

4. Iniciatives i reptes per al català en les tecnologies de la parla

MIREIA FARRÚS¹

Professora de Lingüística Computacional
Centre de Llenguatge i Computació (CLiC), Universitat de Barcelona

1. ALGUNES GENERALITATS SOBRE LES TECNOLOGIES DE LA PARLA

Durant molts anys, i encara ara, els lingüistes computacionals i els que treballen en tecnologies de la parla hem hagut d'explicar a què ens dedicàvem i per a què servia tot el que fèiem. Hem entrat en una etapa, però, en què els desenvolupaments tecnològics recents ens ho han posat força més fàcil: «Coneixeu el *Siri* i l'*Alexa*? Heu fet servir mai Google Translate? Doncs som més o menys els que hi ha darrere de tot això». Amb les tecnologies de la llengua podem desenvolupar traductors automàtics de text, traductors de veu a veu, assistents virtuals i fins i tot sistemes de reconeixement de la parla que poden ser utilitzats com a eines de transcripció i traducció en una aula universitària, o com a eines de subtitulació de pel·lícules, entre moltíssimes altres aplicacions.

L'explosió de les tècniques basades en xarxes neuronals ha fet que els sistemes basats en tecnologies de la llengua —de la mateixa manera que ha passat en molts altres àmbits— fessin un salt qualitatiu altíssim, de manera que el seu ús s'acosta cada dia més a ser una realitat en la nostra vida diària. Darrere d'aquest salt qualitatiu, però, s'amaga un repte que generalment només poden assumir les llengües majoritàries que tenen més recursos: la gran quantitat de dades que es necessiten per a entrenar aquests sistemes, amb el risc de deixar enrere les llengües amb menys parlants —que són la gran majoria.

Poder viure plenament en la llengua pròpia hauria de ser un dret reconegut i universal. No obstant això, ser parlant d'una llengua minoritzada i intentar viure-hi al 100% és una feinada: una inversió de temps i d'esforç, i una reivindicació constant dels nostres drets. Des de l'històric domini .cat fins al cas més recent de les queixes a Google per la pèrdua de prioritació de la llengua catalana en les

1. A/e: mfarrus@ub.edu

cerques fetes en català, els catalanoparlants hem d'estar sempre alerta per a reivindicar els nostres drets digitals —i els no digitals. Des d'una mirada més global, els Objectius de Desenvolupament Sostenible de l'Agenda 2030 de les Nacions Unides² són molt clars pel que fa a salvaguardar el patrimoni cultural de món, i també quant a reduir la bretxa digital per a garantir la igualtat d'accés a la informació i el coneixement. De manera que el desenvolupament d'eines computacionals destinades a facilitar la comunicació ha de ser garantit en totes les llengües, tinguin els parlants que tinguin.

2. LA SITUACIÓ DEL CATALÀ EN L'ÀMBIT TECNOLÒGIC: INICIATIVES SOCIALS I INSTITUCIONALS

Si d'alguna cosa podem presumir els catalanoparlants és de l'alta consciència tecnològica que tenim, que es reflecteix en una forta presència a internet. No per casualitat el català fou la tercera llengua que va tenir la Viquipèdia per darrere de l'anglès i de l'alemany, i la segona que en va crear continguts, només per darrere de l'anglès. Actualment, és la vintena llengua en nombre d'articles, mentre que en nombre de parlants a nivell mundial es troba a la posició 127 segons el rànquing d'*Ethnologue* de 2022.³ No obstant tot això, ens toca conviure amb les dificultats de qualsevol llengua minoritària i minoritzada: el cost de generar recursos lingüístics que puguin donar forma a tots aquests serveis i aplicacions és molt alt.

L'alta consciència tecnològica que tenim els catalans s'ha reflectit en diverses iniciatives a nivell social. D'entre les més rellevants trobem *Softcatalà*⁴ i *CollectivaT*,⁵ que promouen l'ús de programari en català, així com de serveis basats en tecnologies de la llengua, com ara correctors, traductors automàtics o reconeixedors i sintetitzadors de la parla.

Softcatalà es va fundar l'any 1998 com a associació sense ànim de lucre, amb l'objectiu principal de fomentar l'ús i la presència del català a la informàtica, internet i les noves tecnologies. L'associació s'ha dedicat a treballar, amb molta constància, en la traducció al català de programari informàtic, i en el desenvolupament d'eines lingüístiques com ara correctors, traductors automàtics i diccionaris. Fa un parell d'anys va implementar un nou traductor automàtic anglès-català basat en xarxes neuronals, que és la tecnologia més avançada en aquest àmbit.

2. L'Agenda 2030 a les Nacions Unides: <https://cads.gencat.cat/ca/Agenda_2030/nacions-unides/>

3. *Ethnologue: Languages of the World*: <<https://www.ethnologue.com/>>

4. Associació *Softcatalà*: <<https://www.softcatala.org/>>

5. Cooperativa *CollectivaT*: <<https://collectivat.cat/>>

Una mica en la mateixa línia, la cooperativa Col·lectivaT desenvolupa recursos computacionals en traducció automàtica, reconeixement i generació de la parla. Fa també un parell d'anys, Col·lectivaT va posar en marxa el *Catotron*,⁶ el primer sistema de síntesi de la parla lliure i obert, basat també en xarxes neuronals.

Tant Softcatalà com Col·lectivaT aposten clarament pel programari codi i obert, fet que els ha convertit en exemples paradigmàtics de sobirania tecnològica per al català. La sobirania tecnològica rau en la gestió i la creació de recursos propis lliures i oberts, amb independència dels programaris i altres serveis tecnològics de propietat. És un corrent social i polític que defensa i promou l'autogestió de les infraestructures tecnològiques per part dels ciutadans. La sobirania tecnològica té uns avantatges enormes per a les llengües minoritzades com la nostra, perquè ens permet crear i gestionar els recursos lingüístics necessaris per a desenvolupar tots aquests productes basats en tecnologies de la parla, sense haver de dependre constantment de les grans empreses del sector.

Però la sobirania tecnològica no és exclusiva de les aplicacions informàtiques, sinó que és extensible a altres sectors digitals, com ara l'audiovisual. En aquest camp podem destacar la iniciativa del Canal Malaia, que produeix i engloba continguts en català, i aprofita els algorismes d'aprenentatge automàtic de YouTube per a agrupar vídeos i aconseguir que les recomanacions que fa la plataforma es basin també en la llengua de visualització de l'usuari. És una manera indirecta d'utilitzar la tecnologia per a fer visible el català i i afavorir-ne l'ús, especialment entre el públic jove. Hem de ser conscients que, avui en dia, els canals audiovisuals són un puntal per a la difusió de la llengua.

En l'àmbit institucional, una de les iniciatives més recents que podem trobar és el projecte AINA,⁷ impulsat pel Departament de Polítiques Digitals de la Generalitat de Catalunya en col·laboració amb el Barcelona Supercomputing Centre (BSC). L'objectiu principal del projecte AINA és generar recursos per a entrenar algorismes d'intel·ligència artificial i poder desenvolupar tota una sèrie de sistemes basats en xarxes neuronals, com ara traductors automàtics o sistemes de reconeixement de la parla, entre d'altres.

3. ALGUNES IDEES PER A ANAR FENT

Massa sovint ens trobem que les empreses desenvolupen els seus productes tecnològics únicament i exclusivament per a llengües majoritàries i dominants. D'entrada, és una qüestió pràctica que respon a interessos econòmics. No obstant això, les empreses responen tard o d'hora a les demandes dels usuaris, i generaran

6. Sistema de síntesi de veu Catotron: <<https://catotron.cat/>>

7. Projecte AINA: <<https://www.projecteaina.cat/>>

productes en català si hi ha mercat. En aquest sentit, hi ha dues accions que es poden promoure: en primer lloc, invertir i consumir sempre que sigui possible en productes de petites i mitjanes empreses locals, més properes geogràficament i culturalment, i donar suport sobretot a aquelles amb més consciència lingüística. En segon lloc, facilitar que aquestes empreses puguin produir els seus productes en català, mitjançant la creació de recursos que es puguin compartir lliurement, per tal de potenciar cada vegada més la sobirania tecnològica.

Al nivell individual també tenim l'opció de col·laborar en iniciatives com el projecte Common Voice,⁸ una iniciativa de Mozilla que recopila dades de veu en format lliure i obert per a desenvolupar sistemes de reconeixement i síntesi de la parla, amb l'objectiu de reduir les desigualtats entre llengües en l'àmbit de les tecnologies de la parla. De fet, la Common Voice és la base del projecte AINA en la part de tecnologies de la parla: la campanya «La nostra llengua és la teva veu. Fes que la tecnologia parli català» en el marc de l'AINA té com a objectiu principal promoure la contribució a la base de dades Common Voice per part dels ciutadans. D'aquesta manera podem contribuir, des de casa mateix i fàcilment, a incrementar i fer visible la sobirania tecnològica en la nostra llengua. A més a més, la Common Voice té un valor afegit per a moltes llengües, i és que permet contribuir-hi en diferents varietats lingüístiques d'una mateixa llengua, de manera que els sistemes que es desenvolupin a partir d'aquestes dades estaran adaptats a una gran diversitat de parlants.

Col·laborar en projectes com el de la Common Voice és un granet de sorra individual que pot impulsar el desenvolupament de sistemes tecnològics en català: sobretot en sistemes de reconeixement de la parla i de síntesi de la veu. Són dos sistemes que, com dèiem al principi, poden ser la base de sistemes més complexos com assistents virtuals i traductors de veu, que incorporen altres mòduls com sistemes de diàleg o traducció de text. Però per si soles ja són aplicacions amb un gran potencial en molts àmbits. Vegem-ne un parell d'exemples per a l'aula i per al cinema.

Els últims anys hi ha una polèmica recurrent a les universitats catalanes: el canvi de llengua —sempre del català al castellà— per part de professors que havien anunciat que l'assignatura seria en català, davant d'alguns, pocs —a vegades només un— alumnes castellanoparlants o provinents del programa Erasmus que diuen no entendre la llengua autòctona. Deixant de banda l'aspecte politico-administratiu —qualsevol professor d'una universitat de Catalunya ara mateix pot impartir la classe en castellà o en anglès si així ho anuncia al pla docent corresponent—, podríem implementar un sistema de reconeixement de la parla connectat a un sistema de traducció automàtica que proporcionés la transcripció de la classe i

8. Projecte Common Voice de Mozilla: <<https://commonvoice.mozilla.org/ca>>

la corresponent traducció a una de les llengües oficials de la universitat. Fins i tot es podria arribar a personalitzar el sistema i que cada estudiant pogués escollir la seva pròpia llengua com a llengua de traducció. Per als defensors del cinema en versió original subtítulada, això hauria de ser una solució molt més bona que el canvi de llengua.

I parlant de cinema: la tecnologia que ens permet crear un sistema de transcripció traduïda en una aula és la mateixa que ens permet crear un sistema de subtítolació d'una pel·lícula. I anant una mica més enllà, només caldria incorporar-hi un sistema de síntesi de la parla per a aconseguir un sistema automàtic de doblatge. I encara més, podríem aconseguir tant els defensors del doblatge com de la subtítolació, i que cadascú pogués triar la seva fórmula predilecta. La subtítolació automàtica té encara dos reptes molt importants, que són la sincronització i l'expressivitat. La sincronització ha de ser tant temporal d'acord amb la veu original, com foneticovisual amb els llavis de l'actor. A banda d'això, el contingut a subtítular té un alt component expressiu, i a diferència d'una classe magistral, aquesta expressivitat és un aspecte crucial del contingut audiovisual, que hauria de ser transferida a la veu generada automàticament per a mantenir l'originalitat.

Per això, així com la subtítolació automàtica ja podria ser una realitat, per al doblatge automàtic encara hi ha feina a fer. Els detractors del doblatge ho poden veure com un motiu més per a no doblar, juntament amb el cost que representa actualment el doblatge humà. Però —nota al peu—, l'argument de veure pel·lícules en versió original per mantenir la genuïnitat a la vegada que s'aprèn una altra llengua és un argument molt bo en el marc d'una situació sociolingüística que ara mateix no tenim. Per a una llengua minoritzada, és crucial que els nous nadius puguin consumir productes audiovisuals de tota mena en la llengua autòctona perquè la puguin aprendre. Si no, el que fem sobretot és potenciar encara més l'ús i la difusió de les llengües més dominants. Això per si sol ja és un motiu per a invertir recursos com més aviat millor per a fer que el doblatge automàtic sigui una realitat a mitjan termini.

4. REPTES DE FUTUR I CONCLUSIONS

Viure plenament en la pròpia llengua és fer-ho també amb les nostres pantalles. Per a això cal una consciència lingüística individual, però també col·lectiva, amb l'acompanyament de les institucions. Una consciència que posi de manifest la riquesa que representa cada llengua, i que tingui en compte que una llengua que perd el tren de la digitalització en l'era en què ens trobem és condemnada a l'extinció.

Les tecnologies lingüístiques han tingut una evolució exponencial durant l'última dècada amb l'explosió de les xarxes neuronals que n'han incrementat la

precisió i la qualitat a uns nivells que no s'havien vist mai. El recent llançament del ChatGPT de l'empresa OpenAI, per exemple, ha mostrat a la població els nivells de qualitat a què poden arribar els models basats en intel·ligència artificial. Hi ha hagut, a més a més, la sorpresa addicional en veure que el model també funcionava (molt bé) en català. No se n'ha parlat tant, però fa pocs mesos la mateixa empresa ens va sorprendre amb el Whisper, un sistema de reconeixement de la parla que funciona també en català i que té una taxa de reconeixement impressionant. A més, ha estat especialment desenvolupat per a reconèixer la parla en entorns adversos —com ara ambient sorollosos— que normalment en dificulten el funcionament.

Tot i que podem gaudir del llançament d'aquests productes en català, és important anar cap a la sobirania digital i lingüística en la nostra llengua. I aquí les institucions tenen un paper molt important. De fet, llengües molt majoritàries com l'anglès ja disposen d'aquesta mena de recursos generats per institucions que se'n fan càrrec. Així, grans empreses com Google, Microsoft i Apple utilitzen recursos digitals creats per la DARPA —l'Agència de Projectes de Recerca Avançada del Departament de Defensa dels Estats Units d'Amèrica— per a generar els seus models d'intel·ligència artificial. Qualsevol país, nació o territori amb una llengua pròpia, per petit que sigui, ha de tenir aquest suport institucional en el camp digital.

Òbviament, això hauria de ser indispensable sobretot en llengües minoritzades com la nostra. No obstant això, aquesta necessitat contrasta a vegades amb alguns obstacles amb què ens trobem els catalanoparlants. En aquest punt permeteu-me explicar una anècdota personal per a exemplificar-ho: tinc un fill de 12 anys que uns mesos enrere es va entusiasmar, com molts de nosaltres, amb el *Paraulògic* —quina bona idea per a impulsar, visibilitzar i modernitzar el català, per cert!—. El seu entusiasme es va creuar amb el seu interès per la programació, i li va demanar al seu pare que l'ajudés a programar un *Paraulògic*. Per a això necessitaven la llista de paraules del DIEC, i ell mateix els ho va demanar per correu. La resposta per part de l'IEC va ser contundent: no podien cedir la llista. Tant costa, de debò, donar als ciutadans l'opció de descarregar una llista de paraules en un fitxer de text perquè en puguem fer un bon ús? Una llengua com la nostra es pot permetre aquestes restriccions i obstacles addicionals? Parlant, per cert, del DIEC, llengua i digitalització, el *Paraulògic* ens va fer veure que al DIEC no hi ha... *internet*. El DIEC és a internet però no té *internet* (!).

De la consciència lingüística tots en som responsables, i els mitjans audiovisuals no s'escapen d'aquesta responsabilitat. Ja que parlàvem abans de varietats lingüístiques a la Common Voice i de subtitulació, em pregunto en quin moment TV3 va decidir que la forma genuïna del *jo* passava a ser *io* i com si no hagués passat res. Calia aquest canvi? No hi ha moltes altres maneres d'adaptar el

català a les formes col·loquials d'una manera natural? Ningú no s'ha parat a pensar que aquesta adaptació xoca amb el fet que la majoria de varietats dialectals del català mantenen la forma genuïna del *jo* en la seva forma oral i que aquesta col·loquialització és simplement centralista i innecessària? És un petit exemple dels factors que van minant la llengua i assimilant-la a un determinat model dels mitjans de comunicació.

També hem de tenir mirada llarga per a veure els agents que poden ajudar a promoure l'ús social del català. Sovint pensem que els individus, els organismes públics responsables de la llengua i els mitjans de comunicació ho poden fer tot. I hauria de ser així, però recordem el cas de Spotify en català: a vegades pot fer més el Barça o una cançó de la Rosalia que cent campanyes institucionals. Estiguem alerta a qualsevol oportunitat!

Finalment, perdem la por a la tecnologia i aprofitem i gaudim de tot el que ens aporta. Ni la revolució industrial ens va portar a les jornades laborals de tres hores, ni els traductors automàtics han fet desaparèixer les facultats de traducció, ni els cercadors d'internet es van haver de prohibir per a fer treballs acadèmics. Al contrari: ja no consultem enciclopèdies en paper, però cerquem més informació que mai. I per això no té sentit prohibir l'ús del ChatGPT a les escoles, sinó que l'hem d'integrar com una eina més de treball. Prohibir-lo és posar tanques al camp. Quan fa anys feia xerrades sobre traducció automàtica i em preguntaven si els traductors humans es quedarien sense feina, jo deia que no. Fa poc em van preguntar si amb la millora dels sistemes de traducció veu a veu la gent deixaria d'aprendre llengües. I també penso que no. Al món tenim unes 6.000 llengües. Quantes persones n'aprenen més de deu al llarg de la vida? Qui en vulgui aprendre per amor a l'art o per necessitat n'aprendrà igualment. Mentrestant, aprofitem la tecnologia per a facilitar-nos la comunicació amb parlants d'altres llengües i millorar la relació cultural entre nosaltres.