ús i aplicacions

El Projecte AINA facilita el desenvolupament de diversos casos d'ús per al català en l'àmbit de la IA, incloent-hi:

- Assistents de veu: desenvolupament d'assistents virtuals que entenen i responen en català.
- Traductors automàtics: millora de la traducció entre el català i altres llengües.

62 PROJECTE AINA, «Crea eines d'IA per a una administració més àgil, propera i en català».

- Correctors ortogràfics i gramaticals: creació d'eines per millorar la qualitat de l'escriptura en català.
- Motors de cerca: optimització de la cerca d'informació en català a internet.
- Bots de conversa: desenvolupament d'agents conversacionals per a diversos serveis i aplicacions.
- Generació de text: creació automàtica de contingut escrit en català.
- Anàlisi de sentiments: avaluació del to emocional de textos en català.
- Subtitulació automàtica: generació de subtítols per a vídeos en català.

Publicacions científiques

La recerca i els avenços del Projecte AINA es documenten a través de publicacions acadèmiques. S'ha publicat un *Salamandra Technical Report*[63] i s'han detallat els avenços en models de llenguatge com Salamandra 7B, entrenat en 35 llengües europees i codi.[64] A més, s'ha demostrat l'ús dels models AINA en sistemes de generació augmentada per recuperació (RAG),[65] com es mostra en un *notebook* de demostració per a la creació d'un sistema RAG simple en català.

El Projecte AINA il·lustra un model integral i proactiu per a la supervivència digital d'una llengua no majoritària i minoritzada. La combinació d'inversió pública substancial, infraestructura de supercomputació, una estratègia de dades obertes i col·laboració comunitària i la creació de models propis –com Salamandra–, permet al català no dependre exclusivament dels grans actors globals i establir la seva pròpia sobirania tecnològica lingüística.

Aquest model es basa en una estratègia multifacètica que aborda directament els problemes de la manca de dades i recursos per a llengües minoritàries i minoritzades. En primer lloc, el projecte compta amb un finançament públic significatiu –com el milió d'euros per a l'Aina Challenge–[66] que és indispensable per a la recerca i el desenvolupament en aquest camp. En segon lloc, s'aprofita una «infraestructura de supercomputació» d'avantguarda com el MareNostrum 5,[67] que proporciona la capacitat de processament necessària per entrenar models de llenguatge grans. En tercer lloc, es promou una «recollida activa de dades» a través

63 GONZÁLEZ-AGIRRE, «Salamandra Technical Report».

64 HUGGING FACE, «Salamandra Vision Model Card».

65 Procés de millora dels resultats d'un LLM ampliant-los amb informació de fonts autoritzades que no es trobaven entre les dades originals d'entrenament del model.

66 PROJECTE AINA, «Aina Challenge».

67 BSC, «MareNostrum 5».

de plataformes com Common Voice, en què el català ha assolit una posició de lideratge.[68] Aquesta iniciativa és crucial per generar els corpus i recursos necessaris. En quart lloc, el projecte es dedica al «desenvolupament de models propis i oberts», com la família Salamandra,[69] que assegura la disponibilitat de models adaptats a les especificitats del català. Finalment, es fomenta activament un «ecosistema local» d'innovació mitjançant hackatons i reptes per a pimes.

Imatge 4. Família de models Salamandra

Font: Projecte Aina.

Aquesta estratègia holística és una resposta directa als desafiaments identificats en la secció 6, i demostren que la sobirania lingüística digital és assolible amb planificació i inversió estratègiques, i que la col·laboració entre el sector públic, la recerca i la comunitat és la clau per al progrés.

68 METADATA, «El català, la llengua líder de Common Voice».

69 PROJECTE AINA, «Coneix la família de models Salamandra».

7.2. L'experiència islandesa amb OpenAI i ChatGPT

El cas de l'islandès ofereix una perspectiva diferent i complementària a la del català, ja que se centra en la col·laboració directa amb una de les grans empreses tecnològiques globals, OpenAI, per a la preservació de la seva llengua.

Context i motivació

L'islandès, parlat per aproximadament 330.000 persones, s'enfronta a la preocupació que, davant la ràpida digitalització i la integració amb l'anglès i altres llengües europees, pugui patir una extinció de facto en poques generacions si no es manté com a llengua per defecte en l'àmbit digital. La nació islandesa valora profundament la seva llengua com a part de la seva rica herència cultural i identitat, fins al punt que el seu Departament de Planificació Lingüística encunya termes islandesos per a noves idees en lloc d'adoptar manlleus, com és el cas de «*tölva*» –profetessa de nombres– per referir-se a «ordinador».

Col·laboració estratègica

Islàndia va establir una col·laboració estratègica amb OpenAI per utilitzar el seu model GPT-4 en l'esforç de preservació de l'islandès.[70] Aquesta iniciativa va ser concebuda per «*convertir una posició defensiva en una oportunitat per innovar*».[71] La col·laboració va sorgir d'una visita d'una delegació islandesa, que incloïa el president i ministres del Govern, a la seu d'OpenAI, on van expressar el seu interès en la inclusió de la seva llengua.[72] Un dels objectius clau d'aquesta associació era començar a construir un recurs que pogués servir per a promoure la preservació d'altres llengües amb pocs recursos i assegurar la representació de totes les llengües i cultures en les tecnologies digitals.

Millora de models

Inicialment, els models GPT d'OpenAI, entrenats majoritàriament en anglès i altres llengües hegemòniques, no tenien les mateixes capacitats ni comprensió en llengües més petites com l'islandès. Tot i que GPT-4 va mostrar millores significatives en la seva capacitat de resposta en islandès en comparació amb GPT-3.5, encara produïa text amb errors gramaticals, 'translationese' ['traductonès', 'traduccionisme', traduccions que

70 OPEN AI, «Gobierno de Islandia».

71 TÓMAS, «GPT-4 to Aid in the Preservation of the Icelandic Language».

72 OFFICE OF THE PRESIDENT OF ICELAND, «My team and I were intrigued by our conversation with @SamA at @OpenAI this week. Humanity is on the precipice of great change with the proliferation of #AI. This opens up immense possibilities but also ethical questions. Our aim must be to ensure equal opportunities and access».

sonen artificials o massa literals] i «coneixements culturals incorrectes». Per exemple, mentre que GPT-3 inicialment traduïa «Donald Duck» com «Donald el Tonto» i ChatGPT com «Donald Duck té el mateix nom en islandès que en anglès», GPT-4, després de l'entrenament, va ser capaç de traduir-lo correctament com «Andrés Önd». A més, s'ha observat que GPT-4 pot donar respostes diferents a la mateixa pregunta depenent de si es formula en islandès o en anglès, cosa que demostra la necessitat d'una millora en la comprensió contextual específica de la llengua.

Per abordar aquestes deficiències, Vilhjálmur Þorsteinsson, CEO de Miðeind –una empresa islandesa de tecnologia del llenguatge–, va reunir un equip de 40 voluntaris. Aquests voluntaris van utilitzar el procés d'aprenentatge per reforçament a partir del Feedback Humà (RLHF) per refinar el model, proporcionant *prompts* [indicacions] a GPT-4 i seleccionant i editant les millors respostes generades. Aquest procés requereix un nombre relativament petit d'exemples –uns 100– per produir resultats significatius, la qual cosa el fa particularment factible per a llengües amb pocs recursos en què la creació de grans corpus de dades és un desafiament.

L'enfocament Reinforcement Learning from Human Feedback [Aprenentatge per reforç a partir de la retroalimentació humana] (RLHF)

En aquest procés, els verificadors humans proporcionen a GPT-4 una indicació i el model genera quatre possibles respostes. Els verificadors trien la millor resposta de les quatre i la modifiquen per crear una resposta ideal. Aquesta metodologia permet obtenir resultats amb només 100 exemples, la qual cosa la fa molt més factible per a llengües amb dades limitades en comparació amb els grans conjunts de dades tradicionalment necessaris per entrenar models des de zero. Aquesta aproximació permet a l'equip islandès aprofitar les capacitats generals dels grans models d'OpenAI i aplicar-les a la seva llengua, així s'habiliten funcionalitats que abans requerien una gran quantitat de treball manual i preparació de dades per a cada cas d'ús.

Impacte i model per a altres llengües

Aquesta col·laboració és vista com un model prometedor per a la preservació d'altres llengües amb pocs recursos. Demostra que les petites nacions, si han realitzat una «feina» prèvia en la creació d'infraestructura lingüística –com el programa de tecnologia del llenguatge islandès que va precedir aquesta col·laboració–, poden aprofitar la IA de grans empreses. L'objectiu final és permetre que tot el país interactuï amb els models d'OpenAI en la seva pròpia llengua, i es redueix la dependència de l'anglès en aplicacions interactives i bots de conversa empresarials. Això permetrà a empreses islandeses desplegar bots en islandès en lloc de dependre de solucions en anglès.

L'experiència islandesa subratlla que la col·laboració directa amb grans empreses tecnològiques i la integració de la *human-in-the-loop* [supervisió humana] són estratègies clau per millorar el rendiment de la IA en llengües no hegemòniques, fins i tot amb dades limitades. Aquesta aproximació complementa el desenvolupament de models propis i ofereix una via per a la inclusió de llengües amb pocs recursos en plataformes globals dominants.

Tal com hem vist amb aquest exemple, la millora significativa de GPT-4 per a l'islandès es va aconseguir mitjançant una «col·laboració amb OpenAI», una gran empresa tecnològica. Aquesta millora no es va basar exclusivament en la disponibilitat de grans quantitats de dades, sinó en el «refinament amb retroacció humana (RLHF)». Això demostra que la «qualitat» de la intervenció humana, fins i tot amb un nombre limitat d'exemples, pot compensar parcialment la «quantitat» de dades, un punt crucial per a les llengües amb pocs recursos. El fet que OpenAI mostrés interès a «mostrar que el món no és només anglès» suggereix una finestra d'oportunitat per a altres llengües no hegemòniques si s'aproximen proactivament a les grans empreses tecnològiques amb propostes de col·laboració i recursos preexistents. Aquest cas posa de manifest la importància de la diplomàcia lingüística i la capacitat de les nacions per capitalitzar els seus esforços previs en tecnologia del llenguatge per obtenir el suport de grans actors globals.

7.3. Altres casos d'èxit a Europa

Més enllà del català i l'islandès, diverses iniciatives a Europa demostren el potencial de la IA per a la preservació i revitalització d'altres llengües, cadascuna amb un enfocament adaptat a les seves circumstàncies particulars.

7.3.1. Groenlandès

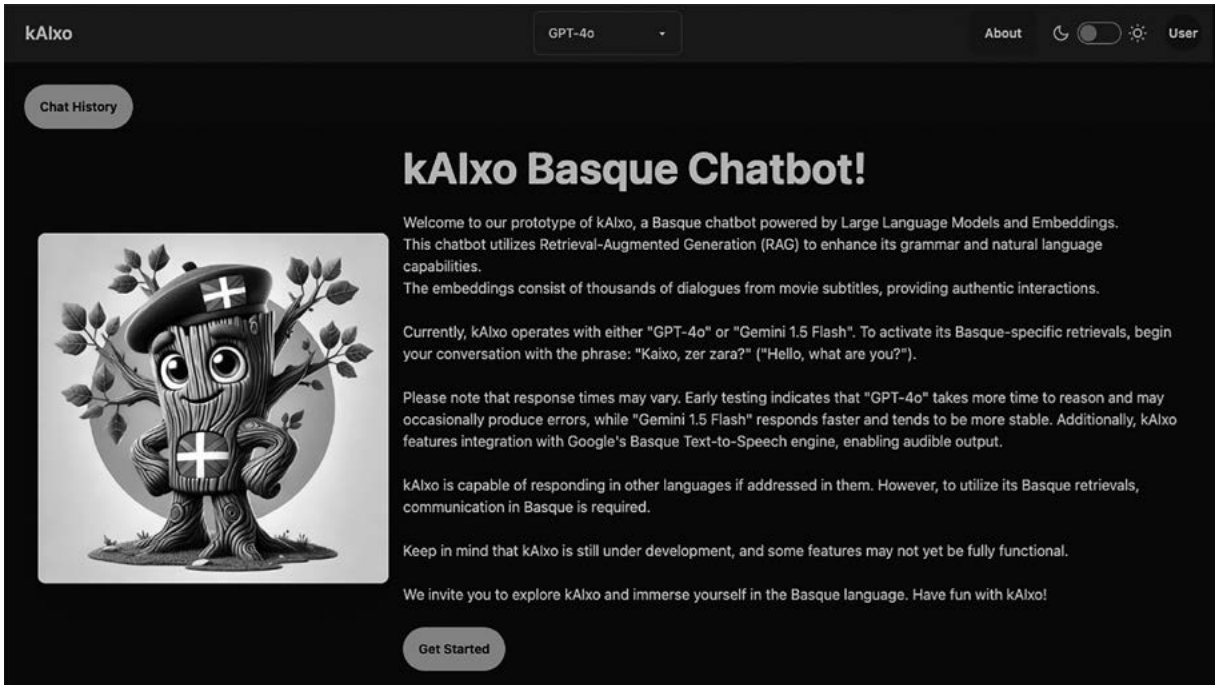
El groenlandès, és un exemple destacat de com la IA pot beneficiar una llengua amb pocs recursos. Sermitsiaq, el principal editor de notícies de Groenlàndia, que per llei ha de publicar contingut tant en danès com en groenlandès, s'enfrontava a retards i costos elevats en les traduccions humanes. El 2023, l'empresa emergent danesa MediaCatch va desenvolupar una eina de traducció d'IA per a Sermitsiaq, entrenada amb més de deu anys de contingut de diaris traduït per professionals humans. Aquesta eina va permetre reduir el temps de traducció de notícies d'hores a minuts, amb una precisió del 80% enfront del 20% de ChatGPT. A més de la millora operativa, Sermitsiaq va desenvolupar una estratègia de subscripció que incloïa l'accés a aquesta eina de traducció, cosa que va permetre al mitjà monetitzar les seves pròpies dades històriques i finançar el periodisme independent. Aquest cas no només va ajudar a la societat groenlandesa, sinó que també va crear un model de negoci sostenible per a un mitjà en un país petit.

7.3.2. Basc

El basc, també ha vist el sorgiment de projectes innovadors basats en IA:

- kAIxo bot de conversa: el prototip de bot de conversa kAIxo[73] promou la llengua basca utilitzant models de llenguatge grans (LLM) i la generació augmentada per recuperació (RAG). Aquest bot està entrenat amb milers de diàlegs extrets de subtítols de pel·lícules, cosa que facilita converses autèntiques i naturals. L'objectiu és preservar la llengua basca fomentant interaccions agradables i significatives, despertant l'interès per aprendre o practicar aquesta llengua.

Imatge 5. kAIxo chatbot



Font: kAIxo.

73 EUSKADI. EUS, «kAIxo Basque Chatbot!».

- Projecte Euskorpus:[74] impulsat pel Govern basc, té com a objectiu la creació d'un corpus digital fonamental per al basc. Aquesta iniciativa, fruit de la col·laboració publicoprivada, busca garantir la presència del basc en el mercat digital en condicions similars a altres llengües. Es preveu la compilació de corpus lingüístics, el desenvolupament de models de codi obert i la creació d'infraestructures per a l'emmagatzematge segur i la validació de dades.

- Itzuli:[75] el projecte Itzuli, implementat pel Govern basc, utilitza la IA per permetre la traducció entre el basc, el castellà, el francès i l'anglès. Aquest sistema representa un pas endavant crucial per a la visibilitat i la funcionalitat del basc en l'entorn digital.

Imatge 6. Itzuli, traductor del basc

Font: Govern basc.

74 EUSKADI. EUS, «Euskorpus, proyecto impulsado por el Gobierno Vasco, llevará el euskera a la revolución de la IA, garantizando de esta manera el futuro de la lengua vasca en una sociedad cada vez más digital»

75 EUSKADI. EUS, «Traductor neuronal».

- Altres projectes: iniciatives com Orai NLP Teknologiak i Gaitu.eus se centren en solucions de processament del llenguatge natural (PNL) i la recollida de dades per al basc, incloent-hi sistemes de dictat adaptats a veus infantils (AskHezi). Aquestes iniciatives demostren un enfocament integral per digitalitzar, investigar i aplicar la IA al basc.

- Llama-eus-8B: s'han desenvolupat models específics per a llengües concretes, com Llama-eus-8B per a la llengua basca, que demostra la viabilitat de crear models lingüístics especialitzats.

7.3.3. Irlandès

La preservació de l'irlandès ha estat una font d'inspiració per a l'ús de la IA en llengües en perill. S'estan explorant mètodes basats en IA per a la seva protecció, i aquests esforços han inspirat projectes similars per a altres llengües indígenes als EUA, com el cherokee.[76] La recerca se centra en com la IA pot anar més enllà de la simple documentació, buscant sistemes interactius que puguin mantenir converses significatives i servir com a eines per a aprenents i educadors.

El projecte E-STÓR[77] a Irlanda per a la llengua irlandesa representa un esforç específic per abordar l'escassetat de dades i desenvolupar recursos lingüístics digitals per a aquesta llengua celta.

7.3.4. Gal·lès

El desenvolupament de Macsen,[78] un assistent de veu en llengua gal·lesa, il·lustra clarament el potencial del reconeixement de veu per a llengües no hegemòniques. A més, hi ha exemples concrets d'aplicacions d'IA que s'estan utilitzant per a l'aprenentatge d'aquest tipus de llengües europees, com Talkpal[79] i SpeakPal[80] per al gal·lès.

---

76 HACKETT, «Tennessee Tech professor uses AI to help preserve Cherokee language».

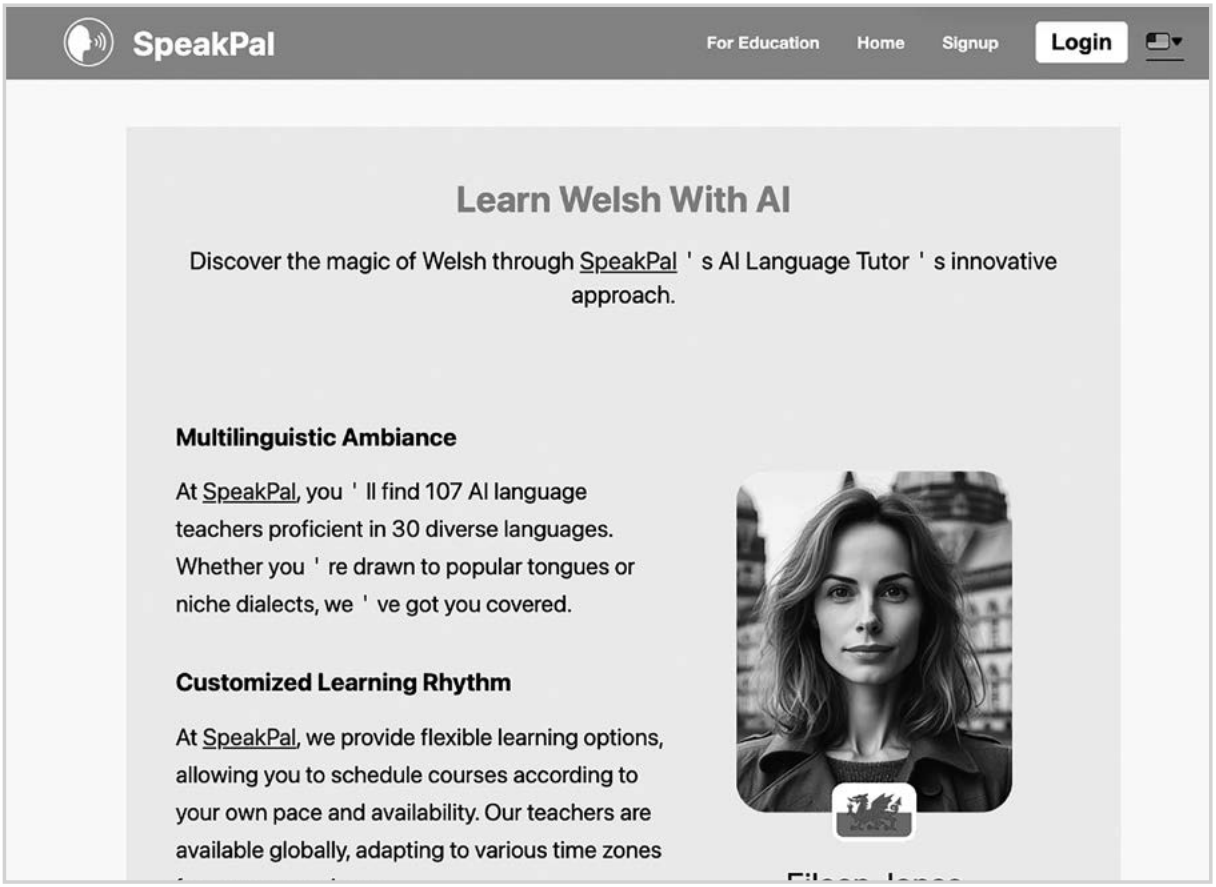
77 ADAPTCENTRE.IE, «Irish Language Technology Resource Marks Growth With Rebrand».

78 WELSH NATIONAL LANGUAGE TECHNOLOGIES PORTAL, «Macsen».

79 TALKPAL.AI, «Learn Welsh».

80 SPEAKPAL.AI, «Learn Welsh With AI».

Imatge 7. Assistent de veu en gal·les SpeakPal

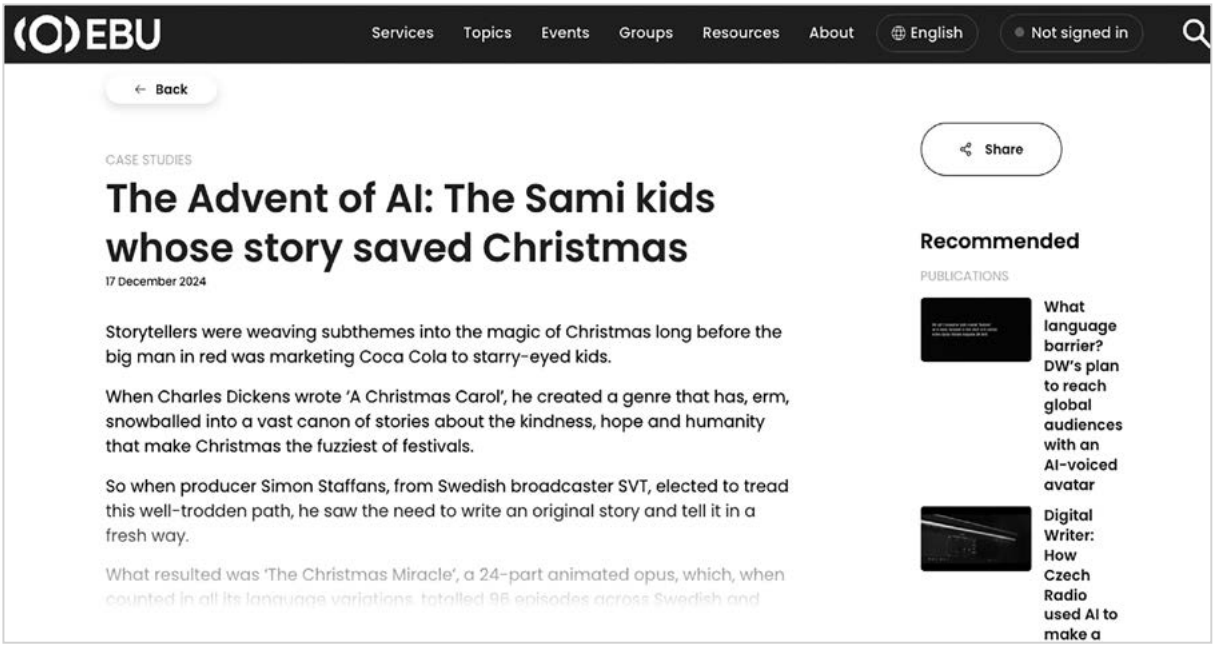


Font: SpeakPal AI.

7.3.5. Sami

Investigadors de diverses institucions estan treballant activament en la creació d'eines de reconeixement i transcripció de veu per a llengües com el sami. Hi ha una aplicació finançada pel Parlament Sami de Noruega per a l'aprenentatge de les llengües sami. Un exemple interessant, com hem apuntat anteriorment, de l'ús de la IA en la creació de contingut és el projecte «The Christmas Miracle» a Suècia, que va utilitzar IA generativa per crear una sèrie de televisió en sami, involucrant nens sami en el procés creatiu.

Imatge 8. Notícia relacionada sobre Christmas Miracle



Font: EBU.

7.3.6. Iniciatives paneuropees

- GraphoGame –Finlàndia–: aquest projecte finlandès, reconegut per la UNESCO, utilitza eines d'aprenentatge basades en jocs per fer que l'educació en alfabetització sigui accessible en més de 30 idiomes a tot el món, incloses les llengües minoritàries.[81] GraphoGame demostra com la IA pot ser aplicada en un format lúdic i accessible per fomentar l'alfabetització en la llengua materna, un factor clau per al desenvolupament cognitiu i l'èxit acadèmic.
- Native Scientists –paneuropeu–: aquest programa paneuropeu, també reconegut per la UNESCO, connecta estudiants migrants amb professionals de les ciències, la tecnologia, l'enginyeria i les matemàtiques (STEM) que comparteixen la seva llengua materna.[82] Operant en dotze països europeus, el pro-

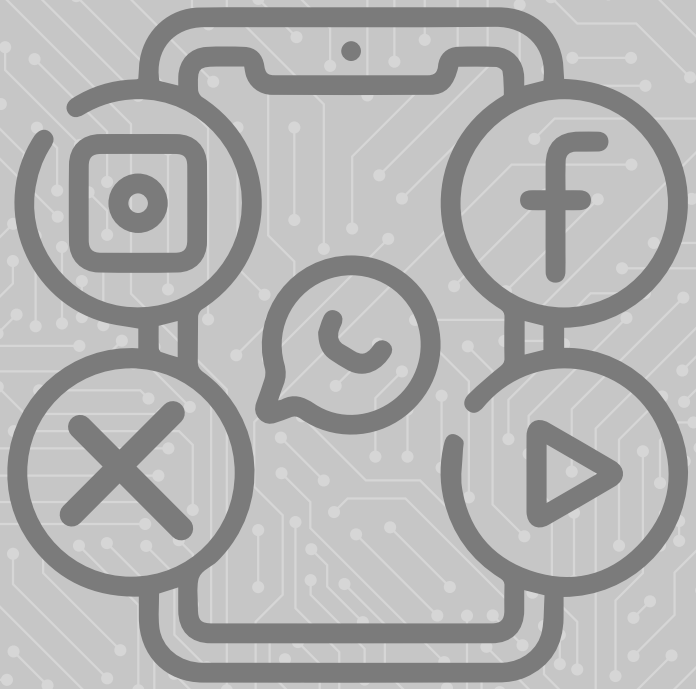
81 UNESCO, «Driving literacy through linguistic diversity and mother-language based learning».

82 «Native Scientists». Veure a: <www.nativescientists.org>.

grama promou la ciència i l'alfabetització en múltiples idiomes, i ajuda a superar la barrera lingüística en l'educació científica per a nens migrants i fomenta la diversitat cultural.

Els casos d'èxit en llengües no hegemòniques europees demostren que l'aprofitament de dades existents –fins i tot si són limitades–, la creació de corpus específics, les aliances publicoprivades i l'enfocament en aplicacions pràctiques –traducció de notícies, bots de conversa, educació– són més efectius que esperar solucions genèriques de grans models. Això posa de manifest la importància de l'agència local i la innovació adaptada.

Aquests exemples revelen que l'èxit no prové de la simple disponibilitat de grans volums de dades, sinó d'una combinació de factors estratègics. En primer lloc, s'observa l'«ús intel·ligent de dades ja existents», com demostra *Sermitsiaq* amb els seus deu-quinze anys de notícies traduïdes professionalment, que van servir com a valuós corpus d'entrenament. En segon lloc, hi ha una clara «inversió en la creació de corpus específics» per a la llengua, com és el cas del projecte Euskorpus per al basc. En tercer lloc, el desenvolupament d'«eines per a casos d'ús concrets» i pràctics, com el bot de conversa kAIxo o la plataforma educativa GraphoGame, demostra que les solucions enfocades són més efectives que les genèriques. Finalment, la «col·laboració entre governs, institucions acadèmiques i empreses» –com el Govern basc amb Euskorpus, o Miðeind amb OpenAI– és un factor comú. Aquest patró suggereix que la resiliència lingüística digital es construeix mitjançant una estratègia combinada «de baix a dalt» –impulsada per la comunitat i les necessitats locals– i «de dalt a baix» –amb suport governamental i aliances amb grans actors–, en què les comunitats i els governs prenen un paper actiu i proactiu en el disseny del seu futur digital.



8

L'ABORDATGE DE LES LLENGÜES A LES GRANS EMPRESES TECNOLÒGIQUES I ELS ÍNDEXS DE REFERÈNCIA COMPARATIUS

L'impacte de la IA en les llengües europees no hegemòniques està profundament influenciat per la manera com les grans empreses tecnològiques aborden la diversitat lingüística i per l'estat actual dels *benchmarks* [índexs de referència] comparatius utilitzats per a avaluar els models de llenguatge grans (LLM) multilingües.

8.1. El domini de les llengües majoritàries en els models de les grans empreses

Les grans empreses tecnològiques, com OpenAI, Google, Anthropic, Microsoft o Meta, són els principals desenvolupadors dels LLM més potents i utilitzats globalment. Aquests models, com GPT-4 o Gemini, són entrenats amb quantitats massives de dades textuals extretes principalment d'internet, llibres i altres fonts digitals. La conseqüència directa d'aquesta pràctica és que la major part del conjunt d'entrenament d'aquests models es troba en anglès i altres llengües majoritàries —com el xinès o el castellà. Aquesta realitat genera una «escletxa lingüística» significativa: els models d'IA moderns processen l'anglès, el mandarí i un grapat de llengües comercialment valuoses amb una proficiència molt superior, mentre que la immensa majoria de les més de 7.000 llengües del món tenen una representació digital mínima.

Aquesta asimetria en les dades d'entrenament es tradueix directament en un rendiment inferior dels LLM per a les llengües amb pocs recursos. Els models principals sovint penalitzen les llengües no angleses i, especialment, les llengües minoritàries i minoritzades, ja que no estan sintonitzats amb els contextos culturals rellevants d'aquestes llengües. Això pot generar errors gramaticals, 'translationese' –traduccions que sonen artificials o massa literals– i referències culturals incorrectes. La dominació de l'anglès en el desenvolupament de la IA ha portat a casos documentats d'assessorament mèdic inexacte en hindi o detencions errònies a causa de males traduccions en àrab.[83]

Les grans empreses tecnològiques prioritzen les llengües amb abundants recursos d'entrenament, mentre que llengües molt parlades però infrarepresentades, com l'hindi, el bengalí o l'urdú, reben comparativament menys atenció.[84] Aquesta situació no és només una qüestió de comoditat, sinó que es converteix en una de les barreres més significatives per a la inclusió global, i fa més profundes les desigualtats lingüístiques existents.

8.2. Rellevància dels índexs de referència comparatius per a llengües no majoritàries

Els índexs de referència (*benchmarks*) són conjunts de dades i tasques estandarditzades que s'utilitzen per avaluar i comparar les capacitats dels models d'IA. Són fonamentals per mesurar el progrés i identificar les àrees de millora. No obstant això, en el context multilingüe i de llengües amb pocs recursos, els índexs de referència actuals presenten deficiències significatives.

- Manca d'índexs de referència complets i rigor científic: una de les limitacions més grans és la manca d'índexs de referència complets que capturin adequadament els matisos de les llengües amb pocs recursos. Les pràctiques d'avaluació de les capacitats generatives dels LLM multilingües encara manquen d'exhaustivitat, rigor científic i una adopció consistent entre els laboratoris de recerca, la qual cosa soscava el seu potencial per orientar significativament el desenvolupament de models multilingües.
- Dependència de la traducció automàtica i biaixos culturals: molts índexs de referència multilingües són simplement traduccions d'índexs de referència anglesos existents. Aquesta pràctica introdueix soroll i errors, i el que és més important, manca de matisos culturals que es troben en altres llengües. Simplement

83 INCIDENTDATABASE.AI, «Incident 72: Facebook translates 'good morning' into 'attack them', leading to arrest».

84 BOSTON INSTITUTE OF ANALYTICS, «NANDA: The Future of Hindi AI and Language Inclusivity».

traduir un índex de referència no funciona, ja que no captura el context cultural clau i pot perpetuar vistes estereotipades o no diverses, resultant una «eliminació cultural». Per exemple, el conjunt d'informació geogràfica –*data set*– SHADES busca abordar aquesta qüestió identificant estereotips i biaixos en setze idiomes i trenta-set regions geogràfiques.[85]

- Rendiment diferencial en llengües amb pocs recursos: les avaluacions sistemàtiques revelen disparitats significatives en les capacitats multilingües dels LLM. Mentre que tenen un bon rendiment en llengües amb molts recursos, el seu rendiment disminueix notablement en llengües amb pocs recursos, amb diferències molt significatives. Estudis indiquen que els LLM tenen un rendiment particularment baix[86] en tasques intensives en sintaxi i en llengües amb múltiples sistemes d'escriptura. La correlació entre el rendiment dels LLM i els judicis humans és més alta en llengües amb molts recursos que en llengües amb pocs recursos.
- Desafiaments en l'avaluació de sistemes d'IA agentius: fins i tot en l'avaluació de sistemes d'IA agentius –sistemes autònoms que poden prendre decisions i fer tasques complexes, com transaccions econòmiques, sense intervenció humana–, els índexs de referència existents se centren exclusivament en l'anglès, deixant inexplorats els entorns multilingües. Això porta a la proposta de nous índexs de referència multilingües com el MAPS[87] per abordar aquesta escletxa.
- Recomanacions de bones pràctiques: per millorar l'avaluació, es recomana triar tasques amb un nivell de dificultat apropiat, evitar índexs de referència traduïts automàticament sense revisió per part d'experts, i utilitzar tasques i mètriques que es correlacionin amb els judicis humans. També es destaca la importància de la transparència en el suport lingüístic dels models, indicant-hi explícitament les llengües d'entrenament i avaluació.

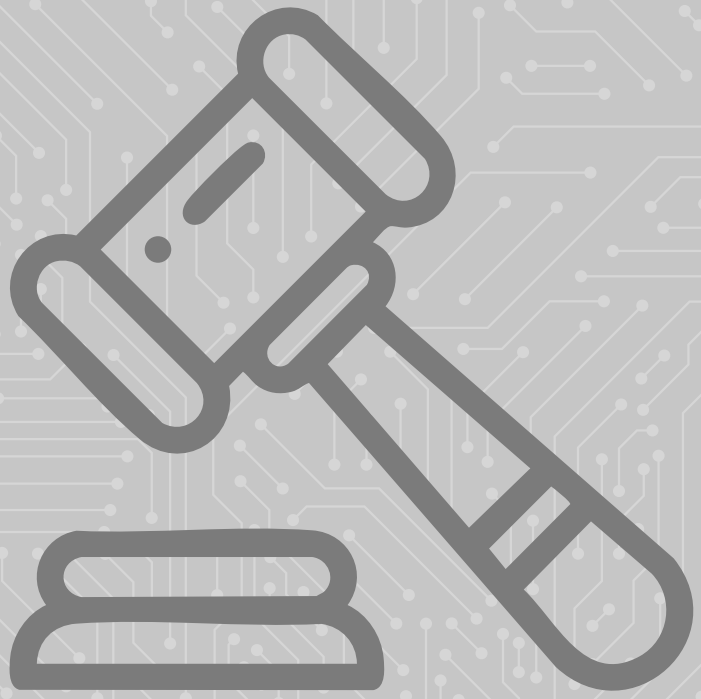
El domini de les llengües majoritàries en el desenvolupament de la IA i les deficiències en els índexs de referència multilingües creen un cicle de retroalimentació negativa. La manca de dades i avaluacions adequades per a les llengües minoritàries i minoritzades perpetua el seu infrarendiment en els models d'IA, la qual cosa desincentiva la inversió i el desenvolupament en aquestes llengües, i consolida així l'escletxa digital i la marginació lingüística.

85 MITCHELL, «SHADES: Towards a multilingual assessment of stereotypes in large language models».

86 ZIHAO, «Qualifying multilingual performance of large language models across languages».

87 HOFMAN, «MAPS: A multilingual benchmark for global agent performance and security».

Aquesta situació es pot considerar un cercle viciós. Els grans models d'IA, entrenats principalment amb dades en anglès i altres llengües hegemòniques, ofereixen un rendiment més baix per a les llengües amb pocs recursos. Aquest rendiment deficient es veu agreujat per la manca d'índexs de referència complets i culturalment rellevants per a aquestes llengües. La conseqüència és que no es pot avaluar amb precisió el rendiment dels models en aquestes llengües, cosa que dificulta la identificació d'àrees de millora i la justificació d'inversions per a la seva adaptació. Aquesta manca d'avaluació fiable i el menor rendiment produeixen una manca d'interès per part de les grans tecnològiques per invertir en llengües amb pocs recursos, ja que no veuen un retorn clar de la inversió. Aquesta absència d'inversió perpetua la manca de dades i el menor rendiment dels models, fet que tanca el cercle i consolida l'escletxa digital que afecta les llengües minoritzades i minoritàries. La situació subratlla que la inclusió lingüística en la IA no és només un problema tècnic, sinó un repte econòmic i estratègic que requereix un canvi de paradigma en la inversió i el desenvolupament.



9

LA LLEI D'IA DE LA UNIÓ EUROPEA I LES LLENGÜES

La UE ha estat pionera en la regulació de la IA amb la promulgació de la Llei d'Intel·ligència Artificial [AI Act] el juny de 2024, que s'ha convertit en la primera legislació integral a nivell mundial en aquest àmbit.[88] Aquesta llei té com a objectiu principal garantir que els sistemes d'IA utilitzats a la UE siguin segurs, transparents, traçables, no discriminatoris i respectuosos amb els drets fonamentals. Tot i que l'AI Act no conté disposicions explícites i detallades sobre la diversitat lingüística o la protecció de les llengües minoritàries, el seu marc general té implicacions indirectes i significatives per a aquestes.

9.1. Marc general i enfocament basat en el risc

L'AI Act adopta un enfocament basat en el risc, classificant els sistemes d'IA en diferents categories[89] –risc inacceptable, alt risc, risc limitat i risc mínim– segons el potencial impacte negatiu que puguin tenir en la seguretat o en els drets fona-

88 PARLAMENT EUROPEU, «Inteligencia artificial: oportunidades y desafíos».

89 COMISSIÓ EUROPEA, «Ley de IA».

mentals dels usuaris. Els sistemes d'alt risc, com els utilitzats en la sanitat, el transport o l'aplicació de la llei estan subjectes a requisits més estrictes d'avaluació, supervisió humana i transparència.

9.2. Implicacions per a la diversitat lingüística

Tot i que la llei no esmenta directament les llengües, diverses de les seves disposicions poden influir en la inclusió de les llengües no hegemòniques:

- No discriminació i biaixos: la llei exigeix que els sistemes d'IA siguin dissenyats, desenvolupats i utilitzats de manera que respectin els principis de justícia, igualtat i dignitat humana, i que no discriminin per motius com l'origen racial o ètnic, la religió o la creença. Aquesta disposició és crucial per a les llengües no hegemòniques, ja que els biaixos en els models d'IA poden reflectir i perpetuar estereotips ètnics o culturals. La manca de dades en aquesta tipologia de llengües pot portar a un rendiment deficient i, per tant, a una discriminació indirecta. La necessitat de mitigar els biaixos algorítmics, com s'ha vist amb el conjunt d'informació geogràfica –*dataset*– SHADES, és un requisit implícit que pot beneficiar la representació lingüística.
- Transparència i publicació de les dades d'entrenament: l'AI Act imposa requisits de transparència[90] per a la IA generativa, incloent-hi l'obligació de revelar que el contingut ha estat generat per IA i de publicar resums de les dades amb drets d'autor utilitzades per a l'entrenament. Aquesta transparència pot obligar les grans empreses tecnològiques a ser més explícites sobre la composició lingüística dels seus conjunts de dades d'entrenament. Si es fa evident la infrarepresentació de certes llengües, podria generar pressió per a una major inclusió.
- Accés a la informació i serveis multilingües: tot i que no és un requisit directe de l'AI Act, la Directiva de serveis digitals (DSA) de la UE, que complementa la regulació de la IA, ja exigeix que les plataformes en línia proporcionin termes i condicions en totes les llengües oficials de la UE, suport al client multilingüe i interfícies localitzades.[91] Aquesta tendència cap al multilingüisme en els serveis digitals pot incentivar el desenvolupament de tecnologies d'IA que suportin més llengües, incloses les no hegemòniques, per complir amb les expectatives d'accessibilitat.
- Foment de la diversitat cultural i lingüística: la Comissió Europea ha expressat el seu suport a iniciatives que promouen la diversitat cultural i lingüística en la IA, com l'Alliance for Language Technologies Euro-

90 EU ARTIFICIAL INTELLIGENCE ACT, «Article 13: Transparency and Provision of Information to Deployers».

91 COMISSIÓ EUROPEA, «Digital Services Act: Questions and Answers».

pean Digital Infrastructure Consortium [l'Aliança per a les Tecnologies Lingüístiques Consorci Europeu d'Infraestructures Digitals] (ALT-EDIC) i el Language Data Space (LDS). Aquests projectes busquen abordar l'escassetat de dades lingüístiques europees per entrenar LLM, amb l'objectiu de revolucionar els sistemes d'IA multilingües i trencar les barreres lingüístiques a la UE. Aquesta orientació política, tot i no ser directament part de l'AI Act, crea un ecosistema favorable.

- Cooperació i governança: la llei preveu la creació d'un Consell Europeu d'Intel·ligència Artificial (EAIB)[92] per proporcionar orientació i facilitar la cooperació entre les autoritats estatals. Aquesta estructura pot ser un canal per a les preocupacions de les comunitats lingüístiques no hegemòniques i per promoure estàndards que tinguin en compte la diversitat.

9.3. Desafiaments i limitacions

Malgrat el marc regulador, hi ha desafiaments. L'AI Act se centra en la regulació de l'aplicació de la IA, no de la tecnologia en si mateixa. Això significa que, si bé es prohibeixen certs usos d'alt risc, la llei no obliga directament al desenvolupament de la IA en totes les llengües. A més, el desplegament complet de l'AI Act és gradual, amb algunes disposicions que no seran aplicables fins al 2026 o 2027.

En resum, la Llei d'intel·ligència artificial de la UE, tot i no abordar explícitament les llengües no hegemòniques, estableix un marc ètic i de seguretat que pot influir indirectament en la promoció de la diversitat lingüística. La seva èmfasi en la no discriminació, la transparència i la responsabilitat pot impulsar les grans empreses tecnològiques a considerar més la inclusió lingüística, especialment si es complementa amb iniciatives de dades i polítiques de foment de la diversitat.

92 COMISSIÓ EUROPEA, «AI Board».

10

LLIÇONS PER A LES LLENGÜES EUROPEES MITJANES

Les llengües europees mitjanes, com l'alemany, el francès o l'italià, per no ser hegemòniques com l'anglès, el castellà o el xinès en l'àmbit digital global, comparteixen alguns reptes amb les llengües minoritzades i minoritàries en l'era de la IA. La seva posició intermèdia les fa vulnerables al domini de l'anglès en el desenvolupament de la IA i, per tant, poden aprendre lliçons valuoses de les estratègies adoptades per les comunitats de llengües minoritàries i minoritzades per assegurar la seva supervivència i rellevància digital.

10.1. Necessitat d'inversió en dades i corpus específics

Una de les lliçons més crucials és la importància de la inversió proactiva en la creació de dades i corpus lingüístics de qualitat. Les llengües no hegemòniques han hagut de fer un esforç considerable per generar els seus propis conjunts de dades, com demostra el Projecte AINA amb el català a Common Voice. Les llengües mitjanes, tot i tenir més dades disponibles, no han d'assumir que els grans models d'IA les cobriran adequadament. La majoria dels models d'IA continuen estant predominantment optimitzats per a l'anglès, amb una caiguda significativa del rendiment en altres idiomes.

Per tant, llengües com l'alemany, el francès i l'italià haurien de:

- Incentivar la creació de corpus especialitzats: més enllà de les dades generals d'internet, calen corpus curats i representatius de diversos dominis –legal, mèdic, tècnic, cultural– per garantir que els models d'IA comprenen els matisos i la terminologia específica de cada llengua.
- Promoure la digitalització del patrimoni lingüístic: digitalitzar llibres, arxius, enregistraments d'àudio i vídeo en aquestes llengües per augmentar la quantitat de dades disponibles per a l'entrenament de models.
- Participar en iniciatives de dades obertes: contribuir a plataformes com Common Voice i altres repositoris de dades lingüístiques obertes per enriquir l'ecosistema global i facilitar el desenvolupament de models multilingües.

10.2. Desenvolupament de models propis i sobirania tecnològica

Les llengües no hegemòniques han demostrat la importància de no dependre exclusivament de les grans empreses tecnològiques. Projectes com AINA amb el model Salamandra o iniciatives basques com Euskorpus i Orai mostren la capacitat de desenvolupar models propis i de codi obert. Aquesta estratègia permet a les llengües mitjanes:

- Fomentar la recerca i el desenvolupament local: invertir en centres de recerca i universitats per desenvolupar models de llenguatge i tecnologies d'IA adaptades a les seves llengües, amb un enfocament en la qualitat i la precisió. Alemanya, per exemple, ja té empreses com Aleph Alpha[93] que busquen ser una alternativa europea als gegants d'IA nord-americans.
- Assegurar la sobirania digital lingüística: tenir models propis garanteix el control sobre la tecnologia i evita la «colonització digital» per part de llengües hegemòniques i dominants. Això és crucial per a la seguretat de les dades i la protecció de la cultura.
- Promoure l'ús de codi obert: alliberar models i eines en forma de codi obert per fomentar la col·laboració i la innovació en l'ecosistema de la llengua.

93 «Aleph Alpha». Veure a: <www.aleph-alpha.com>.

10.3. Col·laboració publicoprivada i foment dels ecosistemes locals

Els casos d'èxit en llengües no hegemòniques sovint impliquen una estreta col·laboració entre governs, institucions acadèmiques i el sector privat. Aquesta és una lliçó clau per a les llengües mitjanes:

- Establir aliances estratègiques: els governs poden impulsar programes de finançament i coordinació, com l'Aina Challenge o els programes d'IA a Alemanya i França, per connectar la recerca amb les necessitats empresarials i fomentar la creació de solucions d'IA en la llengua pròpia.
- Impulsar l'adopció en sectors clau: fomentar l'ús de la IA en la llengua pròpia en sectors com l'administració pública, l'educació, la salut i els mitjans de comunicació, creant una demanda i un mercat per a aquestes tecnologies.
- Crear *living labs* [laboratoris vius] i sandboxes [entorns controlats de proves] reguladors: facilitar entorns de prova on empreses i investigadors puguin desenvolupar i provar solucions d'IA en un context real, amb el suport de la regulació –com l'AI Act.

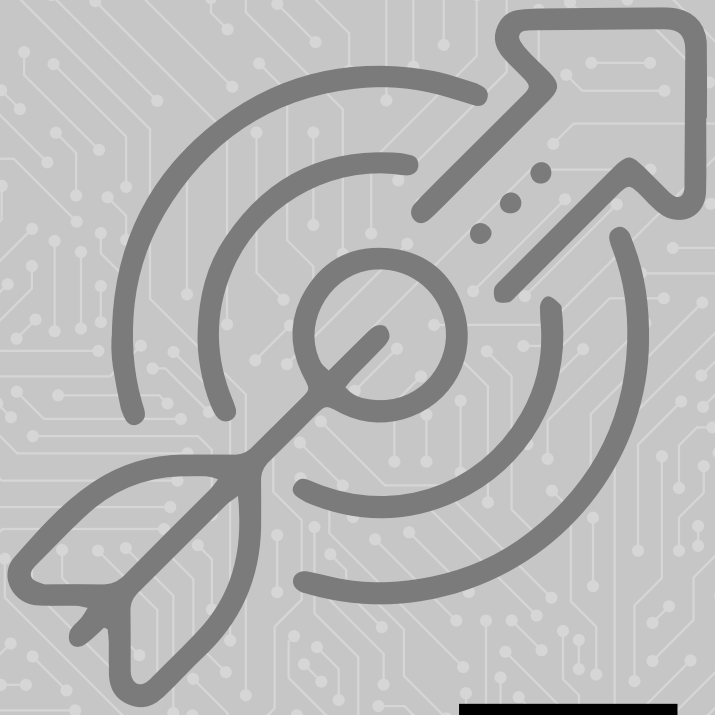
10.4. Adaptació a les necessitats específiques i aplicacions pràctiques

Les llengües no hegemòniques han demostrat que l'èxit no es troba en la creació de models genèrics que competeixin directament amb els gegants globals, sinó en l'adaptació de la IA a les necessitats específiques de la comunitat:

- Enfocament en aplicacions de valor afegit: desenvolupar eines d'IA que resolguin problemes concrets i aportin valor real als parlants, com ara sistemes de traducció de notícies –kalaallisut–,bots conversacionals –basc–, o eines educatives –GraphoGame.
- Ús de la retroacció humana: la col·laboració islandesa amb OpenAI demostra que la revisió humana (RLHF)[94] pot millorar significativament el rendiment dels models fins i tot amb dades limitades. Les llengües mitjanes poden aplicar aquesta metodologia per afinar els models existents.
- Sensibilització i participació ciutadana: fomentar la participació de la ciutadania en la creació de dades –com Common Voice– i en la validació de les tecnologies d'IA, creant un sentiment de propietat i compromís amb el futur digital de la llengua.

94 «Reinforcement learning from human feedback».

En conjunt, les llengües europees mitjanes poden aprendre de les llengües no hegemòniques que la proactivitat, la inversió estratègica en dades i recerca, la col·laboració i l'enfocament en solucions pràctiques són essencials per assegurar-ne la vitalitat en l'era de la IA i evitar caure en una «colonització digital» per part de l'anglès i altres llengües hegemòniques i dominants.



11

CONCLUSIONS I RECOMANACIONS

L'anàlisi detallada dels efectes de la IA en el futur de les llengües europees no hegemòniques revela un panorama complex, marcat per una dualitat inherent d'oportunitats i de desafiaments. La IA, amb el seu potencial transformador, es presenta com una eina de doble tall: pot ser un catalitzador per a la preservació i revitalització lingüística, però també un vector d'homogeneïtzació i marginació.

- Doble tall de la IA: la IA generativa presenta un potencial dual. Per una banda, ofereix eines sense precedents per a la preservació, documentació, revitalització i aprenentatge de llengües, i permet la creació de contingut a escala i la millora de l'accessibilitat. Per altra banda, la seva dependència de grans volums de dades de qualitat significa una amenaça existencial per a les llengües amb pocs recursos, que sovint manquen de la representació digital necessària.
- L'escletxa digital com a problema estructural: la infrarepresentació de les llengües no hegemòniques en els conjunts de dades d'entrenament dels LLM globals condueix a un rendiment inferior, biaixos algorítmics i la propagació

d'estereotips. Aquesta situació pot resultar en un «monocultiu tecnològic» que margina comunitats senceres i erosiona la diversitat cultural i lingüística.

- Agència local i col·laboració com a claus de l'èxit: els casos d'estudi del català –Projecte AINA–, l'islandès amb OpenAI, el basc i el kalaallisut demostren que la proactivitat, la inversió pública i privada en la creació de corpus de dades, el desenvolupament de models propis o l'adaptació de models existents i la col·laboració entre governs, institucions acadèmiques i la comunitat, són factors determinants per a la supervivència digital.
- Deficiències en els índexs de referència: Els índexs de referència o *benchmarks* actuals per avaluar LLM multilingües són sovint insuficients, basats en traduccions automàtiques que introdueixen errors i que no capturen els matisos culturals. Això dificulta una avaluació justa del rendiment dels models en llengües amb pocs recursos i la identificació d'àrees de millora.
- El marc regulador europeu com a oportunitat indirecta: la Llei d'IA de la UE, tot i no centrar-se explícitament en les llengües, estableix principis de no discriminació, transparència i responsabilitat que poden impulsar indirectament una major inclusió lingüística per part dels desenvolupadors d'IA. Iniciatives europees com l'aliança ALT-EDIC[95] i l'European Language Data Space[96] reforcen aquesta orientació.
- Lliçons per a les llengües mitjanes: les llengües europees mitjanes poden aprendre de les llengües minoritàries la importància de la inversió en dades i corpus específics, el desenvolupament de models propis per a la sobirania tecnològica, la col·laboració publicoprivada i l'enfocament en aplicacions pràctiques de valor afegit.

11.1 Recomanacions

Per assegurar un futur digital equitatiu i divers per a les llengües europees no majoritàries, es proposen les següents recomanacions:

1. Inversió estratègica i coordinada en dades lingüístiques:

- Prioritzar la creació de corpus de dades de qualitat: els governs i les institucions han d'invertir significativament en recopilació, curació i digitalització de dades textuals i orals en llengües no hegemòniques. Aquestes dades han de ser representatives de l'ús real de la llengua i no basar-se en traduccions automàtiques de llengües majoritàries en l'àmbit digital.

95 COMISSIÓ EUROPA, «Alliance for language technologies EDIC».

96 COMISSIÓ EUROPA, «European language data space».

- Fomentar les plataformes de dades obertes: promoure la participació activa de les comunitats lingüístiques en iniciatives de *crowdsourcing* [proveïment participatiu] de dades, com Common Voice, i assegurar que els corpus generats siguin accessibles i reutilitzables per a la recerca i el desenvolupament.

2. Desenvolupament de models d'IA adaptats i específics per a llengües minoritzades i minoritàries:

- Suport a la recerca i al desenvolupament local: finançar centres de recerca i empreses locals especialitzades en tecnologies del llenguatge per desenvolupar models de llenguatge (LLM) adaptats a les especificitats de cada llengua, com el model Salamandra per al català.
- Explorar models híbrids i d'adaptació: investigar i implementar estratègies com l'aprenentatge per *multilingual transfer learning* [transferència multilingüe] o *fine-tuning* [l'ajust fi] de models globals amb dades locals, per maximitzar l'eficiència dels recursos existents.

3. Establiment de col·laboracions estratègiques:

- Fomentar la col·laboració publicoprivada: promoure aliances entre governs, institucions acadèmiques i grans empreses tecnològiques per a la inclusió de llengües minoritàries en els seus models, com el cas d'Islàndia amb OpenAI.
- Integrar el *«human-in-the-loop»*: assegurar que els processos de desenvolupament i avaluació dels models d'IA incloguin la supervisió i la retroacció d'experts i parlants nadius per garantir la qualitat i la correcció cultural, especialment en llengües amb pocs recursos.

4. Fer visible la demanda dels usuaris:

- Fomentar la configuració dels dispositius en la llengua d'interès: les plataformes d'IA són capaces d'identificar implícitament la preferència lingüística de l'usuari detectant l'idioma del seu dispositiu, el seu navegador web i el seu perfil de client. Aconseguir que més usuaris configurin el seu entorn digital en el seu idioma incrementarà el cens de parlants visible digitalment, i farà l'idioma més atractiu per a la plataforma.
- Dialogar amb els bots de conversa en la llengua d'interès: molts usuaris tendeixen a fer servir un idioma majoritari en el diàleg amb els bots de conversa d'IA, confiant que els entendrà millor. Aquest comportament ja és conegut dels cercadors web. En realitat, els bots de conversa solen ser capaços d'entendre peticions en molts més idiomes, inclosos molts dels no majoritaris europeus. Fent-los servir en la interacció, l'usuari està manifestant explícitament el seu interès a dialogar-hi en la seva pròpia llengua, un interès que la plataforma o bé atén directament o bé registra com a incidència per a futures ampliacions.

Cal trencar amb aquesta dinàmica de substitució lingüística digital, tal com cal fer també en l'ús normal de la llengua.

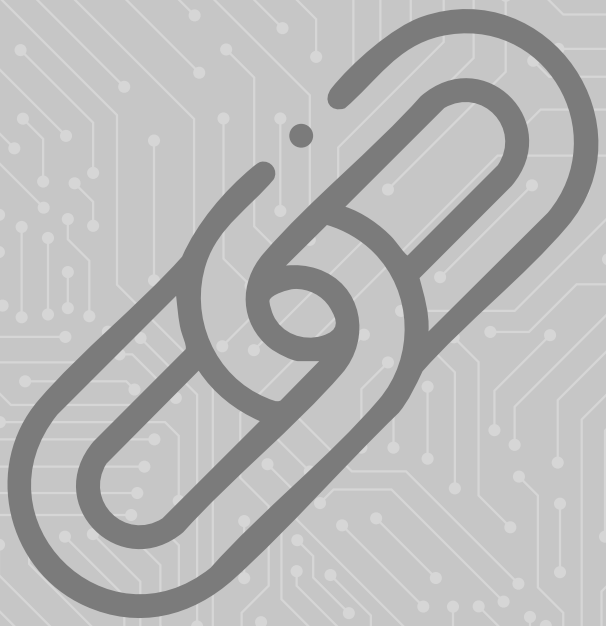
5. Millora dels índexs de referència i mètriques d'avaluació multilingües:

- Desenvolupar índexs de referència culturalment rellevants: crear nous *benchmarks* [índexs de referència] que no siguin meres traduccions, sinó que reflecteixin els matisos culturals i lingüístics de cada idioma, amb la participació d'experts locals.
- Promoure la transparència en l'avaluació: exigir als desenvolupadors de LLM que especifiquin clarament les llengües utilitzades en l'entrenament i l'avaluació dels seus models.

6. Aprofitament del marc regulador europeu:

- Defensar la diversitat lingüística en l'aplicació de l'AI Act: assegurar que els principis de no discriminació i transparència de la Llei d'IA es tradueixin en una major inclusió i un rendiment equitatiu dels sistemes d'IA per a totes les llengües de la UE.
- Impulsar iniciatives europees de dades lingüístiques: donar suport i participar activament en projectes com l'ALT-EDIC i l'European Language Data Space per crear un ecosistema de dades lingüístiques europeu robust i divers.

En última instància, el futur de les llengües europees no hegemòniques en l'era de la IA no és una qüestió de determinisme tecnològic, sinó el resultat de decisions polítiques, inversions estratègiques i capacitat d'innovació i col·laboració de les comunitats lingüístiques. Transformar els desafiaments en oportunitats requereix un compromís ferm amb la diversitat lingüística com a valor fonamental del patrimoni europeu, assegurant que la IA esdevingui una eina per a la inclusió i l'empoderament, i no un factor d'erosió cultural.



12

REFERÈNCIES

ADAPTCENTRE.IE. «Irish Language Technology Resource Marks Growth With Rebrand» [en línia]. De 27 d'octubre de 2022. Disponible a: <www.adaptcentre.ie>.

AMAZON WEB SERVICES. «¿En qué consiste la IA generativa?» [en línia]. Disponible a: <www.aws.amazon.com>.

AMAZON WEB SERVICES. «¿Qué es la IA generativa?» [en línia]. Disponible a: <www.aws.amazon.com>.

APERTIUM. «Apertium. Plataforma lliure i de codi obert per a la traducció automàtica» [en línia]. Disponible a: <www.apretium.org>.

BAPNA, Ankur (et. al). «Building machine translation systems for the next thousand languages» [en línia]. Disponible a: <www.arxiv.org>.

BASTARDES, Albert. «Les polítiques de la llengua i la identitat a l'era 'glocal'» [en línia]. A *Generalitat de Catalunya. Col·lecció Institut d'Estudis Autonòmics*, 50, p. 70. Disponible a: <www.diposit.ub.edu>.

BOSTON INSTITUTE OF ANALYTICS. «NANDA: The Future of Hindi AI and Language Inclusivity» [en línia]. De 12 de setembre de 2024. Disponible a: <www.bostoninstituteofanalytics.org>.

BSC. «MareNostrum 5» [en línia]. Disponible a: <www.bsc.es>.

BSC. «Sobre Aina» [en línia]. Disponible a: <www.bsc.es>.

CHAUVET, Romain. «How using AI translation tools for minority languages can boost subscriptions» [en línia]. A *The Fix*, de 23 de maig de 2025. Disponible a: <www.thefix.media>.

COMMON VOICE MOZILLA. «Tecnologia que parla la vostra llengua» [en línia]. Disponible a: <www.commonvoice.mozilla.org>.

COUNCIL OF EUROPE. «About the European Charter for Regional or Minority Languages» [en línia]. D'1 de març de 1998. Disponible a: <www.coe.int>.

CUESTA, Albert. «La IA de WhatsApp es nega a respondre en català (tot i saber-ne) [en línia]. A *Ara*, 28 de juliol del 2025. Disponible a: <www.ara.cat>.

CUESTA, Albert. «Bon dia. Tinc resposta de @Meta sobre el català. [...]» [en línia]. A *x.com*, de 2 de juliol de 2025. Disponible a: <www.x.com/albertcuesta>.

CUESTA, Albert. «La meva interpretació: si en algun moment heu dialogat en català [...]» [en línia]. A *x.com*, 2 de juliol de 2025. Disponible a: <www.x.com/albertcuesta>.

DEEPL. «DeepL Translator» [en línia]. Disponible a: <www.deepl.com>.

EU ARTIFICIAL INTELLIGENCE ACT. «Article 13: Transparency and Provision of Information to Deployers» [en línia]. De 2 de febrer de 2025. Disponible a: <www.artificialintelligenceact.eu>.

EUROPEAN BROADCASTING UNION. «The Advent of AI: The Sami kids whose story saved Christmas» [en línia]. De 17 de desembre de 2024. Disponible a: <www.edu.ch>.

EUROPEAN COMMISSION. «AI Board» [en línia]. D'1 d'agost de 2024. Disponible a: <www.digital-strategy.ec.europa.eu>.

EUROPEAN COMMISSION. «Alliance for Language Technologies EDIC» [en línia]. Disponible a: <www.language-data-space.ec.europa.eu>.

EUROPEAN COMMISSION. «Digital Services Act: Questions and Answers» [en línia]. D'1 d'agost de 2024. Disponible a: <www.digital-strategy.ec.europa.eu>.

EUROPEAN COMMISSION. «European Language Data Space» [en línia]. Disponible a: <www.language-data-space.ec.europa.eu>.

EUROPEAN COMMISSION. «Ley de IA» [en línia]. De 12 de juliol de 2024. Disponible a: <www.digital-strategy.ec.europa.eu>.

EUROPEAN PARLIAMENT. «Inteligencia artificial: oportunidades y desafíos» [en línia]. De 25 d'abril de 2025. Disponible a: <www.europarl.europa.eu>.

EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS. «Bias in Algorithms. Artificial Intelligence and Discrimination» [en línia]. Disponible a: <www.fra.europa.eu>.

EUSKADI.EUS. «Traductor neuronal» [en línia]. Disponible a: <www.euskadi.eus>.

EUSKADI.EUS. «Euskorpus, proyecto impulsado por el Gobierno Vasco [...]» [en línia]. De 19 de febrer de 2025. Disponible a: <www.euskadi.eus>.

GITHUB. «adefossez/demucs» [en línia]. Disponible a: <www.github.com>.

GITHUB. «Model for Age and Gender Recognition based on Wav2vec 2.0 (24 layers)» [en línia]. Disponible a: <www.huggingface.co>.

GONZÁLEZ-AGIRRE, Aitor (et al.). «Salamandra Technical Report» [en línia]. A *Cornell University*, de 13 de febrer de 2025. Disponible a: <www.arxiv.org>.

HACKETT, Tracey. «Tennessee Tech professor uses AI to help preserve Cherokee language» [en línia]. A *Tennessee Tech*, de 10 de març de 2025. Disponible a: <www.tntech.edu>.

HOFMAN, Omer (et al.). «MAPS: A Multilingual Benchmark for Global Agent Performance and Security» [en línia]. A *Cornell University*, de 21 de maig de 2025. Disponible a: <www.arxiv.org>.

HUGGING FACE. «Projecte-aina/matxa-tts-cat-multiaccent» [en línia]. Disponible a: <www.huggingface.co>.

HUGGING FACE. «Salamandra Vision Model Card» [en línia]. De 20 de maig de 2025. Disponible a: <www.ollama.hf-mirror.com>.

IBANEZ, Frédéric. «L'impact de l'intelligence artificielle sur l'avenir de la traduction» [en línia]. A *Alphatrad France*, de16 de març de 2023. Disponible a: <www.alphatrad.fr>.

INCIDENTDATABASE.AI. «Incident 72: Facebook translates 'good morning' into 'attack them', leading to arrest» [en línia]. De 17 d'octubre de 2017. Disponible a: <www.incidentdatabase.ai>.

IRISH RESEARCH COUNCIL. «Cardamom. Comparative deep models for minority and historical languages» [en línia]. Disponible a: <www.cardamon-project.org>.

KERN TRAINING. «Chancen und Risiken von KI-generierten Lerninhalten im Fremdsprachenunterricht» [en línia]. D'11 de març de 2025. Disponible a: <www.kerntraining.com>.

KLEIBER, Ingo. «Was ist generative künstliche Intelligenz (KI)?» [en línia]. A *Universität zu Köln*. Disponible a: <www.portal.uni-kolen.de>.

KULP, Patrick. «Studies explore challenges of AI for low-resource languages» [en línia]. A *TechBrew*, de 5 de maig de 2025. Disponible a: <www.emergingtechbrew.com>.

LAUMANN, Felix. «Low-resource language: what does it mean?» [en línia]. A *Neural Space*, de 10 de juny de 2022. Disponible a: <www.medium.com>.

LÓPEZ, Beatriz. «Els robots i els sistemes intel·ligents revolucionen la medicina» [en línia]. A *Barcelona Metròpolis* 127, de juliol de 2023, p. 32. Disponible a: <www.barcelona.cat>.

McMONAGLE, Sarah. «Autochtone Minderheitensprache» [en línia]. A *Universität Hamburg*, de 9 de febrer de 2021. Disponible a: <www.mehrsprachigkeit.uni-hamburg.de>.

METADATA «El català, la llengua líder de Common Voice» [en línia]. 16 de maig del 2024. Disponible a: <www.metadata.cat>.

METTA-WINDISCHER, Roberta (et al.). «Reframing Minority Rights Amid Global Challenges: The Role of AI [...]» [en línia]. A *Eurac Research Science Blogs*, de 12 de maig de 2025. Disponible a: <www.eurac.edu>.

MITCHELL, Margaret (et al.). «SHADES: Towards a Multilingual Assessment of Stereotypes in Large Language Models» [en línia]. A *Association for Computational Linguistics*, 2025. Disponible a: <www.aclanthology.org>.

MOSELEY, Christopher. «Atlas of the World's Languages in Danger» [en línia]. A *UNESCO Digital Library*, 2010. Disponible a: <www.unesdoc.unesco.org>.

OFFICE OF THE PRESIDENT OF ICELAND. «My team and I were intrigued by our conversation with @SamA [...]» [en línia]. A *x.com*. Disponible a: <www.x.com>.

OPEN AI. «Gobierno de Islandia» [en línia]. Disponible a: <www.openai.com>.

PANAY, Panos. «Introducing Alexa+, the next generation of Alexa» [en línia]. De 26 de febrer de 2025. Disponible a: <www.abouta.amazon.com>.

PAVA, Juan (et al.). «Mind the (Language) Gap: Mapping the Challenges of LLM Development in Low-Resource Language Contexts» [en línia]. A *Stanford University HAI*, de 22 d'abril de 2025. Disponible a: <www.hai.stanford.edu>.

PROJECTE AINA. «Aina Challenge» [en línia]. Disponible a: <www.projecteaina.cat>.

PROJECTE AINA. «Aina participates in the Deep Learning Barcelona Symposium 2024» [en línia]. De 23 de desembre de 2024. Disponible a: <www.projecteaina.cat>.

PROJECTE AINA. «Coneix la família de models Salamandra» [en línia]. Disponible a: <www.projecteaina.cat>.

PROJECTE AINA. «Crea eines d'IA per a una administració més àgil, propera i en català» [en línia]. Disponible a: <www.projecteaina.cat>.

PROJECTE AINA. «El corpus de Common Voice; optimisme i reptes pendents» [en línia]. De 30 de maig de 2024. Disponible a: <www.projecteaina.cat>.

PROJECTE AINA. «Promoting the use of Catalan in the digital age» [en línia]. Disponible a: <www.projecteaina.cat>.

PROJECTE AINA. «Success in the Aina Challenge call to accelerate AI projects in Catalan» [en línia]. De 25 de març de 2025. Disponible a: <www.projecteaina.cat>.

PROJECTE AINA. «The Aina Challenge, the competition endowed by the Catalan Government with €1M [...]» [en línia]. De 25 de març de 2025. Disponible a: <www.projecteaina.cat>.

RODRÍGUEZ, Carlos (et al.). «Demo notebook to use Aina models in Google Colab to create a simple RAG system» [en línia]. A *Github*. Disponible a: <www.github.com>.

SCHNEIDER, Britta. «Multilingualism and AI: The Regimentation of Language in the Age of Digital Capitalism» [en línia]. A *Cambridge University Press*, d'1 de gener de 2025. Disponible a: <www.cambridge.org>.

SOFTCATALÀ. «Doblatge de vídeos automàtic en català» [en línia]. Disponible a: <www.softcatala.org>.

SPEAKPAL.AI. «Learn Welsh With AI» [en línia]. Disponible a: <www.speakpal.ai>.

SPECTOR, Violette. «AI in Language Preservation: Safeguarding Low-Resource and Indigenous Languages» [en línia]. De 18 de febrer de 2025. Disponible a: <www.welocalize.com>.

TALKPAL.AI. «Learn Welsh» [en línia]. Disponible a: <www.talkpal.ai>.

TATUTRAD TRADUCTORES. «¿Qué idiomas se hablan en Europa?» [en línia]. Disponible a: <www.tatutrad.net>.

MORNINGSIDE. «The Real Cost of Errors in Medical Translations» [en línia]. Disponible a: <www.morningtrans.com>.

TÓMAS, Ragnar. «GPT-4 to Aid in the Preservation of the Icelandic Language» [en línia]. *A Iceland Review* a 15 de març del 2023. Disponible a: <www.icelandreview.com>.

UNESCO. «Driving literacy through linguistic diversity and mother-language based learning» [en línia]. De 20 de febrer de 2025. Disponible a: <www.unesco.org>.

UNITED STATES MILITARY ACADEMY LIBRARY. «Large Language Models» [en línia]. Disponible a: <www.library.westpoint.edu>.

WELSH NATIONAL LANGUAGE TECHNOLOGIES. «Macsen» [en línia]. Disponible a: <www.techiaih>.

ZIHAO, Li (et al.). «Qualifying Multilingual Performance of Large Language Models Across Languages» [en línia]. A *Cornell University*, 16 de juny de 2024. Disponible a: <www.arxiv.org>.


FUNDACIÓ
IRLA


Coppieters
foundation


Accent
obert