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
Barcelona Supercomputing Center
{aitor.gonzalez, montserrat.marimon, carlos.rodriguez1,
javier.aulablasco, irene.baucells, carme.armentano,
jorge.palomar, baybars.kulebi, marta.villegas}@bsc.es

Abstract

Current LLM-based applications are becoming steadily available for everyone with a reliable access to technology and the internet. These applications offer benefits to their users that leave those without access to them at a serious disadvantage. Given the vastly large amount of data needed to train LLMs, the gap between languages with access to such quantity of data and those without it is currently larger than ever. Aimed at saving this gap, the Aina Project was created to provide Catalan with the necessary resources to keep being relevant in the context of AI/NLP applications based on LLMs. We thus present a set of strategies to consider when improving technology support for a mid- or low-resource language, specially addressing sustainability of high-quality data acquisition and the challenges involved in the process. We also introduce a large amount of new annotated data for Catalan. Our hope is that those interested in replicating this work for another language can learn from what worked for us, the challenges that we faced, and the sometimes disheartening truth of working with mid- and low-resource languages.

Keywords: Catalan, dataset, evaluation, infrastructure, sustainability

1. Context and Motivation

The digital world has become an integral part of our lives, reshaping how we connect, communicate, express creativity, or conduct business, among many others. This has empowered individuals and organizations alike, offering unprecedented opportunities for innovation, collaboration and global reach. In this digital age and its diverse applications, language stands as the lifeblood. Language serves as the medium through which humans and machines communicate, share information, and collaborate. As technology continues to advance, the role of language becomes increasingly pivotal, helping to bridge the gap between human users and the digital tools and services that define our current lifestyle.

While English speakers have access to countless Artificial Intelligence (AI) and Natural Language Processing (NLP) applications and services, speakers of other languages find themselves at a disadvantage, as their languages lack adequate digital resources when compared to English. In 2012, the META-NET reports[1] concluded that at least 21 European languages were in danger of digital extinction due to a severe lack of technology support. In a more recent study (Rehm and Way, 2023), the situation of around 30 European languages was analyzed, with the authors concluding that not only the digital inequality continues, but the gap between English and other European languages is widening. In this study, Catalan belongs to the languages with “fragmentary support”, far away from English, with “good support”, and below the exclusive group of languages with “moderate support”.

In this context, the Aina Project was created in an attempt to close the existing gap. Aina is a linguistic infrastructure project that aims to provide Catalan with the necessary resources to assure its presence in AI/LT-based applications and to promote and guarantee the use of Catalan in the digital era. The contributions presented in this paper are: *(i)* a set of strategies to consider when improving technology support for a mid- or low-resource language, *(ii)* four pipelines for sustained high-quality data acquisition, *(iii)* six translated datasets in Catalan to be included in multilingual repositories, *(iv)* eleven original datasets in Catalan created with high-quality data and annotation processes, *(v)* an instruction dataset in Catalan, *(vi)* multiple Catalan fine-tuned models for a variety of relevant tasks.

2. Objective and Strategy

Aina is a 5-year project, funded with €15M by the Catalan Government, which started in 2022. This project essentially serves as a linguistic infrastructure in which the value of data is paramount: technology advances rapidly, but data persists. Having

[1] http://www.meta-net.eu/whitepapers/overview

LREC-COLING 2024, pages 2556–2566
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

a sufficient quantity of high-quality data is a valuable and future-proof asset that ensures that technologies remain up to date.

Large Language Models (LLMs) need increasing volumes of data, and it is difficult to meet this demand for mid- and low-resource languages like Catalan. If we were to train a relatively small model of the GPT-3 family, one with 1.3 billion parameters, following the Chinchilla formula (Hoffmann et al., 2022), we would require a corpus of 26 billion tokens. Considering that the largest Catalan corpus previously available contained 13B tokens and 6.5B words (Xue et al., 2021), this gives an idea of the scale we are dealing with. Concurrent to this paper, we release a larger dataset with 23B tokens and 17.4B words (Palomar-Giner et al., 2024).[2]

To maximize the impact of our work, the strategy followed for the development of new datasets in Catalan is based on the following aspects:

- **Sustainability and long term supply**: To ensure sustained and extended data provision, we implement a system that guarantees the sustainability of data supply in all modalities. Section 3 shows some implemented examples of how data acquisition was operationalized.

- **Openness and FAIR[3] principles**: Whenever possible, datasets are distributed under permissive licenses. The objective is to avoid licensing contamination for the models generated with the project's datasets and to ensure that they comply with all legal requirements. Datasets are always available online in standardized formats.

- **Presence in multilingual reference repositories**: The use of multilingual reference datasets for training and, especially, model evaluation is a common practice. Therefore, the project's goal is to ensure the presence of Catalan in these datasets when it was not originally included, so that cutting-edge research also encompasses Catalan. The details can be found in Section 4.1.

- **Identification of gaps and relevant tasks**: In the case of newly-created datasets, the project's strategy is to focus on the development of datasets that are widely used by the literature in English but are missing in Catalan. Tasks for which translating an English dataset would result in a low-quality and/or artificial dataset such as Conversational Question Answering (CQA), Textual Entailment (TE) and abusive language identification, require the creation of Catalan-exclusive datasets. Section 4.2 includes more details on this matter.

- **Deployment and ready-to-use resources**: All datasets are made accessible through data loaders on Hugging Face in order to optimize their usage across sector-specific platforms and to enable rapid integration. We also offer them through Zenodo. See Section 5 for further details.

3. Operationalizing and Sustaining Data Acquisition

Aina has implemented various data acquisition methods that automate this process and ensure the long-term supply and updating of data in the future. In this section, we present four tools we have developed with that purpose. These tools are available through Hugging Face.[4]

3.1. Parlament Pipeline

One of our primary collaborators is the *Parlament de Catalunya*, the legislature of the regional Catalan government. They openly publish videos and transcriptions of parliamentary sessions on their webpage.[5] These recordings prove to be an important asset for the Catalan language, due to their multi-modality (text and audiovisual), and representation of dialectal variants from all around Catalonia, as evidenced by the use of these recordings in creating ASR datasets (Kulebi et al., 2022). Coincidentally, the *Parlament* has an ongoing process to develop their open data infrastructure, and they chose the Aina project as an early adopter of their open data services. These services consist of two API endpoints which provide near real-time data on the published content on their website, the first endpoint, *session list*, returns a list of recently published sessions with metadata such as URI, type of session, and session id, among others. The second endpoint, *session detail*, takes the session id and returns a list of speeches for that session. This includes the name and details of the speaker, the start and end time of their speech, the reference video URI, and the reference transcription URL. As a part of their open data efforts, the *Parlament* has the intention to open the API for everyone at some point in the future.

Based on the data and metadata gathered, we build a pipeline to process the parliamentary content to generate a speech corpus. The first step of

[2] https://huggingface.co/datasets/projecte-aina/CATalog

[3] https://www.go-fair.org/fair-principles/

[4] https://huggingface.co/projecte-aina

[5] https://www.parlament.cat

the pipeline consists of a recursive download of resources from the API endpoints and modification of the metadata files with the information on the relative paths of the downloaded resources. The rest of the pipeline is similar to the ParlamentParla work (Kulebi et al., 2022). This is, starting from the metadata, it processes the PDF files and audio recordings in order to generate speech segments of 5-20s with their corresponding cleaned transcriptions. Finally, these segments are scored automatically according to the quality of the transcription. The difference between this pipeline and past work is that it has an updated PDF parsing, forced alignment procedure, and further functionality to process both Catalan and Spanish content. As a first step, we have downloaded all recordings of the sessions from 23 January 2008 to 24 October 2023, and currently have 317 sessions, totalling 1,061 hours and over 10M words. Depending on the quality of the segments, we soon expect to have an increment of at least 300 hours to the current ParlamentParla dataset. We are now working on evaluating the segment quality before publishing the new dataset.

Our goal is to automatically run the pipeline every quarter to consistently update the dataset. The details of this will be explained in a future work.

3.2. YouTube Pipeline

YouTube is an important resource for training speech models, as evidenced by the extreme volumes of data that recent speech models need for their effective training. Most famously, the whisper model of OpenAI (Radford et al., 2022) needs 640k hours. Recent TTS models like TorToiSe (Betker, 2023) or Matcha-TTS (Mehta et al., 2023) use around 60k hours of data, which is currently unattainable for Catalan. We thus developed an open software to download videos from YouTube, named *Datapipe*.[6] This tool is forked from the *datapipe* software used by the NGO Softcatalà,[7] to adapt it to our needs. These needs include a convenient deployment for production, integration with Kubernetes[8] development tools, a downloader for user generated subtitles, and a filtering option to ensure that the licenses of the videos allow their download.

The tool does not scan the entirety of YouTube, but takes a channel name or a search term as an input. It then starts to download the list of returned videos if the licenses are open. The application has the option to segment the audios and apply ASR to the segments, in addition to doing gender detection. Later, all the information is stored in an SQL database.

Initially, we focused on downloading Catalan videos with open licenses and subtitles, due to our aim of having relatively high quality data. This enabled us to acquire 1,163 videos with 1,563 hours of recordings. For the future, we would like to work on videos without subtitles and extract the segments with the best ASR result, corroborating the quality of the segments via multiple ASR models.

3.3. DOGC Pipeline

The *Diari Oficial de la Generalitat de Catalunya* (DOGC) is the official government gazette in Catalonia. This publication contains valuable data on everything related to the public or legal notices in Catalonia. To operationalize the acquisition of this data, we use the API provided by *Transparencia de Catalunya*,[9] through which a series of metadata related to all DOGC publications can be obtained. These include, among others, the URL corresponding to all publications, both in Catalan and Spanish. We then extract from the HTML structure obtained from these URLs, which is always the same in all DOGC notices, the plain text that constitutes the body of the publication. Finally, we apply several transformations to the text in order to normalize its paragraph structure so that it is coherent for the training of a LM.

The DOGC updates the API data at the first of each month with the previous month's publications, so this pipeline has been scheduled to consistently run on the 6th of each month. So far (early October 2023), we have extracted 30,503 publications in Catalan, including a total of over 71M words, and 14,258 publications (containing over 54M words) in Spanish.

3.4. Wikipedia Extractor

Wikiextractor-V2[10] is a fork of the WikiExtractor project.[11] This tool provides additional functionality while addressing flaws in the original tool. We tested the quality of our extractor using Wikipedia dumps from 2023. In addition to English, *WikiExtractorV2* can extract high-quality data from a variety of languages, including Catalan. The contributions of this tool include support for list and nested lists, UNICODE characters, and math and chemical formulas. It also returns the output as plain text (.txt) and plain JSON, includes the documents's title by default and fixes several issues with specific templates handled (e.g., segle, coords, etc). There is also an option to ban specific sections, templates, or entire documents by means of a config file. The tool can currently be used both as a single tool or as a module

[6] https://github.com/projecte-aina/datapipe

[7] https://www.softcatala.org

[8] https://kubernetes.io

[9] https://analisi.transparenciacatalunya.cat

[10] https://github.com/langtech-bsc/Wikiextractor-V2

[11] https://github.com/attardi/wikiextractor

ready-to-connect to a pipeline, improving memory and time usage. Thanks to these contributions, and through the use of this new version, we have extracted 692,632 documents (after deduplication and pre-processing), containing over 267M high-quality Catalan words, both in form and content.

4. Annotated Datasets

In this section, we introduce the datasets that have been created within the first two years of the project. We present both datasets generated through human translations of existing ones, and datasets developed independently. These datasets are publicly accessible through Hugging Face,[12] and Zenodo.[13] Through these links, the translation and annotation guidelines for each dataset are also available.

4.1. Adding Presence in Multilingual Repositories

At the beginning of the project, we commissioned professional translations of the English subsets included in several multilingual reference datasets that serve various relevant language understanding tasks. We prefer professional translation over machine translation due to our aim of providing Catalan with high quality datasets. Even though most of the original multilingual datasets included in this section were created from machine translations, professional translations result in datasets that better reflect natural language, and give the option to localize when suitable. This means that our datasets in Catalan are sometimes of better quality than what would have been included in the original datasets.

Detailed translation guidelines were always provided, ensuring that translated datasets are linguistically accurate, contextually relevant, and effective for training and evaluation purposes in Catalan. We achieve this by adapting elements like dates, metric systems and names, maintaining internal logic, and preserving a rich and varied language. We also avoid replicating errors from the source text. In evaluation datasets, translators took care to avoid providing hints for the correct answer and match the response length to the original text, among other guidelines. The translated datasets are:

- XNLI-ca[14] is the translation of the XNLI (Cross-lingual Natural Language Inference) dataset (Conneau et al., 2018), which evaluates cross-lingual language inference in 15 languages.
- COPA-ca[15] is the translation of COPA (Choice of Plausible Alternatives) (Gordon et al., 2012), developed to assess causal commonsense reasoning in English and other 10 linguistically diverse languages via the XCOPA dataset (Ponti et al., 2020).
- PAWS-ca[16] is the translation of PAWS-X (Cross-lingual Paraphrase Adversaries from Word Scrambling) (Yang et al., 2019). This dataset aids paraphrase identification in English, French, Spanish, German, Chinese, Japanese, and Korean.
- The Translated Wikipedia Biographies[17] dataset helps analyzing translation errors in English, German, and Spanish.
- XQuAD-ca[18] is the translation of XQuAD (Cross-lingual Question Answering Dataset), a benchmark for evaluating cross-lingual question answering performance. The dataset consists of a subset of 240 paragraphs and 1,190 question-answer pairs from the development set of SQuAD v1.1 (Rajpurkar et al., 2016) together with their professional translations into eleven languages.
- WNLI-ca[19] is the translation of the Winograd NLI dataset.[20] This dataset presents 855 sentence pairs, in which the first sentence contains an ambiguity and the second one a possible interpretation of it. The label indicates if the interpretation is correct (1) or not (0).

Once each Catalan subset was ready, we prepared it to match the source format and contacted the original authors so that they could add Catalan to the original dataset. This would ensure that Catalan is considered in future work on multilingual tasks. In spite of the overall positive response, we have faced multiple issues in this regard, as some of the contact authors have changed email addresses, some have not found the time or the right person to do so yet, and, in the case of PAWS-X, no answer has been received.

Another problem happened with the Translated Wikipedia Biographies. Even though the Catalan

[12] https://huggingface.co/projecte-aina

[13] https://zenodo.org/communities/catalan-ai

[14] https://huggingface.co/datasets/projecte-aina/xnli-ca

[15] https://huggingface.co/datasets/projecte-aina/COPA-ca

[16] https://huggingface.co/datasets/projecte-aina/PAWS-ca

[17] https://zenodo.org/records/7971398

[18] https://huggingface.co/datasets/projecte-aina/xquad-ca

[19] https://huggingface.co/datasets/projecte-aina/wnli-ca

[20] https://cs.nyu.edu/~davise/papers/WinogradSchemas/WS.html

subset was initially added to the dataset (available in Google Research), we recently discovered that the whole dataset had been removed without notice.

We hope that our translated datasets will be incorporated soon to, at least, the five original datasets we have a confirmation for. Also, we hope that the Translated Wikipedia Biographies come back online at some point in the future. Either way, this highlights a systemic issue in the way current datasets are managed, which we discuss further and provide a possible solution in Section 6.

4.2. Developing Datasets for Relevant Tasks

As mentioned above, we prioritize the creation of datasets that are widely used in the literature and that represent tasks that align with the directions of current research and are demanded by industry trends. We chose tasks that are still unsolved and present a challenge in order to produce relevant datasets. We also prioritize datasets that would be too difficult to translate accurately (e.g., due to localization issues, idiosyncrasies of the Catalan language, etc.), as doing so would return a low-quality dataset far from natural language.

For all datasets mentioned in this section, further information on data sources, distributions across categories, annotation guidelines, annotator details, and usage considerations can be found by following the link to the Hugging Face repository for each dataset (as footnotes). We do not include those details in this paper due to space limitations. However, it is important to note that not only were annotators paid fairly under Spanish national agreements, but they were also professional annotators (see Hugging Face repositories).

4.2.1. Conversational Question Answering

Based on the CoQA dataset (Reddy et al., 2019), CoQCat[21] is a dataset for Conversational Question Answering (CQA) in Catalan. CoQCat comprises 89,364 QA pairs, sourced from conversations related to 6,000 text passages from six different domains (i.e., biographies, literature, news, mythology, short stories and movie plots).

The questions and responses are designed to maintain a conversational tone. During the annotation process, we verified that the questions were varied, referred to different segments of the text (avoiding concentration at the beginning or end), and used rich vocabulary. We achieve this by encouraging the use of synonyms and periphrasis, and avoiding copying the same words used in the paragraph. The grammar of every question was also checked with the help of Language Tool.[22] Finally, a set of 10 randomly chosen contexts from every batch was reviewed by hand and, similarly to CoQA (Reddy et al., 2019), the linguistic phenomena observed in the questions was annotated manually. For each question, more than one linguistic phenomenon could be shown.

In total, 1,715 questions were annotated. Answers are presented in a free-form text format, with evidence highlighted from the passage. For the development and test sets, an additional two responses to each question are included.

4.2.2. Natural Language Understanding

NLUCat[23] is a Natural Language Understanding (NLU) dataset in Catalan. It consists of nearly 12,000 instructions annotated with the most relevant intents and spans. Each instruction is accompanied by the instructions received by the annotator who wrote it.

We took into account 88 different intents, which are the usual ones of a virtual home assistant (e.g., activity calendar, IOT, list management, leisure, etc.), and further specific ones considering social and healthcare needs for vulnerable people (e.g., information on administrative procedures, menu and medication reminders, etc.). Spans are annotated with a tag describing the type of information they contain. They are fine-grained, but can be easily grouped to use them in robust systems.

When writing the examples, annotators were asked to take into account the socio-cultural reality (i.e., geographic points, artistic and cultural references, etc.) of the Catalan-speaking population. They were also asked to be careful to avoid examples that reinforce existing stereotypes. For instance, we asked to be careful with the gender or origin of personal names that are associated with certain activities.

During the process of writing the examples, the grammaticality of the sentences was checked with the help of Language Tool. We also automatically checked that there were no repeated or too similar sentences. In addition, 10% of the sentences were manually reviewed to ensure that they corresponded to what was requested. Also during the annotation process, 10% of the examples from each submission were manually checked.

4.2.3. Summarization

caBREU[24] is a dataset for summarization in Catalan. It consists of 3,000 articles, each averaging

[21] https://huggingface.co/datasets/projecte-aina/CoQCat

[22] https://extensions.libreoffice.org/en/extensions/show/languagetool

[23] https://huggingface.co/datasets/projecte-aina/NLUCat

[24] https://huggingface.co/datasets/projecte-aina/caBreu

about 700 words in length, along with extreme, abstractive and extractive summaries, manually generated by three annotators. The source material for the articles was various Catalan news webpages, including the Catalan News Agency (*Agència Catalana de Notícies*, ACN),[25] VilaWeb[26] and NacióDigital.[27] We use these as they were the only news pages that gave us permission to use their publications.

The summaries adhere to grammatical correctness, standard language conventions, and were composed in accordance with explicit instructions. For extractive summaries, annotators were asked to select four sentences from the original text, encapsulating its most relevant information. In the case of extreme summaries, annotators wrote a concise 15 to 20-word sentence that encapsulated the text's primary theme, addressing the question '*What is this text about?*' Lastly, abstractive summaries required annotators to generate a 50 to 60-word abstract, offering a succinct overview of the text's key information in their own words. It was imperative that these summaries remained clear, objective, and devoid of personal opinions, ideas, or interpretations, while conforming to the text's tense, structure, and avoiding overly-lengthy sentences.

The evaluation of the summaries in the dataset included both automated and manual assessments. Automated assessments revealed a Jaccard index of 0.29 for the similarity between extreme summaries and article titles, and 0.24 for abstractive-extreme summary similarity. Manual reviews gave extractive summaries a Likert score of 4/5 for capturing main ideas. Abstractive summaries scored 4.7/5 on average for coherence, conciseness, interest, readability, and relevance.

4.2.4. Entity Identification and Linking

The Catalan Entity Identification and Linking (CEIL)[28] is a complex Named Entity Recognition (NER) dataset that contains 9 entity types and 52 sub-types annotated on all kinds of short texts, mainly user generated content. The main entity types include the usual person, location and organization, as well as cultural work, geopolitical entities, product, event, building, and other. Example sub-types include painting, building-hospital, event-protest, island, among many others. The granularity achieved makes it useful for applications ranging from media monitoring to document indexing or knowledge discovery. Almost 59,000 documents were annotated, for a total of 380,474 annotations.

In addition, CEIL has been annotated with Wikidata *qlinks*, making it also an entity-linking dataset that can provide precise reference for the 63,062 entities included. This corpus was annotated by three teams of three annotators, and cross-checked to achieve a precision above 90%.

4.2.5. Task-Orientated Conversations

XitXat (Catalan for 'small talk')[29] is a collection of simulated, task-orientated chatbot conversations created under a 'Wizard of Oz' paradigm. The exchanges cover 10 domains and are representative of customer and citizen call center threads dealing with services or product purchases, billing discussions, room/taxi reservations, and public transportation information, among others. Two annotators were instructed to use everyday, natural expressions, and provide fake personal information when prompted for billing or account information, as a secondary objective of this dataset was to test anonymization systems for user-generated content.

User interactions were labeled with relevant intents and entity slots. The 960 conversations cover a wide range of exchanges, including out-of-domain digressions that chatbots have difficulty in handling (Tan et al., 2019), as well as complaints, insults, interruptions and naturalistic turn-taking.

4.2.6. Textual Entailment

TE-ca[30] is a Textual Entailment (TE) dataset in Catalan, which contains 21,163 premise-hypothesis pairs, annotated according to the inference relation they have (i.e., implication, contradiction or neutral).

12,000 sentences from CaText (Armengol-Estapé et al., 2021) and 6,200 headers from the Catalan news site VilaWeb, were randomly chosen. We filtered them by different criteria, such as length and stand-alone intelligibility. For each selected text, we commissioned three hypotheses (one for each entailment category) to be written by a team of native annotators.

4.2.7. Question Answering

CatalanQA[31] is an extractive-QA dataset in Catalan. It is an aggregation and balancing of two previous datasets: VilaQuAD and ViquiQuAD. VilaQuAD[32] contains 2,095 fragments of news articles extracted from VilaWeb, along with 1 to

[25] https://www.acn.cat/

[26] https://www.vilaweb.cat/

[27] https://www.naciodigital.cat/

[28] https://huggingface.co/datasets/projecte-aina/ceil

[29] https://zenodo.org/records/7276036

[30] https://huggingface.co/datasets/projecte-aina/teca

[31] https://huggingface.co/datasets/projecte-aina/catalanqa

[32] https://huggingface.co/datasets/projecte-aina/vilaquad

5 questions referring to each fragment. ViquiQuAD[33] encompasses 3,111 contexts extracted from a set of 597 high-quality, original (no translations) articles in the Catalan Wikipedia,[34] and 1 to 5 questions with their corresponding answer for each fragment. Both datasets follow the SQuAD guidelines (Rajpurkar et al., 2016).

In CatalanQA, splits have been balanced by type of question, and unlike other datasets such as SQuAD, it only contains, per record, one question and one answer for each context, although the contexts can repeat multiple times.

4.2.8. Abusive Language Identification

InToxiCat[35] is a dataset for the detection and span identification of abusive language (see Caselli et al., 2020 for a definition) in Catalan. It consists of 29,809 sentences extracted from online forums in Racó Català[36] by means of keywords, which are available for each instance in the released dataset.

The annotation process was divided into two phases. In the first phase, each sentence was labeled by two annotators as either abusive or non-abusive. These annotators agreed on the classification of 26,497 sentences (88.9%), and a third annotator annotated the rest to resolve the discrepancy. In accordance with the estimations, 6,047 sentences were finally classified as abusive and passed on to a second annotation phase.

The second phase —completed by the same annotators as the first phase— was concerned with the identification of abuse and the target span (i.e., the segments including the abusive message), as well as the type of abuse and target span according to some predefined classes. The third annotator resolved sentences with disagreements on any of these annotations.

Abusiveness was classified as either explicit, when it contains an insult or threat, or implicit, when no offensive language is involved, but sarcasm or a similar mechanism is used. Regarding target spans, in 14.45% of cases the abuse was not explicitly mentioned in-text, and therefore no target span was detected. The target(s), regardless of the presence or absence of a span, were classified as either individual (whose name may be unknown), group (belonging to a particular ideological, social or cultural community), or other (such as a company, situation, etc.), or a combination of these if the target was complex or there were multiple targets. Overall, 37.32% of the abusive messages were labelled as 'individual', 31.49% as 'group', 15.28% as 'other' and 15.91% as having more than one type of target.

4.2.9. Sentiment Analysis

Catalan Structured Sentiment Analysis (CaSSA)[37] is a dataset for the task of Structured Sentiment Analysis (SSA) in Catalan. It consists of 6,400 restaurant reviews gathered from GuiaCat[38] and Racó Català forum messages.[39]

Messages in CaSSA were annotated with all the spans corresponding to the polarity expressions that convey a subjective opinion about an object or service. Each polar expression was, in turn, annotated with the polarity type (positive, neutral or negative) and intensity (strong or standard), as well as the target span(s) (i.e., the object to which the expression is directed) and the source span(s) (i.e., the subject expressing the sentiment). Two annotators completed the entire annotation independently, and a third annotator intervened in case of disagreement between the two.

In total, 25,453 polar expressions were identified in the dataset, with an average of 3.98 per message. 72.34% are positive polar expressions, 12.92% neutral, and 14.70% negative.

4.2.10. Stance and Emotion Detection

CaSET[40] and CaSERa[41] are datasets annotated for the Stance and Emotion Detection tasks. The former was sourced from Twitter (c.k.a. X) posts and encompasses two stance detection tasks, static and dynamic, as well as the emotion detection task. The latter comes from Racó Català messages and includes the dynamic stance and emotion detection tasks. In both datasets, each instance consists of two messages, the 'parent' and the 'reply' (the response to the parent message).[42]

CaSET consists of 6,773 pairs of sentences, while CaSERa has 13,999 pairs.

Static stance detection is the task of deciding whether a message is for or against a given topic. Specifically, for CaSET, messages were annotated as favour, against, neutral or NA (usually,

[33] https://huggingface.co/datasets/projecte-aina/viquiquad

[34] https://ca.wikipedia.org

[35] https://huggingface.co/datasets/projecte-aina/InToxiCat

[36] https://www.racocatala.cat/forums

[37] https://huggingface.co/datasets/projecte-aina/CaSSA-catalan-structured-sentiment-analysis

[38] https://guiacat.cat/

[39] These were selected by a Catalan RoBERTa-base model trained on a binarily classified training corpus to identify messages written in the style of reviews.

[40] https://huggingface.co/datasets/projecte-aina/CaSET-catalan-stance-emotions-twitter

[41] https://huggingface.co/datasets/projecte-aina/CaSERa-catalan-stance-emotions-raco

[42] In the case of CaSET, however, the Twitter messages are left blank to be filled in by the Twitter API using the tweet ID provided.

texts that are off-topic, unintelligible or in a foreign language). The topics included were chosen for their controversial nature at the time of the design of the dataset, and the messages were retrieved using keywords related to them. Topics included are vaccines (mostly related to COVID-19 vaccines), rent regulation, Barcelona airport expansion, surrogate pregnancy, and a potential manipulation of a TV show outcome. For each of these topics, the annotation guidelines specify what it means to be in favor, against, neutral and NA. Two annotators completed the entire annotation independently, and a third annotator intervened in case of disagreement.

Dynamic stance detection is a new annotation scheme proposed to model interactions between messages. Its main advantage, demonstrated in a series of experiments studied in a concurrent work (Figueras et al., 2023), is its portability across topics, which extends its applicability to the analysis of unseen topics. The task consists of determining the positioning of a 'parent' message in relation to the 'reply' message. The possible categories are seven: agree, disagree, elaborate (agrees while adding opinions), query (expresses doubts, questions or asks for more information), unrelated, neutral, and NA. Four annotators worked independently on all pairs of messages, and a fifth annotator intervened when the assigned label was not agreed upon by at least three annotators.

The emotion detection task consists in identifying the main emotions expressed in the text message in a multi-label fashion. There were eight possible labels: anger, anticipation, disgust, fear, joy, sadness, positive surprise, and negative surprise. Each category, properly defined in the annotation guidelines, was annotated by three annotators, and the gold label was then created by aggregating all identified labels.

5. Deployment and Dataset Exploitation

The massive effort and expense taken when creating these datasets might be justified just by their value for corpus linguistics and model evaluation purposes, but their exploitation for training LLMs to perform specific tasks makes them valuable beyond research, opening opportunities for everyone to benefit from our work. All of our datasets are created from scratch with open-source, commercial-use friendly licences. This helps with openness and availability. We now describe how some of the Catalan datasets presented are used for training and instructing specialized models.

5.1. Model Training

The datasets included in this paper were used to fine-tune token and document classification tasks with a base pre-trained RoBERTa, DeBERTa, and Falcon transformer models. These are all available through our Hugging Face.

These powerful, task-oriented models can be used for Named-Entity Recognition and Classification (NERC), document classification, sentiment analysis, and intent detection, among other tasks, either directly through the transformer framework or using spaCy wrappers.[43]

We also used part of the CATalog data (Palomar-Giner et al., 2024) to continually pre-train a BLOOM-7.1B[44] model, resulting in FLOR-6.3B.[45]

At present, we have 56 publicly available models for download, with some Hugging Face Spaces available for FLOR-6.3B,[46] DeBERTa Multi-NER,[47] and a Catalan Text-to-Speech system,[48] among others.

Some datasets are also included in the Catalan Language Understanding Benchmark (CLUB), described in Armengol-Estapé et al. (2021), to help evaluate and compare the performance of different architectures and models available for Catalan. In future work, we will introduce a comprehensive benchmark for downstream-task evaluation in Catalan using Eleuther AI's Evaluation Harness (Gao et al., 2023).

5.2. Generating Instructions

Adapting instructional datasets, used to train LLMs to respond to instructions, for languages not included in well-known collections like Alpaca[49] or Dolly[50] frequently involves translating them, either automatically or by humans. This approach is not always ideal, as transferring short and idiosyncratic contexts is more akin to a localization process than a direct translation, making it very hard to achieve the desired training objectives.

Creating instructions from task-specific datasets requires, as is often the case with LLMs, careful

[43] https://github.com/explosion/spacy-huggingface-pipelines

[44] https://huggingface.co/bigscience/bloom-7b1

[45] https://huggingface.co/projecte-aina/FLOR-6.3B

[46] https://huggingface.co/spaces/projecte-aina/flor-6.3b

[47] https://huggingface.co/spaces/projecte-aina/multiner_demo

[48] https://huggingface.co/spaces/projecte-aina/tts-ca-coqui-vits-multispeaker

[49] https://github.com/tatsu-lab/stanford_alpaca

[50] https://huggingface.co/databricks/dolly-v2-12b

prompt engineering to ensure a successful fine-tuning. For example, converting topic classification or NERC datasets into instructions requires consideration as to whether to include the actual (annotated) labels in the prompt, or just write a generic prompt like: '*What is the subject associated with the following text?*'. Using batch conversion, we created collections of instructions for extractive QA, paraphrasing, sentiment detection, document and entity classification, and summarization, among others, all useful tasks expected to be performed by LLMs in industrial settings. We did this by manually writing prompt templates[51] that were then used to generate instructions from the existing annotated datasets. Test subsets were also converted but never used for model instruction.

The resulting Catalan Instructional set (InstruCat)[52] has approximately 219,000 instructions. We used these to train an instructed version of FLOR-6.3B.[53]

6. Conclusion and Lessons Learned

In this paper, we presented the results of two years of work[54] aimed at providing Catalan, a "fragmentary support" (Rehm and Way, 2023) language, with the technology support needed to improve its relevance in AI/NLP-related industry and research. We offered a first glimpse into our data gathering and processing methods, our dataset translation and creation procedures, our applications into model training and instruction generation, and our compromise with openness and resource availability.

Based on our experience, we advocate for the necessity to offer open-access to datasets in mid- and low-resource languages to maximize applicability and collaboration. We also argue that datasets need to be processed in a way that are ready to use, by means of deploying them as data loaders following standard methods and offering them as instructions for training and as evaluation benchmarks.

During these years we have also learned some valuable lessons that anyone trying to replicate our work for another language should be aware of. First, it may not come as a surprise, but finding texts of enough quality to annotate is particularly challenging in mid-resource languages like Catalan. Licensing rights and copyright laws are sometimes unavoidable roadblocks one should be ready to encounter and manage. We also found surprisingly difficult to add translated subsets to the original datasets. Even though the initial response is, in most cases, positive, materializing this is a slow and laborious process. In one case, the whole dataset was deleted after we had added Catalan to it. This would have implied a waste of resources if it we did not have the original English sentence in our subset. Still, we argue that authors who claim to welcome other languages into their datasets should ensure that the dataset is upkept suitably and that authors of other subsets added *a posteriori* are kept informed of any changes to the original dataset. To minimize the impact of these changes, we strongly suggest self-containing the work done via Hugging Face, GitHub or Zenodo repositories. This can hamper the adoption of the subsets in the language being promoted, but ensures that the work is always publicly available and that any changes that may impact the work always go through the right people.

Regardless of the challenges mentioned, we believe that the contributions presented in this paper are a great starting point in increasing technology support for Catalan. We also hope that our work can help in the development of similar infrastructures for other languages in a similar situation to Catalan.

7. 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.

8. Ethical Considerations

As part of our ongoing efforts on ethics, we present some of the most relevant ethical considerations that impacted this work. Firstly, the main purpose of this paper is enhancing the technology support of a mid-resource language, helping to bring the benefits of AI systems to the speakers of the language. For the translation of datasets and annotation of data, we worked with local companies and professionals, which ensures that translators and annotators were adequately paid under national agreements. The data we use to create new datasets was always freely available online and included no identifiable personal data. Finally, we also openly publish online our datasets, pipelines, models, data sources and guidelines to ensure auditability and traceability.

[51] https://github.com/langtech-bsc/InstruCAT-generation/blob/master/instructions_ca.yaml

[52] https://huggingface.co/datasets/projecte-aina/InstruCAT

[53] https://huggingface.co/projecte-aina/FLOR-6.3B-Instructed

[54] Funded with €3M per year.

9. Bibliographical References

Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. 2021. Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, Online. Association for Computational Linguistics.

James Betker. 2023. Better speech synthesis through scaling. *arXiv preprint arXiv:2305.07243*.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Blanca Figueras, Irene Baucells, and Tommaso Caselli. 2023. Dynamic stance: Modeling discussions by labeling the interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6503–6515, Singapore. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Baybars Kulebi, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2022. ParlamentParla: A speech corpus of Catalan parliamentary sessions. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 125–130, Marseille, France. European Language Resources Association.

Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. 2023. Matcha-tts: A fast tts architecture with conditional flow matching. *arXiv preprint arXiv:2309.03199*.

Jorge Palomar-Giner, Javier Saiz, Ferran Espuña, Mario Mina, Severino Da Dalt, Joan Llop, Malte Ostendorff, Pedro Ortiz Suarez, Georg Rehm, Aitor Gonzalez-Aguirre, and Marta Villegas. 2024. A curated catalog: Rethinking the extraction of pretraining corpora for mid-resourced languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. European Language Resource Association and the International Comittee on Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Georg Rehm and Andy Way. 2023. *European Language Equality: A Strategic Agenda for Digital Language Equality*. Springer Cham.

Ming Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni Potdar, Shiyu Chang, and Mo Yu. 2019. Out-of-domain detection for low-resource text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3566–3572, Hong Kong, China. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.