

L'ús d'eines d'Intel·ligència Artificial generativa per la internacionalització de continguts en català

Karina Gibert, Intelligent Data Science and Artificial Research Center
(Universitat Politècnica de Catalunya IDEAI-UPC)

Xavier Angerri (IDEAI-UPC)

Sergi Ramírez (IDEAI-UPC)



CIDAI - POC-2025-06

Drets reservats. Aquest treball està disponible sota la llicència Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International ([CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/)).

Segons els termes d'aquesta llicència, podeu copiar, redistribuir i adaptar l'obra amb fins no comercials, sempre que l'obra sigui citada adequadament, tal com s'indica a continuació.

En qualsevol ús d'aquest treball, no s'ha de suggerir que el CIDAI doni suport a cap organització, producte o servei específic. No es permet l'ús del logotip CIDAI.

Si adapteu l'obra, heu de llicenciar-la amb la mateixa llicència Creative Commons o equivalent.

Si creeu una traducció d'aquest treball, heu d'afegir la següent exempció de responsabilitat juntament amb la cita suggerida: "Aquesta traducció no la va crear el Centre of Innovation for Data tech and Artificial Intelligence (CIDAI). CIDAI no es fa responsable del contingut ni de l'exactitud d'aquesta traducció. L'edició original en català serà l'edició autèntica i vinculant".

Qualsevol mediació relacionada amb disputes derivades de la llicència es durà a terme d'acord amb les normes de mediació de la World Intellectual Property Organization.

Cita suggerida: CIDAI - POC-2025-06 // L'ús d'eines d'Intel·ligència Artificial generativa per la internacionalització de continguts en Català per Karina Gibert, Xavier Angerri i Sergi Ramírez – IDEAI UPC, CIDAI, 2025. Llicència: CC BY-NC-SA 4.0



1.	ABSTRACT	4
2.	DESCRIPCIÓ DEL PROBLEMA QUE RESOL	8
3.	IMPLEMENTACIÓ DE LA SOLUCIÓ	13
4.	POTENCIAL IMPACTE DE LA SOLUCIÓ	20
5.	BIBLIOGRAFIA	21
6.	AUTORIA I AGRAÏMENTS	23



1. Abstract

1.1. Objectius

Aquest treball enfoca la utilització de grans models de llenguatge per a la transcripció de textos en català dins l'àmbit de la intel·ligència artificial. El català és una llengua mitjana que té molt poca presència en l'àmbit digital i hi ha un ampli marge de millora en la generació i maduresa dels recursos de lingüística computacional per a la llengua catalana [Bassa 2023]. Actualment, en plena transformació digital disruptiva, la presència de totes les cultures i llengües a Internet i altres marcs digitals és crucial. Generar contingut digital en català és una prioritat. Però malauradament, prioritzar la generació de contingut digital en català implica moltes vegades perdre impacte ja que restringeix l'audiència a la que domina aquesta llengua i moltes vegades provoca que l'autor, que podria perfectament produir contingut en català, opti per fer-ho en castellà o en anglès per arribar a un públic més ampli. Amb el desenvolupament de la Intel·ligència Artificial generativa s'obre una finestra d'oportunitat molt estratègica, no només per facilitar la internacionalització dels continguts generats genuïnament en català, sinó per també fer que la llengua catalana sigui coneguda fora del seu territori.

Una forma d'incrementar l'impacte del contingut en català a altres audiències és utilitzar tècniques de lingüística computacional o processament del llenguatge natural per donar accessibilitat a parlants d'altres llengües als continguts en català. En aquest context, un activitat central serà la transcripció de vídeos locutats en català a subtítols en la pròpia llengua catalana, que eventualment permetin la traducció automàtica a altres idiomes. Fins ben recentment, el canal més popular de YouTube encara no oferia la transcripció en llengua catalana d'un vídeo

i no podia, en conseqüència, generar subtítols en Català, i això bloqueja també la possibilitat d'obtenir subtítols traduïts a altres llengües. A Catalunya, s'han posat en marxa algunes polítiques específiques de promoció de la innovació i millora dels recursos computacionals de català [Generalitat 2022], o el projecte AINA [BSC 2024b] que cobreixen aquestes mancances.

En aquest treball, partim d'una anàlisi de l'estat d'art en transcripció de vídeos en català per generar subtítols en tres llengües (Català, Castellà i Anglès). Eventualment, el corpus analitzat conté algun vídeo aïllat en castellà. El treball testreja rigorosament alguns recursos per discernir-ne l'eficàcia i precisió fent transcripcions de veu a text en català i traduccions de català a altres idiomes. Malauradament, les presumptes capacitats d'aquests models no s'ajusten tan com s'esperaria quan parlem de veus reals, l'experimentació plantejada revela unes prestacions inferiors a les esperades, ja que aquests sistemes encara són molt i molt sensibles a les característiques de veu i locució dels parlants. Els principals reptes que hem trobat estan relacionats amb la vocalització de qui locuta, el reconeixement del vocabulari tècnic o especialitzat, especialment en matèria tecnològica, les flexions dialectals o fins i tot la qualitat del senyal, amb més o menys soroll de fons, etc. En aquest article es descriuen els resultats de comparar algunes eines. També es presenta una metodologia estandarditzada per poder subtitular vídeos en català en altres llengües de forma gairebé automàtica. A partir de la comparativa hem pogut extreure algunes conclusions d'interès i hem identificat la diagnosi de limitacions i fortaleeses de les diferents eines a partir de l'aplicació als vídeos reals de dos corpus de diferents característiques relacionats amb accions recurrents dels equips dirigits per Karina Gibert. En concret, es presenten resultats sobre els vídeos de Top Rosies Talent i de Women in Data Science Barcelona

1.2. Visió

La proposta que presentem es basa en la formulació d'una metodologia que permet subtitular vídeos en català a l'anglès utilitzant una combinació de diferents eines de lingüística computacional que permeten internacionalitzar els continguts generats genuïnament en català. Aquest és un enfoc que no s'ha explotat gaire fins a la data perquè el que estem acostumats és a que ens tradueixin al català (en general sota doblatge) els continguts generats en altres llengües, però no és habitual que els continguts generats en català es puguin visionar fora del nostre territori amb subtítols en altres llengües. A part d'ampliar l'audiència, i permetre la internacionalització de continguts generats aquí i locutats en català originalment, l'opció de subtitular aquests continguts a d'altres idiomes permet que la llengua catalana es faci present en el circuit internacional. I que l'audiència estrangera pugui sentir la llengua catalana mentre llegeix els subtítols en la seva pròpia llengua, contribuint així a la difusió del català fora de les nostres fronteres, però també a la presència digital del català i trencant el dilema habitual quan un catalanoparlant es planteja generar continguts digitals entre fer-ho en la pròpia llengua o maximitzar l'impacte del contingut generant-lo en castellà o anglès.

Quan es parla de la internacionalització de continguts hi ha moltes possibilitats però en aquest treball descartarem la de doblar els vídeos catalans a d'altres idiomes, opció viable amb l'actual estat de maduresa de la tecnologia, i optem per l'opció de subtitular-los per mantenir així la possibilitat que quan aquests continguts arribin a persones no catalanoparlants, els puguin consumir llegint els subtítols, però no perdin l'exposició a la llengua catalana perquè continuïn exposats a l'àudio original en català.



1.3. Resum del problema

El treball resol dos problemes simultàniament:

- Permet que tothom que vulgui generi continguts en català sense haver de pensar si fer-ho representa una minimització de l'audiència capacitada per consumir aquell contingut, amb fortes implicacions a nivell d'internacionalització del coneixement que es genera genuïnament en català majoritàriament des de Catalunya
- Per una altra banda permet també que fora de les nostres fronteres es pugui sentir la llengua catalana, augmentant doncs l'espai d'exposició de l'audiència a la llengua catalana, amb totes les implicacions per la supervivència i bona salut del català que això representa. És per aquest motiu que la metodologia que es planteja en aquest treball no contempla en cap moment l'ús de les eines d'IA generativa per doblar la locució a un altre idioma, conscients com som que en aquests moments la tecnologia és prou madura com per a fer-ho fins i tot imitant la pròpia veu de la persona que s'està doblant, i modificant com mou la boca per a que resulti molt i molt realista.

1.4. Solució

La solució que es proposa, com dèiem, està basada en una metodologia que combina diverses peces de lingüística computacional per generar subtítols en castellà o anglès per vídeos en català, explorant la possibilitat d'utilitzar eines de traducció veu-text o transcripció veu-text, traducció text-text.

Eventualment a futur, el text traduït que actualment es situa com subtítols del vídeo original (que manté la locució en català) es podria perfectament connectar a un sintetitzador que pugui llegir el text en veu alta i doblar automàticament els vídeos, tot i que, com hem dit, aquesta és una solució que no forma part de la nostra proposta perquè impedeix que l'audiència d'altres llocs entri en contacte amb la llengua catalana.

La proposta en realitat és escalable a la internacionalització de contingut entre qualsevol parell de llengües, si bé plantejem una utilització clarament enfocada a generar subtítols en castellà de vídeos originalment en català.

La metodologia que es proposa s'ha aplicat a la internacionalització dels materials audiovisuals formatius del programa de generació de talent femení en Intel·ligència Artificial anomenat Top Rosies Talent, per generar una actualització de la plataforma d'e-learning que l'hostatja, on és possible activar els subtítols en castellà de les píndoles formatives, i això obre la possibilitat de realitzar edicions del programa Top Rosies Talent a països centre i sud-americans, o qualsevol altre territori de parla hispana.

D'altra banda, aquesta mateixa metodologia s'ha pogut aplicar amb èxit també a la subtitulació de les 4 edicions ja celebrades del Women in Data Science-Barcelona, WiDS-Barcelona, però en aquest cas l'idioma de subtitulació no és el castellà sinó l'anglès. I això ha permès exportar els interessantíssims debats, xerrades i taules rodones que es celebren en el marc d'aquest esdeveniment a la comunitat internacional del WiDS-WorldWide amb origen a la universitat d'Stanford.

Per últim, s'ha utilitzat aquesta mateixa metodologia en la traducció a l'anglès dels vídeos d'una única professora, locutats en català, d'una assignatura del grau d'intel·ligència artificial de la UPC.

2. Descripció del problema que resol

2.1. Top Rosies Talent

El programa Top Rosies Talent és un programa de [generació de talent femení en Intel·ligència Artificial \(IA\)](#), gratuït, que s'adreça a noies que estan acabant o són recent graduades en una carrera STEM.

Impulsat per les donesIDEAI, de l'Intelligent Data Science and Artificial Intelligence Research Center (IDEAI) de la UPC, en col·laboració amb donesCOEINF, la comissió de gènere del Col·legi Oficial d'Enginyeria Informàtica de Catalunya, compta amb l'inestimable suport de la Facultat d'Informàtica de Barcelona (FIB-UPC) i en les seves 3 edicions ha reunit el recolzament de diferents partners empresarials de l'ecosistema TIC de Catalunya.

El programa inclou una formació en IA en línia dissenyada especialment per les donesIDEAI amb perspectiva de gènere, que es cursa de forma asíncrona al ritme de cada participant, i que actualment compta amb 30 hores de formació. Al llarg d'aquesta formació, que es desenvolupa en un arc temporal d'uns tres mesos, les participants segueixen en paral·lel un cicle de sessions amb role-models, eminentment presencials i algunes activitats de connexió amb l'ecosistema TIC de Catalunya. Acabada la fase formativa, entren en la fase de reptes, proposats conjuntament per la direcció de Top Rosies Talent i les empreses partners de cada edició, que sempre són empreses rellevants del sector, i les professores d'IDEAI ofereixen mentories durant el desenvolupament dels reptes, que dura aproximadament un mes. Les guanyadores de cada repte opten a una beca per cursar el interuniversitari oficial en intel·ligència artificial de la (UPC-UB-URV) coordinat per la FIB-UPC, o una internship a l'empresa que havia proposat el repte. El programa està impulsat per Karina Gibert, donesIDEAI i donesCOEINF.

2.2. El WiDS-Barcelona

Women in Data Science (WiDS) Worldwide és un moviment internacional impulsat des de la Universitat d'Stanford basat en potenciar el paper de la dona en el sector de la ciència de dades i contribuir a la reducció de la preocupant escletxa de gènere del sector. Objectius completament alineats amb donesIDEAI i donesIACat.

Aquest moviment estructura una xarxa internacional de dones de l'àmbit de la ciència de les dades que realitza en diferents activitats per complir la missió del moviment i l'estructura organitzativa es basa en el nomenament d'ambaixadores arreu del món que difonen els objectius, principis i activitats de WiDS. A Barcelona, el 2020 Karina Gibert i Paula Pretone van establir una primera ambaixada WiDS-Barcelona en nom de donesCOEINF i en col·laboració amb la FIB-UPC que te la seu al COEINF i va inaugurar una llinatge d'esdeveniments anuals que es realitzen cada primavera a la FIB des de 2021. A partir de 2022 les ambaixadores del WiDS-Barcelona són Karina Gibert i Núria Castell, ambdues donesCOEINF, donesIDEAI i amb estreta vinculació amb la FIB. WiDS-Barcelona 2025 es celebra el proper 21 de maig en la seva 5a edició.

Els esdeveniments de WiDS estan directament supervisats per la cúpula del moviment WiDS Worldwide que aproven el programa, les ponents i vetllen per a que els principis del WiDS es respectin en tots i cadascun dels esdeveniments de la xarxa a tot el món. El de WiDS-Barcelona no és una excepció. Val a dir que amb el temps, a Barcelona es realitzen més esdeveniments de WiDS organitzats per altres equips i que en aquest treball parlarem únicament de la sèrie WiDS-Barcelona.

Els WiDs Barcelona se celebren sempre a la FIB. Solen constar d'una obertura institucional amb autoritats acadèmiques i polítiques i una o dues taules rodones amb ponents expertes en ciència de les dades de l'ecosistema català o bé una taula rodona i una conferència. L'esdeveniment està obert a tothom i al llarg de les seves 5 edicions ha tocat temes variats, segons l'actualitat del moment com les aplicacions a la salut, l'assetjament a les xarxes, o la ciberseguretat o les organitzacions data-intensive i a l'edició de 2025 es tracta el tema de les dades a la indústria 4.0. En general hem tingut sempre un slot dedicat a temes de caràcter tècnic i professional i l'altre dedicat a qüestions de gènere, com la manca de talent femení, el lideratge femení, l'impacte de sostres de vidre i terres enganxosos a les professions de ciència de dades etc. Les normes de WiDS worldwide són estrictes. Els patrocinadors no poden proposar i portar speakers. Totes les ponents han de ser dones professionals de l'àmbit de la ciència de les dades i per incorporar una ponent que només gestiona aquests equips o realitza funcions directives però no és tècnica, es necessita un permís especial de l'organització WiDS Worldwide.

Per WiDS-Barcelona i les seves ambaixadores Karina Gibert i Núria Castell el WiDS-Barcelona compleix un objectiu prioritari, que és reforçar la xarxa de dones de la ciència de les dades de l'ecosistema local català. Per això totes les edicions del WiDS-Barcelona incorporen espais de networking en forma de cocktails, dinars o sopars i per això també, malgrat que WiDS és un moviment internacional hem tingut molt d'interès a celebrar la jornada majoritàriament en català (més enllà que alguna ponent s'expressa millor en castellà i ha utilitzat aquesta llengua, o que alguna convidada internacional ha fet la conferència en anglès en alguna de les edicions).



Com que el moviment és internacional, i els esdeveniments de les diferents ambaixadores es difonen a tota la xarxa a nivell mundial, la internacionalització dels debats del WiDS-Barcelona és imprescindible.

En un bon nombre d'ocasions el WiDS-Barcelona s'ha celebrat una o 2 setmanes abans del dia de la dona a les TIC i s'ha gravat tot l'esdeveniment (a més de difondre'l en streaming en temps real) per després subtítular-lo en anglès i el dia de la dona a les TIC fer la difusió internacional a través de les xarxes, ja incorporant els vídeos en anglès.

2.3. La internacionalització de continguts de formació reglada

En aquest cas, estem parlant de 15 vídeos de 2h gravats en classes reals d'una única assignatura del grau d'intel·ligència Artificial de la UPC. Hi ha una única veu locutant en tota la sèrie i el llenguatge no només és tècnic, sinó que té un grau important d'especialització en la matèria impartida.

2.4. El problema

El problema que resolem té a veure amb una voluntat de poder generar continguts audiovisuals en català de forma genuïna i fomentar l'ús i presència del català a les xarxes sense que això necessàriament signifiqui confinar el consum d'aquests continguts a la població de parla catalana, concentrada en una petitíssima part del món i per sota dels 8 milions. El problema que es vol resoldre és el de poder subtitular vídeos locutats en català amb altres idiomes, per obrir els continguts en català a altres territoris, però mantenint la locució en català i augmentant el contacte de la resta del món amb la llengua catalana dins l'àmbit digital. En concret per Top Rosies Talent s'ha aplicat la metodologia que proposem a generar subtítols en català i anglès, i per WiDS-Barcelona a generar subtítols en anglès, i per l'assignatura del grau de IA també s'ha volgut la traducció a la qual cosa obre la porta a altres idiomes a partir del pivotatge sobre l'anglès.

2.5. Proposta metodològica

Generació de subtítols en castellà de vídeos en català per Top Rosies Talent:

La metodologia proposada segueix les següents passes:

1. En primera instància, seleccionar una mostra aleatòria dels vídeos en la llengua original que es volen tractar per a provar l'eficàcia de diferents mecanismes i diferents eines disponibles al mercat per transcriure, traduir i subtitular veu. L'estat de l'art evoluciona vertiginosament i a cada nova aplicació d'aquests processos caldria reavaluar quines són les tècniques que millors prestacions mostren per un determinat binomi llengua-origen/llengua.destí.

Pel cas que ens ocupa amb Top Rosies Talent, s'han seleccionat 8 vídeos dels 122 locutats en català, de diferents locutors. I es van fer proves amb 9 eines diferents com Whisper, NeMo, YouTube, SoftCatalà, Sonix, Transkriptor, Happy Scribe i SeamlessM4T.

2. Extreure la pista d'àudio del vídeo

3. Provar dues metodologies diferents, ja que segons evoluciona el sector, pot ser que en un futur proper la traducció veu-text quan el català és la llengua d'origen tingui prou qualitat. En aquest sentit és molt important seleccionar una eina que no faci la traducció pivotant internament a l'anglès perquè els resultats finals presenten una pèrdua molt significativa de qualitat:

- Traducció de veu en català a text en Castellà o Anglès. Per Top Rosies Talent es van provar dues eines (Whisper i Google translation)
 - Transcripció de veu en català a text en català seguit d'una traducció automàtica de text en català a text en Castellà o Anglès, ja que les eines text-a-text estan més desenvolupades i tradueixen amb major qualitat que les que parteixen de l'àudio.
- En aquest cas les eines que es van provar van ser NLLB i AINA.

4. Elecció de l'estratègia a utilitzar de forma generalitzada amb tota el corpus de vídeos.

En el moment de realitzar aquest projecte, l'opció 2.2 va resultar sensiblement millor quan la llengua d'origen va ser el català, mentre que per vídeos amb llengua d'origen en castellà l'opció 2.1 va resultar més eficaç.

5. Escalar la solució a tot el corpus de vídeos a analitzar, dissenyar i implementar el pipeline que combina les diferents tècniques.

6. Supervisió humana: Passar el resultat de la transcripció (de veu en català a text en català) a les pròpies professores per a validació i eventual correcció i completar el procés amb el text transcrit al català corregit, fins a obtenir els subtítols.

7. Crear els subtítols i incorporar-los als vídeos de la plataforma d'e-learning per a poder ser activats en el visionat dels vídeos.

Generació de subtítols en anglès de vídeos en català per WiDS-Barcelona:

La metodologia proposada segueix les següents passes:

8. En primera instància, explorar l'eficàcia de diferents mecanismes i diferents eines disponibles al mercat per transcriure, traduir i subtitular veu.

9. Extreure la pista d'àudio del vídeo

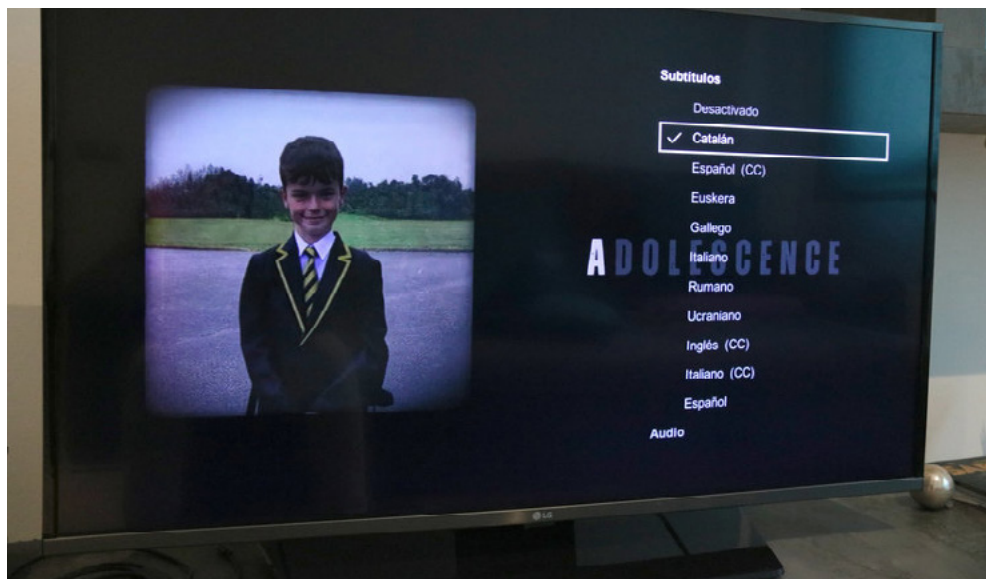
10. Traduir automàticament el text en català a l'anglès. En les 5 edicions de WiDS s'han provat diferents metodologies algunes amb més èxit que d'altres

- En les edicions més antigues s'obria una sessió de meet per fer l'streaming i s'activava la transcripció automàtica dins el propi meet, però el català no estava disponible i la transcripció era molt dolenta i requeria molta feina de supervisió abans de poder-la passar a traducció
- En les darreres edicions hem utilitzant les mateixes eines que per l'opció 2.2 del cas Top Rosies Talent.

11. Supervisió humana: Distribuir la transcripció el vídeo (de 3 o 4 hores) (de veu en català a text en català) entre l'equip organitzador de WiDS per tal que cadascú supervisi i eventualment corregeixi el text per completar el procés amb el text transcrit al català corregit, fins a obtenir els subtítols

12. Crear els subtítols i incorporar-los al vídeo que es publica a YouTube. A partir del subtítol en anglès queda oberta la porta a obtenir traduccions dels subtítols a qualsevol dels idiomes que ofereix YouTube de forma automàtica

L'objectiu en aquest cas és igualment com amb Top Rosies Talent obtenir subtítols en un altre idioma d'un contingut en català. Però l'escenari de partida és bastant diferent. Aquí per cada edició de WiDS-Barcelona hi ha un únic vídeo, però amb molts parlants, no sempre tots en català i a més el llenguatge és en registre col·loquial, si bé l'objecte de les converses segueix sent d'àmbit tecnològic i especialitzat. A la Taula x es veu el nombre de parlants de cada edició.



Generació de subtítols en anglès de vídeos de classes especialitzades on line en català:

La metodologia proposada segueix les següents passes:

1. Extreure la pista d'àudio del vídeo
2. Transcriure de català a l'anglès utilitzant Aina o Softcatalà segons els casos. Per soft català cal una operació prèvia de reduir el pes de la pista d'àudio.
3. Supervisió humana no es fa per diverses raons:
 - La dicció de la parlant és molt nítida i tant la transcripció com la traducció mantenen un alt nivell de qualitat
 - La locució va acompanyada d'unes transparències que permeten identificar si hi ha alguna imprecisió
 - L'audiència és tècnica i pot identificar si hi ha algun error de transcripció en base a les transparències que ja són en anglès
 - S'habilita un canal per a que els estudiants que detecten errors de traducció ho puguin notificar
4. Es puja el text traduït com a suport per estudiants no catalanoparlants.



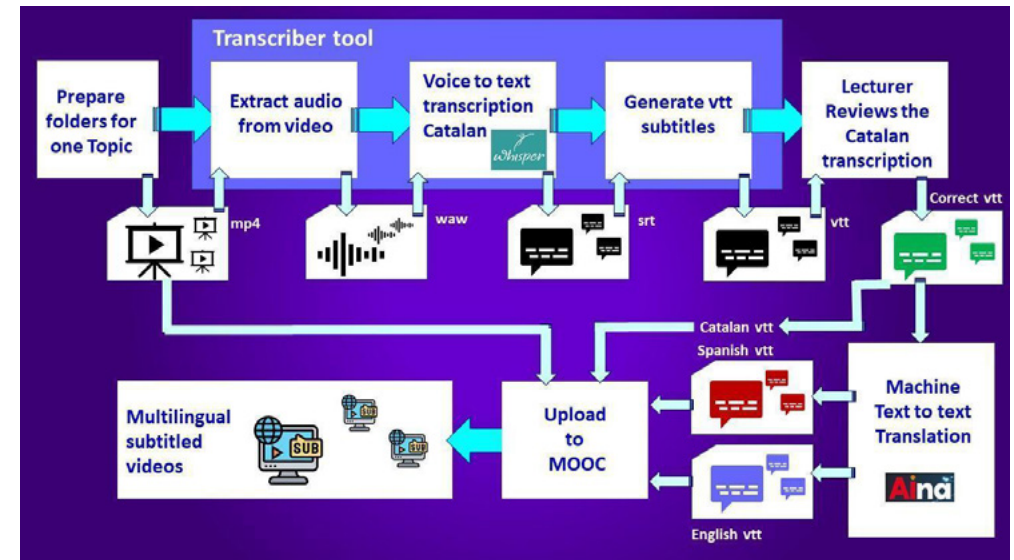
3. Implementació de la solució

A continuació es detallen els resultats obtinguts amb el demostrador.

3.1. Mètriques de rendiment

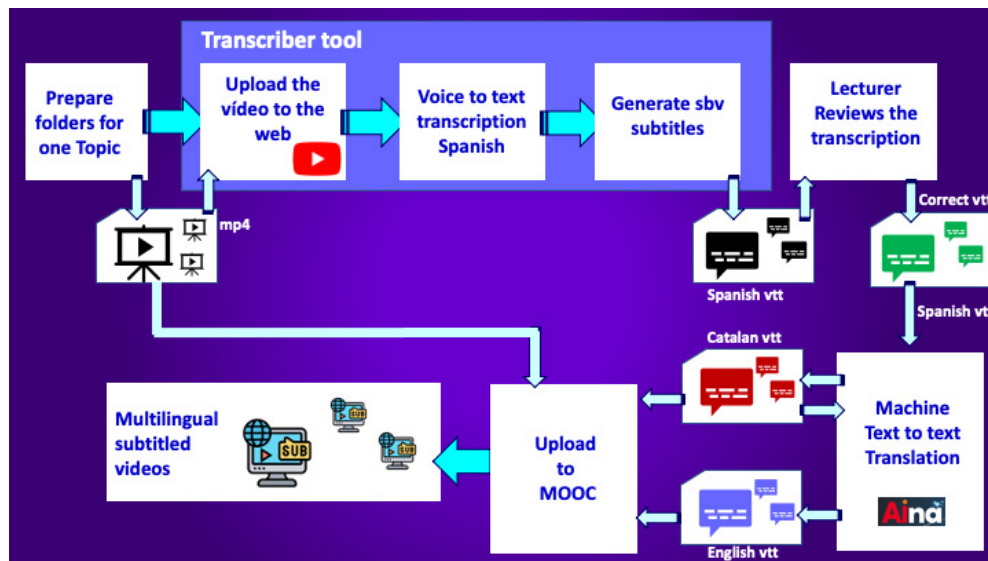
La figura 1 mostra el pipeline del procés proposat si la llengua d'origen és català pel cas del Top Rosies Talent

Figura 1 – pipeline quan la llengua d'origen és el català



La Figura 2 mostra el pipeline del procés proposat si la llengua d'origen és el castellà en el Top Rosies Talent

Figura2 – pipeline proposat quan la llengua d'origen és castellà



L'arquitectura dissenyada permet treballar directament a núvol. Els vídeos es situen en un servidor convenientment protegit, en aquest cas de drive, i s'executa tot el procés en el núvol a través de gitHiub fins a obtenir els textos traduïts i validats. El format que s'ha triat per representar els subtítols és vtt perquè es tracta d'un format editable amb marques temporals que, per ser editable permet que la supervisora humana pugui fer la revisió còmodament, i alhora, la vinculació dels subtítols al vídeo original es pot realitzar en una única passa i el text queda directament vinculat al ritme de la locució.

No s'ha pogut automatitzar aquesta darrera passa de vincular els subtítols en castellà o anglès als vídeos, ja que aquests vídeos estan

incrustats en un curs d'e-learning de Moodle i no es disposa de procediments automàtics per escalar el procés automàticament als 141 vídeos tractats.

Per als altres dos casos d'ús presentats hi ha diferències mínimes respecte els pipelines presentats pel cas de Top Rosies Talent.

Descripció del corpus de vídeos de Top Rosies Talent a tractar

La fase formativa del Top Rosies Talent es desenvolupa a partir d'una plataforma d'e-learning i consisteix a un programa de 10 capítols organitzats en 4 blocs temàtics:

1. Introducció i ètica de la IA,
2. El valor de la dada i l'aprenentatge automàtic,
3. Raonament i sistemes intel·ligents de suport a la presa de decisions,
4. Percepció (que inclou veu i visió)

Aquest programa té per objectiu oferir una visió d'ampli espectre sobre les diferents especialitzacions de la intel·ligència artificial, sempre des d'una perspectiva ètica, que va més enllà de les tècniques més populars, com l'aprenentatge profund (que és un tipus d'aprenentatge automàtic) i la IA generativa. Per cada capítol, l'equip docent del Top Rosies Talent (les donesIDEAI) ha dissenyat transparències, exercicis i pràctiques que es poden anar realitzant des de la pròpia plataforma d'e-learning, Cada capítol està a càrrec d'una especialista en la matèria que forma part de l'equip docent. La locutora és única per cada vídeo del Top Rosies Talent. S'utilitza un llenguatge acadèmic i altament especialitzat de l'àmbit de la Intel·ligència artificial. Les diferents professores presenten els continguts

en píndoles breus que estan gravades, a modus de classe en-línia, i en la que es van presentant les transparències, i indicant com i quan fer les pràctiques i exercicis. En total es tracta d'un corpus de 141 vídeos, tots ells de menys de 20 minuts de durada. D'aquests, 122 estan locutats en català i hi apareixen 5 veus diferents en funció dels diferents capítols i 19 d'ells, de menys de 8 minuts de durada cadascun, estan locutats en castellà amb accent llatinoamericà. Cada vídeo té una única locutora i serà rellevant per la bondat de les tècniques automàtiques que s'han utilitzat la qualitat d'aquesta locució, en termes de volum de l'àudio, nitidesa del senyal acústic (soroll ambiental), claredat en la dicció, etc. A la taula 1 es descriu les característiques d'aquests vídeos.

Taula 1 – Característiques del corpus de vídeos en català del programa Top Rosies Talent

Speaker	nr of videos	average length of	dialect	clarity of pronounce
1	42	10'	Central Catalan	High
2	20	14'	Central Catalan	High
3	11	12'	Central Catalan	Medium
4	18	9'	Central Catalan	Low
5	12	8'	Central Catalan	High

Descripció del corpus de vídeos de WiDS-Barcelona a tractar

Les 5 edicions de WiDS-Barcelona han donat lloc a 5 vídeos que han gravat íntegrament les sessions que hem celebrat. A data de presentació d'aquest treball la 5a edició encara no s'ha celebrat. Els temes i estructures de les 4 edicions de WiDS són:

Ed	Parlants obertura	Idioma	Slot1 Format	Parlants S1	Idioma	Tema S1	Slot2 Format	Parlants S2	Idioma	Tema S2	Slot3 Format	Parlants S3	Idioma	Tema S3	Format S4	Parlants S4	Idioma	Tema S4	Format Slot 5	Parlants S5	Idioma	tema S5	Parlants Closenda
2021	4	mixt	Taula	4	mixt	DS i health and QoL	Impact cases	4	mixt	Impact cases in health	Keynote	1	Es	Leadership	keynote	1	Eng	ethics	taula	4	mixt	female leadership	1
2022	5	cat	taula	5	cat	ecosist catala de DS	taula	5	mixt	porno i control a les xarxes dades i lideratge		x	x	x	x	x	x	x	x	x	x	x	1
2023	3	cat	taula	5	mixt	ciber	taula	4	mixt	ai decisions internacionals		x	x	x	x	x	x	x	x	x	x	x	1
2024	4	cat	taula	5	cat	directives	keynote	1	cat			x	x	x	x	x	x	x	x	x	x	x	1
2025	4	cat	taula	4	cat	indústria	keynote	1	1			x	x	x	x	x	x	x	x	x	x	x	1

WiDS-Barcelona va començar amb una jornada completa amb diferents tipus d'activitats però de seguida es va establir en una jornada d'un matí amb dues sessions separades per un cafè. Segons els anys podem tenir taules rodones o conferències inaugurals. Les ponents poden parlar català o castellà. I el més diferencial respecte el que es dona a Top Rosies Talent o al les classes online del GIA, és que a les taules rodones hi ha multiplicitat de locutors, amb diferents veus, accents, etc, i no essent obligat parlar català a les taules, es donen situacions de bilingüisme durant els debats i pot passar que a la mateixa taula rodona hi hagi algunes persones parlant en català i d'altres en castellà simultàniament. D'altra banda, es fa característic que el llenguatge de les taules rodones és informal i combina el registre col·loquial amb el vocabulari tècnic més especialitzat.

Descripció del corpus de vídeos de les classes online del GIA tractats

El curs es compon de 15 vídeos de dues hores amb una única parlant de dicció clara la locució de la qual es transcriu gairebé a la perfecció amb Whisper per exemple. Totes les sessions visualitzen les transparències ja en anglès. El llenguatge és tècnic, amb molts anglicismes, i també combina amb alguns passatges en llenguatge col·loquial.

3.2. Reptes resolts i resultats obtinguts

La metodologia configurada permet escalar la internacionalització de continguts audiovisuals en català a altres idiomes, mantenint la locució original disponible, de forma que el contingut es pot obrir a altres territoris no catalano-parlants, mentre simultàniament s'exposa a aquests parlants a la llengua catalana.

La proposta plantejada escala el procés a un gran nombre de vídeos, com és el cas dels materials d'un programa formatiu (el Top Rosies Talent en aquesta prova de concepte), tot i que algunes fases encara cal fer-les manualment, com és, per exemple, vincular els fitxers que contenen els subtítols traduïts als diferents idiomes als vídeos incrustats a Moodle.

Un cop transcrit l'àudio en català, la traducció text a text es pot escalar de forma automàtica a qualsevol altre idioma per al que hi hagi un traductor de català a aquest idioma. És preferible que l'eina de traducció faci la traducció directa des de català a l'idioma de destí, sense pivotar internament a l'anglès, ja que es perden molts matisos i significats de les expressions inicials en català.

Els resultats són la correcta traducció dels vídeos en català a subtítols en castellà i anglès pels 141 vídeos del programa Top Rosies Talent, ja disponibles a la plataforma d'e-learning del programa, la qual cosa ha fet possible obrir el programa a dones joves dels camps de refugiats del Sahara, o al mercat llatinoamericà.

Els altres dos casos d'ús aporten simplificacions i matisos de la metodologia plantejada pel Top Rosies Talent relatives al fet que en un únic vídeo apareguin diferents parlants amb diferents diccions i eventualment escenaris de bilingüisme on intervenen simultàniament català i castellà en una mateixa taula rodona per exemple. Això força la necessitat de fer una supervisió acurada de la transcripció per garantir que la traducció automàtica es pugui fer amb prou solvència. En aquest cas, fer una supervisió distribuïda i en paral·lel, agilitza molt l'obtenció dels subtítols definitius en escenaris on no és viable enviar a cada ponent la supervisió de les seves múltiples intervencions durant la jornada.

Per contra, en el cas de les classes en línia del GIA, s'obre la possibilitat de saltar la supervisió de la transcripció, que sens dubte és la fase més costosa del procés, sobre la base que, la locució és molt clara i la transcripció en general és prou bona, i que l'audiència és especialitzada i disposa de transparències en anglès per anar seguint el relat plantejat en el text traduït i detectar les imperfeccions.

Valoració de les eines a utilitzar

Com hem dit es tracta d'un àmbit amb gran desenvolupament en aquests moments i l'estat de l'art evoluciona molt ràpidament. Fins fa poc, el català ni tan sols apareixia com una de les llengües seleccionables per ser transcrita o traduïda en les eines comercials, tipus Youtu-

be. Per tant, la bona notícia és que el mercat ja està atenent el cas del català, tot i tractar-se d'una llengua minoritària.

Taula2 – Comparativa d'eines per la tasca de transcripció de veu en català a text en català

Software	Distributor	Interface	Transcription
Whisper	OpenAI	API	Inadequate transcription quality for Catalan videos, with numerous inaccuracies and mistakes. For Spanish videos we have a suboptimal performance, especially with different accents. Offers accurate punctuation signs and upper/lower cases.
NeMo stt_ca_quartznet15x5	NVIDIA	API	Rapid transcription but may compromise accuracy for both languages
NeMo stt_ca_conformer_ctc_large	NVIDIA	API	Shows superior performance in transcribing Catalan speech with complex linguistic features. For Spanish, the transcription is better from catalan
NeMo stt_ca_conformer_transducer_large	NVIDIA	API	Outperforms others in accuracy but is computationally heavier, for both languages
YouTube	Google (YouTube)	Drag and Drop in GUI	Unavailable for Catalan. Improved transcription accuracy observed for Spanish videos. Up to 5 Spanish accents available. Better transcription than others. Does not include punctuation signs or upper/lower cases in the generated subtitles.
happyscribe	Happy Scribe Ltd	Drag and Drop in GUI	Open version limited 30min/month
Sonix	Sonix	Drag and Drop in GUI	Only 30 minutes free credits
Softcatalà	Softcatalà	Drag and Drop in GUI	It's only available on-demand and limited videos below 2GB for Catalan and not available for Spanish
Buzz [29]	Github open source project	Local App	Use a local app to transcribe the videos but it is required to install this app in your pc. Isn't portable when use multiples persons
oTranscribe+ [6]	BSC	Drag and Drop in GUI	This tool allows for subtitle creation but requires expert review. Is it the same as downloading the Flugging Face model
dictation.io [1]	envato elements	Drag and Drop in GUI	This tools don't transcribe videos
Transkriptor.com	Transkriptor	Drag and Drop in GUI	Only private version is available
SeamlessM4T		API	Computationally expensive
Google Translation	Google	Drag and Drop in GUI	It only transcribes audio using the microphone and does not support uploading video or voice files. Therefore, it cannot be used for video transcription.

No obstant, la transcripció de l'àudio en català amb les diferents eines pateix encara limitacions importants pel que fa als continguts originalment produïts en català, quan parlem de llenguatge tècnic i especialitzat com és el cas de les classes d'intel·ligència artificial del programa Top Rosies Talent o les classes en línia del GIA. La taula 2 mostra els resultats obtinguts en la transcripció sobre una mostra de testing extreta d'aquesta serie.

Per l'escalabilitat del projecte és important no només la qualitat de la transcripció que proporciona l'eina, sinó si és accessible via API o no més admet la càrrega manual d'àudios pel sistema de Drag&Drop o similar.

La Taula 3 mostra la comparativa entre eines que aborten la tasca de traducció de text en català a text en castellà o anglès.

Taula3. Comparativa d'eines per traducció automàtica de text en català a text en castellà

Software	Distributor	Interface	Translation
NNLB	Meta	API	Translation quality impacted by limited integration and representation of Catalan language, resulting in inaccuracies.
Aina	BSC	API	Near-perfect translation accuracy for Catalan to Spanish, Catalan to English and Spanish to Catalan tasks, with a high automatic precision.
Google translate	Google	Drag and Drop in GUI	This tool allows you to upload a file and translate it. However, the uploading and downloading of files is not automatic, so a person is needed to perform this function. Furthermore, translations are not usually accurate and require human oversight for understanding.
happyscribe	Happy Scribe Ltd	Drag and Drop in GUI	The tool also allows for translation, but with a free account, only 30 minutes of video are permitted.
Youtube	Google	Drag and Drop in GUI	It allows for both transcription and translation processes to be carried out simultaneously. However, translations are often inaccurate because it fails to link phrases properly, lacking punctuation marks, verb conjugation, etc.

En el cas de la traducció ha estat difícil trobar una eina que respongués a les necessitats del projecte ja que no va ser fins a gairebé el final del mateix que vam disposar del traductor d'AINA, recent alliberat, i que ha donat molt bons resultats perquè està especialitzat en recursos lingüístics per català i les altres eines o bé tenien poca qualitat o no eren escalables via API.

No obstant, la transcripció de l'àudio en català amb les diferents eines pateix encara limitacions importants pel que fa als continguts originalment produïts en català, quan parlem de ponents amb diferents flexions dialectals, i un llenguatge tècnic i especialitzat com és el cas de les classes d'intel·ligència artificial del programa Top Roses Talent, si bé és molt i molt positiu que existeixi ja un bon nombre d'eines encara hi ha marge de millora.

Les parts que graven taules rodones del WiDS-Barcelona mostren pitjors prestacions, senzillament perquè l'eina de transcripció perd eficàcia si de la mateixa escena surten en paral·lel 3 o 4 veus en idiomes diferents. Cal afegir-hi que a més, la taula rodona és un format que es presta al col·loqui i a les intervencions en paral·lel dels contertulis.

3.3. Limitacions actuals de la tecnologia

Les eines són sensibles a la qualitat del senyal (soroll etc), característiques de la veu (timbre, ritme) de la dicció i presenten resultats que requereixen alts nivells de supervisió encara en aquells parlants que no pronuncien prou clarament o parlen molt de pressa, o han gravat amb massa soroll ambiental.

Si es barregen idiomes en una mateixa escena, baixa molt la qualitat de les transcripcions.

L'escalabilitat del procés descansa sobre la possibilitat de poder cridar les eines de transcripció o traducció per programa (APIs, etc) de manera que es puguin dissenyar algorismes que processin automàticament molts àudios. Si bé existeixen un bon nombre d'eines a mercat que resolen el procés de transcripció o traducció semi-automàticament amb uns estàndards de qualitat molt respectables, moltes d'elles només permeten el Drag&Drop del vídeo/àudio sobre un servei web o una app i això obliga a processar-los manualment d'un en un i impedeix el disseny d'un pipeline d'automatització complet.

Les eines per castellà estan més madures que les de català, i moltes de les disponibles pivoten a través de l'anglès per interactuar amb el català. En aquest sentit AINA ha estat una gran aportació perquè quan es tradueix a través d'una llengua intermèdia és fàcil que es perdin matisos o es tradueixin malament les expressions fetes.

En el cas de la traducció ha estat difícil trobar una eina que respongués a les necessitats del projecte ja que no va ser fins a gairebé el

final del mateix que vam disposar del transcriptor i traductor d'AINA, especialitzat en recursos lingüístics per català i les altres eines o bé tenien poca qualitat o no eren escalables via API.

Les eines veu-veu estan menys madures que les eines text-text i per aquest motiu s'ha abandonat pel moment la traducció automàtica de veu catalana a text en anglès o castellà.

Les eines que mouen textos, veus entre castellà i anglès estan més madures que les que ho fan entre castellà i català i aquestes ho estan més que les que ho fan entre català i anglès.

Els recursos per fer tasques de llenguatge que involucren català estan menys madurs i AINA està oferint solucions d'alta qualitat que permeten grans avenços en aquest àmbit. De fet, hem pogut comprovar com la transcripció i traducció des del castellà ha requerit menys correccions humanes en el procés de supervisió, ja que les eines en llengua castellana estan més desenvolupades i, per tant, la qualitat dels resultats per les diferents tasques és més alta. Així doncs, cal seguir desenvolupant i afinant més els models que estiguin desenvolupats en català. En tots els casos estudiats s'ha treballat amb continguts de caire tecnològic i vocabularis especialitzats, el que incideix en les prestacions de les diferents eines, que poden resultar altament solvents pel text/veu comú o col·loquial i baixar qualitat de forma significativa quan es tracta de terminologia més científica i especialitzada.

Pel que fa a les eines en català, és complicat encara, a excepció feta d'AINA, que es reconeguin apropiadament les flexions dialectals, en especial quan parlem de veu, com el català baleàric o el català occidental (terres de Ponent, franja o el tortosí).

Pel que fa a la valoració conjunta dels 3 casos d'ús presentats, convé realitzar la supervisió humana en el punt del pipeline més feble, per prevenir que els resultats finals no siguin disbarats. Per exemple si s'opta per fer una transcripció d'àudio a text seguida d'una traducció de text a text, si partim de castellà i traduïm a l'anglès es pot fer la supervisió al final. Però si partim de català i volem traduir a castellà, convé supervisar la transcripció, perquè els traductors de text català-castellà ja són molt fiables i les traduccions són molt dolentes si parteixen d'una transcripció que té errors.

En tots els casos mencionats la supervisió i validació humana és necessària, per depurar les imprecisions de les transcripcions i traduccions automàtiques. En el cas de les classes online del grau d'IA s'ha pogut retardar la supervisió humana a un model distribuït on és el propi lector el que pot reportar errors de transcripció o traducció, únicament perquè estem en un entorn amb una audiència perfectament definida i amb alts coneixement tècnics. A excepció feta d'aquests entorns acadèmics i controlats, sempre caldria revisar els textos resultants.



4. Potencial impacte de la solució

El treball que es presenta resol dos problemes simultàniament:

- Permet que tothom que vulgui generi continguts en català sense haver de pensar si fer-ho representa una minimització de l'audiència capacitada per consumir aquell contingut, amb fortes implicacions a nivell d'internacionalització del coneixement que es genera genuïnament en català majoritàriament des de Catalunya i d'impacte del contingut que es vol generar. Aquest no és un fet menor en absolut, ja que durant anys molts autors, professors, escriptors, han optat per renunciar a la llengua pròpia i produir en castellà justament per no haver de pagar aquest preu de minimització de l'impacte, i perquè durant molt temps existien ajuts per traduir continguts del castellà al català però no a l'inrevés. Per tant estaríem parlant d'un impacte potencialment molt important en l'externalització i internacionalització dels continguts en català.
- Per una altra banda, i amb la decisió de no doblar els continguts, sinó mantenir la versió original en català subtitulada a altres idiomes, apareix la possibilitat que fora de les nostres fronteres es pugui sentir la llengua catalana, augmentant doncs l'espai d'exposició a la llengua catalana d'audiències potencialment massives d'altres territoris, amb totes les implicacions per la supervivència i bona salut del català que això representa. És per aquest motiu que la metodologia que es planteja en aquest treball no contempla en cap moment l'ús de les eines d'IA generativa per doblar la locució a un altre idioma, conscients com som que en aquests moments la tecnologia és prou madura com per a fer-ho fins i tot imitant la pròpia veu de la persona que s'està doblant, i modificant com mou la boca per a que resulti molt i molt realista.

5. Bibliografia

- [1] [Agarwal 2024] Agarwal, A. (2024). Voice Dictation - Online Speech Recognition.
- [2] [Anguera 2014] Anguera, X., et al (s/f). Audio-to-text alignment for speech recognition with very limited resources. https://xavieranguera.com/papers/IS2014_phonealignment.pdf
- [3] [Armengol-Estapé 2021] Armengol-Estapé, J., et al. (2021). Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. ACL-IJCNLP 2021 (pp. 4933–4946). Stroudsburg, PA, USA: Association for Computational Linguistics.
- [4] [Bassa 2023] Bassa, M. (2023, julio 4). Almost 45% of Catalan speakers in Catalonia do not use Catalan to search on sites such as Google and YouTube. Fundació .cat
- [5] [Bonafonte 1997] Bonafonte Cávez, A., et al (1997). A bilingual text-to-speech system in spanish and catalan. Procs. of EUROSPEECH '97, 2455–2458. WCL, University of Patras, Grece.
- [6] [BSC] BSC. (s.d). oTranscribe. <https://otranscribe.bsc.es/>
- [7] [BSC 2024] BSC. (2024, mayo 23). <https://www.bsc.es/es/noticias/noticias-del-bsc/aina-impulsa-la-primer-solucion-de-voz-que-incorpora-las-diferentes-variantes-del-catalan>
- [8] [BSC 2024b] BSC. (s/f). Projecte-aina (Projecte Aina). Recuperado el 13 de mayo de 2024, de Aina website: <http://https://huggingface.co/projecte-aina>
- [9] [Devlin 2018] Devlin, J., et al (2018). BERT: Pre-training of deep bidirectional Transformers for language understanding. Recuperado de <http://arxiv.org/abs/1810.04805>
- [10] [Elvira-García 2016] Elvira-García, W., et al(2016). A tool for automatic transcription of intonation: Eti_ToBI a ToBI transcriber for Spanish and Catalan. Language Res. and Evaluation, 50(4), 767–792.
- [11] [Fàbregas 2021] Fàbregas, L (2021, noviembre 2019) El Govern recorre a experts turcos para que Alexa y Siri hablen en catalán. The Objective Media, 19.11.2021
- [12] [Generalitat 2022] Generalitat. (2022). Món digital i tecnologies de la llengua. llengua.gencat.cat/web/.content/temes/pacte-nacional-per-la-llengua/resums-tematics/6-digital.pdf
- [13] [Gulati 2020] Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented Transformer for Speech Recognition (No. arXiv:2005.08100). arXiv. <https://doi.org/10.48550/arXiv.2005.08100>
- [14] [Happy Scribe] Scribe, H. (s/f). Happy scribe: Audio transcription & video subtitles. Recuperado el 14 de mayo de 2024, de Happy Scribe website: <https://www.happyscribe.com/>
- [15] [IBM 1961] IBM. (s/f). (1961) Shoebox - IBM Archives (78-013). Recuperado de [https://mediacenter.ibm.com/media/\(1961\)+Shoebox+-+IBM+Archives+\(78-013\)/O_4m2ynnkk](https://mediacenter.ibm.com/media/(1961)+Shoebox+-+IBM+Archives+(78-013)/O_4m2ynnkk)
- [16] [Kjartansson 2020] Kjartansson, O., et al (s/f). Open-Source High Quality Speech Datasets for Basque, Catalan, and Galician. [Aclanthology.org/2020.sltu-1.3.pdf](https://aclanthology.org/2020.sltu-1.3.pdf)

[17] [Kriman 2019] Kriman, S., Beliaev, S., Ginsburg, B., Huang, J., Kuchaiev, O., Lavrukhin, V., Leary, R., Li, J., & Zhang, Y. (2019). QuartzNet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions (No. arXiv:1910.10261). arXiv. <https://doi.org/10.48550/arXiv.1910.10261>

[18] [Külebi 2024] Külebi, B., et al ParlamentParla: A Speech Corpus of Catalan Parliamentary Sessions., de Aclanthology.org website: <https://aclanthology.org/2022.parlaclarin-1.18.pdf>

[19] [Mallafré 2000] Mallafré, J. (2000). Language Models and Catalan Translation. Selected Papers from the 4th Int'l Congress on Translation, Barcelona, 1998 págs. 141-151, 141-151.

[20] [Meta 2024] Meta. (s/f). Meta AI Research Topic - No Language Left Behind. Recuperado el 13 de mayo de 2024, de Meta.com website: <https://ai.meta.com/research/no-language-left-behind/>

[21] [NeMo 2024] NeMo - Automatic Speech Recognition. (s/f). Recuperado el 13 de mayo de 2024, de NVIDIA NGC Catalog website: https://catalog.ngc.nvidia.com/orgs/nvidia/collections/nemo_asr

[22] [Padro 2012] Padró, L. et al (2012) FreeLing 3.0: Towards wider multilinguality

[23] [Periodico 2023] Periódico, E. (2023, febrero 15). Amazon anuncia que Alexa seguirá sin hablar catalán: “No hay novedades”. El Periódico

[24] [SeamlessM4T 2024] SeamlessM4T. (s/f). Recuperado el 14 de mayo de 2024, de Huggingface.co website: https://huggingface.co/docs/transformers/en/model_doc/seamless_m4t

[25] [Sonix 2024] Sonix. (s/f). A brief history of speech recognition. sonix.ai/history-of-speech-recognition

[26] [Softcatalà 2015] Softcatalà. (2015, agosto 20).: <http://softcatala.org/>

[27] [Transkriptor 2021] Transkriptor: Convert audio or video to text [Transcription]. (2021, abril 30). Recuperado el 14 de mayo de 2024, de Transkriptor! website: <https://transkriptor.com/>

[28] [Vázquez 2020] Vázquez, R., et al (2020, mayo). A Systematic Study of Inner-Attention-Based Sentence Representations in Multilingual Neural Machine Translation. [Aclanthology.org/2020.cl-2.5](https://aclanthology.org/2020.cl-2.5)

[29] [Williams 2024] Williams, C. (2024). Introduction | Buzz. Buzz <https://chidiwilliams.github.io/buzz/docs>

[30] [Whisper] Whisper: Robust Speech Recognition via Large-Scale Weak Supervision. (s/f).

[31] [Youtube 2024] YouTube. (s/f). Recuperado el 13 de mayo de 2024, de <https://www.youtube.com>



6. Autoria i agraïments

Autors:

Karina Gibert
Xavier Angerri
Sergi Ramírez

Agraïments:

L'equip de treball d'aquesta recerca a sigut: Javi Puerta, Dario Sacco,
Kevin Sànchez , Dante Conti. IDEAI-UPC

Socis



Membres

