Do Language Models Care About Text Quality? Evaluating Web-Crawled Corpora Across 11 Languages

Rik van Noord♢, Taja Kuzman♣, Peter Rupnik♣, Nikola Ljubešić♣
Miquel Esplà-Gomis♠, Gema Ramírez-Sánchez♡ and Antonio Toral♢
♢University of Groningen, ♣Jožef Stefan Institute, ♠Universitat d'Alacant, ♡Prompsit
rikvannoord@gmail.com

Abstract

Large, curated, web-crawled corpora play a vital role in training language models (LMs). They form the lion's share of the training data in virtually all recent LMs, such as the well-known GPT, LLaMA and XLM-RoBERTa models. However, despite this importance, relatively little attention has been given to the quality of these corpora. In this paper, we compare four of the currently most relevant large, web-crawled corpora (CC100, MaCoCu, mC4 and OSCAR) across eleven lower-resourced European languages. Our approach is two-fold: first, we perform an intrinsic evaluation by performing a human evaluation of the quality of samples taken from different corpora; then, we assess the practical impact of the qualitative differences by training specific LMs on each of the corpora and evaluating their performance on downstream tasks. We find that there are clear differences in *quality* of the corpora, with MaCoCu and OSCAR obtaining the best results. However, during the extrinsic evaluation, we actually find that the CC100 corpus achieves the highest scores. We conclude that, in our experiments, the quality of the web-crawled corpora does not seem to play a significant role when training LMs.

Keywords: Monolingual corpora, Corpus evaluation, Large language models

1. Introduction

The field of natural language processing has witnessed a paradigm shift with the emergence of large language models (LLMs) that exhibit impressive capabilities in various language understanding tasks (Zhang et al., 2022; OpenAI, 2023; Touvron et al., 2023). Monolingual corpora have played a pivotal role in this data-driven revolution, serving as the foundational resource for training these LLMs. A growing number of monolingual corpora have been published in the last years, many of them specifically conceived to train LLMs (Conneau et al., 2020; Xue et al., 2021), by implementing different methodologies to collect and curate data. However, despite their importance, the content of these corpora has been given modest attention. Since these corpora are compiled using automatic tools, with limited and varying quality control, it is unclear (i) how these corpora are qualitatively different and (ii) if and how the differences actually affect the models in terms of downstream performance.

In this paper, we aim to shed light on these issues. In particular, we evaluate the well-known OSCAR (Ortiz Suárez et al., 2019), CC100 (Conneau et al., 2020), mC4 (Xue et al., 2021) and MaCoCu (Bañón et al., 2022) corpora across eleven non-English European languages. First, we hire professional linguists to manually evaluate the *quality* of the corpora, irrespective of the size. Then, we train a language model on each language-corpus combination for a subset of five languages, to evaluate whether these differences in quality transfer to differences in downstream performance. All code, models and annotations are made publicly available.[1]

In terms of quality, our findings indicate that the MaCoCu and OSCAR corpora are superior options. They contain a greater number of documents that consist of (publishable) running text, while exhibiting a significantly lower number of documents that are either in an incorrect language or lack running text. The mC4 corpus seems to be the one of the lowest quality in our evaluation, with especially concerning results for Maltese, as over 75% of the corpus was in another language.

However, despite the evident differences in quality, the findings do not seem to directly transfer to actually training the LMs on the corpora in question. For a subset of five languages, we continue training XLM-RoBERTa (XLM-R, Conneau et al., 2020) on each of our corpora and evaluate performance across a number of structured prediction and natural language understanding tasks. We find that CC100 actually obtains the best performance, while OSCAR is the worst performing corpus. We intentionally refrained from controlling for data set size, as the size of the data sets is associated with the extent of data cleaning conducted, which is undeniably linked to the quality of the preserved texts. However, even if we do control for size, we do not find any indication that the quality of the data significantly influences performance, counter-intuitive as it may be.

[1] https://github.com/RikVN/Corpus_Eval/

LREC-COLING 2024, pages 5221–5234
20-25 May, 2024.

2. Related work

Given the important role web-crawled corpora play in training LLMs, and given that they are known to be noisy (Junczys-Dowmunt, 2019; Luccioni and Viviano, 2021), it is curious that there are only a few papers that actually aimed to evaluate the quality of these corpora. Caswell et al. (2020) analysed the performance of their automatic language identification system and found serious issues for lower-resource languages. Dodge et al. (2021) documented the C4 data set used for training mC4 (Xue et al., 2021) and found that a significant number of texts came from unexpected sources, such as US military websites. Moreover, they found a substantial amount of machine-generated text and texts from common NLP evaluation benchmarks. Closer to our work is the study by Kreutzer et al. (2022), in which they run a human evaluation on a large number of both monolingual and parallel corpora. The authors focus mostly on the languages for which less data is available in each corpus, but they also include other languages with more data for a more representative evaluation. They find serious quality issues for the evaluated corpora, especially for low resource languages, but do not run an automatic evaluation.

Basque The closest to our work is the study of Artetxe et al. (2022). They perform both human and automatic evaluation on monolingual corpora for Basque, where the automatic evaluation is extrinsic, by training LMs on these corpora and then evaluating them on several downstream tasks. They conclude that there is no clear correlation between either the size or the quality of data and the performance of the LMs trained on them. Our work follows Artetxe et al. (2022), also evaluating the *quality* of monolingual corpora and the performance of LMs trained on them, but we do so for several languages and include a larger number of web crawled corpora.

Cleaning It should be noted that the corpora under evaluation have undergone a preliminary filtering and cleaning process prior to their release (see Section 3). However, these (or similar) corpora often go through an additional cleaning process before being used to train an LM (Rae et al., 2021; Touvron et al., 2023; OpenAI, 2023). However, these cleaning practices are all (slightly) different and are usually not explained in enough detail to ensure reproducibility. Since we have no access to each individual cleaning process, in this paper, we explicitly **do not** attempt to find the best cleaning methods. We simply evaluate existing monolingual corpora by using them *as is* and only performing the simple cleaning steps as instructed by their creators.

Monolingual LMs In our automatic evaluation, we train encoder-only monolingual language models. Even though many current studies focus on large decoder-only models, we believe there is still a need for smaller, monolingual LMs. This is evidenced by the use of such models to *enrich* corpora at scale in a computationally-efficient manner (Kuzman et al., 2023), but also simply by the popularity of such models (de Vries et al., 2019; Sanh et al., 2019; Martin et al., 2020; Le et al., 2020; Souza et al., 2020; Schweter, 2020; Ljubešić and Lauc, 2021; Snæbjarnarson et al., 2022; Seker et al., 2022) on the HuggingFace hub.[2]

Continued training In this paper, we continue training an existing LM, instead of training from scratch. The main advantage of this approach is that it is a lot more efficient, while results remain competitive or even improve (Gururangan et al., 2020; Wang et al., 2020; Chau et al., 2020; Muller et al., 2021; Snæbjarnarson et al., 2022). A similar option is *adapting* LMs to a target language (Pfeiffer et al., 2020, 2021) by learning language-specific representations. However, Ebrahimi and Kann (2021) found that continued pretraining provided the best results on low-resource languages, while also being the simplest method to apply, which is why we use this method in our paper.

3. Corpora

Four corpora were included in the evaluation described in this work. In this section, we provide a general overview of each of them.

CC100 (Conneau et al., 2020). Corpus created to train the popular XLM-R language model (Conneau et al., 2020). The corpus covers 100 languages and was built through cleaning twelve Common Crawl dumps,[3] using one of them to extract only text in English, and the remaining eleven dumps to extract data in the rest of languages. The per-language size of the corpus ranges from 55.6 billion tokens in English to 10 million tokens in Sundanese. This corpus was built by following the approach of Wenzek et al. (2020), which consists of three main steps: (a) deduplication of paragraphs on Common Crawl dumps; (b) language identification with fastText (Grave et al., 2018) and (c) providing, for each paragraph in the corpus, the perplexity as provided by a language model as a proxy of the quality of the text.

[2] For example, CamemBERT (Martin et al., 2020) had over 2.5 million downloads last month (as of March 2024).

[3] https://data.commoncrawl.org/

mC4 (Xue et al., 2021). Corpus created to build the mT5 language model (Xue et al., 2021). The corpus covers 101 languages and was built by processing all the Common Crawl dumps available at that time. The tool cld3[4] was used for language identification, deduplication was performed at paragraph level and pages with too few or too short paragraphs or with bad words were filtered out. The final size of the mC4 corpus is approximately 6.3 trillion tokens in total.

OSCAR (Ortiz Suárez et al., 2019). This corpus is also built through a cleaning and curation process on Common Crawl data. It is developed incrementally, with regular releases of new versions including data from the last Common Crawl dumps. According to the documentation of the last version of the corpus at the time of writing (v23/01),[5] OSCAR uses fastText (Grave et al., 2018) for language identification and the TLSH (Oliver and Hagen, 2021) fuzzy hashing method to identify near duplicates. This version of OSCAR also provides the perplexity provided by a KenLM (Heafield, 2011) language model trained on harmful content identified in previous versions of OSCAR. OSCAR is the corpus that covers the most languages among the collections evaluated in this paper, with a total of 152 languages in its latest version. However, it is worth noting that for about 50 of them, the corpus includes less than 1 million tokens.[6]

MaCoCu (Bañón et al., 2022). In contrast to the rest of corpora compared in this paper, MaCoCu corpora are not obtained by processing Common Crawl data. Instead, a strategy consisting of crawling relevant internet top-level domains directly for the targeted languages is followed (e.g., .al for Albanian). Several studies claim that Common Crawl over-represents English while under-represents other languages (Bender et al., 2021; Ranathunga and de Silva, 2022); the strategy adopted is aimed at avoiding this effect and, at the same time, granting access to more up-to-date data than that stored in Common Crawls. The MaCoCu corpus covers 11 low-resourced European languages, and consists of a total of about 17.3 billion tokens, with Turkish being the largest (4.3 billion tokens) and Montenegrin being the smallest (200 million tokens). During cleaning, deduplication and a set of heuristics is applied to fix or remove evidently problematic text fragments, such as badly encoded or too short ones.

[4] https://github.com/google/cld3

[5] https://oscar-project.github.io/documentation/versions/oscar-2301/

[6] Some languages have extremely little data, such as Kalmyk (27 tokens) or Quechua (13 tokens).

4. Manual Evaluation

In this section, we describe the methodology and results of the manual evaluation of the CC100, MaCoCu, mC4 and OSCAR corpora. We evaluate all languages that are present in MaCoCu, the corpus with the smallest amount of languages present. It should be noted that certain languages found in the MaCoCu corpus are exclusive to this particular corpus and are not represented in any other corpora, namely Bosnian and Montenegrin. While we include these languages in the human evaluation, we mostly focus on the following nine languages that are present in multiple corpora: Albanian, Bulgarian, Croatian, Icelandic, Macedonian, Maltese, Serbian, Slovenian and Turkish.

4.1. Annotation Scheme

We perform annotation at paragraph level as this is the common format that each corpus has available. For OSCAR, we follow the best practice standard of only selecting the paragraphs that are recognized as being in the correct language.The other corpora are used as they are released. For each corpus and language combination, we randomly select 200 paragraphs for annotation. Annotators are asked to rank each paragraph using the following scale:

1. **Wrong language or not language (WL).** The text is not in the correct language, or is not in a natural language (e.g., links, html tags).
2. **Not running text (NR).** The text makes no sense, it is just a concatenation of words or a bunch of words together. Note that short sentences can still be running text.
3. **Partially running text (PR).** More than 50% is running text, but some parts are not. For example, the text is cut-off or has additional elements in brackets. A substantial part of the text should be cut-off for this to apply.
4. **Running text, but slightly non-standard (RT).** More than 90% is running text, but the text contains small mistakes, such as grammatical errors, typos, and missing punctuation. This category includes titles, headers and bullet points.
5. **Publishable text (PT).** 100% running text which is of publishable quality and contains no (formatting) mistakes. You could read this in a blog post, news article, recipe, magazine, etc. Note that the content itself does not have to be formal for this to apply.

This schema is inspired by those in two recent works (Kreutzer et al., 2022; Artetxe et al., 2022), which were taken as a starting point, combined, and refined by means of a pilot annotation conducted on

	% agreement	κ coef.
Albanian	68.0	0.51
Bosnian	86.5	0.59
Bulgarian	49.5	0.32
Croatian	72.1	0.62
Icelandic	81.0	0.39
Macedonian	48.5	0.27
Maltese	66.1	0.55
Montenegrin	47.5	0.23
Serbian	78.0	0.65
Slovenian	70.5	0.36
Turkish	52.5	0.32

Table 1: Inter-annotator agreement between the two annotators for each language for the evaluation of monolingual data. The second column shows the percentage (**%**) of annotations for which both annotators were in exact agreement; the third column shows Cohen's kappa coefficient (κ) between both annotators.

Slovenian and English samples. Each annotator is also provided with a number of example paragraphs in English[7], for each category. These are compiled in Table 11 of the Appendix. The annotators are instructed to select only one option, and to pick the lower number on the scale in case of uncertainty.

Details We hire two professional linguists per language to annotate subsets of all corpora in each of the 11 languages. The evaluation samples comprise 200 instances from each evaluated corpus. The size of the samples depends on the number of corpora in which the evaluated language is present. Specifically, the size ranges from 200 instances when only one corpus is available to 800 instances when all four corpora are included in the comparison. From the sample, we select 200 instances that are provided to both annotators to be able to calculate inter-annotator agreement. We balance the instances for each corpus per annotator, meaning that each annotator sees 100 instances of each corpus. The instances are shown to them in random order and blind fashion, i.e., they do not know which instance belongs to which corpus. For the 200 instances that have double annotations, we select one of them randomly to be included in the analysis (balanced per annotator). The annotators received a fair wage for their efforts that differed per language, but always exceeded minimum wage.

[7]All our annotators were also fluent in English. The advantage of using this language as instruction is that the instructions are the same across all languages.

Albanian	**WL**	**NR**	**PR**	**RT**	**PT**	**RT+PT**
MaCoCu	4	4	48	73	71	144
CC100	1	3	44	62	90	152
mC4	18	12	58	47	65	**112**
OSCAR	1	2	32	70	95	**165**
Bosnian	**WL**	**NR**	**PR**	**RT**	**PT**	**RT+PT**
MaCoCu	2	0	3	38	157	**195**
Bulgarian	**WL**	**NR**	**PR**	**RT**	**PT**	**RT+PT**
MaCoCu	5	8	18	78	91	**169**
CC100	2	23	18	66	91	157
mC4	17	49	25	47	62	**109**
OSCAR	2	26	26	58	88	146
Croatian	**WL**	**NR**	**PR**	**RT**	**PT**	**RT+PT**
MaCoCu	10	23	23	61	83	**144**
CC100	18	15	25	71	71	142
OSCAR	1	37	32	65	64	**129**
Icelandic	**WL**	**NR**	**PR**	**RT**	**PT**	**RT+PT**
MaCoCu	2	4	6	15	173	188
CC100	2	6	9	19	164	183
mC4	24	15	16	15	130	**145**
OSCAR	2	1	4	4	189	**193**
Macedonian	**WL**	**NR**	**PR**	**RT**	**PT**	**RT+PT**
MaCoCu	5	5	26	76	88	**164**
CC100	1	11	41	76	71	147
mC4	10	20	30	59	81	**140**
OSCAR	2	7	31	64	96	160
Maltese	**WL**	**NR**	**PR**	**RT**	**PT**	**RT+PT**
MaCoCu	9	101	38	16	36	52
mC4	164	17	4	1	14	**15**
OSCAR	7	32	30	17	98	**115**
Montenegrin	**WL**	**NR**	**PR**	**RT**	**PT**	**RT+PT**
MaCoCu	25	4	18	49	104	**153**
Serbian	**WL**	**NR**	**PR**	**RT**	**PT**	**RT+PT**
MaCoCu	1	1	14	82	102	**184**
CC100	0	5	24	60	111	171
mC4	5	14	47	65	69	**134**
OSCAR	0	1	24	69	106	175
Slovenian	**WL**	**NR**	**PR**	**RT**	**PT**	**RT+PT**
MaCoCu	3	4	14	33	146	179
CC100	1	13	29	15	142	157
mC4	11	22	23	30	114	**144**
OSCAR	0	2	12	13	173	**186**
Turkish	**WL**	**NR**	**PR**	**RT**	**PT**	**RT+PT**
MaCoCu	5	27	19	90	59	**149**
CC100	0	33	30	95	42	137
mC4	8	62	30	67	33	**100**
OSCAR	3	38	26	84	49	133

Table 2: Summary of the human evaluation of the web-crawled corpora. Paragraphs were annotated as Wrong Language (WL), Not-running text (NR), Partially Running Text (PR), Running Text (RT) or Publishable Text (PT).

	WL	NR	PR	RT	PT
WL	**101**	—	—	—	—
NR	29	**94**	—	—	—
PR	10	56	**108**	—	—
RT	14	38	117	**275**	—
PT	8	32	77	371	**847**

Table 3: Confusion matrix of the labels assigned by annotators to each instance with aggregated results for all languages. For each column, the highlighted value corresponds to cases in which both annotators agreed on the annotation label. Subsequent values in each column correspond to cases in which one annotator chose the label corresponding to the column and the other one chose a different label.

4.2. Annotation Results

The inter-annotator agreement in terms of exact annotation overlap and Cohen's kappa coefficient (κ) scores are shown in Table 1. Generally, the annotators agree a fair amount of the time. It is not surprising that there is some disagreement: for example, the difference between "Running Text" and "Publishable Text" is partially subjective. These are indeed the two annotation categories that annotators disagree most often about, as can be seen in Table 3. The instances where annotators disagree about a paragraph being in the wrong language generally come exclusively from the Serbo-Croatian languages (Bosnian, Croatian, Montenegrin and Serbian).

Full results The full results of the annotation are shown in Table 2. We distinguish between "Running Text, but slightly non-standard" (RT) and "Publishable Text" (PT) in our annotation scheme but, for the purpose of training LMs, we consider both categories appropriate. In fact, language models might actually benefit from also observing non-standard language use during training. Therefore, we also show the aggregated score of these two categories. Similarly, the other three categories are considered problematic for language model training. One exception is the category "Wrong Language" for Serbian, Croatian, Bosnian and Montenegrin: annotators were asked to distinguish between them, but a paragraph in Montenegrin instead of Serbian is very likely to be considered useful when training a Serbian LM. Actually, in the next section we only train a single language model for Croatian and Serbian. Generally speaking, we observe that MaCoCu and OSCAR contain paragraphs that are most often at least running text. MaCoCu has the highest score for 5 out of the 9 languages for which a comparison can be made, while OSCAR has the highest quality corpus for the other 4 languages. MC4 is the most problematic corpus: it has the least amount of useful paragraphs for all 8 languages it was included in. Especially Maltese seems to have issues: 164 out of 200 instances were in the wrong language for mC4, while half the MaCoCu instances did not contain running text.

	WL	NR	PR	RT	PT	RT+PT
MaCoCu	1.8	3.6	10.4	29.9	54.3	**84.2**
CC100	0.4	6.9	13.3	26.2	53.2	79.4
mC4	6.4	13.9	16.0	22.4	41.3	**63.7**
OSCAR	0.6	5.4	9.9	24.7	59.4	84.1

Table 4: Percentage of annotations for each of the annotation categories, averaged over corpus across the **seven** languages included in all evaluated corpora.

Average scores To get a clearer overview of the quality of each corpus, we also show an averaged score of the corpora involved. For a fair comparison, we only average over the seven languages (Albanian, Bulgarian, Icelandic, Macedonian, Serbian, Slovenian and Turkish) included in all of the four evaluated corpora. We do not show the total counts but the percentage of each annotation and average across the seven languages. This is shown in Table 4. In this scenario, MaCoCu and OSCAR still seem to be the highest quality corpora, with CC100 not far behind. The mC4 corpus is clearly of lower quality than the other three. Generally speaking, the results of this annotation do paint a slightly worrying picture about web-crawled monolingual data. For example, for mC4, around 1 out of every 5 paragraphs has serious issues: being in the wrong language or not (completely) consisting of running text. What might be even worse is that, for all corpora, only around half the paragraphs are of publishable quality, while the standards for this category were not particularly strict.

Document length We now examine the impact of the varying document lengths within each corpus. It can be argued that a corpus with numerous lengthy documents might still offer greater utility for language model training, even if it contains a lower percentage of high-quality documents. Moreover, longer documents are perhaps more likely to not be of publishable quality (since there is simply more text that can have issues), while still potentially preferable over very short, but high-quality documents. In Figure 1, we plot the percentage of documents that fit into a certain document length range for each corpus. For each value, we also indicate what percentage was annotated as not fully running text (i.e., WL, NR or PR). The percentages shown here are averaged over all seven languages that had data for each of the four corpora. We can

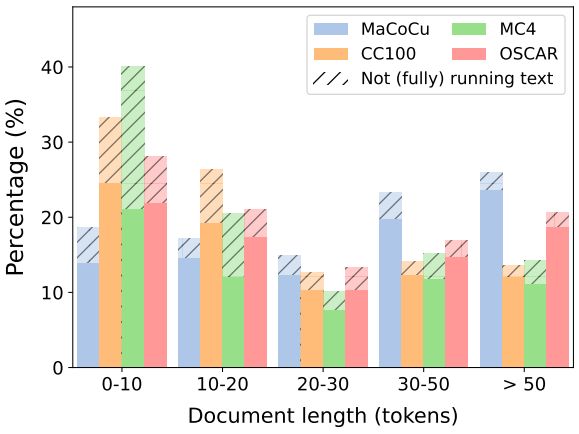

Figure 1: Percentage of annotated documents that are of certain length for each corpus, averaged over the seven languages that had data available in each corpus. For each bar we indicate the percentage of documents that did not fully contain running text, i.e., were annotated as Wrong Language, Not-running Text or Partially Running Text.

clearly see that the two corpora that had the highest amount of running and publishable text (MaCoCu and OSCAR) also have the highest percentage of large documents. Therefore, we are confident that the two potential issues identified above did not unfairly influence our annotation results.

5. Automatic Evaluation

This section focuses on the automatic evaluation of the corpora. We evaluate the corpora extrinsically by training general purpose encoder-only LMs. As our budget is limited, we evaluate on a subset of the languages in the manual evaluation, including: Albanian, Croatian, Icelandic, Serbian and Slovenian.

Continued training & Data We do not start training models from scratch, but continue training XLM-R (Conneau et al., 2020) for each corpus and language, as this is a more realistic usage of the often limited available data, and it is more computationally efficient. We opt for the base model instead of the large variant as the performance differences were small in initial experiments, while running the large model is approximately twice as expensive. We also train a system per language that concatenates all four available corpora as a comparison. Data sizes per corpus and language are shown in Table 5. We train only a single model for Croatian and Serbian, as the languages are very similar and mutually intelligible, though we do transliterate all Cyrillic Serbian and Croatian data to Latin.[8] Note

[8]The Croatian corpora had < 0.1% of Cyrillic data.

Language	CC100	MaCoCu	mC4	OSCAR	Comb
Albanian	2.1	1.4	4.6	0.9	9.0
Icelandic	1.3	1.6	2.9	0.6	6.3
Serbo-Croatian	10.8	12.1	4.9	1.4	29.2
Slovenian	4.2	4.7	11.0	0.4	20.1
Croatian	8.6	5.9	—	0.003	—
Serbian	2.2	6.2	4.9	1.4	—

Table 5: Data set sizes in GB of compressed text for the included corpora in LM experiments. Serbian and Croatian individual figures are included for reference although only a single model is trained.

that the mC4 corpus does not treat Croatian as a separate language, while OSCAR has very little Croatian data.

Details As previously stated, we continue training the XLM-R-base model. Each model is trained for 50,000 steps. We use a batch size of 1,024, a max learning rate of 1e-4 and 5,000 steps as warm-up. We do not make any modifications to the vocabulary. In total, we train 20 different LMs, one for each corpus-language combination (with Serbo-Croatian being treated as single language). A single experiment (50,000 steps) took around 4 days on a single Google Cloud TPU.

Fine-tuning The trained LMs are evaluated by fine-tuning them on downstream tasks. Even though we trained a single model for Serbo-Croatian, we evaluate Serbian and Croatian separately. We use the same tasks across all five languages: language-specific Part-of-Speech tagging (XPOS), Named Entity Recognition (NER), Choice of Plausible Alternatives (COPA, Roemmele et al., 2011) and Commitment Bank (CB, De Marneffe et al., 2019). The first two are classic evaluation tasks in NLP, while the latter two are part of the well-known SuperGLUE benchmark for English (Wang et al., 2019). For XPOS and NER, we use data from the Universal Dependencies project[9], with the exception of Icelandic NER, for which we use the MIM-GOLD-NER set (Ingólfsdóttir et al., 2020). The COPA data set was originally created for just English (Roemmele et al., 2011), but has gold standard translations available for Croatian, Serbian and Slovenian (Ljubešić, 2021; Ljubešić et al., 2022; Žagar et al., 2020). For Albanian and Icelandic, we translated the English data with Google Translate. For the CB task, we used Google Translate for all languages.[10] We use the suggested train, development and test splits for each task. For each language-task combination, we tune the learning rate of XLM-R-base on the development set and subsequently use this learning rate across other

[9]https://universaldependencies.org/

[10]All translations were obtained in June 2023.

experiments. For XPOS/NER we report weighted F-score, while for COPA and CB we report accuracy. Further details regarding data set sizes, training regimes and hyper-parameter settings are described in the Appendix.

5.1. Results

First, we evaluate performance after training for 50,000 steps. We average each of these evaluations over 10, 20 and 30 different random seeds for XPOS/NER, COPA and CB, respectively. For each language, we also report the *position* in the ranking of each corpus. This gives a clearer overview of the performance of the corpus despite the fact that such a ranking does not show the relative differences in the scores. Two models are considered to have a different position if they differ significantly according to the Mann-Whitney test (Mann and Whitney, 1947). Note that a lower number for position means that the model performed better.

Full results All results are shown in Table 6. Since XPOS is a relatively easy task, the differences between the corpora here are, as expected, quite small. For the other tasks, there is a bit more variation. One thing that stands out is that we improve on the XLM-R baseline in virtually all settings. Continuing training a multi-lingual language model on a specific language of interest is a simple and (relatively) cheap way of improving performance. There is also quite some variance in the results. For example, the Serbo-Croatian model trained on mC4 obtains the best performance on CB for both languages, while getting the worst performance on the COPA task (even worse than the baseline).

Averaged results Nevertheless, to obtain a clearer overview of performance per corpus, we show the averaged relative rankings in Table 7. We observe a surprising result: the best performing model seems to be based on the CC100 corpus, even ahead of the combined model. Though differences are small, the OSCAR corpus seems to be the worst corpus for LM training. This can likely be attributed to the fact that it is generally the smallest (see Table 5).

Importance of data set size So far, we explicitly did not control for data set size. The size of the released data is in many ways a design choice: stricter filtering leads to higher quality data, but this might not be preferable when training data-hungry LMs. It is therefore not entirely fair on a given corpus to run experiments in which data set size is controlled for. Nevertheless, we want to investigate how much the results are influenced simply by the amount of unique data available for each corpus,

		Scores				Positions				
Corpus	**Ep.**	**NER**	**POS**	**CP**	**CB**	**NER**	**POS**	**CP**	**CB**	**Avg.**
Croatian										
XLM-R	—	89.4	93.4	58.4	77.8	5	6	1	5	4.25
CC100	3.6	90.9	94.4	59.8	80.1	1	1	1	1	1.00
MaCoCu	3.2	90.7	94.3	59.2	77.9	1	1	1	5	2.00
mC4	7.4	89.8	94.2	55.8	80.1	4	4	5	1	3.50
OSCAR	24.9	89.3	94.1	56.2	78.9	5	5	5	1	4.00
Combined	1.3	90.8	94.3	58.6	79.6	1	1	1	1	1.00
Serbian										
XLM-R	—	93.5	91.0	57.6	76.6	5	6	3	5	4.75
CC100	3.6	94.6	92.6	61.2	80.3	1	1	1	1	1.00
MaCoCu	3.2	94.5	92.4	58.3	76.2	1	3	3	5	3.00
mC4	7.4	93.8	92.4	53.8	80.9	4	3	6	1	3.50
OSCAR	24.9	93.5	92.4	58.0	77.9	5	3	3	3	3.50
Combined	1.3	94.6	92.4	60.4	78.3	1	1	1	3	1.50
Albanian										
XLM-R	—	92.7	93.9	54.9	77.5	3	5	6	1	3.75
CC100	17.2	92.9	94.1	60.8	78.2	3	1	1	1	1.50
MaCoCu	26.2	93.2	94.0	57.8	76.5	1	2	3	5	2.75
mC4	7.8	92.8	94.0	56.4	77.8	3	2	4	1	2.50
OSCAR	37.2	92.8	94.0	55.7	79.5	3	2	4	1	2.50
Combined	4.2	93.1	93.9	59.7	76.6	1	5	2	5	3.25
Icelandic										
XLM-R	—	83.9	92.0	54.6	75.1	6	6	6	1	4.75
CC100	28.1	88.1	93.6	59.1	74.8	1	4	1	1	1.75
MaCoCu	23.1	88.2	93.8	58.5	73.9	1	1	1	1	1.00
mC4	11.5	87.8	93.8	55.8	75.1	1	1	4	1	1.75
OSCAR	53.1	87.6	93.5	59.4	74.1	4	4	1	1	2.50
Combined	5.6	88.1	93.7	58.2	74.6	1	1	1	6	2.25
Slovenian										
XLM-R	—	88.8	94.0	54.7	77.0	6	6	4	1	4.25
CC100	8.8	90.7	95.6	56.6	76.0	1	1	1	1	1.00
MaCoCu	6.6	90.1	95.3	53.9	76.0	4	4	4	1	3.25
mC4	3.3	90.4	95.6	54.5	77.3	1	1	4	1	1.75
OSCAR	96.3	89.8	95.4	56.8	76.5	4	4	1	1	2.50
Combined	1.8	90.7	95.6	56.7	75.6	1	1	1	6	2.25

Table 6: Evaluation results for our 5 languages for XPOS, NER, COPA (CP) and CB. For Albanian, only UPOS data was available. Reported scores are averaged over 10, 20 and 30 runs for POS/NER, COPA and CB, respectively. Ep. denotes the number of epochs corresponding to 50,000 steps. We consider a position to be different if we find a significant difference between two systems when using the Mann-Whitney test (Mann and Whitney, 1947).

	hr	**sr**	**sq**	**is**	**sl**	**Avg.**
XLM-R	4.25	4.75	3.75	4.75	4.25	4.35
CC100	1.0	1.0	1.5	1.75	1.0	1.25
MaCoCu	2.0	3.0	2.75	1.0	3.25	2.40
mC4	3.5	3.5	2.5	2.75	1.75	2.80
OSCAR	4.0	3.5	2.5	2.75	2.5	3.05
Comb	1.0	1.5	3.25	1.75	2.25	1.95

Table 7: Results for each corpus when averaging the position for each language (i.e., over 4 tasks, see Table 6), and finally (last column) averaging over all the languages. Each model was trained for **50,000 steps**.

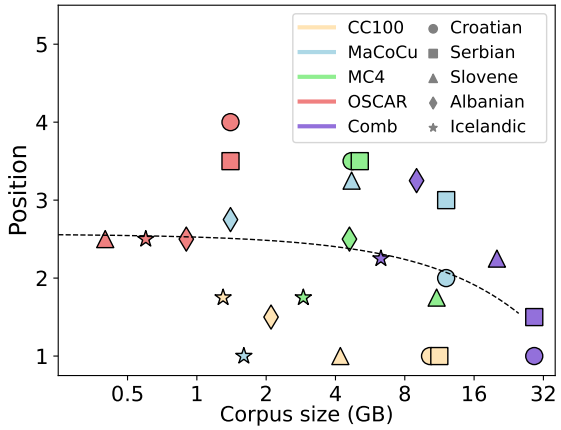

Figure 2: Average position across the four evaluation tasks plotted over the data set size (GB), for each language-corpus combination. Dotted line is the linear regression line. Note the log scale of the X-axis.

irrespective of data quality. Therefore, we plot the average position (where lower position means better performance) over the size of the data set used for each corpus in Figure 2. As expected, models trained on smaller corpora tend to underperform, though there only seems to be a small effect. However, note that there is a clear confounding factor here: the only corpus that is noticeably smaller for all languages is OSCAR, making it unclear whether the relatively bad performance is due to the corpus or the amount of unique data.

Controlling for size Therefore, we recreate the results in Table 6 and 7, but now for the results after just 10,000 steps instead of 50,000. At this stage, the models did not see much data yet, so the effect of size should not play a (large) role here, as was also previously shown in Muennighoff et al. (2023). For space reasons, we only show the averaged results in Table 8. The CC100 and combined corpus still have the best performance, while the other three corpora are similar. Surprisingly, even in the data-controlled setting, the perceived quality of the data by humans does not seem to influence the downstream performance of LMs. Nevertheless, the performance of OSCAR is now similar to that of mC4 and MaCoCu, suggesting that it indeed was disadvantaged by its size in the previous experiments.

XLM-R pretraining corpus The CC100 corpus was in fact already used for pretraining XLM-R. This could potentially be a disadvantage for this corpus, as the data will never be completely new when continuing to train XLM-R. However, given the fact that our languages only make up a very small part of the full corpus[11], we did not expect that this would make a large difference. In fact, given the excellent performance of CC100, we can be reasonably sure that this did not negatively affect the results of the CC100 corpus. On the contrary: it is looking more like it *positively* affected the results of CC100, especially since it also outperformed the combined corpus. We leave investigating why and how this could be the case for future work.

	hr	sr	sq	is	sl	Avg.
XLM-R	3.5	3.5	3.5	4.75	4.0	3.85
CC100	1.0	1.0	1.5	2.5	1.0	1.40
MaCoCu	1.5	2.5	2.0	2.75	4.25	2.60
mC4	2.25	2.5	2.25	3.0	2.5	2.50
OSCAR	3.75	2.5	2.0	3.0	2.25	2.70
Comb	1.0	1.0	2.25	1.5	2.25	1.60

Table 8: Results for each corpus when averaging the position for each language (i.e., over 4 tasks, see Table 6), and finally (last column) averaging over all the languages. Each model was trained for **10,000 steps**.

6. Conclusion

Even though large, curated, web-crawled corpora form the lion's share of the training data of all popular language models, the quality of these corpora has been given relatively little attention. In this paper, we compared four of the currently most relevant large, web-crawled corpora (CC100, MaCoCu, mC4 and OSCAR) across eleven lower-resourced European languages. We first performed a human evaluation by hiring professional linguists to annotate the (rudimentary) quality of the corpora. We found clear differences between the corpora: MaCoCu and OSCAR were of higher quality than CC100 and mC4. We then performed an automatic evaluation of the corpora on a subset of five languages by training language models on each language-corpus combination and evaluation performance on downstream applications. Surprisingly, we found that CC100 is the corpus that has the best performance, even if we control for data set size. We therefore conclude that data set quality (as judged by humans) of web-crawled corpora does not seem to play a significant role in training language models.

[11] Croatian is the language with the most data, occurring in the 30th position of the 100 languages involved.

7. Limitations

Clearly, our work has a number of limitations. For one, our annotation scheme is quite rudimentary: annotators mainly have to determine whether a given paragraph consists of running text or not. However, we believe this is the most vital characteristic of the text when training language models. Moreover, the use of a more complex annotation scheme has major challenges in making sure the results are comparable across different annotators and languages. Secondly, we train encoder-only language models and not generative ones such as GPT4 or LLAMA. It is conceivable that data set quality plays a larger role when models actually have to generate text. We plan to investigate this in future work. Thirdly, we only evaluate on European languages. This was driven by the fact that we wanted to compare multiple corpora, with MaCoCu only working with European languages. Since we do have quite some variety within our languages, we are confident that our results still generalize. Lastly, our models are evaluated on just four tasks, only two of which concern natural language understanding. The main limitation here is simply availability: there are not many evaluation tasks available across a large number of languages. Since we wanted to compare corpora across languages, we opted to only include tasks that had data available for all languages used in our study.

8. Acknowledgements

The MaCoCu project has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341. This communication reflects only the authors' views. The Agency is not responsible for any use that may be made of the information it contains. This work was also funded by the Slovenian Research Agency within the national projects J7-4642 and L2-50070, as well the research programme P6-0411. This research was supported with Cloud TPUs from Google's TPU Research Cloud (TRC). We thank the Center for Information Technology of the University of Groningen for providing access to the Hábrók high performance computing cluster. We are grateful to all our MaCoCu colleagues for the fruitful discussions. We thank Wietse de Vries for his feedback on the annotation scheme and for agreeing to do a pilot annotation run. Finally, we thank the reviewers for their comments.

9. Bibliographical References

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.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

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.

Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. 2020. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1324–1334, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4555–4567, Online. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Svanhvít Lilja Ingólfsdóttir, Ásmundur Alma Gudhjónsson, and Hrafn Loftsson. 2020. MIM-GOLD-NER – named entity recognition corpus (21.09). CLARIN-IS.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. Association for Computational Linguistics.

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.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models. *Machine Learning and Knowledge Extraction*, 5(3):1149–1175.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Nikola Ljubešić and Davor Lauc. 2021. BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, Kiyv, Ukraine. Association for Computational Linguistics.

Alexandra Luccioni and Joseph Viviano. 2021. What's in the box? An analysis of undesirable

content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online. Association for Computational Linguistics.

H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. 2023. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 50358–50376. Curran Associates, Inc.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Jonathan Oliver and Josiah Hagen. 2021. Designing the elements of a fuzzy hashing scheme. In *2021 IEEE 19th International Conference on Embedded and Ubiquitous Computing (EUC)*, pages 1–6.

OpenAI. 2023. Gpt-4 technical report. Technical report, OpenAI.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673.

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*.

T. C. Rajapakse. 2019. Simple transformers. https://github.com/ThilinaRajapakse/simpletransformers.

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.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Stefan Schweter. 2020. BERTurk - BERT models for Turkish.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Greenfeld, and Reut Tsarfaty. 2022. AlephBERT: Language model pre-training and evaluation from sub-word to sentence level. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–56, Dublin, Ireland. Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Vilhjalmur

Thorsteinsson, and Hafsteinn Einarsson. 2022. A warm start and a clean crawled corpus - a recipe for good language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4356–4366, Marseille, France. European Language Resources Association.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLAMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. 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.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

10. Language Resource References

Ljubešić, Nikola. 2021. *Choice of plausible alternatives dataset in Croatian COPA-HR*. Slovenian language resource repository CLARIN.SI.

Ljubešić, Nikola and Starović, Mirjana and Kuzman, Taja and Samardžić, Tanja. 2022. *Choice of plausible alternatives dataset in Serbian COPA-SR*. Slovenian language resource repository CLARIN.SI.

Žagar, Aleš and Robnik-Šikonja, Marko and Goli, Teja and Arhar Holdt, Špela. 2020. *Slovene translation of SuperGLUE*. Slovenian language resource repository CLARIN.SI.

A. Appendix

A.1. Annotation details

The specific annotation examples that were used to instruct all annotators are shown in Table 11. Since all annotators were also fluent English speakers, we opted to show the annotation examples as English texts. The main advantage of this approach is that the annotation instructions were the exact same across languages.

A.2. Evaluation details

The specific train, development and test set sizes per language and task are reported in Table 9. Specific hyper-parameter settings (that differ from the default) are shown in Table 10. For NER and POS, we use the NERmodel implementation of the Simpletransformers package (Rajapakse, 2019). For

COPA, we use the ModelForMultipleChoice from the Transformers (Wolf et al., 2020) library. For CB, we use the SequenceClassification model from the same library.

COPA For the COPA task, the training is not always stable. There are often runs where the training loss simply does not go down. These are considered to be failed runs. Failed runs are simply discarded, i.e., when averaging over 20 runs we do not take the failed runs into account.

	Train	Dev	Test
POS			
Croatian	6,914	960	1,136
Serbian	3,328	536	520
Albanian	5,307	611	708
Icelandic	8,896	4,865	5,157
Slovenian	10,903	1,250	1,282
NER			
Croatian	19,792	2,487	2,487
Serbian	3,329	537	521
Albanian	5,000	1,000	1,000
Icelandic	10,651	6,479	5,889
Slovenian	7,625	921	942
COPA	400	100	500
CB	250	56	250

Table 9: Train, development and test set sizes for the four tasks used during our automatic evaluation. For COPA and CB, the sizes are the same across languages.

	POS	NER	COPA	CB
Learning rate	1e-05	1e-05	1e-05	3e-05
Batch size	8	8	8	4
Epochs	8	8	15	10
Max length	512	512	100	512
Runs	10	10	20	30

Table 10: Hyper-parameter settings used across our evaluation experiments. Settings not mentioned are left at default.

Wrong language / Not language
1: STRAND 240 242 ECO:0000244|PDB:1FMK.
2: box-shadow: 1px 1px 3px 2px #121e03;
3: Ísland er fallegt land til að heimsækja.
Not running text
1: Archives Select Month December 2022 (2) November 2022 (11) October 2022 (14) September 2022 [...]
2: #fish #koi #carpe #carpekoi #poisson #japon
3: September 23, 2018
4: Sheraton Signature Sleep Experience® | Sheraton Tirana Hotel | Official Website
Partially running text
1: bomb blew up in her face on Christmas Eve. Police refused to speculate about the
2: in approximately 13,000 women every year in the United States, and kills almost 5,000 American
3: Complete Your Bachelors Degree or Associate Degree | Charter . . .
4: The premium is the amount you'll pay for the huge benefits protected
Running text, but (slightly) non-standard
1: What The Riga Elections Say About Latvian Politics – Analysis – Eurasia Review
2: Demonstrated critical thinking and decision-making competencies
3: Friday.4th: At Noon, the Detachment of Marines fired 3 vollies in honour of the Day.
4: HOW DO I WRITE A BUSINESS REPORT?
5: (c) It is because of God, then, that we have language and words
6: I get a long line of numbers (where you can extract the windspeed from), but the outcome is strange. This is shown in my log file:
Publishable text
1: You don't mean that, said Bones hoarsely.
2: Sounds delicious. My daughter makes it with a puréed jalapeño swirl that looks and tastes amazing.
3: Does your business need an interactive website or app?
4: The Caucasian rugs are made in the regions located in the mountain chain of the Caucasus, an area situated between the Black Sea and the Caspian Sea. The area is spanned across Georgia, Russian, Armenia and Azerbaijan.

Table 11: Example English texts for each annotation category that were given to all annotators. All examples are actual examples taken from the English versions of the respective corpora.