Towards a Cleaner Document-Oriented Multilingual Crawled Corpus

Julien Abadji[1], Pedro Ortiz Suarez[1, 2], Laurent Romary[1], Benoît Sagot[1]

Inria[1], Sorbonne Université[2]

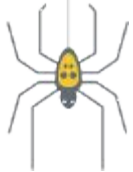
LREC 2022 - XX June 2022 - Marseille, France

Introduction OSCAR

- Series of multilingual corpora (2019, 21.09, 22.01)
- Very large dataset size (~8To)
- Widely used (140+ citations of the original paper, 15k+ downloads on HuggingFace)
- Based off of CommonCrawl dumps.

https://oscar-corpus.com/

Introduction CommonCrawl



- Open repository of web crawl data
- (2–)Monthly web crawls available to download
- WARC (extracted HTML) or WET (extracted text) formats
- Numerous corpora are based off of CommonCrawl: mC4, CCAligned..
- Each dump is comprised of records (≃ a web page/document)

Introduction Context

- OSCAR is a widely used corpus in NLP
- Models are getting bigger and needs more data…
- … and data has to be of a better quality.

Previous OSCAR Corpora have been destroying source documents integrity, providing line-based corpora. We aim to provide a document-oriented corpus that can still be used (with some work) as a line-oriented one, while adding quality-related annotations.

Introduction OSCAR 2019

- First release
- Simple filtering, text-only dataset
- 166 languages
- Source material from CC November 2018 text extracts.

Source data size	7.4TB (compressed)
Corpus size	6.3TB
Number of languages	166
English corpus	2.3TB
Icelandic corpus	1.5GB
Occitan corpus	5.8M

[PJO Suarez 2019]

Introduction OSCAR 21.09

- Addition of metadata imported from source material
- Backward compatible (metadata in side files)
- Update of source material to include recent events (Crawl of February/March 2021)

Source data size	8.1TB (compressed)
Corpus size	7.2TB+1.2TB(metadata)
Number of languages	168
English corpus	2.9TB
Icelandic corpus	2GB
Occitan corpus	12MB

[Abadji 2021]

Introduction Line-oriented corpora

- Both 2019 and 2109 are line-oriented.
- Lines from a same record are separated depending on the language identification

Consequences:

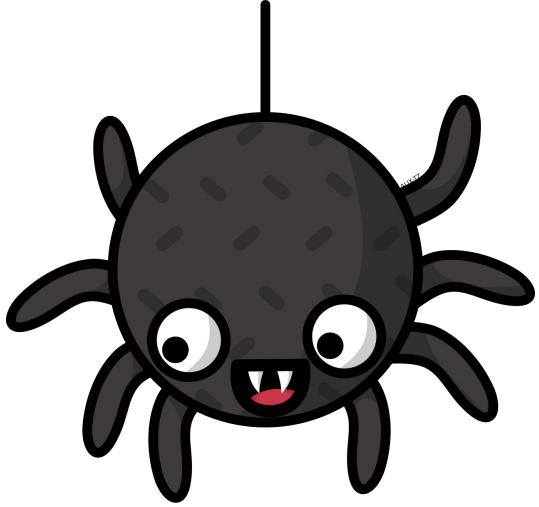
- A single record can be split into multiple subcorpora
- Lines that have not been properly identified (confidence not sufficient) are discarded

Introduction OSCAR 22.01 Rationale

- Go from line-oriented to document-oriented
- Try to preserve document integrity
- Add quality-related annotations to indicate potentially bad data
- Preserve line-level identifications to provide line-oriented corpus extraction

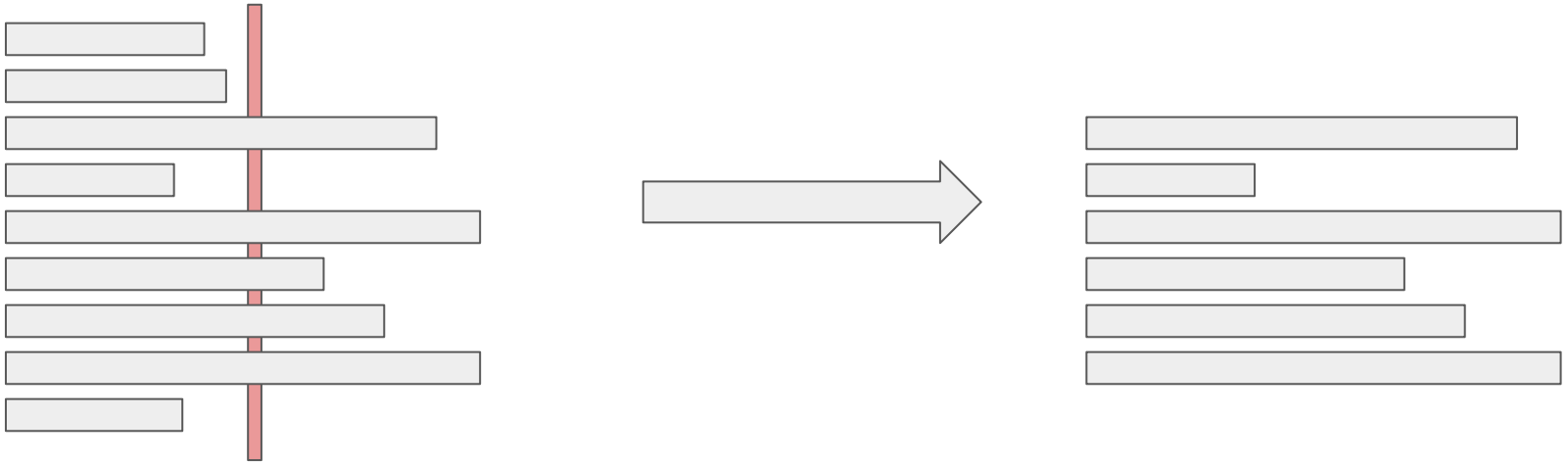
Generation pipeline Outline

- Go from November/December 2021 CC Crawl dump
- Iterate on records
  - Filter out records on a length criterion
  - Filter out short sentences at top/bottom
  - Identify lines in record (fastText)
  - Figure out the main language from the line identifications
  - Annotate and add metadata from CommonCrawl
  - Write documents in language subcorpora
- Push on Huma-Num/HuggingFace



Generation pipeline Filtering

- short line := less than 400 bytes
- Remove documents that mainly have content in short lines
- Remove short lines at beginning/end of documents



Generation pipeline Identification

- fastText as the identification model/library
- Line-based identification, use confidence and size as guides to identify document language
- Identification process:
  - Identify lines (confidence threshold at 0.8, no labeling if under)
  - Count content per language (in bytes), and compute a confidence score
  - Multilingual check (if there are N languages with enough content)
  - Compute the overall weighted confidence of the document

Generation pipeline Annotations

- Annotations are quality-related tags that are added to documents.
- They enable filtering the corpus from the users' side.
- One document can have multiple annotations, or none.

Annotation name	Description
`tiny`	Low number of lines
`short_sentences`	High number of short lines
`header/footer`	High number of short lines at start/end of document
`noisy`	High punctuation/letters ratio
`adult`	URL/domain is on an adult blocklist

Corpus Format

- Each document is embedded in a JSON object
- One subcorpus per language
- Subcorpora in JSONLines
- Distributed in 1GB, gzipped splits.

Field	Description
`content`	Textual data
`warc_headers`	Crawl metadata
`metadata`	Identifications and Annotations

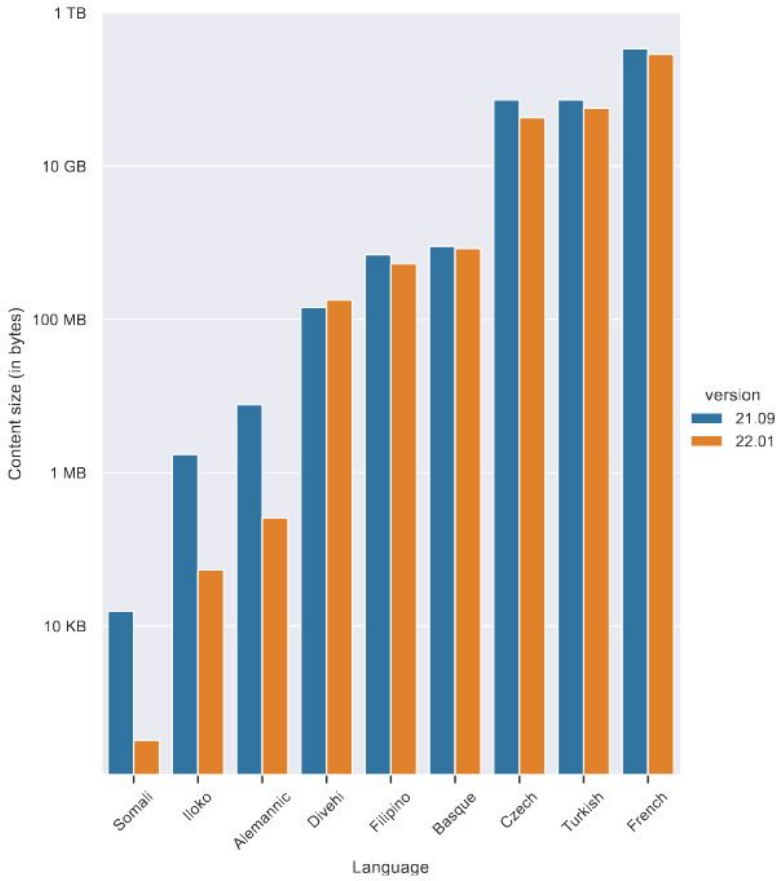
```
{
  "content":"newline\nseparated\ncontent", //content itself
  // Headers from the crawler
  // Note that nothing is changed, so the content length may be incorrect.
  "warc_headers":{
    //...some fields omitted
    "warc-date":"2021-09-16T11:07:14Z",
    "warc-identified-content-language":"eng",
    "content-length":"1694",
    "warc-target-uri":"https://foo.bar",
  },
  "metadata":{
    // Document-wide identification.
    // The "prob" is the weighted average of the identified lines.
    "identification":{
      "label":"en",
      "prob":0.6268374
    },
    // Annotations. Can be null if no annotations have been added.
    "annotation":[
      "short_sentences",
      "footer"
    ],
    // Line-by-line identifications
    // Can have null values for lines that did not get an identification.
    "sentence_identifications":[
      {
        "label":"en",
        "prob":0.93925816
      },
      null,
      {
        "label":"en",
        "prob":0.9606543
      }
    ]
  }
}

```

Corpus Size I

We take 3 low (<5MB), 3 mid (~100MB) and 3 high (>10GB) resource languages and compare their size evolution.

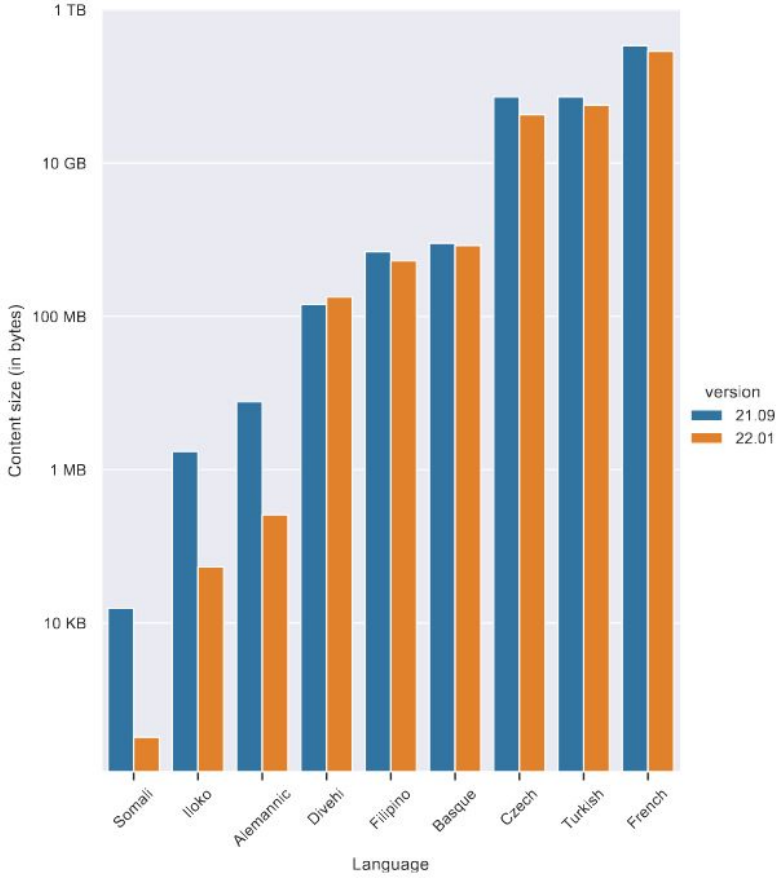
We notice that low-resource languages exhibit an important size decrease.



Corpus Size II

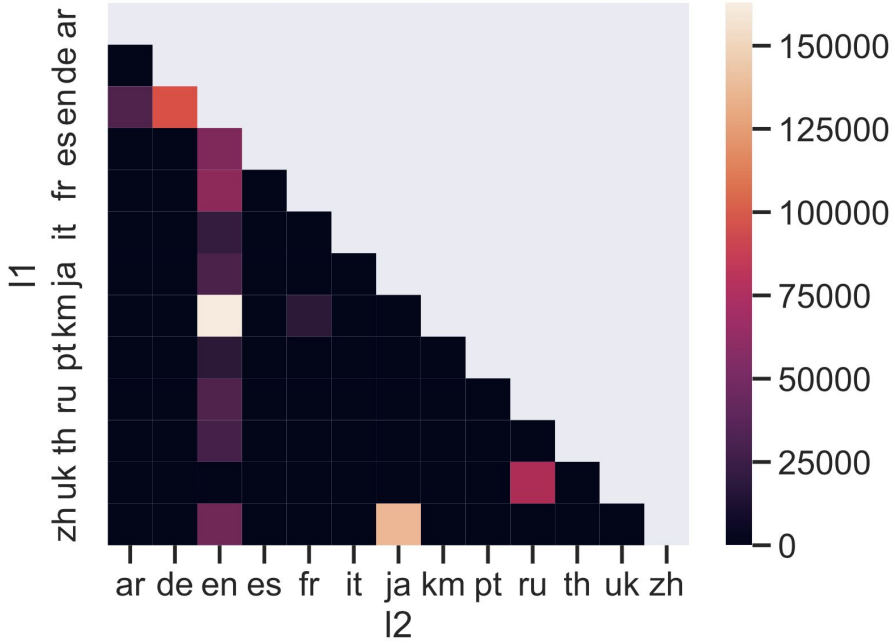
When looking for Alemannic in the German corpus, we find around 30MB of data, which once quality-filtered could be comparable to the 21.09 Alemannic corpus size (380KB versus 5MB).

Low-resource language corpora could be rebuilt by looking into higher-resource language corpora.



Corpus Multilingual

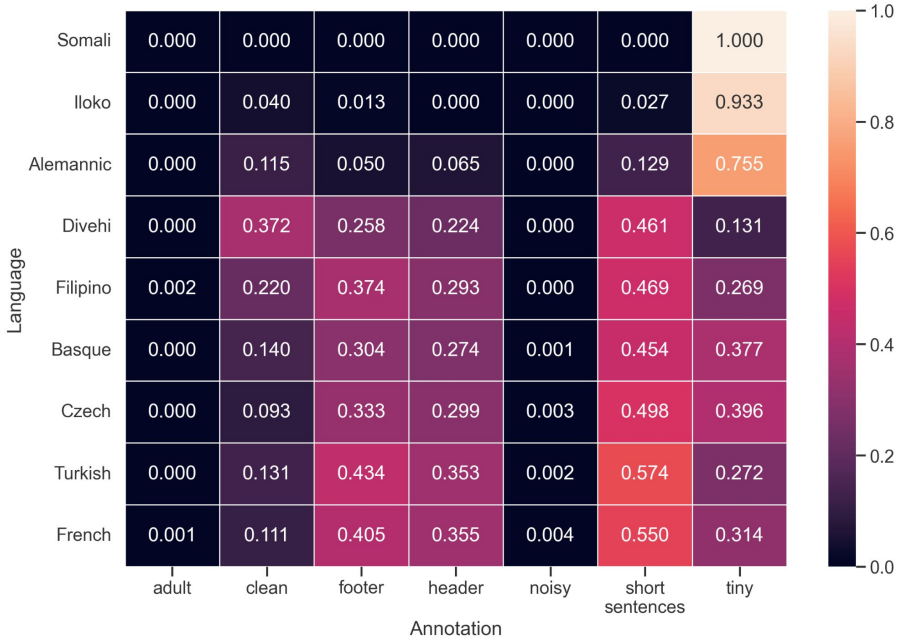
- We also distribute a multilingual corpus for documents that exhibit a homogen content distribution in numerous languages
- Corpus weighs around 12GB
- Strong presence of English in all documents: Boilerplate?
- Other interesting overlaps to explore: ID problem? Linguistic proximity? Truly multilingual?



Corpus Annotations I

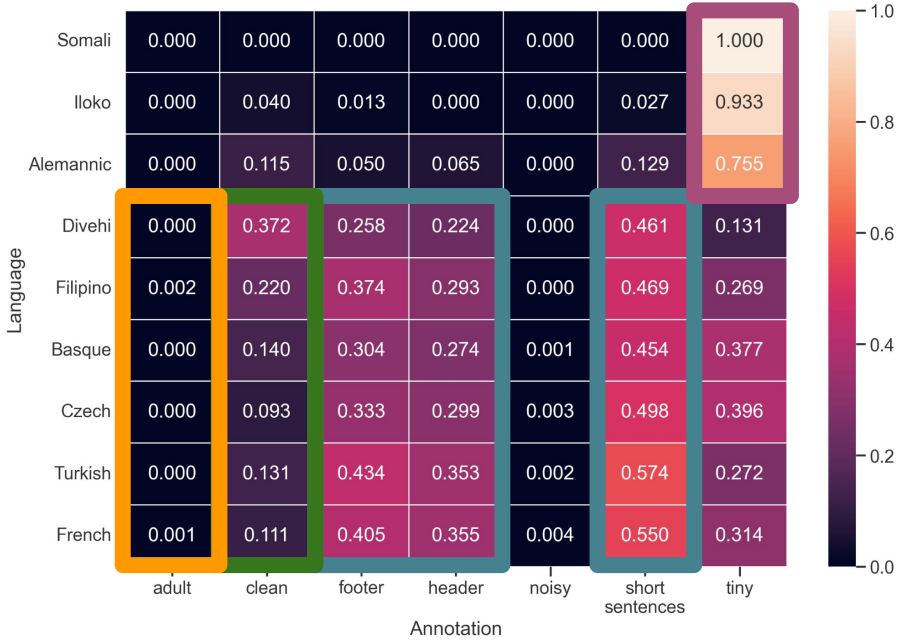
- Distribution of annotations on selected languages
- `clean` := no annotation

Annotation name	Description
`tiny`	Low number of lines
`short_sentences`	High number of short lines
`header/footer`	High number of short lines at start/end of document
`noisy`	High punctuation/letters ratio
`adult`	URL/domain is on an adult blocklist



Corpus Annotations II

- `adult` documents usually less than 0.1%, but with a potentially good precision
- Low amount of `clean`, but enough in terms of size on big corpora (French=32GB).
- High number of documents with `footer`/`header` or `short_sentences`
- Documents with `tiny` mainly in low-resource languages

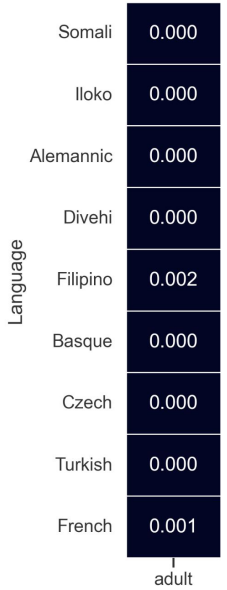


Corpus Adult annotation

- URL/Domain filtering from blocklist (UT1)
- On French corpus, 32k adult documents (out of 52M).

- Filter on French LGBTQI+ website tetu.com shows ~3.2% of annotated documents comes from this website.
- Manual inspection of 100 documents reveals 21FP + 2 LGBTQI+ FP.

Adult annotation relatively precise but very bad recall and risk of flagging non-adult LGBTQI+ content.



Discussion Corpus size

- Important size changes for low-resource languages.
  - Low-resource languages content found in other subcorpora
  - Need for tooling to extract this content?
- Assert quality of low-resource languages.
  - Are size changes related to quality filtering changes?
  - Inform on quality using users feedback

cbk	Chavacano	521B	521B	168B	168B	CBK 1 OPEN

Discussion Annotations

- Some annotations widespread (>50% of documents) in the corpus, reducing relevancy.
- Adult annotation is blocklist-based and has important drawbacks:
  - Only URL/Domain based
  - Limited multilinguality
  - High number of False Negatives
- Annotation methods have to be improved:
  - Model based yet lightweight (adult ngrams?)
  - Better thresholds

Main difficulty is to keep annotators simple to keep corpus generation fast enough.

Conclusion

- OSCAR 22.01 marks a new, breaking step for OSCAR Corpora
- Both document- and line-oriented processing are easily possible
- Annotations provide a first step towards better labeled and filtered data

- Change of format requires more tooling to access the data
- Annotations relevance has to be assessed and refined for future versions

Thank you for your time.