parlam.omnium.cat

Parla'm

Fem que la tecnologia parli català

GUIA DE GRANS APORTACIONS

INTRODUCCIÓ

Estem fent una gran recollida de textos i àudios en català, per ensenyar a les tecnologies a entendre i parlar català.

PER QUÈ?

Les empreses tecnològiques utilitzen grans bancs de veus i textos per entrenar les tecnologies i incorporar cada llengua a Alexa, Siri, Google o les intel·ligències artificials.

Per aconseguir que el català també estigui disponible, **és imprescindible construir un gran banc de textos i veus en català.**

QUI POT APORTAR VEUS O TEXTOS?

- **Persones individuals**, des del mòbil o ordinador, a través del web del projecte Parla'm: parlam.omnium.cat

- **Mitjans locals, ràdios, revistes, autors/es, streamers, pòdcasts, universitats**... en grans quantitats.

QUÈ FAREM AMB AQUESTES VEUS I TEXTOS?

Els compartirem anonimitzats amb el **Barcelona Supercomputing Center (col·laboradors de l'Aina)**, que els utilitzarà per entrenar reconeixedors i sintetitzadors de la parla i per crear models de llenguatge en català.

COM PODEU CONTRIBUIR-HI?

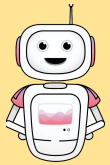
Mitjans nacionals
Mitjans locals
Revistes
Blogs
Editorials
Autors/es
Arxius
Streamers
Pòdcasts
Universitats
Entitats
...

Podeu fer aportacions de grans volums dels vostres àudios i textos en català, per aconseguir l'objectiu de recollir 100.000 hores i 5.000 milions de paraules en català per accelerar la incorporació del català a les tecnologies.

Les aportacions individuals de textos i veus són molt valuoses, i per això també les estem recollint a través del web parlam.omnium.cat. Tanmateix, el ritme de recollida és lent i costós. Per exemple, per aconseguir 20 hores de veus enregistrades a la plataforma CommonVoice, cal la participació d'unes 1.500 persones (de mitjana cada persona s'enregistra llegint 5-10 frases).

Mentrestant, cada aportació massiva de textos o d' àudios en català multiplica ràpidament el volum total de textos i veus disponibles per ensenyar català a les tecnologies.

EXEMPLE



Una ràdio local de Tortosa, aportant els àudios i guions només del programa de notícies dels últims 3 anys, si grava 3 edicions al dia (matí, tarda i nit), **podria aportar de cop 3.204 hores de textos i veus aparellades en la varietat local de català**, que accelerarien substancialment l'ensenyament del català a les tecnologies.

PASSOS PER FER L'APORTACIÓ

1. Omplir aquest formulari, indicant tipus d'àudio o text, format i quantitat aproximada.

2. Una persona del BSC es posarà en contacte amb vosaltres per acordar les condicions i la forma d'entregar els àudios i/o textos.

3. Signar el document legal, confirmant que teniu els drets sobre els àudios i/o textos que entregareu, triant en quines condicions el voleu compartir:

- ❏ Només per entrenar models de llenguatge dins del BSC.
- ❏ Per a integrar el material en un corpus que, a més de permetre entrenar els models de llenguatge dins del BSC, quedarà disponible per a l'estudi de la llengua i la creació de nous models, d'acord amb les tecnologies que es desenvolupin en el futur.

4. Entregar els àudios i textos pel canal acordat.

Tipus d'àudios i textos que podeu aportar:

a) Àudios:
- Amb transcripció literal o completa.
- Amb transcripció aproximada.
- Amb descripció general sobre el contingut
- Sense transcripció

b) Textos:
- Contingut web
- Format editable (doc/odt/txt/rtf...)
- PDF*

*Per ara NO recollim escanejats que requereixin tractament de reconeixement de text

PARLA'M

QUÈ IMPLICA L'APORTACIÓ?

- **Els textos o àudios han d'haver estat obtinguts i transmesos de forma lícita i l'entitat que les aporta ha de tenir ple dret jurídic per a lliurar-los al BSC:** compta amb tots els drets i llicències de propietat intel·lectual i d'explotació econòmica que els són relatius. Amb l'aportació, no perdreu cap dret de propietat intel·lectual ni drets d'explotació econòmica, distribució, etc.

- **L'aportació és unilateral i sense contraprestació de cap mena:** són aportacions lliures en pro del desenvolupament tècnic i el progrés científic, amb l'objectiu de fer possible la presència de la llengua catalana en el terreny digital i de la Intel·ligència Artificial.

- **El BSC tractarà prèviament els àudios i textos aportats**, per excloure dades personals, triar els fragments útils per generar models del llenguatge en català, com també per eliminar o anonimitzar qualsevol informació personal.

- El BSC adoptarà les mesures tècniques i organitzatives adients per al **tractament i la protecció de les dades** aportades. El text o àudio original aportat mai serà publicat pel BSC ni per Òmnium, ni es podrà reconstruir a partir dels fragments utilitzats, impedint qualsevol ús indegut d'aquests.

- **Òmnium i el BSC agrairan públicament l'aportació de l'entitat**, en el format acordat entre les tres parts.

Qui ha aportat textos o àudios en grans volums fins ara?

➜ Enciclopèdia Catalana (400.000 articles i 32 diccionaris)

➜ ACN (113.376 articles)

➜ VilaWeb (58.000 frases)

➜ Màrius Serra i Maria Carme Marí Vila (tota l'obra).

➜ Racó Català (10 milions de missatges del fòrum)

➜ Ateneu Barcelonès (gravacions de col·loquis)

➜ GuiaCat (ressenyes)

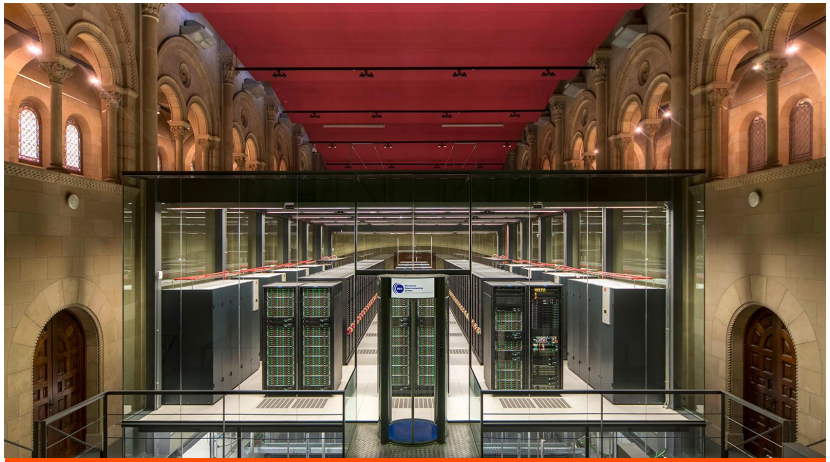
I molts més!

AGRAÏMENTS PÚBLICS

Amb totes les entitats
que facin aportacions
massives:

Foto de representants d'Òmnium i de l'entitat, agafant símbol del Parla'm

Agraïment a XXSS
Menció al butlletí
Logo al web



MAIG

Gran acte d'entrega dels textos i àudios al BSC

ESTEU INTERESSATS/DES EN FER L'APORTACIÓ?

Ompliu aquest formulari abans del dia 20 de març:

ja.cat/aportacionsParlam

Òmnium / el BSC ens posarem en contacte amb vosaltres per efectuar l'entrega.

PARLA'M

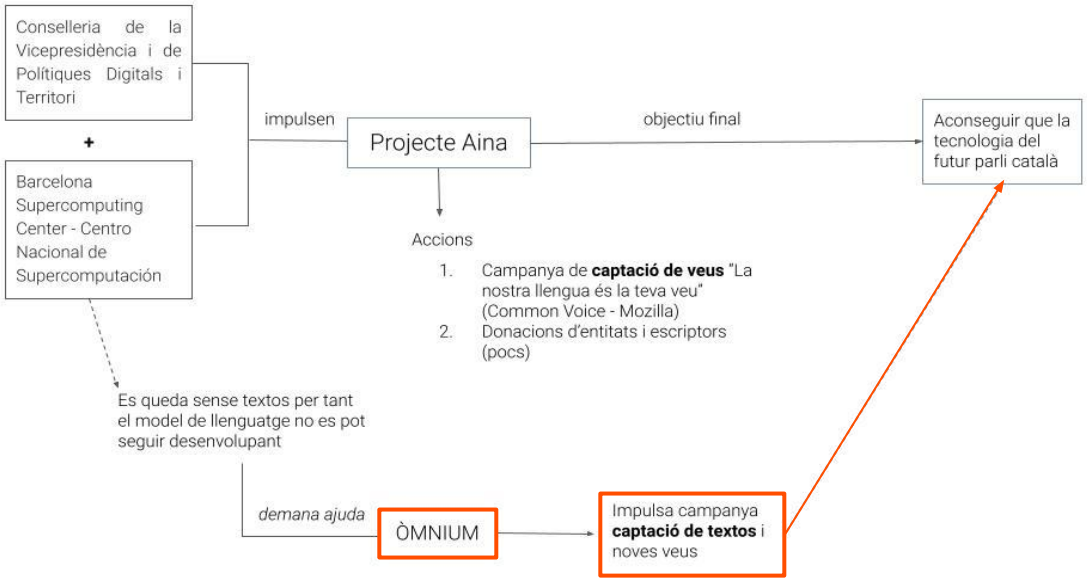
FAQS

ES FARÀ AGRAÏMENT O VISIBILITZACIÓ A LES ENTITATS QUE FACIN APORTACIONS MASSIVES?

Sí, agrairem públicament i visibilitzarem les aportacions massives a les entitats que les hagin fet. Ho farem a les xarxes socials, a través de correu electrònic als nostres socis i la nostra comunitat, com també al web.

Un cop assolim una xifra considerable d'aportacions, realitzarem un acte central de visibilització i agraïment a tots els mitjans, entitats i persones que han fet aportacions massives al projecte Parla'm.

COM ES RELACIONA EL PARLA'M AMB EL PROJECTE AINA?



FAQS

COM ES GESTIONARAN ELS ÀUDIOS I TEXTOS QUE RECOLLIREM?

Compartim els continguts recollits amb el BSC, que els farà servir per entrenar reconeixedors i sintetitzadors de la parla i crear models de llenguatge. Els continguts cedits se segmenten en frases, se seleccionen les que compleixen certes característiques (ni molt curtes, ni molt llargues, sense sigles, paraules o noms estrangers, etc.) i es barregen amb frases d'altres textos o àudios. Aquest contingut es converteix en codi informàtic. D'aquesta manera, el text o àudio original no es pot reconstruir.

HI HA RISC DE PLAGI?

No. En cap cas es fa servir el text o àudio sencer ni es publica. El mecanisme descrit a la pregunta anterior fa impossible reconstruir el text original.

ES PUBLICARÀ A ALGUN LLOC? S'UTILITZARÀ TOT L'ÀUDIO O TEXT?

Els textos i àudios cedits no es publicaran enlloc. Aproximadament s'utilitza un 15% del contingut recollit, processat, fragmentat i barrejat amb altres textos, per convertir-lo en codi apte per generar models de llenguatge.

QUI ELS PODRÀ FER SERVIR?

El Barcelona Supercomputing Center. Si en el formulari marques l'opció *"Vull que els meus textos o àudios, a part de servir per entrenar models de llenguatge dins del Barcelona Supercomputing Center (BSC), passin a formar part del corpus del català que es distribuirà de forma lliure per a entrenar models de llenguatge i estudiar la llengua catalana"*, també centres de recerca i empreses tecnològiques per desenvolupar aplicacions en català i per entrenar els seus propis models per poder oferir els seus productes en català.

FAQS

QUINS TIPUS DE TEXTOS I ÀUDIOS SÓN ÚTILS?

Tots. Des de notícies, pòdcasts, streams en català, programes de ràdio, tertúlies, articles d'opinió, documents oficials, textos infantils, obres literàries, treballs de fi de grau, tesis doctorals, poemes, etc. Pel que fa al format:

Àudios: el format més útil és l'àudio aparellat amb la seva transcripció. També ens resulten útils les transcripcions aproximades, descripcions o inclús els àudios sense text que els acompanyi, perquè podem utilitzar programes de transcripció automàtica.

Textos: els que ens resulten més útils són els que estan en format editable (doc/docx/odt/txt/rtf...). També recollim PDFs on es pugui seleccionar el text. Tenim eines d'extracció de contingut de webs. Per ara no recollim escanejats fotogràfics que requereixin un tractament de reconeixement del text a partir d'imatge.

S'HA DE SIGNAR UN DOCUMENT LEGAL?

Sí. Signaràs amb el BSC un document legal de cessió de les dades per entrenar models de llenguatge, a partir d'aquesta plantilla. En el moment d'entregar els textos o àudios et demanarem quin ús vols que es faci amb el teu àudio o text. En cap cas perdràs l'autoria ni els teus drets.

QUIN PAPER TÉ ÒMNIUM EN LA CAPTACIÓ D'ÀUDIOS I TEXTOS?

Òmnium posa la seva xarxa relacional i la seva estructura nacional i territorial a disposició de la missió d'aconseguir incorporar el català a les tecnologies. L'entitat realitzarà una gran recollida de textos i àudios, i els entregarà al Barcelona Supercomputing Center, que rebrà els textos i els processarà per entrenar models de llenguatge que permetran que les tecnologies i intel·ligències artificials incorporin el català.