

Awal – Tecnologia Lingüística per a l'Amazic Impulsada per la Comunitat

Alp Öktem^{1,2}, Farida Boudichat²¹Col·lectivaT
alp@collectivat.cat

²Equip d'Awal
awal@collectivat.cat

Abstracte.

En aquest article es presenta Awal (اڤال), una iniciativa impulsada per la comunitat per desenvolupar recursos de tecnologia lingüística per a la llengua amaziga. Proporcionem un estudi exhaustiu del panorama del processament del llenguatge natural en amazic, investigant els progressos recents en recursos computacionals i l'emergència de les iniciatives comunitàries per abordar l'escassetat persistent de dades. La plataforma awaldigital.org, llançada el 2024, aborda la infrarepresentació de l'amazic als espais digitals a través d'una plataforma col·laborativa que permet contribuir a les persones parlants amb traduccions i dades de veu. Analitzem 18 mesos d'implicació comunitària, revelant barreres significatives davant la participació, inclosa la confiança limitada amb l'amazic escrit i els reptes d'estandardització en curs. Tot i la recepció positiva generalitzada, la contribució real de les dades ha quedat concentrada entre lingüistes i activistes. La modesta escala de les contribucions de la comunitat—6.421 parells de traducció i 3 hores de dades de la parla—posa de manifest les limitacions d'aplicar els enfocaments estàndards de proveïment participatiu en llengües amb contextos sociolingüístics complexos. Estem treballant en models de traducció automàtica de codi obert millorats a partir de les dades recollides.

†.XEUξ†

[illegible]

[illegible]

1. Introducció

La representació digital de les llengües al món segueix sent profundament desigual, i els avenços tecnològics beneficien principalment a un petit subconjunt de llengües àmpliament parlades, mentre que milers d'altres queden marginades en els espais digitals. Aquesta bretxa lingüística digital és particularment pronunciada per a les llengües d'Àfrica, on, malgrat representar més de 2.000 llengües parlades per més de mil milions de persones, els recursos computacionals i les tecnologies lingüístiques segueixen sent escassos (Nekoto et al., 2020; Kreutzer et al., 2022). Entre aquestes llengües infrarepresentades, l'amazic (dit també tamazight o berber) és un convincent estudi de cas tant dels reptes com de les possibilitats del desenvolupament de tecnologia lingüística impulsada per la comunitat parlant.

L'amazic, parlat per més de 40 milions de persones a tot el nord d'Àfrica i per les comunitats de la diàspora a tot el món, ha experimentat un notable viatge des de la marginació fins al reconeixement oficial en diversos països des del 2011. No obstant això, aquest reconeixement polític no s'ha traduït automàticament en una presència digital sòlida ni en recursos computacionals. Tot i els progressos recents—inclosos el llançament de l'amazic a la Viquipèdia el 2023, la incorporació de l'amazic al Traductor de Google i la disponibilitat d'eines digitals facilitada per part de l'Institut Reial

de la Cultura Amaziga (IRCAM) i d'altres lingüistes i desenvolupadors/es— la llengua encara no té presència suficient al web per impulsar les tecnologies basades en intel·ligència artificial (IA), especialment els models de llenguatge extens (MLE, també conegut per les sigles en anglès LLM, de *Large Language Model*). Si bé les iniciatives actuals se centren en l'estandardització i la promoció de les habilitats d'escriptura, es necessiten esforços paral·lels per crear continguts digitals d'alta qualitat i dades estructurades que puguin entrenar sistemes d'IA eficaços.

En aquest article es presenta el projecte Awal (اڤال), una iniciativa impulsada per la comunitat, llançada el 2024 per tres organitzacions de Catalunya amb l'objectiu de fomentar el desenvolupament de tecnologies lingüístiques per a l'amazic mitjançant la recollida de dades obtingudes a través de proveïment participatiu. Awal recopila parells de traducció i dades de la parla transcrita que es poden utilitzar directament per entrenar models d'IA, com ara el reconeixement automàtic de la parla (RAP), la traducció automàtica (TA) i els MLE. La plataforma anima la participació de la comunitat a través d'elements de gamificació, alhora que manté la qualitat del procés mitjançant mecanismes de validació entre iguals.

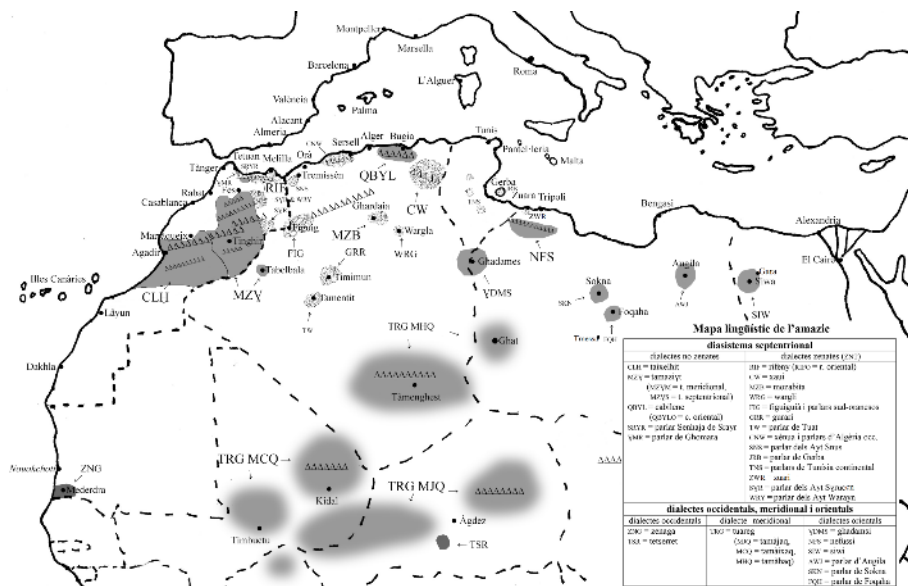
Aquest article s'estructura de la següent manera. A la secció 2, proporcionem un estudi exhaustiu del panorama del processament del llenguatge natural (PLN, també conegut per les sigles en anglès NLP, de *Natural Language Processing*) en amazic, estudiant-ne el perfil lingüístic i l'estatus sociopolític (secció 2.1) i els reptes i progressos en el PLN en amazic (secció 2.2), i els enfocaments impulsats per la comunitat (secció 2.3). La secció 3 detalla el projecte Awal, explicant-ne el començament i objectius (secció 3.1), la plataforma awaldigital.org (secció 3.2), la recollida de dades de veu a través de Common Voice (secció 3.3), les campanyes de divulgació i participació comunitària (secció 3.4), les estadístiques actuals i l'accés a les dades (secció 3.5), i les limitacions, els reptes i el camí a seguir (secció 3.6). Finalment, acabem amb les conclusions en la secció 4.

2. El panorama del PLN en amazic

Aquesta secció analitza l'estat actual de l'amazic en l'àmbit digital. En primer lloc, presentem una visió de conjunt de la diversitat lingüística i l'evolució sociopolítica d'aquesta llengua, després revisem els treballs i recursos computacionals existents abans de debatre sobre l'aparició d'iniciatives que promouen la implicació comunitària.

2.1. *El perfil lingüístic i l'estatus sociopolític de l'amazic*

L'amazic, també conegut com a tamazight o berber, pertany a la família lingüística afroasiàtica. El parlen més de 40 milions de persones en una àmplia regió del nord d'Àfrica, principalment al Marroc, Algèria, Tunísia, Líbia, així com les comunitats [amazigues] a Espanya, Mauritània, Níger, Mali i Egipte, i també les que formen part de la diàspora. La imatge 1 mostra la distribució geogràfica dels dialectes amazics al nord d'Àfrica. L'amazic és una de les llengües oficials al Marroc des del 2011, a Algèria des del 2016, a determinats districtes de Líbia des del 2017, i a Mali des del 2023. La llengua presenta una diversitat significativa, amb dialectes importants com el taixelhit (shi) parlat al sud-oest del Marroc i l'Alt Atles, el rifeny o tarifit (rif) parlat a la regió del Rif del Marroc, el tamazic central de l'Atles (tzm) parlat al centre i sud-est del Marroc, el cabilenc o taqbaylit (kab) parlat a Algèria, el shawiya o tacawit (shy) també a Algèria, el tamahaq (thv) a les regions del sud d'Algèria, el Níger i Mali, i el tamasheq (taq) a Mali i el Níger (Lafkioui, 2018).



Imatge 1: Distribució geogràfica dels dialectes amazics al nord d'Àfrica (Múrcia, 2017)

L'amazic ha afrontat marginació sota els efectes de la colonització i l'arabització, cosa que ha portat una decadència de l'estatus i l'ús d'aquesta llengua. L'activisme amazic, particularment durant la Primavera Amaziga, va tenir un paper crucial a l'hora de destacar aquestes qüestions i defensar

els drets lingüístics i culturals (Roque, 2009; CIEMEN i Casa Amaziga de Catalunya, 2019). Aquest activisme va portar a un progrés significatiu al Marroc, començant amb la fundació de l'IRCAM el 2001, que va marcar l'inici de la institucionalització de l'amazic. La seva forma estandarditzada, l'amazic estàndard marroquí (codi ISO-639: zgh), va ser desenvolupada per l'IRCAM, combinant característiques dels tres dialectes principals shi, tzm i rif (Boukous, 2014), així com altres variants com el tuàreg. Com a part dels esforços d'estandardització, l'escriptura neo-tifinag, desenvolupada per l'IRCAM i basada en l'escriptura tradicional tifinag, va ser escollida com a escriptura oficial de l'amazic l'any 2003. Aquest sistema gràfic modern, conegut com a *tifinag-IRCAM*, es va preferir a les escriptures llatines i àrabs utilitzades històricament, ja que ofereix un sistema d'escriptura estandarditzat i fonològicament precís per a les diverses varietats de l'amazic marroquí (Soulaimani, 2016; Ataa-Allah i Boulaknadel, 2012).

2.2. *Reptes i progressos en el PLN en amazic*

La complexitat de l'amazic en el PLN sorgeix de la seva escriptura, morfologia i alta variació dialectal. L'existència de diversos alfabetes, incloent el tifinag, el llatí i l'àrab, complica l'estandardització i el processament computacional. A més, la rica morfologia de la llengua, que implica procediments tant de flexió com de derivació, juntament amb diferències significatives dialectals, planteja reptes per desenvolupar aplicacions coherents de PLN. Aquests factors fan que tasques com la desambiguació lèxica, l'anàlisi sintàctica i la traducció automàtica resultin particularment difícils (Ataa-Allah i Boulaknadel, 2012).

Tot i la seva estandardització i entrada al món digital relativament recent, exemplificada pel llançament de la Viquipèdia amaziga el 2023, en els darrers anys el PLN en amazic ha mostrat un progrés considerable. Els treballs notables inclouen la construcció d'un corpus de l'amazic estàndard (Boulaknadel i Ataa Allah, 2013), el desenvolupament d'un corpus anotat morfosintàcticament (Amri et al., 2017) i els avenços en la incrustació de mots en amazic (Faouzi et al., 2023). A més a més, s'ha desenvolupat una eina per a la conjugació dels verbs en amazic, per Ataa-Allah i Boulaknadel (2014), i un programari de concordança i un cercador adaptat al tifinag, per Ataa-Allah i Boulaknadel (2012). Algunes d'aquestes eines i recursos estan disponibles al portal de l'IRCAM¹.

Els treballs de TA de codi obert més notables inclouen el projecte NLLB (*No Language Left Behind*) [Cap Llengua Es Queda Enrere], que incorpora

¹ <https://tal.ircam.ma/talam/>

l'amazic a les 200 llengües que desenvolupa (NLLB Team et al., 2022). Aquest projecte compta amb els conjunts d'entrenament i avaluació de FLORES, que, com també pretén abordar el projecte Awal, requereixen correccions per una millor precisió (Goyal et al., 2022). SIB-200, una extensió de FLORES, ofereix un nou punt de referència per la classificació de temes en 205 llengües, incloent-hi l'amazic (Adelani et al., 2024). El projecte MADLAD, un altre desenvolupament significatiu, ofereix un gran conjunt de dades multilingüe i auditat a nivell de document basat en CommonCrawl, que amplia encara més els recursos disponibles per a llengües de pocs recursos, com ara l'amazic (Kudugunta et al., 2023). A més, el projecte GlotCC proporciona un conjunt de dades monolingüe de domini general derivat de CommonCrawl (Kargaran et al., 2024), que aborda més de 1000 llengües i introdueix GlotLID, un model d'identificació de llengües de codi obert que admet més de 2000 etiquetes i ambdós inclouen l'amazic estàndard marroquí així com altres varietats de l'amazic (Kargaran et al., 2023).

2.3. *Enfocaments impulsats per la comunitat*

El repte fonamental que afronten l'amazic i altres llengües infrarepresentades en l'era digital és clar: sense dades no hi ha la IA. Els models de llenguatge extens i altres tecnologies d'IA depenen de quantitats massives de dades per funcionar de manera eficaç. Per a llengües infrarepresentades, com l'amazic, la solució requereix tant la creació orgànica de continguts—blocs, vídeos i recursos en línia—com els esforços directes de recollida de dades dissenyats específicament per a l'entrenament de la IA.

Les iniciatives impulsades per comunitats de parlants han sorgit com a resposta a la infrarepresentació de les llengües en els espais digitals. Com que aquestes llengües tenen menys presència en línia, molt poques vegades apareixen en rastrejos de web a gran escala que formen la base de les tecnologies lingüístiques modernes. La participació comunitària en la recollida de dades, semblant al model col·laboratiu de la Viquipèdia, ha mostrat èxits en la creació de recursos lingüístics substancials. Alguns exemples importants són Common Voice (Ardila et al., 2020); Tatoeba²; NaijaVoices per l'igbo, el haussa i el ioruba (Emezue et al., 2025); i la recollida de dades del dàrija³.

² <http://tatoeba.org>

³ <https://www.atlasia.ma/>

Aquest tipus d'iniciatives han demostrat que fins i tot les llengües marginalitzades poden aconseguir representació tant en els serveis de tecnologia lingüística comercials com en els models de codi obert quan les comunitats es comprometen seriosament en la tasca de recollida de dades (Emezue et al., 2025; Gonzalez-Agirre et al., 2024). A diferència d'aquesta forma de fer participativa, el desenvolupament de tecnologies lingüístiques per a llengües infrarepresentades sol seguir un abordatge de dalt a baix, impulsat per institucions acadèmiques o empreses tecnològiques amb aportacions limitades per part de la comunitat (Moshagen et al., 2024; Bird, 2020; Schwartz, 2022). Aquesta manera d'abordar-ho corre el risc de perjudicar significativament les llengües i les comunitats en donar una falsa imatge de les llengües mitjançant tecnologies construïdes sense la participació valuosa de les comunitats de parlants. Com adverteix Bird (2020), en tractar el coneixement indígena com una mercaderia, els tecnòlegs lingüístics i de la parla corren el risc de "privar de drets de veu a les autoritats locals del coneixement, reproduint les causes de les amenaces lingüístiques". La qualitat de les dades d'entrenament es fa particularment crucial per a les llengües marginalitzades, en què continguts inexactes poden distorsionar les representacions digitals i propagar errors a través dels sistemes d'IA (Kreutzer et al., 2022; Lau et al., 2025).

Concretament per l'amazic, han sorgit diverses iniciatives comunitàries per omplir aquest buit. La inclusió de l'amazic al Traductor de Google va marcar una fita important, tot i la baixa qualitat inicial de la traducció. No obstant això, els comentaris de retroalimentació i la participació de la comunitat van anar millorant el rendiment del sistema, demostrant la importància de la implicació de les persones parlants en refinar les tecnologies lingüístiques.

La comunitat de PLN en amazic a Hugging Face⁴ representa encara un altre esforç d'organització de base per avançar les tecnologies lingüístiques. Aquesta comunitat de 40 membres té com a objectiu desenvolupar models i eines per l'amazic tant en PLN com en tecnologies de la parla. En l'ecosistema de la Viquipèdia, l'amazic estàndard marroquí va ser llançat el 2024⁵, unint-se a les varietats amazigues existents, incloent-hi el cabilenc (llançada el 2007) i el taixelhit/tassussit (llançada el 2021), amb altres varietats amazigues en fase d'incubació.

⁴ <https://huggingface.co/Tamazight-NLP>

⁵ <https://zgh.wikipedia.org/>

Comprendre aquestes dinàmiques va influir en el disseny del projecte Awal, que busca contribuir a les lliçons apreses a partir dels esforços comunitaris existents i abordar algunes de les seves limitacions.

3. Projecte Awal

El projecte Awal (اولال, “parla” o “discurs” en amazic) representa una iniciativa comunitària centrada en preservar i promoure l’amazic en el món digital. Posat en marxa el 2024 a través de la col·laboració entre entitats catalanes sense ànim de lucre que promouen l’amazic i les tecnologies lingüístiques obertes⁶, el projecte té com a objectiu enfortir les tecnologies lingüístiques en amazic amb un enfocament en la recollida de dades lingüístiques amb proveïment participatiu de persones parlants nadiues.

En aquest capítol es detalla la metodologia del projecte Awal, l’arquitectura de la plataforma, els assoliments actuals i les seves limitacions. Analitzem tant el sistema de recollida de dades de traducció implementat a awaldigital.org com els esforços de recollida de dades de veu a través de Common Voice, i estudiem la seva eficàcia en la creació de recursos lingüístics per al desenvolupament del PLN en amazic.

3.1. Començament i objectius del projecte

El projecte Awal sorgeix a Catalunya, on viuen més de 100.000 amazigoparlants en un territori que té una història de promoció de les seves llengües marginalitzades: el català i l’aranès. Aquest context compta amb el suport institucional i la infraestructura comunitària necessaris per a les iniciatives de desenvolupament de tecnologies lingüístiques, com es pot veure en els projectes Aina⁷ i Araina⁸.

Els objectius principals d’Awal són:

1. preservar i promoure l’amazic a través d’eines digitals innovadores,
2. motivar la participació comunitària en la creació de dades,
3. desenvolupar tecnologies de suport de codi obert, incloent-hi els sistemes de traducció automàtica i reconeixement automàtic de la parla adaptats a l’amazic,
4. proporcionar recursos educatius per a l’aprenentatge d’idiomes, i
5. enfortir la comunicació entre parlants per reforçar la identitat cultural.

⁶ CIEMEN, Casa Amaziga de Catalunya, i Col·lectivaT

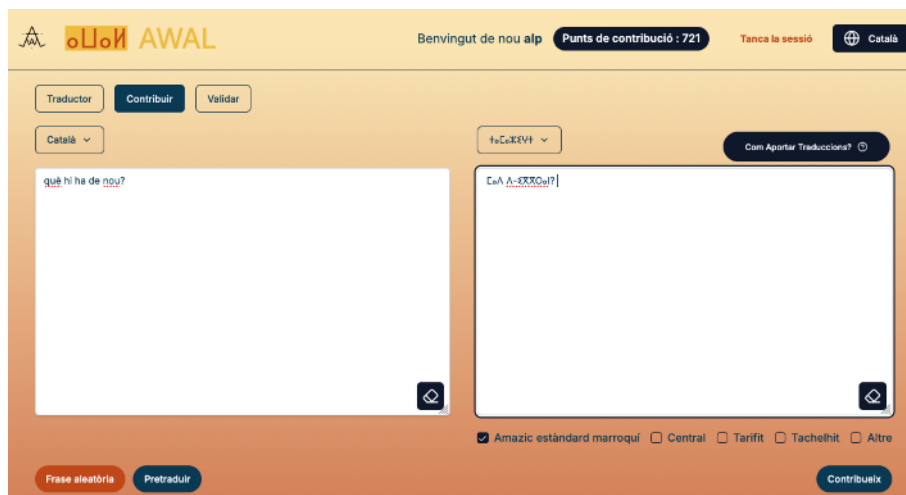
⁷ <https://projecteaina.cat/>

⁸ <https://www.projecte-araina.org/>

Els esforços inicials es van centrar en la recopilació manual de les frases traduïdes, però això va resultar insuficient per ampliar l'escala als milers de parells de frases requerides per una traducció automàtica eficaç. Per tant, el projecte va adoptar un enfocament de plataforma impulsada per la comunitat, que fomenta una participació més àmplia alhora que manté la qualitat mitjançant mecanismes de validació entre iguals.

3.2. *Plataforma awaldigital.org*

La plataforma awaldigital.org⁹ serveix com a espai central per al projecte. Qualsevol persona usuària pot accedir a la informació sobre el projecte i pot utilitzar l'aplicació de traducció automàtica integrada. La imatge 2 mostra la pàgina "Contribuir" on els i les usuàries registrades poden aportar traduccions on la frase de partida apareix a l'esquerra i la d'arribada a la dreta. Els i les usuàries trien la llengua per les dues bandes, i almenys una d'elles ha de ser l'amazic, que es pot escriure tant en escriptura tifinag com en alfabet llatí.



Imatge 2: La pàgina "Contribuir" permet a les usuàries registrades afegir traduccions

Opcionalment, els i les usuàries poden indicar el dialecte en què escriuen. La funció de frases aleatòries carrega al quadre de partida frases d'una base de dades de textos amb llicència Creative Commons.

⁹ <https://awaldigital.org/>

L'opció de pretraduir tradueix automàticament el text de partida a la llengua d'arribada mitjançant el model de traducció automàtica integrat. Després, les persones usuàries han de corregir aquesta traducció abans d'enviar-la, creant un flux de treball de postedició que millora l'eficiència.

Un sistema de gamificació atorga punts per cada entrada de caràcters tant en el quadre de partida com en el d'arribada, i les persones usuàries poden veure la seva qualificació en una taula de classificació. Aquest mecanisme crea un ambient de competència amistosa i fomenta la participació sostinguda.

La interfície de traducció admet la traducció bidireccional entre l'amazic i diverses llengües, incloent-hi el català, el castellà, el francès, l'àrab marroquí i l'anglès. També es permet la contribució de parells de frases en aquestes llengües.

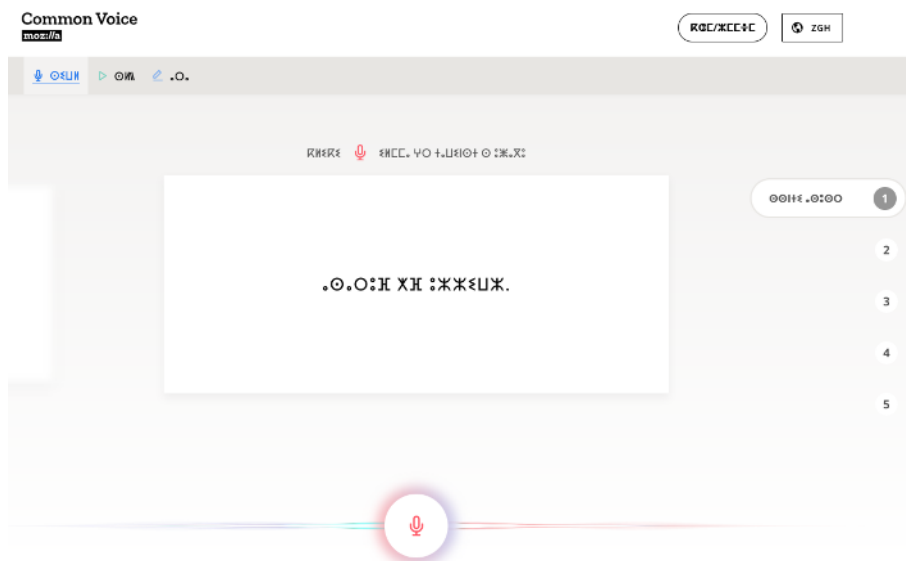


Imatge 3: Pantalla de validació.

El control de qualitat es manté mitjançant la validació entre iguals (Imatge 3), on els i les usuàries revisen les traduccions enviades per altres persones. Les persones validadores utilitzen pautes d'acceptabilitat centrant-se en el significat, la fluïdesa i la precisió gramatical, i dues aprovacions de validació traslladen les entrades al corpus validat a través de l'avaluació entre iguals. Dues aprovacions de validació traslladen les entrades al corpus validat.

La identitat visual de la plataforma reflecteix la identitat cultural amaziga i respon als requisits específics de la llengua, en particular, la seva diversitat dialectal i les variacions d'escriptura. En lloc d'imposar una estandardització

La recollida de dades de veu es va realitzar a través de la plataforma Common Voice de Mozilla¹¹. Aquesta integració va requerir traduir la interfície de la plataforma a l'amazic i poblar-la amb una col·lecció de frases adequada per l'enregistrament. La imatge 4 mostra la pantalla d'enregistrament. Les persones que contribueixen a la recollida de dades llegeixen en veu alta les frases mostrades a la pantalla, creant un corpus de la parla per al desenvolupament del reconeixement automàtic de la parla. A Common Voice també existeix un sistema semblant al de validació entre iguals.



Imatge 4: Aportació de dades de veu en amazic a Common Voice.

¹⁰ La plataforma classifica les contribucions en cinc variants: l'amazic estàndard marroquí, el central, el tarifit, el taixelhit, i altres. Aquestes categories representen un compromís pragmàtic entre distintes perspectives de la comunitat sobre l'estandardització dialectal i exclouen algunes varietats per força, particularment les de fora del Marroc.

¹¹ <https://commonvoice.mozilla.org/zgh>

3.4. *Campanyes de divulgació i participació comunitària*

La difusió d'aquesta iniciativa va incloure campanyes a les xarxes socials a través d'Instagram, Facebook, LinkedIn i canals de Telegram, juntament amb presentacions en esdeveniments culturals, i una datatò emblemàtica.

La Datatò d'Awal, celebrada el 17 de febrer de 2024 a Barcelona, va marcar l'acte de llançament del projecte. Organitzada tant virtualment com presencialment, la datatò, que va durar un cap de setmana, va recollir amb èxit més de 3.500 frases traduïdes i 1 hora d'enregistraments de veu en amazic amb la participació de gent de Catalunya, el Marroc i diferents parts d'Espanya. Més enllà d'aquest esdeveniment, el projecte va organitzar tallers i presentacions dirigits, celebrats a Bilbao, Tortosa i Vic.

3.5. *Les estadístiques actuals i l'accés a les dades*

A data de 13 de juny de 2025, la plataforma compta amb 286 usuaris i usuàries registrades, tot i que només 66 (23%) han contribuït a les traduccions a la plataforma. La resta d'usuàries es van registrar però no van aportar cap contingut. Des de la seva creació el gener de 2024, la plataforma ha recollit un total de 6.421 contribucions, de les quals 2.182 (34%) estan validades. Entre els i les usuàries que hi han contribuït, la mitjana és de 86 aportacions per persona.

El nombre de traduccions recollides per cada parell es detalla a la taula 1. Català-amazic (llatí) apareix com el parell més productiu amb 1.718 aportacions, seguit d'anglès-amazic (tifinag) amb 1.539 aportacions i castellà-amazic (llatí) amb 1.203 aportacions.

A més de les traduccions de text, Awal ha recollit 3 hores de dades de veu a través de la plataforma Common Voice, 2 hores de les quals estan validades¹².

	tifinag	llatí	TOTAL
ar	11	10	21
ary	311	136	447
ca	395	1718	2113
es	90	1203	1293
en	1539	693	2232
fr	83	232	315

¹² Publicació de Common Voice Corpus 22.0

TOTAL	2429	3992	6421
--------------	------	------	------

Taula 1: Contribució en cada parell de llengües a dia del 13.06.2025.

Compartim totes les dades de text paral·lel i monolingüe a través del nostre repositori de Hugging Face¹³ mentre que a les dades de veu s'hi pot accedir directament des de la plataforma Common Voice¹⁴. Les extraccions periòdiques de dades de la plataforma Awal garanteixen que el repositori es mantingui actualitzat amb noves contribucions. Les mètriques de dades en temps real estan disponibles a la pàgina d'inici d'Awal.

3.6. *Limitacions, reptes i camí a seguir*

Després de divuit mesos de participació comunitària, el projecte Awal ha revelat diversos reptes que afecten la participació i l'eficàcia de la recollida de dades. Mitjançant entrevistes amb promotores i col·laboradores, hem identificat barreres clau que expliquen per què la iniciativa no ha arribat als nivells de participació esperats inicialment. La nostra anàlisi revela sis conclusions principals:

1. **El públic general amazigoparlant no confia en les seves pròpies habilitats de lectura i escriptura.** La majoria de les persones no saben escriure en alfabet tifinag i mai no han tingut l'oportunitat d'aprendre'n. Malgrat parlar amb fluïdesa, moltes membres de la comunitat expressen una profunda incertesa sobre les seves habilitats d'escriure. Fins i tot aquelles que escriuen regularment en amazic en alfabet llatí en grups de xat amb la família o a les xarxes socials consideren que la seva redacció informal és inadequada per a la recollida de dades. Els i les participants dels tallers sovint van expressar que "no escri[uen] correctament" o "no cone[ixen] la forma estàndard", cosa que les porta a excloure's a si mateixes de contribuir a un espai on la seva escriptura seria registrada i vista per altri.
2. **Les tasques de micro-traducció resulten una manera prometedora per a la creació de dades.** Els esforços activistes-lingüistes actuals se centren en animar la gent a escriure, i la traducció resulta prometedora per afavorir la creació de continguts. En lloc de demanar a les persones participants que generin textos

¹³ <https://huggingface.co/datasets/collectivat/amazic>

¹⁴ <https://commonvoice.mozilla.org/ca/datasets>

originals, proporcionar material de partida en altres llengües—o fins i tot contingut generat per la IA—elimina la barrera creativa que impedeix la contribució de molta gent. Aquest plantejament ha mostrat èxit en plataformes com Tatoeba, on la traducció col·laborativa va resultar en un contingut substancial de l'amazic en múltiples varietats. La funcionalitat de càrrega automàtica de frases de la plataforma Awal aborda aquest problema eliminant l'obstacle de "què hauria d'escriure?", cosa que bloqueja la participació.

3. **La majoria de les contribucions provenen de membres de la comunitat acadèmica, lingüística i activista.** Tot i que vam rebre moltes respostes positives per part del públic general durant els esdeveniments, el major impacte en termes de volum de dades va provenir del món activista i acadèmic que compren la importància de la digitalització lingüística—persones que ja treballen en associacions culturals, investigadores o persones amb educació lingüística formal. El model de proveïment participatiu suposa un nivell d'alfabetització i estandardització que no coincideix amb la realitat actual de la comunitat general amazigoparlant, particularment en la diàspora. El projecte ha de centrar-se en aquest públic especialitzat si desitja generar impacte en termes d'extensió de dades, alhora que manté la difusió al públic en general perquè més persones coneguin la tecnologia i estiguin motivades per desenvolupar les seves habilitats de lectura i escriptura.
4. **No hi ha consens sobre la representació de dialectes a la plataforma.** El projecte tenia com a objectiu acollir la diversitat dialectal tot mantenint la qualitat de dades mitjançant un sistema d'etiquetatge de contribucions que permetia a les usuàries indicar el seu dialecte. Però aquest plantejament va revelar tensions dins l'ecosistema lingüístic amazic. L'amazic estàndard marroquí segueix sent desconegut per a moltes parlants de la diàspora i per a aquelles que van aprendre la llengua oralment a casa. Mentrestant, la comunitat acadèmica recomana no utilitzar cap etiqueta dialectal, tenint en compte que ja hi ha unitat en la forma escrita.
5. **L'alternança de codis (*code-switching*) i la puresa lingüística presenten reptes per a la qualitat de dades.** Les comunitats de la diàspora sovint incorporen mots de les seves llengües de residència, mentre que els i les parlants al Marroc sovint barregen el dàrija amb el seu amazic. Trobar equivalents reals en amazic requereix una àmplia investigació transversal de dialectes i regions, incloent-hi consultar variants d'altres països—una tasca que exigeix més expertesa lingüística que coneixement casual. Vam observar que les dones grans de la diàspora mantenien formes dialectals més pures en

comparació amb els i les parlants més joves, cosa que suggereix que els efectes del contacte lingüístic s'intensifiquen transversalment entre les generacions i la distància geogràfica de les comunitats d'origen.

6. **Un abordatge més internacionalista podria generar més impacte.** Els materials promocionats, principalment en català, limitaven abastar la diàspora amaziga internacional en països com Bèlgica, França i altres regions, i també impedié arribar a la població al Marroc, la qual cosa hauria ajudat a ampliar la participació de manera massiva. Això requeriria la col·laboració amb entitats més enllà de Catalunya, cosa que el camp d'acció actual no permet.

Malgrat aquests reptes, el projecte ha demostrat valuosos camins a seguir. La recepció positiva i la consciència generalitzada creada per Awal indiquen un fort interès de la comunitat de parla amaziga en el desenvolupament de tecnologies lingüístiques. Moltes persones, gràcies a aquesta iniciativa, es van sorprendre i motivar amb la idea que les tecnologies avançades també poden existir en la seva llengua materna.

Com a camí a seguir, l'enfocament més prometedori implica concentrar els esforços en les comunitats acadèmiques i activistes que disposen tant d'expertesa lingüística com d'alfabetització tecnològica per a la preservació digital de llengües. Aquestes col·laboradores poden construir conjunts de dades fonamentals mentre el projecte desenvolupa components pedagògics complementaris que generin confiança amb l'escriptura en la comunitat en general. A més, adaptar la plataforma perquè accepti l'entrada d'àudio juntament amb el text podria reduir la complexitat de tenir contribucions en dues plataformes diferents. Els reptes identificats aquí també apunten cap a la necessitat de col·laboracions institucionals a més llarg termini amb institucions educatives al Marroc, on els esforços d'estandardització són més avançats i la formació d'escriptura està integrada en l'educació formal¹⁵.

4. Conclusió

El projecte Awal demostra que el desenvolupament comunitari de tecnologies lingüístiques per a llengües infrarepresentades requereix l'adaptació a les realitats lingüístiques locals i no pot replicar simplement els models dissenyats per a llengües estandarditzades. Si bé el nostre experiment de 18 mesos va recollir dades significatives i va augmentar la consciència de

¹⁵ Malgrat els múltiples intents d'establir una col·laboració amb l'IRCAM, no vam rebre resposta per part d'aquesta institució.

la comunitat, va revelar tensions fonamentals entre els supòsits de proveïment participatiu i el complicat context sociolingüístic de les amazigoparlants. Paral·lelament, estem organitzant i corregint conjunts de dades estàndards de TA i afinant els models de TA a partir de les dades recollides. Com a iniciativa finançada amb fons públics a la diàspora, Awal destaca la necessitat d'obrir col·laboracions transfrontereres que reflecteixin la realitat transnacional de les comunitats amazigoparlants. Exigim un major suport institucional que combini els esforços comunitaris de base amb la planificació lingüística formal, i que vagi més enllà d'iniciatives aïllades cap a un desenvolupament coordinat de les tecnologies lingüístiques que serveixin a les persones parlants i permetin travessar fronteres i generacions.

5. Agraïments

Aquest article s'ha redactat com a resultat d'una recerca duta a terme en el marc del projecte Som Part del CIEMEN i la Fundació pels Drets Col·lectius dels Pobles, que va rebre finançament de l'Agència Catalana de Cooperació al Desenvolupament (ACCD) i de l'Ajuntament de Barcelona.

Agraïm sincerament a Lalla Ghizlan Baryala, Brahim Essaidi, Mohamed Aymane Farhi i Naceur Jabouja per l'assessorament i la perspectiva d'inestimable valor que han aportat sobre la realitat amaziga al desenvolupament d'aquest article.

També donem les gràcies a tots i totes les voluntàries que van contribuir a la plataforma Awal i van participar en els nostres tallers, compartint les seves experiències i perspectives sobre el seu coneixement de l'amazic i la tecnologia.

Finalment, també volem agrair a Pelin Dogan per la traducció al català d'aquest article escrit originalment en anglès.

Referències

Adelani D., Liu H., Shen X., Vassilyev N., Alabi J., Mao Y., Gao H. i Lee E. (2024). SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects. En Graham Y. i Purver M., editors, *Proc. of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Vol. 1: Long Papers)*, pp. 226-245. Association for Computational Linguistics.

Amri S., Zenkouar L. i Outahajala M. (2017). Build a Morphosyntactically Annotated Amazigh Corpus. En *Proceedings of the 2nd International Conference on Big Data, Cloud and Applications*, article 8. Association for Computing Machinery.

Ardila R., Branson M., Davis K., Kohler M., Meyer J., Henretty M., Morais R., Saunders L., Tyers F. i Weber G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. En Calzolari N. et al., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218-4222. European Language Resources Association.

Ataa-Allah F. i Boulaknadel S. (2012). Toward Computational Processing of Less Resourced Languages: Primarily Experiments for Moroccan Amazigh Language. En Sakurai S., editor, *Theory and Applications for Advanced Text Mining*, capítol 9. IntechOpen.

Ataa-Allah F. i Boulaknadel S. (2014). Amazigh Verb Conjugator. En *International Conference on Language Resources and Evaluation*.

Bird S. (2020). Decolonising Speech and Language Technology. En Scott D., Bel N. i Zong C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3504-3519, Barcelona, Espanya (en línia). International Committee on Computational Linguistics.

Boukous A. (2014). The Planning of Standardizing Amazigh Language: The Moroccan Experience. *Iles d'Imesli*, vol. 6(1): 7-23.

Boulaknadel S. i Ataa Allah F. (2013). Building a Standard Amazigh Corpus. En Kudělka M., Pokorný J., Snášel V. i Abraham A., editors, *Proceedings of the Third International Conference on Intelligent Human Computer Interaction (IHCI 2011)*, Prague, Czech Republic, August, 2011, pp. 91-98. Springer Berlin Heidelberg.

CIEMEN i Casa Amaziga de Catalunya (2019). *El poble amazic a Catalunya*. CIEMEN.

Emezue C., NaijaVoices Community, Awobade B., Owodunni A., Emezue H., Emezue G. M. T., Emezue N. N., Ogun S., Akinremi B., Adelani D. I. i Pal C. (2025). The NaijaVoices Dataset: Cultivating Large-Scale, High-Quality, Culturally-Rich Speech Data for African Languages. En *Proc. of Interspeech 2025*, Rotterdam, Països Baixos.

Faouzi H., El-Badaoui M., Boutalline M., Tannouche A. i Ouanan H. (2023). Towards Amazigh Word Embedding: Corpus Creation and Word2Vec Models Evaluations. *Revue d'Intelligence Artificielle*, vol. 37(3): 753-759.

Goyal N., Gao C., Chaudhary V., Chen P., Wenzek G., Ju D., Krishnan S., Ranzato M., Guzmán F. i Fan A. (2022). The Flores-101 Evaluation Benchmark for Low-

Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, vol. 10.

Kargaran A., Imani A., Yvon F. i Schuetze H. (2023). GlotLID: Language Identification for Low-Resource Languages. En *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

Kargaran A. H., Yvon F. i Schütze H. (2025). GlotCC: an open broad-coverage CommonCrawl corpus and pipeline for minority languages. En *Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24)*, Vol. 37, pp. 16983–17005, Red Hook, NY, EEUU. Curran Associates Inc.

Kreutzer J., Caswell I., Wang L., Wahab A., van Esch D., Ulzii-Orshikh N., Tapo A., Subramani N., Sokolov A., Sikasote C., Setyawan M., Sarin S., Samb S., Sagot B., Rivera C., Rios A., Papadimitriou I., Osei S., Ortiz Suarez P., Orife I., Ogueji K., Niyongabo Rubungo A., Nguyen T. Q., Müller M., Müller A., Muhammad S. H., Muhammad N., Mnyakeni A., Mirzakhlov J., Matangira T., Leong C., Lawson N., Kudugunta S., Jernite Y., Jenny M., Firat O., Dossou B. F. P., Dlamini S., de Silva N., Çabuk Balı S., Biderman S., Battisti A., Baruwa A., Bapna A., Baljekar P., Azime I. A., Awokoya A., Ataman D., Ahia O., Ahia O., Agrawal S. i Adeyemi M. (2022). Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, vol. 10: 50-72.

Kudugunta S., Caswell I., Zhang B., Garcia X., Xin D., Kusupati A., Stella R., Bapna A. i Firat O. (2023). MADLAD-400: a multilingual and document-level large audited dataset. En *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, pp. 67284–67296, Red Hook, NY, EEUU. Curran Associates Inc.

Lafkoui M. B. (2018). Berber Languages and Linguistics. *Oxford Bibliographies*.

Lau M., Chen Q., Fang Y., Xu T., Chen T. i Golik P. (2025). Data Quality Issues in Multilingual Speech Datasets: The Need for Sociolinguistic Awareness and Proactive Language Planning. En *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, Viena, Àustria, 27 de juliol–1 d'agost 1. Association for Computational Linguistics.

Moshagen, S. N., Antonsen, L., Wiecheteck, L., i Trosterud, T. (2024). Indigenous language technology in the age of machine learning. *Acta Borealia*, 41(2), 102–116.

Múrcia, Carles. "La llengua amaziga." *amazic.cat*, 2017. <https://www.amazic.cat/wp-content/uploads/2017/02/1.-La-llengua-amaziga.pdf>

Nekoto W., Marivate V., Matsila T., Fasubaa T., Fagbohungebe T., Akinola S. O., Muhammad S., Kabenamualu S. K., Osei S., Sackey F., Niyongabo R. A., Macharm

R., Ogayo P., Ahia O., Berhe M. M., Adeyemi M., Mokgesi-Seling M., Okegbemi L., Martinus L., Tajudeen K., Degila K., Ogueji K., Siminyu K., Kreutzer J., Webster J., Ali J. T., Abbott J., Orife I., Ezeani I., Dangana I. A., Kamper H., Elsahar H., Duru G., Kioko G., Espoir M., van Biljon E., Whitenack D., Onyefuluchi C., Emezue C. C., Dossou B. F. P., Sibanda B., Bassey B., Olabiyi A., Ramkilowan A., Öktem A., Akinfaderin A. i Bashir A. (2020). Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. En *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp 2144–2160, en línia. Association for Computational Linguistics.

NLLB Team, Costa-jussà M. R., Cross J., Çelebi O., Elbayad M., Heafield K., Heffernan K., Kalbassi E., Lam J., Licht D., Maillard J., Sun A., Wang S., Wenzek G., Youngblood A., Akula B., Barrault L., Mejia-Gonzalez G., Hansanti P., Hoffman J., Jarrett S., Sadagopan K. R., Rowe D., Spruit S., Tran C., Andrews P., Ayan N. F., Bhosale S., Edunov S., Fan A., Gao C., Goswami V., Guzmán F., Koehn P., Mourachko A., Ropers C., Saleem S., Schwenk H. i Wang J. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:1902.01382*.

Roque M. (2009). *Els amazigs avui, la cultura berber*. Pagès editors/IEMed.

Schwartz L. (2022). Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. En *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 2: Short Papers)*, pp. 724–731, Dublín, Irlanda. Association for Computational Linguistics.

Soulaimani D. (2016). Writing and rewriting Amazigh/Berber identity: Orthographies and language ideologies. *Writing Systems Research*, vol. 8(1): 1–16.