15 DESEMBRE 2022 | De 10:00h a 11:30h

Jornada AINA, resultats 2022

https://www.youtube.com/user/BSCCNS

https://politiquesdigitals.gencat.cat/ca/economia/catalonia-ai/aina/

Generalitat de Catalunya
Departament d'Empresa i Treball

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Recordem

OBJECTIUS

1

Proveir el català de la **infraestructura** necessària **per al desenvolupament d'aplicacions basades en IA/TL,** (assistents de veu, traductors automàtics, agents conversacionals, etc)

2

Fer que **la inclusió del català** a les aplicacions de IA/TL sigui **rendible i atractiva per a les empreses del sector**, tant a nivell local com global.

3

Aconseguir que el ciutadà de Catalunya pugui **participar en català** en el món digital **al mateix nivell que un parlant d'una llengua global**, com ara l'anglès o el castellà.

DEFINICIÓ i FONAMENTS

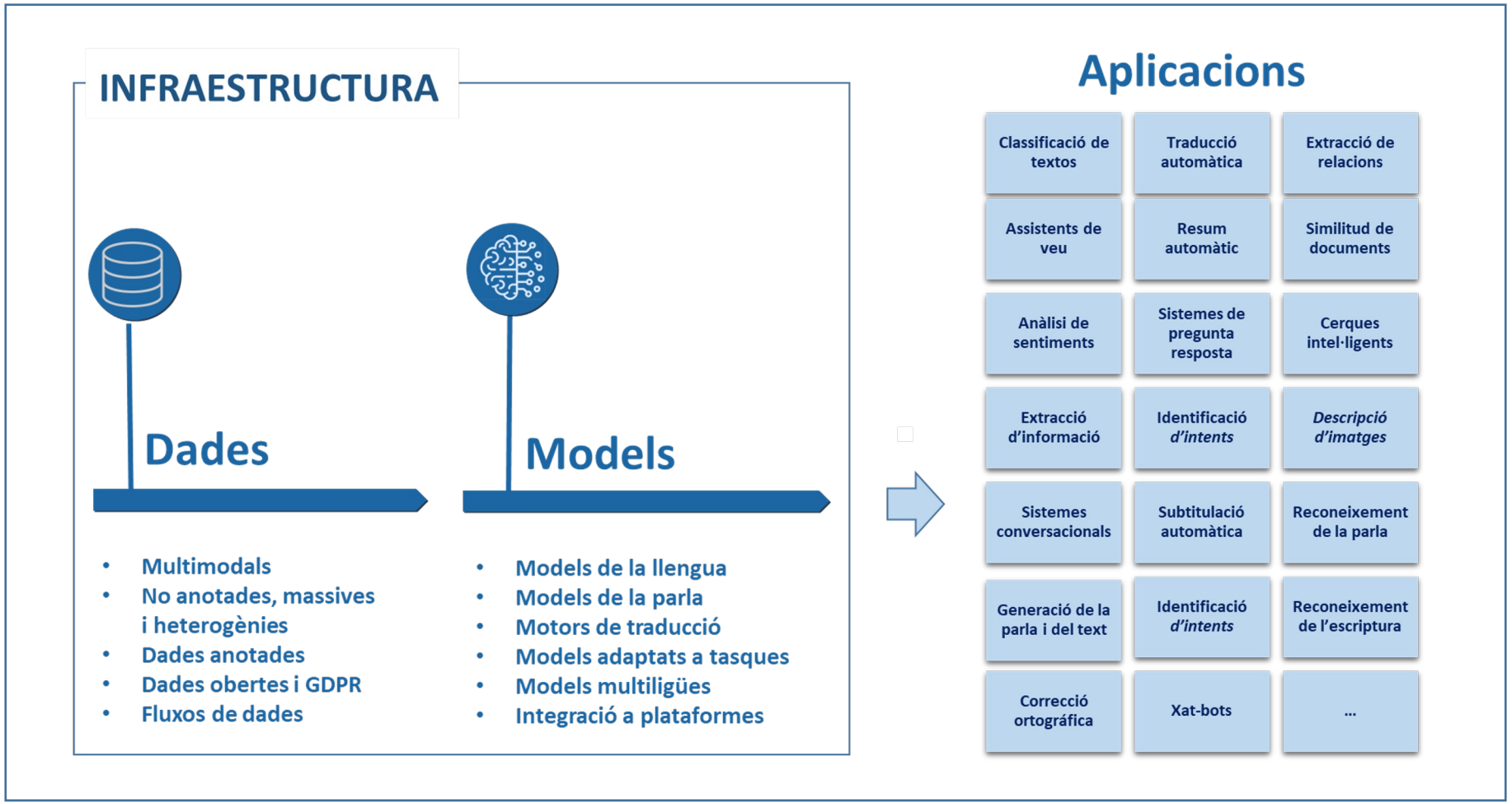
- AINA és essencialment **infraestructura** lingüística.
- El valor de les **dades**.
  - La tecnologia avança molt ràpidament però **les dades són persistents**.
  - Únicament des de la **iniciativa pública**, el català es pot garantir el subministrament de dades suficients.
  - Disposar de **dades de qualitat suficients és un actiu segur i de futur** que garanteix l'actualització de la tecnologia.

ESTRATÈGIA

- AINA implementarà una infraestructura de recollida i neteja de dades amb la **implicació de grans actors.**
- AINA reaccionarà ràpidament als **avenços tecnològics** mitjançant la vigilància tecnològica.
- AINA detectarà i donarà resposta a **noves necessitats** de les empreses i de la societat mitjançant la vigilància sectorial i de mercat.

Barcelona, 15 de febrer del 2022

Recordem


INFRAESTRUCTURA
Dades
Multimodals
No anotades, massives i heterogènies
Dades anotades
Dades obertes i GDPR
Fluxos de dades
Models
Models de la llengua
Models de la parla
Motors de traducció
Models adaptats a tasques
Models multilingües
Integració a plataformes
Aplicacions
Classificació de textos
Traducció automàtica
Extracció de relacions
Assistents de veu
Resum automàtic
Similitud de documents
Anàlisi de sentiments
Sistemes de pregunta resposta
Cerques intel·ligents
Extracció d'informació
Identificació d'intents
Descripció d'imatges
Sistemes conversacionals
Subtitulació automàtica
Reconeixement de la parla
Generació de la parla i del text
Identificació d'intents
Reconeixement de l'escriptura
Correcció ortográfica
Xat-bots
...

Objectius específics

1. Desenvolupament de **serveis lingüístics bàsics i transversals** que serveixin com a baselines i/o mòduls bàsics sobre els quals desenvolupa aplicacions complexes.
2. **Viver de dades**: compilació i preparació de dades massives i de qualitat per a entrenar models genèrics de llengua i models per a tasques específiques.
3. **Entrenament de models pre-entrenats de llengua**, generals i adaptats a domini i/o a tasca, llestos per servir de base per crear noves aplicacions.
4. Entrenament de **models de reconeixement i síntesi de la parla** de qualitat per al català, que puguin ser incorporats als assistents de veu més comuns del mercat
5. Entrenament de **motors de traducció automàtica** de qualitat, tant genèrics com adaptats a domini, amb totes les llengües d'interès.
6. Sistemes intel·ligents, **implementació de prototip(s)** d'impacte que permeti(n) posar en producción els recursos generats i visibilitzar el seu ús i impacte.
7. **Difusió i adopció de la tecnologia**. tasques de difusió i promoció de l'adopció de la tecnologia.

Serveis lingüistics bàsics

- Servei d'anonimització
- Classificació de textos
- Identificació d'entitats i conceptes

Servei d'anonimització

- Anonimitzador de continguts generats per usuaris
- Dades de XitXat (nou dataset!) i de l'Ajuntament de Barcelona
- Accés: https://github.com/TeMU-BSC/AnonymizationPipeline

Classificació de textos

- Classificador entrenat amb dades de l'Agència Catalana de Notícies
- Accés: https://huggingface.co/projecte-aina/roberta-base-ca-v2-cased-tc
- Classificador entrenat amb dades de la Viquipèdia Catalana
- Accés https://huggingface.co/projecte-aina/roberta-base-ca-v2-cased-wikicat-ca

Identificació d'entitats i conceptes

- Model de reconeixement i classificació d'entitats nombrades
- S'ha incorporat a Spacy i SparkNLP
- Accés: https://huggingface.co/projecte-aina/roberta-base-ca-v2-cased-ner

5

Viver de dades

- **Generació de dades anotades**
  - Sistema de benchmarking
- Segona versió del corpus textual
- Corpus de veu
  - Campanya de recollida de veu (CM)
  - Corpus de veu amb transcripció
  - Corpus de veu no alineat
  - Corpus de traducció parla a parla
- Corpus paral·lels per TA
- Provisió de dades

Licitacions per valor de 1,4M d'euros, dividit en 9 lots diferents

1. Traducción de corpus de referencia multilingües.
2. Anotación y vinculación de entidades nombradas
3. Creación de un conjunto de preguntas y respuestas.
4. Anotación de polaridad, emociones y opinión.
5. Creación y anotación de un corpus de NLU para el catalán.
6. Creación y anotación de un corpus de lenguaje abusivo.
7. Creación de un corpus de resúmenes extractivos y abstractivos.
8. Adquisición y traducción de un corpus bilingüe
9. Segmentación, alineación y traducción de corpus de voz.

✓ Guies d'anotació

✓ Preparació de dades d'anotació

Viver de dades

- **Generació de dades anotades**
  - Sistema de benchmarking
- Segona versió del corpus textual
- Corpus de veu
  - Campanya de recollida de veu (CM)
  - Corpus de veu amb transcripció
  - Corpus de veu no alineat
  - Corpus de traducció parla a parla
- Corpus paral·lels per TA
- Provisió de dades

11 nous conjunts de dades anotades per a 8 tasques diferents

- ***CatalanQA**:* amb un total de 21.426 parells preguntes/resposta sobre Wikièdia i notícies..
  - Tasca: Pregunta resposta (QA).
  - Accés: https://huggingface.co/datasets/projecte-aina/catalanqa
- ***WikiCAT_ca:*** Corpus català per a tasques de classificació temàtica de textos no periodístics.
  - Tasca: Classificació de documents.
  - Accés: https://huggingface.co/datasets/projecte-aina/WikiCAT_ca
- ***CAT ManyNames**:* Versió catalana del conjunt de dades ManyNames, orientat a models de Llenguatge i Visió. El corpus consisteix en més de 23.000 imatges i les seves anotacions corresponents, traduïdes automàticament de l'anglès, més un conjunt de 1.072 imatges anotades a mà directament en català.
  - Tasca: Identificació d'imatges.
  - Accés: https://huggingface.co/datasets/projecte-aina/cat_manynames
- ***WNLI-ca**:* Dataset del benchmark GLUE dissenyat per avaluar la capacitat de comprensió i raonament dels models. N'hem encarregat l'adaptació al català a una traductora professional.
  - Tasca: Inferència textual.
  - Accés: https://huggingface.co/datasets/projecte-aina/wnli-ca
- ***XitXat**:* Corpus de 950 converses anotades entre xatbots i usuaris, de 10 dominis diferents.
  - Tasca: NLU (classificació d'intents), detecció d'entitats associades i entrenament/ avaluació de sistemes conversacionals.
  - Accés: https://zenodo.org/record/7276036#.Y2zMn4LMITU

Viver de dades

- **Generació de dades anotades**
  - Sistema de benchmarking
- Segona versió del corpus textual
- Corpus de veu
  - Campanya de recollida de veu (CM)
  - Corpus de veu amb transcripció
  - Corpus de veu no alineat
  - Corpus de traducció parla a parla
- Corpus paral·lels per TA
- Provisió de dades

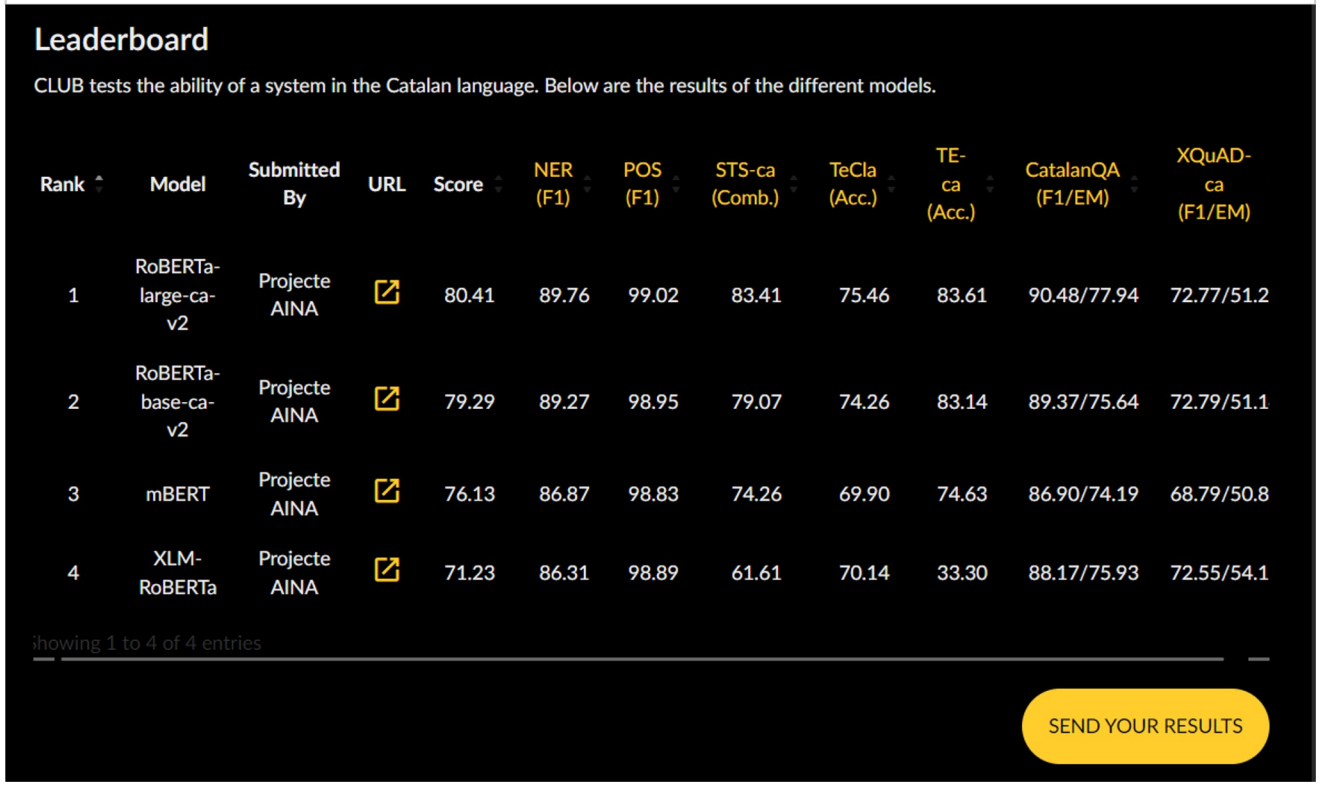
- ***Parafraseja:*** Corpus de 21.984 parells de frases anotades segons si són paràfrasis o no.
  - Tasca: Paràfrasi.
  - Accés: https://huggingface.co/datasets/projecte-aina/Parafraseja
- ***GuiaCat:*** Corpus de 5.750 ressenyes de restaurants en català de la plataforma GuiaCat.
  - Tasca: Anàlisi de sentiments.
  - Accés: https://huggingface.co/datasets/projecte-aina/GuiaCat
- ***NoNiRes:*** Conjunt de dades en les que s'ha anotat les expressions de negació de 20.541 frases en català. S
  - Tasca: Negació.
  - Accés: https://zenodo.org/record/7319487#.Y3S_uL7MLOs
- ***TeCla v2:*** Segona versió del dataset TeCla, corpus de notícies en català per a tasques de classificació de textos multiclasse..
  - Tasca: Classificació de textos.
  - Accés: https://huggingface.co/datasets/projecte-aina/tecla
- ***ANCORA_ca v2:*** Segona versió del dataset ANCORA_ca, corpus d'entrenament de cadenes de processament, afegint la columna NER a la versió CONLLU de UD versió 9, per fer *multitask learning* dins d'Spacy. .
  - Tasca: Anotació d'entitats nombrades i dependències.
  - Accés: https://doi.org/10.5281/zenodo.5036650
- ***MASSIVE 1.1:*** Per suggeriment de l'equip d'AINA, s'ha aconseguit que l'equip d'Alexa a Amazon incorporés el català al dataset multilingüe d'intents per assistents virtuals més gran, el corpus **MASSIVE,**
  - Tasca: Intent classification.
  - Accés: https://github.com/alexa/massive

Viver de dades

- Generació de dades anotades
  - **Sistema de benchmarking**
- Segona versió del corpus textual
- Corpus de veu
  - Campanya de recollida de veu (CM)
  - Corpus de veu amb transcripció
  - Corpus de veu no alineat
  - Corpus de traducció parla a parla
- Corpus paral·lels per TA
- Provisió de dades

Plataforma d'avaluació contínua de models amb avaluació extrínseca sobre 7 tasques diferents

https://club.aina.bsc.es/



Viver de dades

- Generació de dades anotades
  - Sistema de benchmarking
- **Segona versió del corpus textual**
- Corpus de veu
  - Campanya de recollida de veu (CM)
  - Corpus de veu amb transcripció
  - Corpus de veu no alineat
  - Corpus de traducció parla a parla
- Corpus paral·lels per TA
- Provisió de dades

Segona versió del corpus textual, pre-processat i publicat

Corpus	v1 (GB)	v2 (GB)	Estat	Disponible a	Llicència
DOGC	0.78	0.78	v1	OPUS	CC0 4.0
Catalan Open Subtitles	0.02	0.02	v1	OPUS	Oberta
Catalan Oscar	4.00	4.00	v1	OSCAR	CC0 4.0
CaWaC	3.60	3.60	v1	CaWac	CC-BY-SA 4.0
Cat. General Crawling	2.50	2.50	v1	Zenodo	CC-BY 4.0
Cat. Goverment Crawling	0.24	0.24	v1	Zenodo	CC0 4.0
Viquipèdia	0.98	1.10	Actualitzat (04/2022)	-	CC-BY 4.0
ACN	0.45	0.45	v1	Zenodo	CC-BY-NC-ND 4.0
NacióDigital	-	0.45	Nou		
VilaWeb	-	0.06	Nou		
CaCrawlat	-	13.00	Nou	En vies d'exploració	
Padicat	-	0.65	Nou	Zenodo	CC-BY 4.0
RacoCatalà	-	8.10	Nou	HuggingFace	CC-BY-NC 4.0
Total	**12,57**	**34,95**			

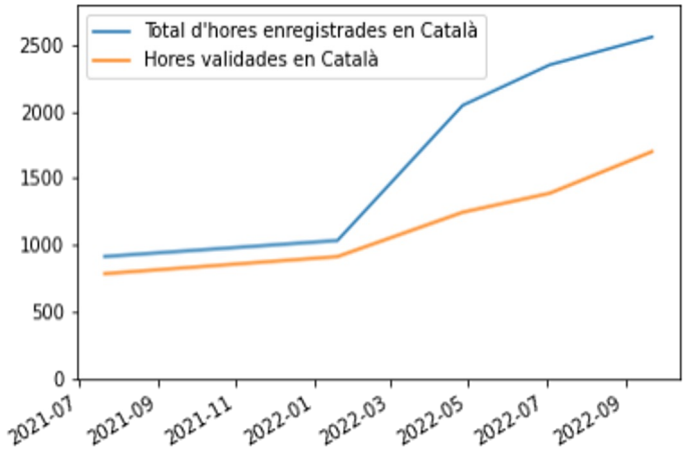
Viver de dades

- Generació de dades anotades
  - Sistema de benchmarking
- Segona versió del corpus textual
- Corpus de veu
  - **Campanya recollida de veu CM**
  - Corpus de veu amb transcripció
  - Corpus de veu no alineat
  - Corpus de traducció parla a parla
- Corpus paral·lels per TA
- Provisió de dades

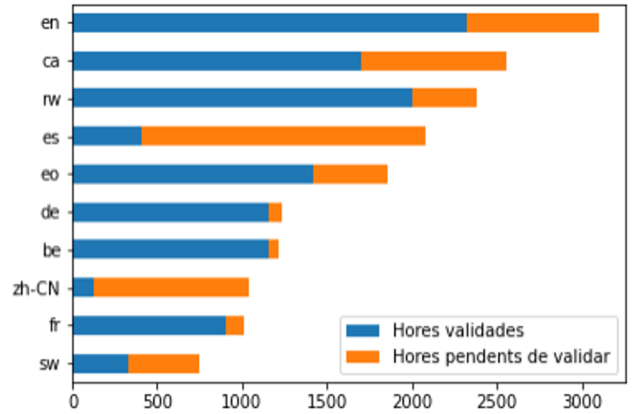
S'ha superat amb escreix l'objectiu de doblar les hores enregistrades **Common Voice**:

- s'ha passat de les **1000h** d'inicis d'any a més de **2700h a principis de desembre**.
- s'ha passat de **6.665 voluntaris** a més de **26.000**

El català s'ha situat com a la s**egona llengua amb més hores enregistrades** a Common Voice i la **tercera pel que fa a hores validades**.



Evolució d'hores enregistrades i validades



Llengües amb més hores enregistrades a la v. 11

- Recopilació de més de **320k frases amb llicència CC0**
- **Servei de monitorització** de la campanya

Viver de dades

- Generació de dades anotades
  - Sistema de benchmarking
- Segona versió del corpus textual
- Corpus de veu
  - Campanya recollida de veu CM
  - **Corpus de veu amb transcripció**
  - **Corpus de veu no alineat**
  - Corpus de traducció parla a parla
- Corpus paral·lels per TA
- Provisió de dades

Datapipe

Una eina per recollir dades de veu

Datapipe és una eina per generar un corpus de la parla a partir del continguts audiovisuals amb llicències obertes a la web. Projecte iniciat per comunitat de programari lliure, que hem adaptat i millorat.

L'objectiu és facilitar la generació de datasets per a ASR, procesant continguts de manera automàtica.

https://github.com/projecte-aina/datapipe

Actualment, l'usem en tres contextos diferents:

- **Canals de youtube**: extreu audioclips dels videos i genera transcripcions candidates en dos tecnologies diferents (una amb Vosk (Kaldi) i l'altra amb Wav2Vec2 model.

  S'han identificat i descarregat continguts amb subtítols (~ 300 hores)
- **CCMA** (~ 4000 hores)
- **IB3** (~ 60 hores)

El conjunt de dades s'utilitzarà en la generació de corpus de veu amb transcripció per a entrenar models de reconeixement de la parla.

Viver de dades

- Generació de dades anotades
  - Sistema de benchmarking
- Segona versió del corpus textual
- Corpus de veu
  - Campanya recollida de veu CM
  - Corpus de veu amb transcripció
  - Corpus de veu no alineat
  - Corpus de traducció parla a parla
- **Corpus paral·lels per TA**
- Provisió de dades

Corpus paral·lels creats per AINA

	Corpus	Font	Llengües	Frases	Domini	Disponible a	Llicència
1	GEnCaTa anotat	crawling	ca, en	51.908	Administratiu	Hugging Face	CC0 4.0
2	GEnCaTa filtrat	crawling	ca, en	38.595	Administratiu	ELRC-Share	CC0 4.0
3	Corpus bilingüe CA-EN de la CE	documents bilingües	ca , en	46.048	Administratiu	ELRC-Share	CC-BY 4.0
4	Col·lecció de corpus CA-EN de l'AP	documents bilingües	ca, en	36.116	Diversos	ELRC-Share	CC-BY 4.0
5	Col·lecció de corpus CA-ES de l'AP	documents bilingües	ca , es	63.773	Diversos	ELRC-Share	CC-BY 4.0
6	Ca-Zh Wikipedia	corpus comparable (Viquipèdia)	ca, zh	111.455	General	Hugging Face	CC-BY 4.0
7	Cyber MT test set	traducció	ca, en, es	1.715	Ciberseguretat	ELRC-Share	CC-BY-NC-SA 4.0
8	Catalan WMT2013	traducció	several	3.000	Notícies	ELRC-Share	CC-BY 4.0

	Corpus	Llengües	Frases
1	Català-Anglès	ca, en	11,58 M
2	Català-Castellà	ca,es	92,58 M

Corpus pre-existents compilats per l'entrenament dels motors de TA

Viver de dades

- Generació de dades anotades
  - Sistema de benchmarking
- Segona versió del corpus textual
- Corpus de veu
  - Campanya recollida de veu CM
  - Corpus de veu amb transcripció
  - Corpus de veu no alineat
  - Corpus de traducció parla a parla
- Corpus paral·lels per TA
- **Provisió de dades**

- **Campanya de recollida de dades adreçada a l'Administració Pública**, mitjançant una plataforma online per pujar les dades i recollir metadades. Ara per ara, hi han participat els següents Departaments de la Generalitat de Catalunya: Economia i Hisenda, Empresa i Treball, Justícia, Drets Socials i Entitat Autònoma del Diari Oficial i de Publicacions. Està previst ampliar-la a totes les Administracions catalanes, també les d'àmbit local.
- Protocol i **API amb el Parlament de Catalunya**
- Treballs amb la **Corporació Catalana de Mitjans Audiovisuals.** per tal d'assegurar el proveïment de dades tenint en compte tots els aspectes legals a considerar.
- Col·laboració inicial amb **Ens Públic de Radiotelevisió de les Illes Balears (IB3)**:
- **Generalitat de Catalunya, departaments de Cultura, Economia, Empresa, Habitatge, Residus:** han cedit frases dels seus webs via crawler per a publicar sota llicència CC0 pel corpus Common Voice.
- **Patrimoni Digital de Catalunya de la Biblioteca de Catalunya**: Ens han cedit els crawlings massius per entrenar els models.

Viver de dades

- Generació de dades anotades
  - Sistema de benchmarking
- Segona versió del corpus textual
- Corpus de veu
  - Campanya recollida de veu CM
  - Corpus de veu amb transcripció
  - Corpus de veu no alineat
  - Corpus de traducció parla a parla
- Corpus paral·lels per TA
- **Provisió de dades**

- **Enciclopèdia catalana**
- **Agència Catalana de Notícies**
- **Diari digital VilaWeb**
- **Revistes Catalanes amb Accés Obert /Consorci de Serveis Universitaris de Catalunya**
- **Mitjà de comunicació Racó Català**
- **Màrius Serra (Escriptor)**
- **Maria Carme Marí Vila (Escriptora)**
- **Joan Pujolar**
- **Associació Cultural El Cérvol**
- **Plataforma per la llengua**
- **Operadora de telecomunicacions Parlem**
- **Empresa 1MillionBot**
- **Diari La Veu**
- **Òmnium Cultural**
- **Coordinadora d'Associacions per la Llengua Catalana**
- **Plataforma GuiaCat**
- **…**

Models de la llengua

Generació de models de la llengua

- **Models Transformers de la llengua**
- Implementació mòduls de català per a frameworks d'impacte
- Participació en models multilingües i massius en col·laboració amb altres iniciatives

Models Transformers

https://huggingface.co/projecte-aina

HUGGING FACE

- RoBERTa-base v2
  - Descripció: segona versió del model RoBERTa-base (12-layer, 768-hidden, 12-heads, 125M parameters) del català entrenat amb la segona versió del corpus textual català
  - Accés: https://huggingface.co/projecte-aina/roberta-base-ca-v2
- RoBERTa-large
  - Descripció: model RoBERTa-large (24-layer, 1024-hidden, 16-heads, 355 M parameters) del català entrenat amb la segona versió del corpus textual català
  - Accés: https://huggingface.co/projecte-aina/roberta-large-ca-v2
- Lomgformer
  - Descripció: model Longformes basat en RoBERTa-large que permet inputs de fins a 4096 tokens.
  - Accés:
- Destilat
  - Descripció: model destilat basat en RoBERTa-large que permet disposar d'un molde més petit facilitant-ne la producció.
  - Accés:

Models de la llengua

Generació de models de la llengua

- **Models Transformers de la llengua**
- Implementació mòduls de català per a frameworks d'impacte
- Participació en models multilingües i massius en col·laboració amb altres iniciatives

Models Adaptats a tasques específiques

https://huggingface.co/projecte-aina

HUGGING FACE

- **RoBERTa-base v2 fine-tuned for POS**
  - Tasca: anotació morfosintàctica.
  - Accés: https://huggingface.co/projecte-aina/roberta-base-ca-v2-cased-pos
- **RoBERTa-base v2 fine-tuned for QA**
  - Tasca: Pregunta/resposta (QA)
  - Accés: https://huggingface.co/projecte-aina/roberta-base-ca-v2-cased-qa
- **RoBERTa-base v2 fine-tuned for TE**
  - Tasca: Implicació textual
  - Accés: https://huggingface.co/projecte-aina/roberta-base-ca-v2-cased-te
- **RoBERTa-base v2 fine-tuned for STS**
  - Tasca: Similitud textual semàntica
  - Accés: https://huggingface.co/projecte-aina/roberta-base-ca-v2-cased-sts
- **RoBERTa-base v2 fine-tuned for Paraphrase Detection**
  - Tasca: Paràfrasi
  - accés: https://huggingface.co/projecte-aina/roberta-large-ca-paraphrase
- **Word embeddings Floret per al català, v1.0**
  - Accés: https://zenodo.org/record/733033

Models de la llengua

Generació de models de la llengua

- Models Transformers de la llengua
- **Implementació mòduls de català a frameworks d'impacte**
- Participació en models multilingües i massius en col·laboració amb altres iniciatives

spaCy

- S'ha desenvolupat una cadena de processament per al català a **Spacy 3.4** que fa servir el model ***Transformer*** català més recent (https://huggingface.co/projecte-aina/roberta-large-ca-v2) i que, a més de millorar les prestacions del components existents, afegeix la funcionalitat de classificació temàtica de textos, que les versions oficials d'Spacy no tenen. La nova versióes pot instal·lar directament des de HuggingFace.
  - Accés: https://huggingface.co/projecte-aina/ca_bsc_demo_trf
  - Accés a la demo: https://spacydemo.aina.bsc.es/
- També en aquesta plataforma, s'ha desenvolupat una cadena de processament que fa servir els **embeddings floret** generats amb el corpus general de català, amb una mida més lleugera, sense perdre prestacions, i amb la funcionalitat afegida de fer estimacions de similitud semàntica entre paraules.
  - Accés https://huggingface.co/projecte-aina/ca_bsc_demo_md

John Snow LABS Spark

- S'ha desenvolupat d'una cadena de processament per al català dins de la plataforma **SparkNLP**, amb funcionalitats de tokenització, detecció d'entitats, reducció de dimensionalitat, lematització, embeddings, chunking y normalització. Aquesta plataforma és altament escalable i de robustesa industrial.
  - Accés: https://nlp.johnsnowlabs.com/2022/07/11/pipeline_md_ca_3_0.html

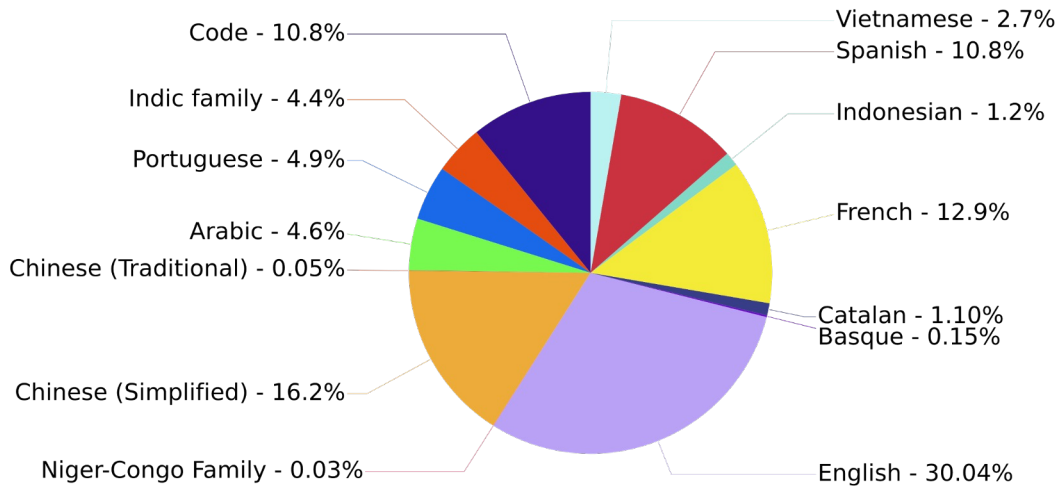
Models de la llengua

Generació de models de la llengua

- Models Transformers de la llengua
- Implementació mòduls de català per a frameworks d'impacte
- **Participació en models multilingües i massius en col·laboració amb altres iniciatives**

BigScience és un projecte col·laboratiu promogut per Hugging Face (+1000 investigadors, +60 països, +250 institucions)

- https://huggingface.co/bigscience/bloom
- s'han recollit grans quantitats de dades multilingües
- S'ha creat BLOOM un **model generatiu multilingüe** entrenat durant 4 mesos al superordinador Jean Zay a París.



Models de la parla

Generació de models de la parla

- **Models de tecnologies de la parla**
- La presència dels models als ecosistemes i plataformes d'impacte

S'ha aprofitat els enregistraments fets per **Common Voice** per crear dos models de la parla

Reconeixement de Parla - Nvidia Nemo: Alta precisió, mida mitjana, fàcilment desplegable amb el framework de **Nemo**.

- 36.5M paràmetres, WER of 6.684.
- https://huggingface.co/projecte-aina/stt-ca-citrinet-512

Síntesi de la Parla - Multiparlant: Model de síntesi de la parla d'alta qualitat i d'alt rendiment. S'ha entrenat amb 257 veus amb diverses variants dialectals. A més de CV, aprofita els corpus de Festcat i de Google TTS-ca. Fàcilment desplegable amb el framework de **Coqui**.

  - https://huggingface.co/spaces/projecte-aina/VITS_ca_multispeaker
  - integrat a la demo amb Bookline i al bot d'AINA

Models de la parla

Generació de models de la parla

- Models de tecnologies de la parla
- **Presència dels models als ecosistemes i plataformes d'impacte**

COQUI is a library for advanced Text-to-Speech generation

https://github.com/coqui-ai/TTS

Rasa is the leading platform for transforming how people interact with organizations through extensible conversational AI

https://rasa.com/

NVIDIA NeMo, part of the NVIDIA AI platform, is a framework for building, training, and fine-tuning GPU-accelerated speech and natural language understanding (NLU)

https://developer.nvidia.com/nvidia-nemo

Motors de Traducció

Pendents de disposar de més dades d'entrenament pels motors entre anglès i català (licitació en curs), s'han generat uns primers models molt competitius:

	Aina MT	SoftCatalà	Google TR
CA → EN	44,0	39,2	**45,0**
EN → CA	41,0	38,8	**41,1**
CA → ES	**56,8**	53,4	56,7
CA → ES legal*	**85,5**	80,9	81,4

Avaluació (mètrica: BLEU)

*Aquest motor s'ha avaluat sobre text d'especialitat administratiu-legal. La resta s'han avaluat sobre text general i de diversos àmbits.

- ***Motor de traducció genèric ca-en***
  - Entrenat amb corpus de 11 milions de frases
  - Accés: https://huggingface.co/projecte-aina/mt-aina-ca-en.
- ***Motor de traducció genèric en-ca***
  - Entrenat amb corpus de 11 milions de frases
  - Accés: https://huggingface.co/projecte-aina/mt-aina-en-ca
- ***Motor de traducció genèric ca-es***
  - Entrenat amb corpus de 92,5 milions de frases
  - Accés: : https://huggingface.co/projecte-aina/mt-aina-ca-es
- ***Motor de traducció de textos legals ca-es***
  - Adaptat a domini amb un corpus de 62.773 unitats de traducció
  - Accés: https://huggingface.co/projecte-aina/mt-aina-ca-es-adm

- Motor de traducció ca-zh
- Motor de traducció zh-ca
- Motor de traducció de-ca

On trobar-ho?

HUGGING FACE

https://huggingface.co/projecte-aina

27 models i 23 datasets

GitHub

https://github.com/projecte-aina

15 repositoris

zenodo

https://zenodo.org/communities/catalan-ai

24 recursos

European Language Resource Coordination
Connecting Europe Facility
ELRC-SHARE

https://elrc-share.eu/repository/search/

Prototips i demostradors (https://aina.bsc.es/)

Natural Language Processing

Demostració de les capacitats de la plataforma Spacy entrenada amb models i dades d'AINA per fer **comprensió d'un text, detectant tema, entitats, relacions, etc**. Les cadenes de processament son a la base de moltes aplicacions com ara xatbots i aplicacions de monitorització de mitjans. https://spacydemo.aina.bsc.es

Pregunta/Resposta a la Viquipèdia

Demostració del **model extractiu de pregunta/resposta** "roberta-base-ca-v2-cased-qa". Donat un tema o pàgina, i una pregunta, pot trobar el fragment on hi ha la resposta a la Wikipèdia. https://viquiqa.aina.bsc.es

Traducció Automàtica

Traductors entre català i anglès i de català a castellà, en text genèric i d'especialitat administratiu-legal.
https://traductor.aina.bsc.es/

Prototips i demostradors (https://aina.bsc.es/)

Integració de la veu d'AINA a un assistent virtual de mòbil

tts-api és un software que permet una integració fàcil de les veus d'AINA a diverses aplicacions. Aquest demostrador està desenvolupat juntament amb **bookline**, que van integrar una de les veus al seu assistent virtual de mòbil. A més es pot sentir les veus d'AINA aquí https://huggingface.co/spaces/projecte-aina/VITS_ca_multispeaker

La síntesi de la parla permet parlar a les màquines és la interfície natural entre humans i màquines.

Transcripció automàtica

oTranscribe+ és una eina de transcripció automàtica que facilita també l'edició. Transcriu els enregistraments automàticament, sense compartir els àudios amb un servei extern, i fent servir els **models de reconeixement de la parla** emmagatzemats en local.

L'eina està disponible sota llicència oberta MIT i es pot provar a https://otranscribe.bsc.es/ o com a aplicació d'escriptori.

Xatbot de veu

Hem desenvolupat un xatbot per respondre a preguntes sobre AINA i hi hem afegit funcionalitats de veu mitjançant una aplicació de web. El software està publicat amb una llicència lliure per l'ús obert per crear altres experiències conversacionals en català.

Comunicació

- ***AINA: novetats, objectius inicial i expectatives: https://youtu.be/zeXZUUZErXI***
- ***Jornades Juliol 2022:***
  - ***AINA en acció: resultats, experiències i expectatives de la indústria : https://youtu.be/aFn3ids_Avw***
  - ***La veu dels usuaris i les empreses: https://youtu.be/4_uyk6UAIpk?t=23***

- ***https://twitter.com/projecte_aina*** **2.814** Seguidors

- Una **Newsletter** (vegeu aquí), de publicació bimensual, amb 400 substriptors:
  - **4 de març:** Notícies del Projecte Aina!
  - **3 de maig:** Nova versió del corpus Common Voice i altres notícies
  - **11 de maig:** Fem que la tecnologia hi vegi en català!
  - **4 de juliol:** AINA en acció: resultats, experiències i expectatives de la indústria
  - **11 de juliol:** AINA en acció. Recordatori i enllaç actualitzat.
  - **22 de juliol:** AINA en acció: resum de la jornada i enllaços
  - **5 d'agost:** Publicada la licitació de creació i anotació de dades per a la IA en català, dins del projecte Aina.
  - **24 d'octubre:** Una tardor plena de novetats


SC SOFTCATALÀ
Corporació Catalana de Mitjans Audiovisuals
Ens Públic de Radiotelevisió de les Illes Balears
Enciclopèdia Catalana
AteneuBarcelonès
A B C D E F G H I J K
L M N O P Q R S T U
V W X Y Z
UPC
alexa
1million bot
nuclia
M47 LABS
parlem telecom
VilaWeb
EL CÉRVOL
CERVIÀ DE LES GARRIGUES
Associació Cultural El Cérvol
PLATAFORMA PER LA LLENGUA
bookline
basetis
Guiacat
"la nostra guia"
*STeL
CLiC
verificat
padicat
PATRIMONI DIGITAL DE CATALUNYA
Sabadell
PARLAMENT DE CATALUNYA
OMNIOS
CaixaBank
Generalitat de Catalunya
CSUC
acn
RACÓ CATALÀ
LLEGEIXPENSAOPINA
ÒMNIUM
RACO
Revistes Catalanes amb Accés Obert
CAL
pangeanic
BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
Aina