A CURATEd CATalog: Rethinking the Extraction of Pretraining Corpora for Mid-Resourced Languages

Jorge Palomar-Giner*[1], José Javier Saiz*[1], Ferran Espuña*[1], Mario Mina*[1], Severino Da Dalt[1], Joan Llop[1], Malte Ostendorff[2], Pedro Ortiz Suarez[2], Georg Rehm[2], Aitor Gonzalez-Agirre[1], Marta Villegas[1]

[1]Barcelona Supercomputing Center, Spain

[2]DFKI GmbH, Germany

Abstract

We present and describe two language resources in this paper: CATalog 1.0, the largest text corpus in Catalan to date, and CURATE (Corpus Utility for RAting TExt), a modular, parallelizable pipeline used for processing and scoring documents based on text quality that we have optimised to run in High Performance Computing (HPC) environments. In the coming sections we describe our data preprocessing pipeline at length; traditional pipelines usually implement a set of binary filters such that a given document is either *in* or *out*. In our experience with Catalan, in lower-resource settings it is more practical to instead assign a document a soft score to allow for more flexible decision-making. We describe how the document score is calculated and highlight its interpretability by showing that it is significantly correlated with human perception as obtained from a comparative judgement experiment. We additionally describe the different subcorpora that make up CATalog 1.0.

Keywords: Catalan, curation, dataset, mid-resourced, preprocessing

1. Introduction

Nowadays, with the ever-increasing demand for data for NLP applications, the need for feasible solutions to obtain high quality training data has never been higher. The largest source of which are typically crawls of the internet, rather than curated and cleaned sources, housing terabytes of raw textual data. The content of these crawls typically varies greatly in terms of topic, but also in terms of general quality, which can range from pristine to completely undesirable.

Most work includes heavy filtration as a preprocessing step (Ortiz Suárez et al., 2019; Xue et al., 2021; Rae et al., 2021; Laurençon et al., 2023; Kudugunta et al., 2023) to separate desirable data from undesirable data; relevant or clean documents are separated from irrelevant or unclean documents using a specific metric or combination thereof such as reaching a minimum number of sentences or not containing a certain number of specific strings or substrings (e.g. `Facebook`, `cookies`, `Lorem Impsum`, etc). Typically, these features aim to determine if a document is well-formed or contains enough textual data to be relevant, for example, for the purposes of training a language model. The result of this filtration yields documents that are either *in* or *out*. However, text quality can be viewed as a continuous spectrum taking into account many different aspects. For instance, a text may contain several sentences but be oddly punctuated, or it may be well-formed while having very low lexical richness.

These considerations regarding text quality are especially relevant in low-to-mid resource scenarios (Artetxe et al., 2022) as it can allow for a more flexible way to build a dataset and more easily examine the interaction between data quantity and quality. For instance, in a low-resource setting, it might be beneficial to be more forgiving in the case of a text that is oddly punctuated, but that is otherwise of sufficiently high quality.

We highlight the importance of efficient processing methods when dealing with large corpora; given the current data requirements of language models, even in mid-resource scenarios, we typically deal with large amounts of data in absolute terms. For Catalan, Galician and Basque, internet crawls can contain up to millions of documents (Xue et al., 2021; Abadji et al., 2022; Kudugunta et al., 2023), and therefore preprocessing steps such as deduplication or filtration can be time-consuming or even intractable if the applied methods are inefficient.

This paper presents CATalog 1.0 and CURATE: A large, multi-domain corpus in Catalan, and the pipeline used to obtain it from its raw form. We develop the pipeline with specific principles of **adaptability**, **continuity** and **modularity** in mind. In the upcoming sections we describe the original data in terms of domain and size. We additionally describe in detail the preprocessing steps we perform to clean and score documents, as well as the effect this has on the final output. We highlight that our pipeline is designed to handle large volumes of data efficiently in HPC environments, and that our decision to assign the documents continuous scores and to determine a threshold *a poste-*

*These authors contributed equally to this work

LREC-COLING 2024, pages 335–349

riori allows for a more efficient use, such that the pipeline does not need to be completely rerun if a more lenient filtration needs to be applied in order to maximise training data.

In addition, to demonstrate the accuracy of the applied soft-labels, we perform human evaluation on the scored document and examine the correlation between the document scores and human evaluator judgement.

Human evaluation is done through a comparative judgement experiment involving native and near-native Catalan speakers. Although the data used in the experiment is significantly smaller compared to the corpus size, the results are still informative. They *do* suggest that the pipeline's evaluation framework effectively captures high-level aspects of text quality.

2. Context and Motivation

Recent years have seen many advancements in language modelling, with larger language models seemingly around every upcoming corner (Lieber et al., 2021; Smith et al., 2022; Thoppilan et al., 2022). This tendency of creating language models of ever-increasing sizes comes with greater dependency on computational resources and data, as highlighted in Hoffmann et al. (2022). The value of adequate data in this context cannot be overstated.

Furthermore, there has been a major shift towards scrutinising several aspects of the content of the data used for training language models. For instance, there is heavy debate revolving the usefulness of deduplication with some works highlighting its advantages (Lee et al., 2022; Rae et al., 2021; Hernandez et al., 2022), while others demonstrate its drawbacks (Wenzek et al., 2019). Other efforts are centred around the actual preprocessing of data. In general, the different approaches proposed are extremely useful for obtaining large volumes of data in the most widespread languages, but present several inequalities when dealing with mid or under-resourced languages (Ranathunga and de Silva, 2022). Problems such as not correctly identifying the language (Caswell et al., 2020), not obtaining sufficient amounts of data (Rehm and Way, 2023) or not processing the information properly due to the generality of the heuristics used in the cleaning or because they are designed on higher-resourced languages (Kreutzer et al., 2022). In this section, we highlight the different works and contributions that also attempt to tackle this task of creating corpora from different sources and how their approaches tie in with ours, specifically with the principles of **modularity**, **continuity**, and **adaptability**.

Modularity. We have developed CURATE as a set of independent modules designed to be executed sequentially, either manually or by some *higher-level* program, while storing the output of intermediate steps in common formats and in agreed upon locations. This allowed our team to work in parallel, while making it easy to track bugs or inefficiencies when something went wrong. Additionally this makes the pipeline more flexible, allowing us to *hot-swap* or even remove steps depending on the desired application.

Although modularity is a widespread aspect in software development since it is a good practice, there are different opinions about the modules that are useful in document-oriented data processing pipelines. Some of the most common are deduplication (Lee et al., 2022), reformatting (Jennings et al., 2023), language identification (Kudugunta et al., 2023), quality evaluation (Abadji et al., 2021), quality filtering (Xue et al., 2021), downstream task data removal (Smith et al., 2022) and sampling (Brown et al., 2020). We include many of these in our pipeline as shown in Section 3.

Continuity. As mentioned in the introduction, the general approach in most textual data cleansing is based on performing a binary filtering through a series of heuristics (Xue et al., 2021; Rae et al., 2021; Laurençon et al., 2023; Kudugunta et al., 2023; Nguyen et al., 2023). In contrast, CURATE offers a continuous output, which aims to cover the full spectrum of quality that a collection of texts can contain. In this sense, there are very few alternatives that offer this approach to the best of our knowledge. An example where this binary filtering is not done is described in Abadji et al. (2021), where it is explained how some annotations are assigned to documents based on some heuristics. In a sense, CURATE aims to combine some of these (or similar) heuristics, among many others, and compute them to assign a continuous score between 0 and 1 that is intuitive for sampling.

Adaptability. We refer to CURATE's ability to integrate multiple sources with very different formats in a simple way, since it first normalizes all these sources and converts them into a document-oriented representation that is interpretable by the rest of the pipeline modules, similarly to the approach used in Gao et al. (2020), where the final dataset is composed of 22 sub-datasets. However, there are not many precedents that unify so many data sources, and that offer the possibility of combining web sources with higher quality data. Most of the available tools are mainly focused on processing Common Crawl dumps (Abadji et al., 2021; Wenzek et al., 2019).

Catalan is considered by many sources to be a mid-resourced (Ortiz Suárez et al., 2020) or even under-resourced language (Bañón et al., 2022). In

fact, in Rehm and Way (2023), which analyzes the situation of around 30 European languages, Catalan is placed in the group of languages with *fragmentary support*, far away from English, with *good support*, and below the exclusive group of languages with *moderate support*. For this reason, Catalan also presents some of the problems described in this section. So far the largest Catalan datasets that have been released are the Catalan part of the mC4 (Xue et al., 2021), with approximately 7B words by our count (13B tokens according to the paper), and the recently released Colossal OSCAR 1.0, with about 15B words. However, considering that the English counterparts of the same corpora are orders of magnitude larger, it is clear that the lack of data can be a limiting factor for the development of more robust language models, especially when considering the scaling laws presented in Hoffmann et al. (2022). Moreover, these datasets contain only web-sourced data as they come from Common Crawl dumps, which are very useful for obtaining large volumes of data but lack variety and quality.

With this mind, we follow three mentioned design principles to create CURATE, and use it to build CATalog 1.0, the largest dataset in Catalan to date, which contains parts of Colossal OSCAR 1.0 and mC4 in Catalan, plus many others, for a total of 17.45B words. It should be noted that CATalog 1.0 is only one example of the application of CURATE, and that this can be replicated with other languages with similar characteristics. The details of the dataset are described in Section 4. We publicly release the data[*] under a set of permissive licenses; and the code[*], under the Apache 2.0 license[*].

3. Methods

In this section we describe the different methods CURATE employs for data processing, and how they align with the three principles described in the previous section. It has been approached from a modular perspective, so each subsection focuses on one module.

3.1. Data Management and Reformatting

As mentioned in Section 2, the adaptability of the pipeline is one of the principles that govern its way of being. To this end, it is prepared to integrate different data sources, so that each input format has a reading method that transforms the original format into a particular document representation (see Section 3.3). For new input formats not covered, adding them is as simple as adding a reading method; the rest of the work will be done by CURATE. In addition, to facilitate the management of various sources, datasets are kept along with a set of metadata that allow data governance and facilitate task automation. Finally, the metadata module itself splits the original dataset into parts of 1GB, or alternatively the most atomic size of its source files, in order to facilitate parallel processing of each of these parts.

3.2. Deduplication

The pipeline applies a simple three-step process for (exact) deduplication of the dataset. In the first step, hash deduplication is applied within each part (always $<=$ 1GB) of the dataset, and the hash values are saved for the second step. During the second step, these hash values are used to deduplicate documents between parts. Finally, in the third step, the documents are reorganized again, so that they belong again to the original subpart of the dataset. Results are saved in the specified output format. This method of processing documents allows deduplication to scale easily and quickly since the first and third step allow a parallel computation between different parts of the dataset, while the second step compares independently hash-values ranges over the entire dataset.

3.3. Document Representation

Each input format takes in the raw text of a document D and splits it into *paragraphs* $P^1, \ldots P^{N_p}$. This is done differently depending on the source of the documents, as they use different representations for paragraph separation (for instance, a book might just use a line break, whereas an HTML file might have different tags indicating it). Then, each paragraph P^j is split into *sentences* $S^{b_j}, S^{b_j+1}, \ldots, S^{e_j-1}, S^{e_j}$. We obtain that the document contains the sentences $S^1, \ldots, S^{e_{N_p}}$. For some input formats we also remove leading or tailing paragraphs of a document that contain certain common repetitive phrases specific to web crawlings (such as `cookie policy` or `follow us`).

3.4. Language Identification

We run each sentence of each document through the FastText language identification pipeline (Kudugunta et al., 2023) to obtain the probabilities of the languages in which each sentence is written, and then calculate an average (weighted by number of words) of these probabilities to obtain an approximation of the fraction of each language in the document. The language for

[*] https://huggingface.co/datasets/projecte-aina/CATalog/tree/main

[*] https://github.com/langtech-bsc/CURATE/tree/main

[*] https://www.apache.org/licenses/LICENSE-2.0

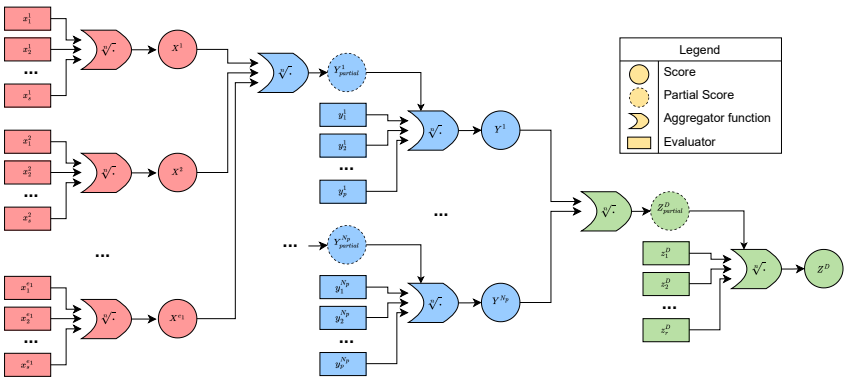

Figure 1: Document evaluation diagram.

which this estimate is largest will be considered the *main language* of the document (provided that the estimate exceeds a certain threshold, which in our experiments is fixed to 0.5). When we run the pipeline, we specify a list of *languages of interest*. The documents whose main language is in this list will be written to separate files (one for each language) for easy access in the sampling step. These language percentage estimates are also used in some of the document evaluators (see Section 3.7).

3.5. Preprocessing

We apply a series of filters to each sentence S^i of each document D. These are designed to make small adjustments such as ensuring the encoding is correct, removing parts of the sentence that are clearly not desirable such as surrounding whitespace, etc. They should not affect the actual content of the text.

3.6. Document Evaluation

Once we have a document D in our internal representation, and it has been preprocessed, we wish to assign it a *score* between 0 and 1, which we will later use to filter the whole dataset more or less aggressively, depending on human evaluation and the task at hand. For this, we define several individual evaluators at the sentence $(x_1, \ldots, x_s)$, paragraph $(y_1, \ldots, y_p)$ and document $(z_1, \ldots, z_d)$ levels that judge the quality of the text contained in the appropriate object and assigns it a score between 0 and 1. These evaluators can be adjusted individually, and then are combined in the process described below (for a more visual explanation of the hierarchical scoring process, see Figure 1). Note that the process is applied independently to each document, so we have omitted the indexes indicating the document number to avoid cluttering the notation.

- For each sentence, S^i we get sentence scores $x_1^i = x_1(S^i), \ldots, x_s^i = x_s(S^i)$. We aggregate these scores into a single score X^i, which is their geometric mean.
- For each paragraph P_j we get a set of scores $y_1^j = y_1(P^j), \ldots, y_p^j = y_p(P^j)$, as well as the total scores of its sentences $X^{b_j}, X^{b_j+1}, \ldots, X^{e_j-1}, X^{e_j}$. We aggregate the scores X^i of the sentences of the paragraph into their geometric mean $Y^j_{partial}$, and again we aggregate the paragraph scores y_k^j together with $Y^j_{partial}$, obtaining Y^j.
- Finally, the document D has its own scores $z_1^D = z_1(D), \ldots, z_d^D = z_d(D)$, as well as the total scores of its paragraphs $Y^1, \ldots, Y^{N_p}$. We combine them analogously to the process we used to obtain the paragraph scores Y^j, obtaining partial ($Z^D_{partial}$) and final (Z^D) scores for the document D.

One might think that using a linear model for the scores (just letting them take any real value and adding them up, scaled by constants) is equivalent to this process. The reason for sticking to scores between 0 and 1 and using multiplication (in our case, geometric means) is to encourage the creators of the evaluators to think of them like a *probability* that the sentence, paragraph or document has of being *desirable*.

3.7. Evaluators

The evaluators are functions that, given some sentence, paragraph or document properties (such as number of words), return a score between 0 and 1. In order to adjust the shape of the function, a list of interpolation points are given to each function, and to obtain the score for that function, we linearly interpolate between the two nearest prescribed points. Here we describe the evaluators that we have used to score each document:

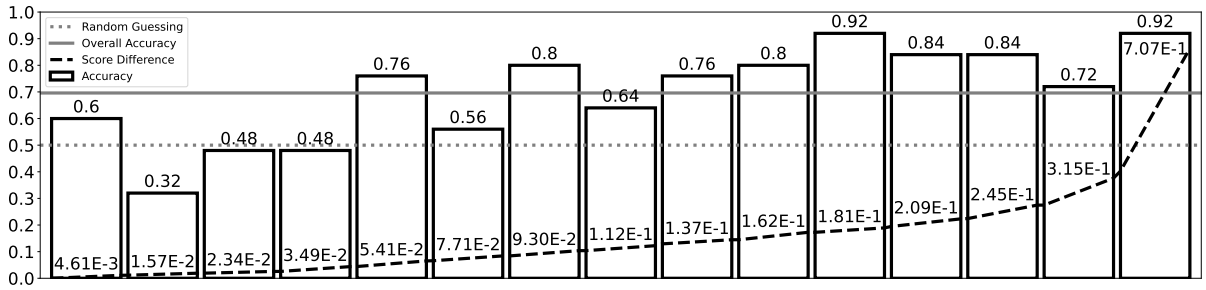

Figure 2: Given two documents with different CURATE scores, a native speaker will prefer the document with highest score 70% of the time, represented by the solid gray line. Each bar along the x axis represents 25 pairs of documents within a certain *score difference* range (dashed line, with the number above it indicating the average score difference in that range), and the y axis represents the number of times the document with the highest score is preferred by a human.

Minimum Words per Document A simple evaluator that penalizes a document based on a minimum number of words. We empirically determine that any document containing under 300 words should be penalized. Penalties are assigned via the interpolation-based method described in this section such that the score decreases linearly with the number of missing words with respect to the threshold.

Average Word per Sentence This evaluator computes the average number of words per sentence according to the following formula:

$$\overline{W} = \frac{\sum_{j=1}^{N_p} \sum_{i=b_j}^{e_j} |S^i|}{\sum_{j=1}^{N_p} |P^j|} \quad (1)$$

That is, the total number of words it contains divided by the total number of sentences. We empirically determine two different cut-offs, penalizing values that are too high or too low. This reflects the intuition that a text with sentences that are very short are undesirable, whereas one with very long sentences may be poorly punctuated.

Punctuation per Word Rate With this evaluator, we aim to examine the relationship between punctuation signs and the number of words in a document. We compute the punctuation per word rate by dividing the number of punctuation characters by the number of total words. For simplicity, let D' be the document D viewed as a sequence of M characters $c_1, \cdots, c_m$. We then simply divide the number of punctuation characters (i.e. belonging to a set of specific punctuation characters $Punc$) by the number of words as shown in:

$$PWRate = \frac{|\{1 \leq i \leq m \mid c_i \in Punct\}|}{m} \quad (2)$$

Similarly to the average words per sentence evaluator, we also establish an upper bound a lower bound from which we start penalizing; little or no punctuation in a document may indicate reduced document richness, while an excess thereof might be a sign that the document is malformed.

Unique Sentences Ratio This evaluator examines repeated sentences in the same document. We do this by dividing the number of unique sentences by the number of total sentences. We consider the presence of repeated sentences to be a generally negative trait, and therefore begin to penalize the document if any repeated sentences are detected.

Stopword Ratio We consider stopwords to be strong indicators of text quality such that their absence may suggest a document is a word list, not suitable for NLP development, or mostly numerical data. We compute the stopword-to-content-word ratio by dividing the number of stopwords by the number of total words. Penalization in this case is single-sided (i.e. we apply a penalty to the document if there are too few stopwords)

Brunet Index The Brunet W index is a measure of textual richness originally described in Brunet et al. (1978) that is computed using the following formula:

$$W = N^{V^{-0.165}} \quad (3)$$

where N is the total number of words and V is the number of distinct words. Many applications of this index are applied to stemmed tokens (Khodabakhsh et al., 2015; Slegers et al., 2018). Because the pipeline is meant to deal with large volumes of data, performing extensive token preprocessing such as stemming or lemmatization would

render it much less efficient. Penalisation is single-sided.

Bad Language ID We follow Kudugunta et al. (2023) and use the results from the language identification step (see Section 3.4) to penalize documents that contain other languages besides the expected languages if they exceed a specific percentage.

Cursed Regex We again follow Kudugunta et al. (2023) and apply the cursed regex evaluator they present that penalizes documents for containing certain keywords. We penalize and use the same regular expressions they use in their paper.

Too Long Word We consider documents that contain words that are too long to be undesirable. This evaluator penalizes a document if it has one or several words that exceed a maximum length, as long as these words only consist of alphabetic characters. The penalty increases for each long word in the document. The penalty is one-sided.

Too Frequent Character The too frequent character evaluator checks the characters of the most frequent word in the corpus to see if that word is over-represented in a given document in documents that pass a certain length. The penalty is one-sided.

Weird Streak This evaluator examines well-formedness in documents by applying a penalty if several non-alphanumeric characters appear consecutively. This penalty is one-sided.

3.8. Sampling

At the end of the pipeline, if one wants to obtain a dataset with certain properties, this can be framed in terms of selecting or rejecting certain documents based on properties like their main language (see Section 3.4), final evaluation score (see Section 3.6). FFor the evaluation score, we suggest setting a minimum threshold score to be selected, based either on human evaluation or the desired number of documents/tokens for the task. Alternatively, one can do a *stratified* sample where selection probabilities is manually selected for different score ranges.

3.9. Experimental Setup

To evaluate the correlation between the scores assigned by the pipeline and the documents that are actually desirable, we set up a comparative judgement experiment with 15 native Catalan speakers. We selected 375 random pairs of documents from our dataset that had been passed through the pipeline with the same parameters. We showed 25 of these pairs to each person and, for each pair, asked them to choose the document that they considered *better*, according to the following criteria:

- **Written in Catalan**: The document primarily uses Catalan, with minimal other languages.
- **Error-free**: No spelling, grammar, or structural mistakes.
- **Adheres to Catalan rules**: It follows standard Catalan writing conventions.
- **Fluency**: The language is understandable and natural.
- **Naturalness**: The text seems human-written, not artificial.
- **Well-structured**: The document presents information effectively.

We did not give any further instructions to the reviewers in order not to bias them towards any particular criteria, such as the heuristics used in CURATE.

4. CATalog 1.0

As a result of the application of CURATE, described in Section 3, and together with a great effort in the compilation of data sources, we present CATalog 1.0, the largest Catalan dataset that has been published to date. This dataset is constituted by combining several sources, whose retrieval methods can be classified into two groups:

1. Web-sourced datasets with some preprocessing available under permissive licence or domain-specific raw crawls.
2. Manually curated data obtained through collaborators, data providers (by means of legal assignment agreements) or open source projects.

The CURATE evaluators have only been applied to the data in the first group, while those in the second group have been added to the dataset with the highest score (one), as their quality has been carefully reviewed and manually curated.

The first group includes the Catalan sections of the following datasets: mC4 (Xue et al., 2021), OSCAR 22.01 and 23.01 (Abadji et al., 2022), caWaC (Ljubešić and Toral, 2014), MaCoCu (Bañón et al., 2022) and some minor Wikimedia initiatives (Wikibooks[*], Wikisource[*], Wikinews[*], Wikiquote[*]), as

[*] https://www.wikibooks.org/
[*] https://en.wikisource.org/
[*] https://www.wikinews.org/
[*] https://www.wikiquote.org/

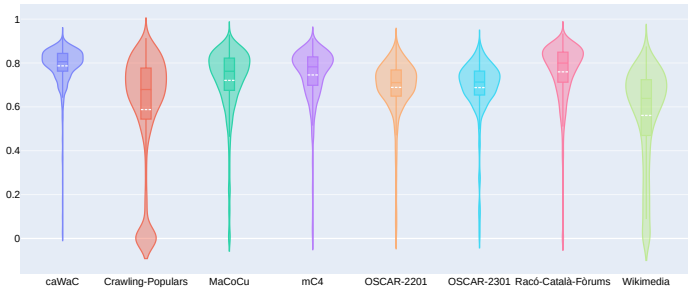

Figure 3: Final document score distribution for web-sourced corpora.

well as a crawling of the most popular websites of the .cat domain (Armengol-Estapé et al., 2021) and the Racó Català Fòrums[*]. These sources add up to a total of 15,251.18 million words, which represents 87.4% of the total resulting dataset.

The second group contains a wide variety of high quality but smaller sources:

- On the one hand, it contains data provided by media that carry out their main activity in Catalan. This data has been obtained in all cases by legal agreement with the data providers. These texts are mainly made up of journalistic articles of a very diverse nature, dealing with current affairs, especially in the region of Catalonia, but also national and international. These media are IB3[*], Grup El Món[*], Vilaweb[*], Nació Digital[*], ACN[*], Racó Català[*] (RC) and Aquí Berguedà[*]. Other collaborators have also contributed data, such as TDX[*], in its case providing scientific texts (doctoral theses in Catalan).

- On the other hand, it contains openly licensed texts extracted *ad hoc* from the web, but in a very cautious and manually revised way. This is the case of the Diari Oficial de la Generalitat de Catalunya[*] (DOGC), the transcripts of the plenary sessions of the Parliament of Catalonia[*] (PC), the Catalan Wikipedia[*], and the books in Catalan of the Gutenberg project[*]. The operationalization of the data acquisition of the first three sources is described in Gonzalez-Aguirre et al. (2024).

- In addition, some sources of the Valencian variant have been integrated in order to obtain a certain representation of the whole Catalan-speaking region. These data have been obtained through collaboration with the VIVES project on language technologies in the Valencian Community. These sources include the transcriptions of the sessions of the Valencian Parliament[*] (DSCV), the Diari Oficial de la Generalitat Valenciana[*] (DOGV) and the Butlletí Oficial de la Universitat d'Alacant[*] (BOUA).

This second group totals 1,601.27 million words, which represents 9.18% of CATalog 1.0.

Finally, a third group should be added to the classification. This is a particular filtering of the recently released Colossal OSCAR 1.0 (CO), carried out by the DFKI on the basis of the annotations resulting from the application of the Ungoliant pipeline (Abadji et al., 2021). Due to overlapping issues, only the sections corresponding to the cleaning of three CC dumps have been integrated: 2023-23, 2023-14 and 2022-27. Because of prior harsh filtering, this contribution, which makes up 3.43% of the dataset and comprises a total of 598.04 million words, has been combined into the dataset with the highest default score (one).

The sum of these three groups gives CATalog 1.0 a total of **17,450,496,729 words** (about 23B tokens) distributed in 34,816,765 documents. The dataset is composed of 26 subdatasets. Each subdataset can consist of one or multiple parts (each part corresponds to a file) depending on its original size, but can always be identified separately from the rest. Some datasets have specific usage restrictions, so CATalog 1.0 has been released with the particularity that some of its components are covered by a different licence, which is detailed in

[*] https://www.racocatala.cat/forums
[*] https://ib3.org/
[*] https://grupmon.cat/
[*] https://www.vilaweb.cat/
[*] https://www.naciodigital.cat/
[*] https://www.acn.cat/
[*] https://www.racocatala.cat/
[*] https://www.aquibergueda.cat/
[*] https://www.tesisenred.net/
[*] https://dogc.gencat.cat/ca/
[*] https://www.parlament.cat/
[*] https://ca.wikipedia.org/
[*] https://www.gutenberg.org/
[*] https://www.cortsvalencianes.es/
[*] https://dogv.gva.es/
[*] https://www.boua.ua.es/

the official repository. In this way, we do not prejudice the data with more permissive licenses. Each part contains thousands of documents, and each document includes six fields:

- **document**: document identifier. Each document is identified by the subdataset code, the part number and this ID.
- **text**: the plain text of the document, with paragraphs separated by two newlines escape sequences.
- **score**: Each of the documents is associated with a corresponding score, i.e. there is no filtering beyond the preprocessing. We leave the choice of sampling strategy to the data consumers. The distribution of documents by score and source can be found in Figure 3.
- **strategy**: strategy used to evaluate the document. *curate* if it is a subdataset of the first group and *perfect* if it is a subdataset of the second or third group.
- **languages**: dictionary of languages identified in the document.
- **url**: link to the document (if available).

In order to provide a general idea of the contents of each of the subdatasets that make up CATalog 1.0, we have classified these sources into the following categories according to their domain: Web, Journalistic, Scientific, Social, Legal, Political and Books. Additionally to their particular domain, the subdatasets of the Valencian Parliament, the Catalan Parliament and Racó Català Fòrums mainly consist of dialogues. Moreover, it includes the source curation method (Manual, CURATE, Ungoliant), as well as the absolute and relative contribution of each source to the final dataset.

5. Discussion

As mentioned in Section 3.9, we performed a human evaluation of the pipeline's output to determine if it could actually assess the quality of documents and thus produce a high quality dataset by using an appropriate sampling strategy (see Section 3.8). We have obtained 375 pairs of documents, each with the preferred one by a human evaluator. We also have access to the scores given by the pipeline to each of these documents. Our goal with this is to determine whether scores given by the pipeline can accurately predict text quality, in terms of well-formed Catalan text produced by a human.

With the scores provided by the pipeline, we are able to correctly predict which document is preferred 261 times, or 70% of the time. This gives a *Kendall's* coefficient τ of 0.39. Overall, this means that we can say with confidence that the pipeline scores are positively correlated with the human scores ($p < 10^{-10}$), and that we can be 95% confident that the score assigned by our pipeline will predict the human preference more than 65% of the time in a pairwise comparison. We note that the accuracy of these predictions is highly dependent on the actual score difference given by the pipeline. This is illustrated in Figure 2. We see that with a score difference over 0.1 the document with the highest score is preferred over 80% of the time. We acknowledge limitations on the scope of this testing. It is based on few human evaluators of similar backgrounds, and texts only in Catalan.

Further testing and maybe tuning of the preprocessing filters and evaluators will be needed in order to adapt the pipeline to other domains. However, for the purposes of publishing the CATalog 1.0 we are certain that the scores given reflect actual text quality according to the criteria specified in Section 3.9. Therefore, they can be used to curate high-quality pretraining datasets in Catalan as proposed in Section 3.8.

6. Conclusion and Future Work

We presented CATalog 1.0, an extensive Catalan corpus of 17.45B words resulting from the interdisciplinary development of CURATE, our text processing tool. Our methodology deviates from the conventional binary approach to corpus filtering that is commonly used to process large corpora. Instead, we evaluate corpus documents along a continuous spectrum, allowing us to fully exploit the corpus for training and analysis, which is particularly beneficial for low- and medium-resource languages. The human evaluation in Section 5 demonstrates the effectiveness of our scoring system, as documents of higher perceived quality are mostly scored higher than those of lower quality. In addition, we release both the corpus and the code developed for CURATE under permissive licenses that allow use by researchers and commercial entities alike.

CURATE includes a topic labeling module that filters specific content such as adult material, abusive language, or boilerplate from a corpus. It categorizes documents asynchronously with a fine-tuned Transformer Encoder model that determines their relevance to the topic and generate a separate score. However, these filters are missing in CATalog 1.0. For this reason, we plan to include different topic classifier models in the future to better assess corpus quality from different perspectives.

7. Acknowledgements

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

This work is the result of the project reference 2022/TL22/00215337 funded by the Ministerio de Asuntos Económicos y Transformación Digital and by the Plan de Recuperación, Transformación y Resiliencia founded by EU - NextGenerationEU.

We thank the VIVES Plan for language technologies of the Valencian community, https://vives.gplsi.es/, from the CENID Digital Intelligence Center of the University of Alicante.

8. Ethical Considerations and Limitations

Efficiency vs. Reliability In developing CURATE, we faced a trade-off between performance and latency. We benchmarked sentence tokenizers for efficiency, including the Europarl sentence splitter by Philipp Koehn and Josh Schroeder[*], `spaCy`'s tokenizer (Honnibal and Montani, 2017), and the Punkt Sentence Tokenizer from the `nltk` package (Bird et al., 2009). The Punkt Tokenizer was the most efficient, being 6.5 times faster than Europarl and 18 times faster than `spaCy`. Given the large data volume, we chose this efficient implementation, aware it may occasionally produce suboptimal tokenizations. However, our evaluators, as explained in Section 3.7, are designed to produce a score that does not depend solely on correct tokenization. Improving tokenization accuracy usually requires more time and computing resources, which would increase text processing latency. Balancing high performance with accuracy is a task for future development, not only regarding tokenization, but within the whole pipeline.

Dialectal variety When collecting content from web-crawled sources, we often find that standard language varieties are over-represented (Dunn, 2020), which can significantly impact the performance of language models. While topic and domain variety in the training data is known to improve the generalization capabilities of LLMs (Gao et al., 2020; Artetxe et al., 2022), we must also emphasize the importance of language diversity. If our training data is dominated by the standard language variant, this will negatively affect the model's accuracy in encoding and generating non-standard dialects. Consequently, this could lead to the exclusion of certain demographic groups (Weidinger et al., 2021). Indeed, NLP applications tend to perform poorly when trained on one demographic sample and tested on another (Plank, 2016; Hovy and Prabhumoye, 2021).

Our corpus consists mainly of web-crawled content in Central Catalan, which is the standard variant with the most demographic weight, spoken in the province of Barcelona and most of Tarragona and Girona. However, we aim to include other dialects as well. In Section 4, we discussed the manual inclusion of sources from Valencian (BOUA, DOGC, DOGV, Les Corts Valencianes), as spoken in the Valencian Community; and Balearic Catalan (IB3), as spoken in the Balearic Islands. These dialects account for 0.98% and 0.1% of the words in the whole corpus, respectively. In addition, we include posts and conversations from the Racó Català Fòrums, which reflect more diverse sociolects. While the dependence on large web-crawled sources tends to favor dominant voices and language varieties in our corpus, we hope that our focus on diverse data sources across different domains and geographical regions will contribute to a more representative Catalan dataset in the future.

Data Ownership and Copyright Web-scraped data presents legal uncertainties in different jurisdictions that affect data creators and users. While academic researchers enjoy fair use rights, these protections may not apply to commercial use. We only release datasets under permissive licenses such as CC-BY[*], CC-BY-SA[*], CC0[*], and Open Data Agreements. However, data creators should adopt best practices to protect the privacy of individuals within their datasets. Some suggested approaches include the use of anonymization tools such as regular expressions or NER tagging (Laurençon et al., 2023), while others allow contributors to opt out of sharing their data (Kocetkov et al., 2022). Given the nature of web-scraped data, ensuring privacy is challenging because personally identifiable information (PII) appears in multiple documents, is intertwined with useful general knowledge, and is difficult to remove effectively at scale (Wallace et al., 2020). While our efforts for this initiative are not yet sufficient, we believe it is important to identify and protect potential PII to address privacy concerns in future steps.

9. Bibliographical References

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a

[*] https://pypi.org/project/sentence-splitter/

[*] https://creativecommons.org/licenses/by/4.0/

[*] https://creativecommons.org/licenses/by-sa/4.0/

[*] https://creativecommons.org/public-domain/cc0/

very large-scale multilingual web corpus. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event)*. Leibniz-Institut für Deutsche Sprache.

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. *arXiv preprint arXiv:2201.06642*.

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.

Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de Viñaspre, and Aitor Soroa. 2022. Does corpus quality really matter for low-resource languages? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7383–7390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marta Bañón, Miquel Esplà-Gomis, Mikel L. Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, Tobias van der Werff, and Jaume Zaragoza. 2022. MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 303–304, Ghent, Belgium. European Association for Machine Translation.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the Natural Language ToolKit*. "O'Reilly Media, Inc.".

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Étienne Brunet et al. 1978. *Le vocabulaire de Jean Giraudoux structure et évolution*. Slatkine.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jonathan Dunn. 2020. Mapping languages: the corpus of global language use. *Language Resources and Evaluation*, 54(4):999–1018.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling.

Aitor Gonzalez-Aguirre, Montserrat Marimon, Carlos Rodriguez-Penagos, Javier Aula-Blasco, Irene Baucells, Carme Armentano-Oller, Jorge Palomar-Giner, Baybars Kulebi, and Marta Villegas. 2024. Building a data infrastructure for a mid-resource language: The case of catalan. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. European Language Resources Association and the International Committee on Computational Linguistics.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al. 2022. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*.

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.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8).

Joseph Jennings, Mostofa Patwary, Sandeep Subramanian, Shrimai Prabhumoye, Ayush Dattagupta, Mohammad Shoeybi, and Bryan

Catanzaro. 2023. Curating Trillion-Token Datasets: Introducing NVIDIA NeMo Data Curator. [Accessed 21-03-2024].

Ali Khodabakhsh, Fatih Yesil, Ekrem Guner, and Cenk Demiroglu. 2015. Evaluation of linguistic and prosodic features for detection of alzheimer's disease in turkish conversational speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):1–15.

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro von Werra, and Harm de Vries. 2022. The stack: 3 tb of permissively licensed source code.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, et al. 2023. Madlad-400: A multilingual and document-level large audited dataset. *arXiv preprint arXiv:2309.04662*.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2023. The bigscience roots corpus: A 1.6tb composite multilingual dataset.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. 2021. Jurassic-1: Technical details and evaluation. *White Paper. AI21 Labs*, 1.

Nikola Ljubešić and Antonio Toral. 2014. caWaC – a web corpus of Catalan and its application to language modeling and machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1728–1732, Reykjavik, Iceland. European Language Resources Association (ELRA).

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in nlp.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 823–848, Online only. Association for Computational Linguistics.

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

Antoine Slegers, Renee-Pier Filiou, Maxime Montembeault, and Simona Maria Brambati. 2018. Connected speech features from picture description in alzheimer's disease: A systematic review. *Journal of Alzheimer's disease*, 65(2):519–542.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Eric Wallace, Florian Tramèr, Matthew Jagielski, and Ariel Herbert-Voss. 2020. Does gpt-2 know your phone number?

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data.

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.

10. Language Resource References

Abadji, Julien and Ortiz Suarez, Pedro and Ismail, Rua and Takeshita, Sotaro and Nagel, Sebastian and Sagot, Benoît. 2022. *OSCAR* 23.01. *Hugging Face* repository. PID https://huggingface.co/datasets/oscar-corpus/OSCAR-2301

Abadji, Julien and Ortiz Suarez, Pedro and Romary, Laurent and Sagot, Benoît. 2022. *OSCAR* 22.01. *Hugging Face* repository. PID https://huggingface.co/datasets/oscar-corpus/OSCAR-2201

Armengol-Estapé, Jordi and Carrino, Casimiro Pio and Rodriguez-Penagos, Carlos and de Gibert Bonet, Ona and Armentano-Oller, Carme and Gonzalez-Agirre, Aitor and Melero, Maite and Villegas, Marta. 2021. Catalan General Crawling *CaText*. *Zenodo* repository. PID https://zenodo.org/records/5483031

Bañón, Marta and Chichirau, Malina and Esplà-Gomis, Miquel and Forcada, Mikel L. and Galiano-Jiménez, Aarón and García-Romero, Cristian and Kuzman, Taja and Ljubešić, Nikola and van Noord, Rik and Pla Sempere, Leopoldo and Ramírez-Sánchez, Gema and Rupnik, Peter and Suchomel, Vít and Toral, Antonio and Zaragoza-Bernabeu, Jaume. 2023. Catalan web corpus *MaCoCu*-ca 1.0. *ILC-CNR* for *CLARIN-IT* repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa. PID http://hdl.handle.net/11356/1837

Ljubesic, Nikola and Toral, Antonio. 2014. *caWaC*. *Hugging Face* repository. PID https://huggingface.co/datasets/cawac

Ortiz Suarez, Pedro and Abadji, Julien and Ismail, Rua and Sagot, Benoît and Takeshita, Sotaro and Nagel, Sebastian. 2022. *Colossal OSCAR* 1.0. *Hugging Face* repository. PID https://huggingface.co/datasets/oscar-corpus/colossal-oscar-1.0

Raffel, Colin and Shazeer, Noam and Roberts, Adam and Lee, Katherine and Narang, Sharan and Matena, Michael and Zhou, Yanqi and Li, Wei and Liu, Peter. 2019. *mC4*. *Hugging Face* repository. PID https://huggingface.co/datasets/mc4

A. Appendix

A.1. CATalog 1.0 details

Source	Domain	Curation	Parts	Words	Percentage
mC4	Web	CURATE	19	6,377.99	36.55%
OSCAR 23.01	Web	CURATE	10	2,171.68	12.45%
OSCAR 22.01	Web	CURATE	14	1,397.77	8.01%
caWaC	Web	CURATE	13	1,394.81	7.99%
MaCoCu	Web	CURATE	11	1,724.07	9.88%
Populars .cat	Web	CURATE	53	838.42	4.81%
CO 2023-23	Web	Ungoliant	1	207.59	1.19%
CO 2023-14	Web	Ungoliant	1	195.43	1.12%
CO 2022-17	Web	Ungoliant	1	195.03	1.12%
Wikimedia	Web	CURATE	1	3.90	0.02%
TDX	Scientific	Manual	3	323.60	1.85%
Wikipedia	Scientific	Manual	1	266.69	1.53%
IB3	Journalistic	Manual	1	15.82	0.09%
Grup El Món	Journalistic	Manual	1	85.27	0.49%
Vilaweb	Journalistic	Manual	1	46.90	0.27%
Nació Digital	Journalistic	Manual	1	216.27	1.24%
ACN	Journalistic	Manual	2	81.25	0.47%
RC Articles	Journalistic	Manual	4	358.57	2.06%
Aquí Berguedà	Journalistic	Manual	1	8.27	0.05%
PC	Political	Manual	1	10.09	0.06%
DSCV	Political	Manual	2	26.88	0.15%
DOGV	Legal	Manual	1	76.48	0.44%
DOGC	Legal	Manual	1	70.51	0.40%
BOUA	Legal	Manual	1	12.42	0.08%
RC Fòrums	Social	CURATE	1	1,342.53	7.69%
Gutenberg	Books	Manual	1	1.29	0.01%
Total	-	-	-	**17,450.50**	**100%**

Table 1: Classification of the 26 subdatasets that make up CATalog 1.0 along with their domain, their method of curation, their number of parts, their total number of words (in millions) and the percentage contribution to the total dataset.

A.2. CURATE's dataflow

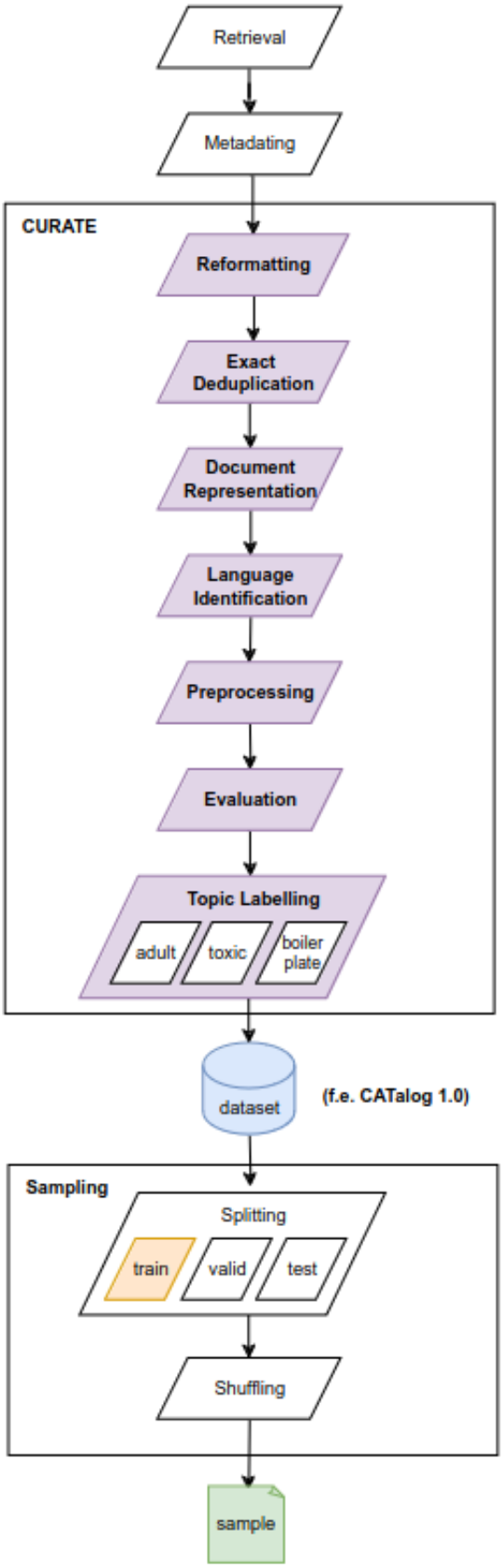

Figure 4: Dataflow.