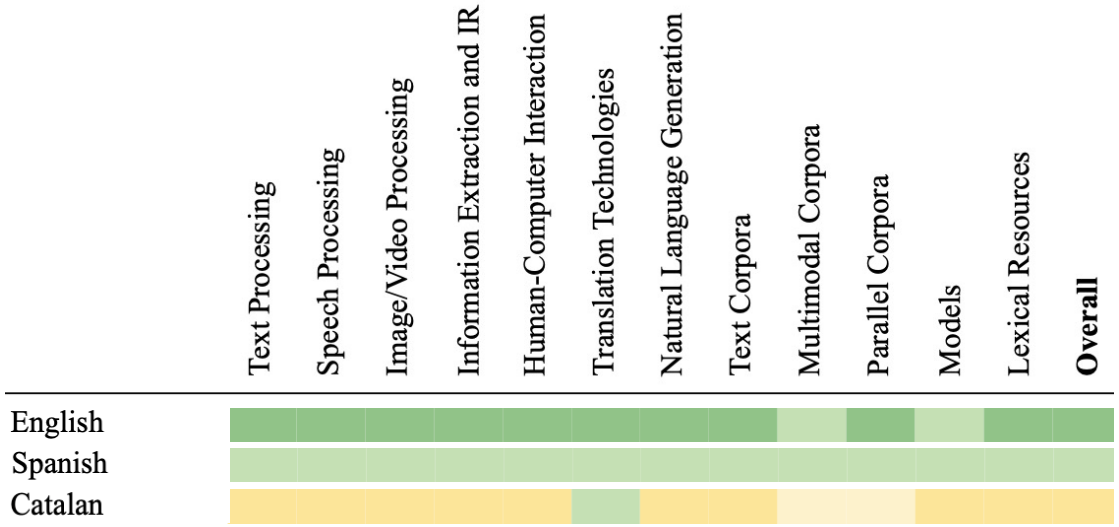
Building a Data Infrastructure for a Mid-Resource Language: The Case of Catalan

Aitor Gonzalez-Agirre, Montserrat Marimon, Carlos Rodriguez-Penagos,
Javier Aula-Blasco, Irene Baucells, Carme Armentano-Oller,
Jorge Palomar-Giner, Baybars Kulebi, Marta Villegas

LREC-COLING 2024

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Background



State of technology support for selected languages (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support).

Source: Giagkou et al., 2023: 80

English linguistic dominance has transformed into a digital dominance.

Leaving languages with “weak” and “fragmentary” support behind widens the gap.

There is an urge to provide languages other than English with technology support.

Objective and Strategy

5-year project funded with €15M by the Catalan Government.

Technology advances rapidly, but data persist.

Strategy focused on:

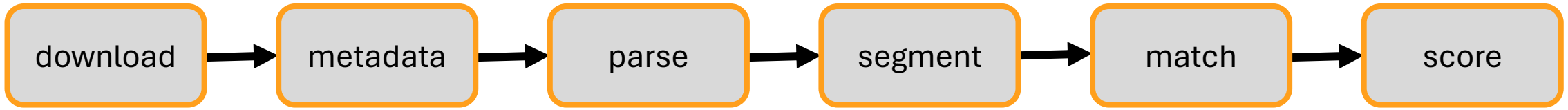
- Sustainability and long-term supply.
- Openness and FAIR principles.
- Presence in multilingual reference repositories.
- Identification of gaps and relevant tasks.
- Deployment and ready-to-use resources.

Aina
CATalog

Parlament Pipeline

Use *Parlament de Catalunya* parliamentary sessions.

Access via two API endpoints: session list + session detail.



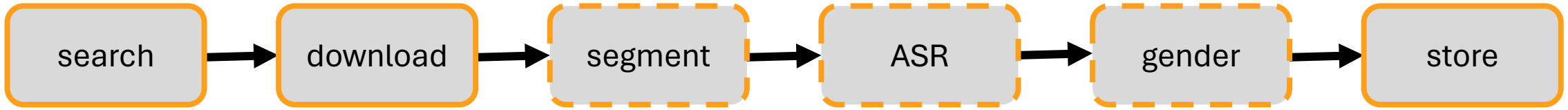
Currently gathered 1,061 hours and over 10M words.

PARLAMENT DE CATALUNYA

YouTube Pipeline

Use YouTube videos with permissive license.

Adapted from existing software.



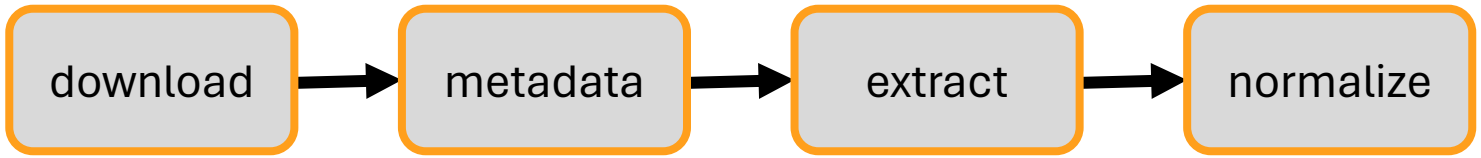
Currently gathered 1,163 videos with 1,563 hours of recordings.

SOFTCATALÀ YouTube

DOGC Pipeline

Use *Diari Oficial de la Generalitat de Catalunya* publications.

Access via API endpoint.



Currently gathered 30,503 publications with over 71M words in Catalan, and 14,258 publications with 54M words in Spanish.

DOGC
Diari Oficial
de la Generalitat de Catalunya

transparència
catalunya

WikiExtractor-V2

Use Wikipedia articles. Multilingual.

Adapted from original WikiExtractor tool.

Improvements in processing, output, configuration + modularity.

Currently gathered 692,632 documents and over 267M words.

Translated Datasets

Professional translations of English subsets in several multilingual reference dataset.

Focus on language naturalness, with option to localize if required, and ability to fix problems existing in English datasets.

Detailed guidelines are always provided, and translation are revised manually in small batches to check the quality.

Matched source format and contacted original authors.

Aina

Developed Datasets

Prioritized tasks that align with the directions of current research and industry trends, that are considered unsolved, and for which translation would be too expensive or unreliable.

Datasets include CQA, NLU, summarization, NER, conversations, TE, EQA, abusive language identification, SSA + SED.

Aina

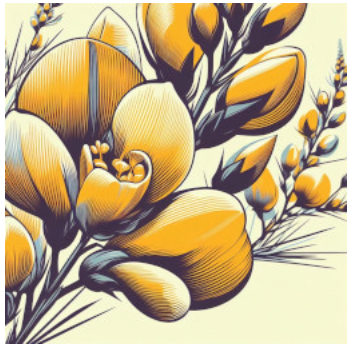
Deployment and Dataset Exploitation

Using part of the CATalog data, we continually pre-trained a BLOOM-7.1B model, resulting in FLOR-6.3B.

We use the datasets mentioned to fine-tune other models.

We currently offer 56 publicly available models for download.

We present a >200k instruction tuning dataset.



Aina

InstruCAT

Conclusions and Lessons Learned

Openness and standardization of formats is key.

Finding suitable texts with permissive licenses that can be used to annotate for some tasks is the biggest challenge.

“Things move fast”: reference multilingual benchmarks vanish or are not properly maintained.

Self-contain your own work to avoid wasting resources.

Acknowledgements

This work has been promoted and financed by the *Generalitat de Catalunya* through the Aina project.

This work is funded by the *Ministerio para la Transformación Digital y de la Función Pública* and *Plan de Recuperación, Transformación y Resiliencia* - Funded by EU – NextGenerationEU within the framework of the project ILENIA with reference 2022/TL22/00215337, 2022/TL22/00215336, 2022/TL22/00215335 and 2022/TL22/00215334.

Building a Data Infrastructure for a Mid-Resource Language: The Case of Catalan

Aitor Gonzalez-Agirre, Montserrat Marimon, Carlos Rodriguez-Penagos,
Javier Aula-Blasco, Irene Baucells, Carme Armentano-Oller,
Jorge Palomar-Giner, Baybars Kulebi, Marta Villegas

LREC-COLING 2024

Barcelona Supercomputing Center
Centro Nacional de Supercomputación