El projecte AINA busca milions de veus perquè la tecnologia entengui i parli el català

written by David Folch | 16 de febrer de 2022

Sota el lema **'La nostra llengua és la teva veu'**, el Govern de Catalunya llança aquest 17 de febrer una **campanya de captació de veus per generar el primer corpus o "diccionari" de veu del català.** La campanya s'inscriu en el **projecte AINA**, impulsat pel Departament de la Vicepresidència i de Polítiques Digitals i Territori en col·laboració amb el Barcelona Supercomputing Center-Centro Nacional de Supercomputación (BSC-CNS) per fer que la tecnologia entengui i parli la nostra llengua.

El projecte **AINA** està construint els corpus (conjunts massius de dades) i els models de la llengua catalana de manera que qualsevol empresa o organització pugui utilitzar-los per desenvolupar les seves solucions o serveis específics (traductors, assistents personals, sintetitzadors de veu, classificadors de textos, etc.) per tal de poder-nos relacionar amb les màquines en català.

En definitiva, per ensenyar català a les màquines de manera que la ciutadania pugui relacionar-se amb elles i participar en el món digital en català al mateix nivell que els parlants d'una llengua global, com ara l'anglès, i evitar, així, l'extinció digital de la llengua catalana.

La participació ciutadana a la campanya de recollida de veus 'La nostra llengua és la teva veu' es farà a través de la iniciativa de Common Voice de Mozilla pel català, una plataforma on tothom que ho vulgui podrà llegir i enregistrar un nombre il·limitat de frases (agrupades de 5 en 5 però sense

límit) per ajudar les màquines a aprendre com parlem les persones.

Tot i que aquesta col·laboració es pot fer de manera totalment anònima i sense cap registre previ, conèixer els paràmetres de gènere, edat i variant dialectal de la persona "donant" de veu facilita molt la feina de classificar les dades de veu obtingudes i, alhora, permet saber si s'està contemplant tota la diversitat lingüística del català. Per això, la campanya anima la ciutadania a registrar-se i crear un perfil a la plataforma per avançar més ràpidament en els objectiu del projecte AINA.

Ensenyar català a les màquines, tot un repte

"Ensenyar" una llengua a les màquines de manera que siguin capaces no només d'entendre'ns quan els parlem sinó de respondre'ns de manera coherent a allò que els hem preguntat o demanat és avui un repte.

Si volem que els ordinadors, assistents de veu i altres sistemes informàtics parlin i entenguin el català, cal aconseguir dades massives de la llengua (en format de text i de veu). Aquestes dades es passen a una xarxa neuronal profunda que va aprenent com es combinen les paraules fins a generar un mòdel de la llengua capaç, per exemple, de distingir els diferents significats de la paraula "banc" gràcies als diferents contextos en què es fa servir.

Per construir el corpus de la llengua (conjunts de dades) que necessita una màquina, és necessari tenir milions de textos i milions d'hores d'àudio i vídeo en aquella llengua i, a més a més, que aquests milions i milions de dades representin tota la riquesa de la llengua incloent, per exemple, gravacions de veu de persones de diferents gèneres, diferents franges d'edat i diferents variants dialectals i registres.

Obtenir aquest volum i concreció de dades és especialment difícil per a les llengües minoritàries a escala mundial com

el català, ja que llengües majoritàries com l'anglès tenen fàcilment a disposició tota aquesta informació: només cal anar a Internet per trobar milions i milions de textos, àudios i vídeos en anglès.

Per aquest motiu, la campanya 'La nostra llengua és la teva veu' convida la ciutadania de parla catalana de totes les edats, gèneres, condicions i procedències a "donar" la seva veu, amb l'objectiu d'obtenir uns continguts de veu que copsin tota la riquesa del català oral, amb tots els seus registres i varietats dialectals. Actualment, el perfil de veu majoritari a la plataforma Common Voice de Mozilla és la d'homes d'entre 30 i 50 anys parlants de català central.

Crear el primer corpus de veu en català, fita de l'AINA per al 2022

La **creació de la primera versió del corpus de veu** del català és una de les principals fites del projecte AINA per aquest 2022. Aquest corpus es nodrirà dels continguts obtinguts a través de la plataforma de Common Voice de Mozilla, però també de l'aportació del repositori documental de la Corporació Catalana de Mitjans Audiovisuals (CCMA) o el Consell de l'Audiovisual de Catalunya (CAC), entre d'altres.

En paral·lel, el projecte es marca també com a objectiu d'aquest any la **creació de la segona versió del corpus de text** del català. A dia d'avui, el projecte disposa d'un primer corpus textual, consistent en 1.770 milions de paraules reunides en 95 milions de frases, que s'ha obtingut a base de descarregar textos de diferents fonts digitals en català (planes web, arxius, etc.), netejar-los i esborrar duplicitats. Ara, es continuarà treballant en aquest corpus de text per generar-ne una segona versió millorada i enriquida que reculli tots els matisos de la llengua escrita, ja siguin variants dialectals o registres lingüístics, com ara el col·loquial, el literari o l'administratiu.

Altres objectius destacats en el full de ruta del projecte AINA per aquest 2022 són:

- Crear tres **serveis lingüístics bàsics** (d'anonimització, de classificació de documents i d'identificació d'entitats i conceptes clau) necessaris per construir futures aplicacions i solucions per a l'usuari final
- Crear **models ("cursos") de la llengua especialitzats** en un àmbit concret (per exemple el de la salut o el jurídic) o en una tasca concreta (per exemple, traducció de textos), per ajudar a les màquines a entendre i a analitzar millor els matisos i el context de les paraules en un text o conversa.
- Crear un **motor de traducció català-castellà** per millorar la qualitat dels motors actualment disponibles
- Implementar un **cas d'ús d'impacte a l'Administració Pública** catalana per mostrar el potencial i la integració a aplicacions reals de les diferents peces desenvolupades per l'AINA.

3M€ de pressupost per al 2022 per a un projecte estratègic

Per fer possible aquest full de ruta, el Departament de la Vicepresidència i de Polítiques Digitals i Territori **destinarà aquest any 3 M€ del seu pressupost al projecte AINA** mitjançant una subvenció directa al BSC, que serà l'encarregat d'executar-lo. Amb aquesta aportació, que **multiplica per 12** el pressupost destinat per la Generalitat al 2021, el Govern reforça la seva ferma aposta per aquest projecte estratègic que té com a objectiu últim garantir que la ciutadania pugui parlar i interactuar en català en el món digital al mateix nivell que els parlants d'altres llengües com l'anglès o el castellà, llengües que, ara per ara, tenen garantida la seva supervivència digital perquè darrere han tingut Estats que han invertit per dotar-les de recursos suficients pel que fa a les tècniques d'aprenentatge i xarxes neuronals en Intel·ligència Artificial.

El projecte AINA, presentat el desembre de 2020, s’emmarca en l’estratègia digital del Govern, a través de dues iniciatives liderades pel Departament de la Vicepresidència: l’Estratègia d’Intel·ligència Artificial de Catalunya (Catalonia.AI), aprovada el febrer del 2020, i el Consell de Direcció interdepartamental per a la promoció del català a Internet i en les tecnologies digitals avançades, aprovat el desembre del 2018.