CEF Telecom

Project title: MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages

Evaluation of data release 2

Milestone number: 14

Version 1.0

Co-financed by the Connecting Europe Facility of the European Union

Funded by the European Union's CEF Telecom programme
under the Grant Agreement No. 2020-EU-IA-0078

Project Acronym: MaCoCu
Project Full Title: MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languag
Year of the Call: 2020
Grant Number: 2020-EU-IA-0078
Project URL: https://macocu.eu

Document title:	Evaluation of data release 2
Lead author:	Rik van Noord (Rijksuniversiteit Groningen)
Contributing authors:	Rik van Noord, Malina Chichirau, Antonio Toral (Rijksuniversiteit Groningen) Miquel Esplà-Gomis (Universitat d'Alacant) Gema Ramírez-Sánchez (Prompsit) Peter Rupnik, Taja Kuzman, Nikola Ljubešić (JSI)
Milestone number:	14
Activity associated to the milestone:	Evaluation
Dissemination level:	Public (PU)
Contractual Delivery Date:	July 31, 2023
Actual Delivery Date	
Number of pages:	23

Document history

Version	Date	Changes
1.0	July 31, 2023	Original Submission

Abstract

This report describes the methodology and results of the evaluation process carried out to assess the usefulness of the second data release of the MaCoCu project. This data release contains monolingual data for Albanian, Bosnian, Bulgarian, Croatian, Icelandic, Maltese, Macedonian, Montenegrin Serbian, Slovenian and Turkish, and parallel data for the same languages aligned with English. The corpora have been evaluated in a number of natural language processing (NLP)-related tasks in order to compare their performance to the one of similar publicly available monolingual and parallel corpora. Both, monolingual and bilingual corpora are also evaluated by professional translators that perform direct assessment of MaCoCu and other web-crawled corpora. The MaCoCu corpora compare favorably to the other corpora, in particular during the human evaluation. Moreover, results are indicative of an improvement in the quality of the MaCoCu parallel corpora over the course of the project. This is not always reflected in the automatic metrics used to evaluate NLP tasks, but is clear from human evaluation.

Contents

1. Executive summary

1.1. Introduction

This deliverable, submitted as Milestone 14: *Evaluation of data release 2*, corresponds to the verification means for Activity 7: *Evaluation* for the second year of the project. The main goal of Activity 7 is to evaluate the usefulness of the data produced through Activities 5: *Curation of bilingual data* and 6: *Curation of monolingual data*. Through them, this action produced the corpora that were released as part of the second data release in April 2023.

The report is divided into two main blocks: evaluation of monolingual data and evaluation of parallel data. The data is evaluated by comparing MaCoCu corpora to other well-known, large, publicly available corpora. We run automatic and human evaluation for both monolingual and bilingual corpora.

For automatic monolingual evaluation, we continue training pre-trained language models (LMs) based on XLM-R [1] on the different monolingual corpora for 4 languages, and compare the performance on a range of downstream tasks. For automatic parallel evaluation, state-of-the-art neural machine translation models are trained on different corpora, including both the MaCoCu corpora released. These models are automatically evaluated on available test sets from the FLoRes data set [2], using automatic evaluation metrics such as COMET [3] and BLEU [4].

Regarding human evaluation, for both monolingual and parallel data and for all 11 languages, we have professional annotators directly assess the quality of the texts. Specifically, we compare MaCoCu data to CC100 [1], MC4 [5] and OSCAR [6] for monolingual data. For parallel, we compare the second release to the first release of the MaCoCu data to see the effect of our improved processing pipeline and we also compare to CCAligned [7], CCMatrix [8] and ParaCrawl [9] corpora.

1.2. Brief summary of the MaCoCu action

The objective of MaCoCu is to gather, clean and enrich monolingual and parallel corpora for several European languages with scarce resources. This will be achieved by crawling, cleaning and adding extra info to data in multiple languages.

Four European partners from academia and industry, all highly specialised, take part in this action. All of them play a crucial role in the development of the objectives of the action.

The four partners are:

- University of Alacant (UA): responsible for code and project management.
- Jožef Stefan Institute (JSI): responsible for data crawling and monolingual curation and enrichment.
- Rijksuniversiteit Groningen (RUG): responsible for DSI-specific content enrichment and evaluation.
- Prompsit Language Engineering, SL (Prompsit): responsible for bilingual data curation and dissemination and outreach.

Partners RUG, JSI and Prompsit were involved in Activity 7, on which this report is focused.

This report covers the second data release of the project, which includes: Albanian, Bosnian, Bulgarian, Croatian, Icelandic, Maltese, Macedonian, Montenegrin, Serbian, Slovenian and Turkish.

2. Evaluation of monolingual corpora

2.1. Automatic evaluation of monolingual corpora

This section focuses on the automatic evaluation of the monolingual corpora that were part of the second MaCoCu release. We evaluate the corpora extrinsically by training general purpose language models (LMs). Pre-trained LMs are state-of-the-art in virtually all NLP tasks, but it is often unclear which corpora are better suited for training such models. We conduct a thorough analysis on 4 popular monolingual corpora (CC100, MaCoCu, MC4 and Oscar), across 4 languages (Albanian, Serbo-Croatian, Slovene and Icelandic). We choose the first three languages since they were not part of the first release, and therefore did not already have pre-trained LMs available. We pick Icelandic since it is a low-resource language on which we spend extra time during the data collection phase.

We continue training XLM-R [1] for each corpus and language. We also train a system per language that concatenates the 4 available corpora. All code and models are made publicly available.[1] Data sizes, as well as number of epochs per corpus and language, are shown in Table 2.1 for the MaCoCu data and the other models that we trained. We transliterate all Cyrillic Serbian and Croatian data to Latin.[2] Note that the MC4 corpus does not treat Croatian as a separate language, while OSCAR only has very little Croatian data.

	CC100		MaCoCu		MC4		OSCAR		Comb	
Language	**GB**	**Ep.**	**GB**	**Ep.**	**GB**	**Ep.**	**GB**	**Ep.**	**GB**	**Ep.**
Albanian	2.1	17.2	1.4	26.2	4.6	7.8	0.9	37.2	9.0	4.2
Icelandic	1.3	28.1	1.6	23.1	2.9	11.5	0.6	53.1	6.3	5.6
Serbo-Croatian	10.8	3.6	12.1	3.2	4.9	7.4	1.4	23.9	29.2	1.3
Slovene	4.2	8.8	4.7	6.6	11.0	3.3	0.4	96.3	20.1	1.8
Croatian	8.6	—	5.9	—	—	—	0.003	—	—	—
Serbian	2.2	—	6.2	—	4.9	—	1.4	—	—	—

Table 2.1: Data set sizes (in GB, of compressed text) and number of epochs during 50k steps for the included corpora in LM experiments. Serbian and Croatian individual figures are included for reference although a single model combining both is trained for practical matters.

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. Where possible, 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, [10]) and Commitment Bank (CB, [11]). For XPOS and NER we use data from the Universal Dependencies project[3], with the exception of Icelandic NER, for which we use the MIM-GOLD-NER set [12]. The COPA data set was originally created for just English [10], but has gold standard translations available for Serbian, Croatian and Slovene. For Albanian and Icelandic, we translated the English data with Google Translate. For the CB task, we used Google Translate for all languages.

2.1.1. Training

One of the conclusions of the first evaluation report produced by this action was that it is more cost-efficient to *continue* training existing LMs than to train them from scratch. At the same time, it was clear that the performance did not improve much after 50,000 steps. Therefore, for the evaluation described in this report, we only continue training XLM-R and stop training after 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. The only exception is the combined model, i.e. a model combining all languages, which we train until 100,000 steps, as this is

[1] https://github.com/macocu/LanguageModels/
[2] The Croatian corpora had less than 0.1% of Cyrillic data.
[3] https://universaldependencies.org/

the most likely to be useful to the community. A single experiment (50,000 steps) took around 4 days on a single Google Cloud TPU.

2.1.2. Results

First, we evaluate performance across different time steps of training the model. For each model, language and task combination (e.g. Icelandic trained on OSCAR for NER) we calculate performance after 10,000, 20,000, 30,000, 40,000 and 50,000 steps. We average each of these evaluations over 5 different random seeds for XPOS and NER, and over 30 random seeds for COPA and CB, as the latter data sets are quite small. For each language, we show the average performance over the different amount of steps. In other words, for example, Icelandic trained on OSCAR for NER has a score of 87.4, which is the average of 25 different runs (5 for 10,000 steps, 5 for 20,000 steps, etc). Similarly, COPA and CB scores are averages of 150 runs. Even though fine-tuning has a lot of variance (especially for CB and COPA), we believe that this gives us a fair overview of the performance of the models. For each language, we also report the *position* in the ranking of each corpus. This gives a more clear overview of the performance of the corpus despite the fact that such a ranking does not show the relative differences in the scores.

All results are shown in Table 2.2. Since XPOS is a relatively easy task, the differences between the corpora here are, as expected, quite small. For Albanian, all corpora even obtain the same score, which is just 0.1 higher than XLM-R-base. 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). Nevertheless, to obtain a more clear overview of performance per corpus, we show the averaged relative rankings in Table 2.3. We can see a bit of trend here: the best performing corpora seem to be the combined corpus (which was to be expected), but also the CC100 corpus. The OSCAR corpus seems to be the worst corpus, which likely can be attributed to the fact that it was generally the smallest (see Table 2.1). Since data set size plays a large role in LM training, this does not tell us much about the general quality of each corpus. Therefore, we perform a manual evaluation of the corpora in the next section.

Corpus	Epochs	Scores				Positions				
		XPOS	NER	COPA	CB	XPOS	NER	COPA	CB	Avg.
Croatian										
XLM-R-base	—	93.4	89.4	55.8	77.8	6	5	3	5	4.75
CC100	3.6	94.3	90.8	59.5	79.4	1	1	2	3	1.75
MaCoCu	3.2	94.2	90.6	55.5	77.8	2	3	4	5	3.50
MC4	7.4	94.1	89.5	52.6	80.0	4	4	6	1	3.75
OSCAR	24.9	94.0	89.1	55.5	79.5	5	6	4	2	4.25
Combined	1.3	94.2	90.7	60.0	78.9	2	2	1	4	2.25
Serbian										
XLM-R-base	—	91.0	93.5	55.6	76.6	6	5	5	6	5.50
CC100	3.6	92.3	94.3	60.6	78.6	2	1	1	2	1.50
MaCoCu	3.2	92.3	94.3	56.7	76.9	2	1	3	5	2.75
MC4	7.4	92.3	93.8	52.9	79.6	2	4	6	1	3.25
OSCAR	24.9	92.3	93.4	56.6	77.9	2	6	4	3	3.75
Combined	1.3	92.4	94.3	60.0	77.7	1	1	2	4	2.00
Albanian										
XLM-R-base	—	93.9	92.7	54.4	77.5	6	6	6	5	5.75
CC100	17.2	94.0	93.0	59.0	78.8	1	2	2	2	1.75
MaCoCu	26.2	94.0	92.9	55.9	76.7	1	4	3	6	3.50
MC4	7.8	94.0	93.0	55.9	79.3	1	2	3	1	1.75
OSCAR	37.2	94.0	92.8	54.5	78.2	1	5	5	3	3.50
Combined	4.2	94.0	93.1	59.1	77.3	1	1	1	4	1.75
Icelandic										
XLM-R-base	—	92.0	83.9	55.2	75.1	6	6	6	1	4.75
CC100	28.1	93.5	87.8	58.7	73.9	3	2	2	5	3.00
MaCoCu	23.1	93.4	87.9	57.7	73.9	4	1	4	5	3.50
MC4	11.5	93.6	87.5	56.4	74.9	1	4	5	2	3.00
OSCAR	53.1	93.4	87.4	58.2	74.1	4	5	3	4	4.00
Combined	5.6	93.6	87.6	60.2	74.8	1	3	1	3	2.00
Slovene										
XLM-R-base	—	94.0	88.8	54.1	77.0	6	6	4	2	4.50
CC100	8.8	95.5	90.4	58.5	76.1	1	1	1	4	1.75
MaCoCu	6.6	95.2	90.0	53.4	75.9	5	3	6	5	4.75
MC4	3.3	95.4	89.8	54.1	77.7	3	4	4	1	3.00
OSCAR	96.3	95.3	89.8	54.2	76.8	4	4	3	3	3.50
Combined	1.8	95.5	90.4	57.2	75.6	1	1	2	6	2.50

Table 2.2: Evaluation results for our 5 languages for XPOS, NER, COPA and CB. For Albanian, only UPOS data was available. Reported scores are averaged over 5 checkpoints (10k, 20k, 30k, 40, 50k) and 5 runs per checkpoint for POS/NER, while we use 30 runs for COPA and CB. Epochs denotes the number of epochs corresponding to 50,000 steps. We consider a position different if the average score differs by 0.1 or more.

	hr	sr	sq	is	sl	Avg.
CC100	1.75	1.50	1.75	3.00	1.75	1.95
MaCoCu	3.50	2.75	3.50	3.50	4.75	3.60
MC4	3.75	3.25	1.75	3.00	3.00	2.95
OSCAR	4.25	3.75	3.50	4.00	3.50	3.80
Combined	2.25	2.00	1.75	2.00	2.50	2.10

Table 2.3: Results for each corpus when averaging the position for each language (i.e. over 4 tasks, see Table 2.2), and finally (last column) averaging over all the languages.

2.2. Human evaluation of monolingual corpora

In this section, we describe the methodology and results of the manual evaluation carried out for the MaCoCu monolingual corpora along with other popular ones. We do the comparison on all languages covered in second release (MaCoCu-V2), that is, 11 languages. We compare MaCoCu-V2, CC100, MC4 and OSCAR data sets. 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 and filter out other paragraphs. The other corpora are used as they are released. For each corpus and language, we randomly select 200 paragraphs for annotation. Annotators are asked to rank each paragraph using the following scale:

1. **Wrong language or not language (WNL).** The text is not in the correct language, or is not in a natural language (e.g. links, html tags).
2. **Not running text (NRT).** 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 (PRT).** 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 (RTE).** 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 scale is inspired by those in two recent works [13, 14], which were taken as a starting point, combined, and refined by means of a pilot annotation conducted among project members and other colleagues on Slovene and English. Each annotator is also shown a number of example paragraphs in English, for each category. These are compiled in Appendix A. They are instructed to always pick only one option, and to pick the lower number in the scale, when in doubt. Annotators work using online annotation tool KEOPS,[4] which is adapted to the proposed scale as shown in Figure 2.1.

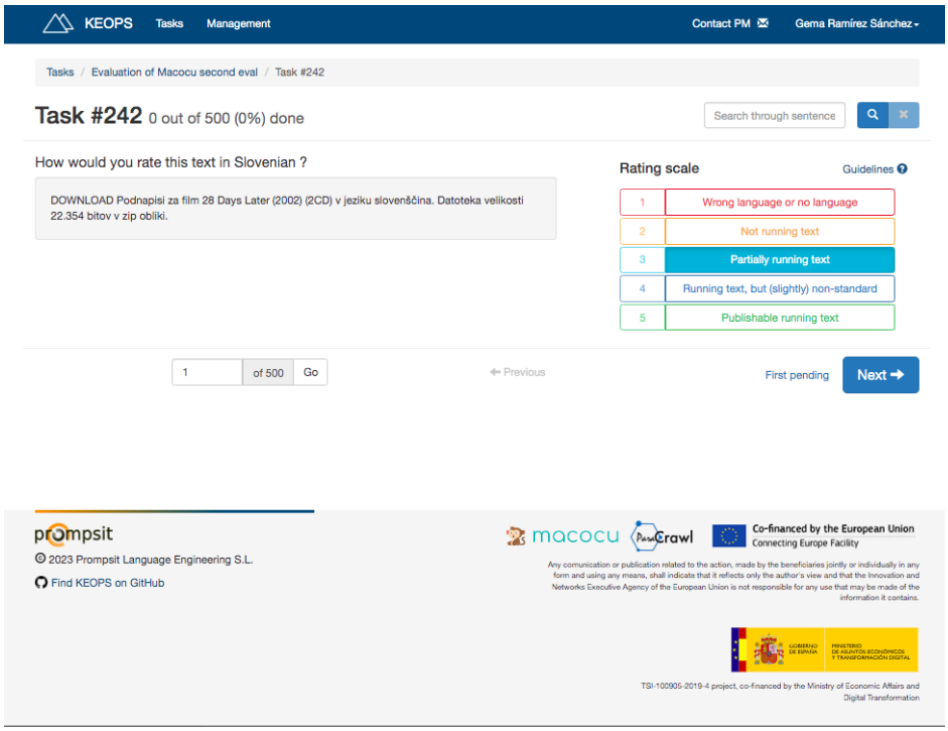

Figure 2.1: Monolingual annotation screen in KEOPS.

[4] https://keops.prompsit.com/

We hire two professional linguists per language to annotate subsets of all corpora in each of the 11 languages. We select 200 instances that overlap between the 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 a 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 inter-annotator agreement in terms of exact annotation overlap and Cohen's kappa coefficient (κ) scores are shown in Table 2.4. 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.

	Monolingual data	
	% exact agreement	**κ coefficient**
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
Slovene	70.5	0.36
Turkish	52.5	0.32

Table 2.4: 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.

The full results of the annotation are shown in Table 2.5. We distinguish between "Running Text, but slightly non-standard" and "Publishable Text" in our annotation scheme but, for the purpose of training LMs, we consider these categories both appropriate. 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. In fact, in the previous section we only trained a single language model including all of them.

Generally speaking, we observe that MaCoCu-V2 and OSCAR contain paragraphs that are most often at least running text. In fact, MaCoCu-V2 has the highest score for 5 out of 9 languages, while OSCAR has the highest quality corpus for the other 4 languages for which a comparison can be made. 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.

For Montenegrin and Bosnian there are no other corpora to compare to, but both corpora seem to be of high quality. In particular Bosnian only has five instances that would be considered problematic by our standards. Montenegrin has 25 paragraphs that were annotated as "Wrong language", but a manual analysis revealed that these were mostly in Serbian, Bosnian or Croatian, which, as stated before, is not necessarily an issue for training LMs.

To get a more clear 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, Slovene 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 2.6.

Albanian	**WLN**	**NRT**	**PRT**	**RTE**	**PT**	**RTE+PT**	**Bulgarian**	**WLN**	**NRT**	**PRT**	**RTE**	**PT**	**RTE+PT**
MaCoCu	4	4	48	73	71	144	MaCoCu	5	8	18	78	91	**169**
CC100	1	3	44	62	90	152	CC100	2	23	18	66	91	157
MC4	18	12	58	47	65	**112**	MC4	17	49	25	47	62	**109**
OSCAR	1	2	32	70	95	**165**	OSCAR	2	26	26	58	88	146
Bosnian	**WLN**	**NRT**	**PRT**	**RTE**	**PT**	**RTE+PT**	**Montenegrin**	**WLN**	**NRT**	**PRT**	**RTE**	**PT**	**RTE+PT**
MaCoCu	2	0	3	38	157	**195**	MaCoCu	25	4	18	49	104	**153**
Croatian	**WLN**	**NRT**	**PRT**	**RTE**	**PT**	**RTE+PT**	**Maltese**	**WLN**	**NRT**	**PRT**	**RTE**	**PT**	**RTE+PT**
MaCoCu	10	23	23	61	83	**144**	MaCoCu	9	101	38	16	36	52
CC100	18	15	25	71	71	142	MC4	164	17	4	1	14	**15**
OSCAR	1	37	32	65	64	**129**	OSCAR	7	32	30	17	98	**115**
Icelandic	**WLN**	**NRT**	**PRT**	**RTE**	**PT**	**RTE+PT**	**Macedonian**	**WLN**	**NRT**	**PRT**	**RTE**	**PT**	**RTE+PT**
MaCoCu	2	4	6	15	173	188	MaCoCu	5	5	26	76	88	**164**
CC100	2	6	9	19	164	183	CC100	1	11	41	76	71	147
MC4	24	15	16	15	130	**145**	MC4	10	20	30	59	81	**140**
OSCAR	2	1	4	4	189	**193**	OSCAR	2	7	31	64	96	160
Serbian	**WLN**	**NRT**	**PRT**	**RTE**	**PT**	**RTE+PT**	**Turkish**	**WLN**	**NRT**	**PRT**	**RTE**	**PT**	**RTE+PT**
MaCoCu	1	1	14	82	102	**184**	MaCoCu	5	27	19	90	59	**149**
CC100	0	5	24	60	111	171	CC100	0	33	30	95	42	137
MC4	5	14	47	65	69	**134**	MC4	8	62	30	67	33	**100**
OSCAR	0	1	24	69	106	175	OSCAR	3	38	26	84	49	133
Slovene	**WLN**	**NRT**	**PRT**	**RTE**	**PT**	**RTE+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**							

Table 2.5: Human evaluation of monolingual corpora. The highest value for column RT+PT per language is shown in green, and the lowest value in red.

	Langs	**WLN**	**NRT**	**PRT**	**RTE**	**PT**	**RTE+PT**
MaCoCu-V2	7	1.8%	3.6%	10.4%	29.9%	54.3%	**84.2%**
CC100	7	0.4%	6.9%	13.3%	26.2%	53.2%	79.4%
MC4	7	6.4%	13.9%	16.0%	22.4%	41.3%	**63.7%**
OSCAR	7	0.6%	5.4%	9.9%	24.7%	59.4%	84.1%

Table 2.6: Percentage of annotations for each of the annotation categories, averaged over corpus across the seven languages included in all evaluated corpora.

In this scenario, MaCoCu-V2 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. Interestingly enough, this was not reflected in the automatic monolingual evaluation, in which MC4 outperformed both MaCoCu and OSCAR (see Table 2.3). 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.

3. Evaluation of bilingual corpora

In this section, we describe the evaluation of the quality of our bilingual corpora. We aim to answer three main research questions:

- **RQ1**: How does the parallel corpora of our second data release compare to that of our first release?
- **RQ2**: How do our parallel corpora compare to other popular web-crawled parallel corpora?
- **RQ3**: Does adding the MaCoCu data improve on the current best open-source MT systems available?

To answer these three questions, we perform both automatic and manual evaluation. In the automatic evaluation, we train neural machine translation (NMT) systems on the different parallel corpora and compare their performance. In the manual evaluation, we ask professional translators to directly annotate the quality and issues of the automatically crawled and aligned sentence pairs in each corpus. These two evaluation methods are described below.

3.1. Automatic evaluation of bilingual corpora

We build NMT systems from English into the 11 languages targeted in MaCoCu's second release. We compare the performance of the MaCoCu parallel corpora to three other well-known, large publicly available parallel corpora: ParaCrawl [9], CCAlign [7] and CCMatrix [8]. We consider two experimental settings: we either train models from scratch or we continue training strong existing NMT systems.

When training from scratch, we use a Transformer [15] model implemented in Marian [16], with settings used in the previous evaluation of parallel MaCoCu corpora. We train a Transformer-base model with 6 layers for the encoder and decoder and 8 attention heads, with a hidden size of 2,048. For each language, we train a vocabulary of 32,000 pieces through byte-pair encoding [17, 18]. We truncate the input to a maximum of 200 of such pieces. During training, we automatically use a batch size that fits into our memory (32GB on a GPU). We use a learning rate of 0.0003, with a warm-up of 16,000 steps. During training, we apply label smoothing with a value of 0.1. Training is either stopped using early stopping, calculated with BLEU after each epoch (with a patience of three), or after 21 epochs. We use the same settings in all our experiments.

When continuing training systems, we choose language-specific OPUS models that are publicly available on HuggingFace and we use this platform's API to conduct fine-tuning. For most of the Slavic languages (Croatian, Slovenian, Bosnian, Serbian and Montenegrin) we used the same multilingual model. Since this model was pre-trained using particular language-identifiers added to the source, the Bosnian texts originally written in Cyrillic were transliterated into Latin script, and the Montenegrin texts were either assigned Serbian Latin or Serbian Cyrillic tokens. For Icelandic, Albanian, Bulgarian, Macedonian, Maltese we used language-specific models, and for Turkish we used a multilingual model for Turkic languages. For training, we used a batch size of 16, with 2 gradient accumulation steps, and an initial learning rate of 1e-5. Due to memory limitations, we used the Adafactor optimizer and gradient checkpointing to reduce memory requirements. Finally, we fine-tuned our models for 10 epochs maximum, with an early stopping strategy after 2 epochs. We used the same settings for all experiments.

To evaluate performance, we use the following well-established MT metrics: COMET [3], BLEU [4], CHRF [19], BERTscore [20] and BLEURT [21]. For brevity, we only show performance of the traditionally most important metric in MT (BLEU), as well as the current best performing metric (COMET). We evaluate performance on the FLoRes dev and devtest data sets [2], showing only the devtest scores for brevity. For Montenegrin we evaluate on a random subset of 5,000 sentences from the Opus-Subs data set [22], since this language was not present in FLoRes. Each corpus we use is pre-processed by the script used for training models in the Tatoeba Translation challenge [23], after which duplicate and near-duplicate sentence pairs are removed. The sizes in terms of sentence-pairs of the corpora we use throughout our experiments are shown in Table 3.1. If we combine corpora, we redo strict and near-deduplication.

We show the results of all our experiments in Table 3.2. The first 5 rows of results are mainly for reference. It shows that the MaCoCu corpora offer competitive performance to the other web-crawled corpora, despite often being of smaller size. The English-Icelandic MaCoCu data was too small for training a stable NMT system from scratch. Here, we can already compare performance of the first and second release of MaCoCu data (**RQ1**), for the 6 language pairs that them both have available. Going by COMET scores, we improve on 4 out of the 6 language pairs, despite actually having less data for

Language pair	MaCoCu-V1	MaCoCu-V2	CCAligned	CCMatrix	ParaCrawl
English-Albanian	—	493,528	1,328,419	9,513,370	—
English-Bosnian	—	464,300	154,119	—	—
English-Bulgarian	2,155,589	1,676,464	5,828,652	24,213,749	10,235,100
English-Croatian	1,920,118	2,147,340	4,793,411	8,326,907	2,596,513
English-Icelandic	291,307	257,535	848,022	3,398,623	2,103,053
English-Macedonian	401,819	358,862	1,168,161	4,262,556	—
English-Maltese	979,424	867,107	—	—	988,548
English-Montenegrin	—	203,853	—	—	—
English-Serbian	—	1,663,593	1,589,959	13,427,449	
English-Slovenian	2,157,771	1,788,210	2,646,945	14,573,801	7,046,791
English-Turkish	3,801,207	1,533,376	8,541,001	31,015,593	—

Table 3.1: Data set sizes (sentence pairs) per corpus per language pair. Not each language is present in each corpus. Sizes are after normalization and near-deduplication.

three of these pairs. Especially Turkish is interesting: we go from 3.8M to 1.5M sentence pairs, but actually obtain a 2.0 increase in COMET score.

However, instead of training only on MaCoCu data, a more realistic use case is adding the MaCoCu data to existing corpora or models. We distinguish two settings here. In the first, we continue training strong baseline models on both the MaCoCu data sets. This is shown in the "Cont. training" subset of Table 3.2. We find that we clearly improve in COMET score for all languages when we add either the first or second release of the MaCoCu data **RQ3**, except for Macedonian, in which we obtain the same score when adding MaCoCu-V2. Here, we also compare MaCoCu V1 and V2 (**RQ1**) and find that MaCoCu-V2 receives a higher COMET score in 4 out of 7 cases, with 1 language obtaining equal scores. Only Macedonian and Maltese obtain higher COMET scores for the first release.

In the second, we train NMT models from scratch, and compare adding the MaCoCu data to the CCAligned, CCMatrix and ParaCrawl data, individually. For some of the smaller language pairs, we also compare a setting in which we train on all available corpora. These results are shown in the "Adding MaCoCu" part of Table 3.2. We find that we generally improve in performance for adding either of the MaCoCu corpora (**RQ3**), though the differences can be small for the languages with more resources available. The difference between MaCoCu-V1 and MaCoCu-V2 is not as pronounced here. We find that MaCoCu-V2 obtains a higher performance than V1 in only 10 of the 20 comparisons. In other words, we find no difference in usefulness between the two releases here (**RQ1**).

However, we are also interested in comparing the quality of the data sets (**RQ2**). The comparison in row 1-5 of Table 3.2 is heavily influenced by data set size. Therefore, we perform an experiment in which we limit the size of each corpus to be the same as the second MaCoCu release. This is shown in the last subset of results of Table 3.2. We find that MaCoCu-V2 does very well in this setting: we outperform MaCoCu-V1 for 4 out of 6 language pairs, ParaCrawl for 3 out of 4 pairs, CCAligned for 5 out of 6 pairs and CCMatrix for 4 out of 6 pairs. Generally, this seems to indicate that MaCoCu-V2 has higher quality sentences for training NMT systems, but reaches this quality by excluding sentences (from MaCoCu-V1) that are still useful, though of lower general quality.

	en-bg		en-bs		en-hr		en-is		en-mk		en-mt		en-sl		en-sq		en-sr		en-tr	
	BL	CO	BL	CO	BL	CO	BL	CO	BL	CO	BL	CO	BL	CO	BL	CO	BL	CO	BL	CO
From scratch																				
MaCoCu-V1	34.7	86.2	—	—	26.3	85.9	—	—	23.6	77.3	36.0	71.5	25.4	83.9	—	—	—	—	25.0	83.2
MaCoCu-V2	35.3	86.5	21.4	76.9	28.1	87.3	—	—	21.3	73.0	35.0	71.2	25.8	84.3	25.3	82.3	30.1	85.2	24.4	85.2
δ MCC2 - MCC1	**+0.6**	**+0.3**	—	—	**+2.2**	**+1.4**	—	—	**-2.2**	**-4.3**	**-1.0**	**-0.3**	**+0.4**	**+0.4**	—	—	—	—	**-0.6**	**+2.0**
CCAligned	37.5	86.2	2.3	38.0	26.4	84.2	18.1	71.7	23.5	76.8	—	—	24.3	81.1	23.9	80.4	28.2	81.6	27.1	84.3
CCMatrix	42.8	90.0	—	—	28.2	86.8	24.3	79.7	33.9	87.3	—	—	29.9	86.9	31.2	87.4	31.5	86.4	31.6	88.6
ParaCrawl	36.4	86.7	—	—	31.2	89.0	—	—	—	—	34.2	70.6	25.0	82.4	—	—	—	—	—	—
Cont. training																				
Baseline	40.1	88.0	27.2	86.8	25.6	85.7	18.3	76.0	29.5	85.5	29.7	69.8	24.3	83.6	27.5	85.5	27.4	83.1	10.1	71.7
+ MaCoCu-V1	**-1.6**	**+0.3**	—	—	**+2.4**	**+2.1**	**+4.0**	**+4.7**	**+0.6**	**+0.2**	**+9.3**	**+2.6**	**+2.0**	**+2.1**	—	—	—	—	**+7.4**	**+8.1**
+ MaCoCu-V2	**-1.1**	**+0.9**	**+0.6**	**+1.0**	**+3.0**	**+2.3**	**+4.3**	**+4.7**	**+0.9**	**+0.0**	**+8.3**	**+2.3**	**+2.5**	**+2.8**	**+2.0**	**+1.1**	**+3.1**	**+3.4**	**+8.8**	**+9.5**
δ MCC2 - MCC1	**+0.5**	**+0.6**	—	—	**+0.6**	**+0.2**	**+0.3**	**+0.0**	**+0.3**	**-0.2**	**-1.0**	**-0.3**	**+0.5**	**+0.7**	—	—	—	—	**+1.4**	**+1.4**
Adding MaCoCu																				
CCM + MCC1	43.1	90.2	—	—	30.7	89.0	24.7	81.0	34.1	87.4	—	—	29.8	87.4	—	—	—	—	31.4	88.0
CCM + MCC2	43.0	90.3	—	—	30.9	89.2	25.0	81.5	34.2	87.5	—	—	29.6	87.2	31.6	87.8	—	—	32.2	88.7
δ MCC2 - MCC1	**-0.1**	**+0.1**	—	—	**+0.2**	**+0.2**	**+0.3**	**+0.5**	**+0.1**	**+0.1**	—	—	**-0.2**	**-0.2**	—	—	—	—	**+0.8**	**+0.7**
CCA + MCC1	39.2	87.8	—	—	28.8	87.1	21.4	76.2	30.1	83.2	—	—	28.3	84.8	—	—	—	—	27.4	85.6
CCA + MCC2	38.5	87.4	—	—	29.3	87.5	21.5	76.1	30.0	82.8	—	—	27.9	84.5	27.2	84.2	32.1	86.1	28.2	85.7
δ MCC2 - MCC1	**-0.7**	**-0.4**	—	—	**+0.5**	**+0.4**	**+0.1**	**-0.1**	**-0.1**	**-0.4**	—	—	**-0.4**	**-0.3**	—	—	—	—	**+0.8**	**+0.1**
Para + MCC1	40.6	89.2	—	—	28.8	88.0	23.6	79.9	—	—	37.7	71.9	29.2	86.9	—	—	—	—	—	—
Para + MCC2	40.4	89.0	—	—	29.4	88.3	23.3	79.7	—	—	38.2	72.0	29.3	86.7	—	—	—	—	—	—
δ MCC2 - MCC1	**-0.2**	**-0.2**	—	—	**+0.6**	**+0.3**	**-0.3**	**-0.2**	—	—	**+0.5**	**+0.1**	**+0.1**	**-0.2**	—	—	—	—	—	—
All corpora	—	—	—	—	—	—	23.6	79.7	34.3	87.1	37.6	71.7	—	—	31.2	87.2	—	—	—	—
All + MCC1	—	—	—	—	—	—	24.4	79.8	34.5	87.2	39.9	72.6	—	—	—	—	—	—	—	—
All + MCC2	—	—	—	—	—	—	24.8	80.0	34.3	87.1	39.2	72.3	—	—	31.0	87.2	—	—	—	—
δ MCC2 - MCC1	—	—	—	—	—	—	**+0.4**	**+0.2**	**-0.2**	**-0.1**	**-0.7**	**-0.3**	—	—	—	—	—	—	—	—
Equal size to MCC2																				
MaCoCu-V2	35.3	86.5	—	—	28.1	87.3	—	—	21.3	73.0	35.0	71.2	25.8	84.3	25.3	82.3	30.1	85.2	24.4	85.2
MaCoCu-V1	**-0.4**	**-1.0**	—	—	**-1.8**	**-1.4**	—	—	**+0.7**	**+2.1**	**+0.2**	**+0.1**	**-0.6**	**-0.7**	—	—	—	—	**-5.1**	**-6.5**
ParaCrawl	**+1.1**	**+0.2**	—	—	**-0.9**	**-0.3**	—	—	—	—	**-0.8**	**-0.6**	**-0.8**	**-1.9**	—	—	—	—	—	—
CCAligned	**-2.7**	**-4.2**	—	—	**-2.9**	**-4.6**	—	—	**+0.9**	**+2.1**	—	—	**-1.7**	**-3.9**	**-5.8**	**-8.0**	**-1.9**	**-3.6**	**-3.1**	**-5.7**
CCMatrix	**+2.7**	**+0.9**	—	—	**+0.1**	**-0.5**	—	—	**-10.5**	**-17.5**	—	—	**+1.3**	**+0.0**	**-4.6**	**-5.3**	**+0.1**	**-1.0**	**+1.3**	**+0.0**

Table 3.2: Performance on FLoRes devtest for our machine translation systems trained on various corpora. CO and BL are short for COMET and BLEU. The rows starting with δ MCC2-MCC1 denote the relative performance of the MaCoCu-V2 corpus versus the MaCoCu-V1 corpus, whether training from scratch or being used as an additional corpus.

3.2. Human evaluation of bilingual corpora

Besides automatic evaluation, human evaluation of five parallel publicly available web-crawled parallel corpora is performed: MaCoCu-V1, MaCoCu-V2, CCAligned, CCMatrix and ParaCrawl. In the previous section, we showed that the size of the data sets had a big influence on the results. In this section we are interested in the quality of the sentence pairs in the data sets, regardless of the total size.

Evaluating web-crawled corpora for the purpose of training NMT systems is quite different than evaluating the output of NMT systems. We are not necessarily interested in fine-grained error analysis of the translations. A reasonable translation, though not perfect, is still useful for training NMT systems. Web-crawled corpora are created by automatically aligning potential sentence pairs, and this is alignment is far from perfect. We want to identify 1) the number of times this process is indeed imperfect and 2) what issue do the wrong sentence pairs suffer from. In other words, we want to detect where our automatic tools made mistakes (as this can be improved), rather than detecting imperfect translations by human web editors (as we cannot reasonably expect our tools to capture this). We created an annotation scheme based on this philosophy, which is shown below:

1. **Level 1.** Is the content written in the expected languages?
   a) **Wrong Language (WL).** The content of one of the two sentences is not in the expected language. Example: English expected in a sentence but it's in Icelandic. Annotation is finished.
   b) **Mixed Languages (ML).** The content of one of the two sentences is written in a mix of languages, one of which is the expected one. Example: English expected on one of the sentences but it is a mix of English and Icelandic. Annotation is finished.
   c) **CL (Correct Languages).** The content of both sentences is in the expected languages. We move to level 2.
2. **Level 2.** Is the content on both sides roughly the same?
   a) **Missing Content (MC).** The content in one sentence is missing a substantial part of the content from the other sentence. Example: a whole sentential clause is present on one sentence but missing on the other. Annotation is finished.
   b) **Replaced Content (RC).** The second sentence looks like a reasonable translation of the first, but one or more content words seem to be replaced by a wrong word or phrase. Common examples are different dates, proper nouns and numbers. Example: in one sentence it's written Madrid and on the other it's written London.
   c) **Misalignment (MA).** The content of both sentences is completely different. Annotation is finished.
   d) **Same Content (SC).** The content of both sentences is roughly the same. We move to level 3.
3. **Level 3.** Is the translation reasonably correct?
   a) **Low Quality Translation (LQT).** The content of both sentences is roughly the same but there are serious translation errors. Example: a mistranslation or overly literal translation. Annotation is finished.
   b) **Correct, but boilerplate translation (CBT).** The content of both sentences is roughly the same, but the content is boilerplate. Boilerplate includes pieces of website text that are unrelated to the content (e.g. HTML, cookies, website navigation). It can also include sentences that look automatically generated, instead of being written by a human. Annotation is finished.
   c) **Reasonable Translation (RT).** The content of both sentences is roughly the same and the translation is at least reasonable. Annotation is finished.

In level 1 the most serious issues are annotated: are the sentences in the correct language, or are there clear issues? If this is the case, annotation stops. In level 2, alignment issues are annotated. Parallel corpora often suffer from two issues. The first is that a translation of a single sentence is split in two sentences on the target side, but our automatic tools align only one of them. This should be annotated as "Missing Content". The second is that two sentences are similar, but a small part (often a name, number or noun phrase) is different. This is usually a mistake that a (professional) translator would never make: e.g. the number 27 if translated by 28, or "Thursday" is translated by "Wednesday". This should be annotated as "Replaced Content".

Subsequently, in level 3, translation issues are annotated. Some translations are about the same content, but are simply of very low quality. When dealing with web crawls, the content is often badly machine translated. A different issue is boilerplate. Websites contain a lot of standard boilerplate texts, in which we are not interested. Examples are shown in Appendix B. If none of the previous options apply, annotators automatically have to pick "Reasonable Translation". As stated previously, translations do not have to perfect to be useful for training MT systems: a reasonable translation is good enough for us.

Finally, independently of the rest of the annotation, we ask annotators to identify two other issues:

- Does the source or target contain offensive or pornographic content? At least one of the sentences in the sentence pair is offensive or is likely to be offensive to a subgroup of the population, and/or the sentence is pornographic in nature.
- Is the source or target not running text? This means that a substantial part of the text is just a bunch of words together, for which it does not make sense to judge the translation.

Again, the KEOPS online annotation tool is used to perform this task. A previous similar task based on a different annotation scheme initially planned as the default for MaCoCu but later discarded for the problems discussed in [24] inspires a change in the annotation procedure based on the above discussed levels. Only if a sentence pair is considered good to keep moving to the next level, labels for this particular level are shown. The full screen is shown in Figure 3.1:

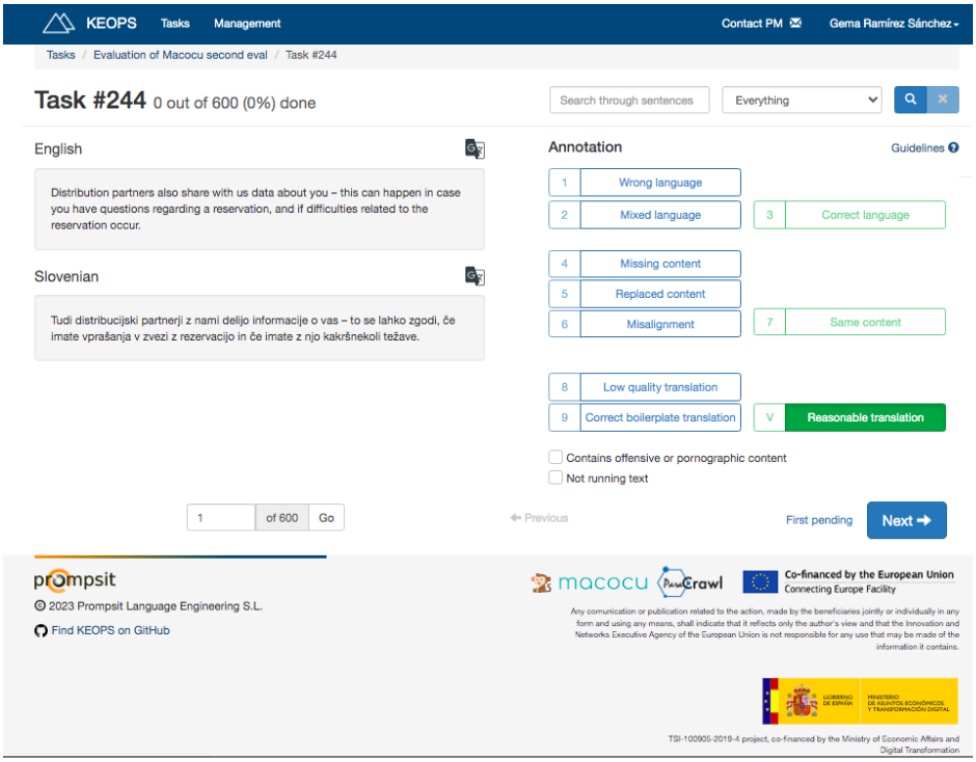

Figure 3.1: Parallel-data annotation screen in KEOPS.

We hired professional annotators for the 11 languages and 5 corpora under consideration. Similar to the monolingual annotation, we annotate 200 instances per corpus-language combination, with an additional 200 annotations to assess inter-annotator agreement. We hired two annotators per language, meaning that each annotator does between 200 and 600 annotations. The instances per annotator are balanced by corpus and given to the annotators in a randomized and blind fashion. When an instance has two annotations, we pick one of the annotations at random. The inter-annotator agreement in terms of exact annotation overlap and Cohen's kappa coefficient are shown in Table 3.3.

	Parallel data	
	% exact agreement	**κ coefficient**
English-Albanian	60.0	0.48
English-Bosnian	53.5	0.30
English-Bulgarian	53.0	0.39
English-Croatian	65.5	0.56
English-Icelandic	70.5	0.60
English-Macedonian	52.5	0.31
English-Maltese	55.0	0.22
English-Montenegrin	49.0	0.23
English-Serbian	80.5	0.71
English-Slovene	68.0	0.43
English-Turkish	45.0	0.34

Table 3.3: Inter-annotator agreement between the two annotators for each language pair for the evaluation of parallel data. The second column sho79ws the percentage (%) of annotations for which both annotators were in exact agreement; the third column shows Cohen's kappa coefficient (κ) between both annotators.

The detailed results of the annotation process across 5 corpora and 11 languages are shown in Table 3.4.

Bulgarian	**WL**	**ML**	**MC**	**RC**	**MA**	**LQT**	**CBT**	**RT**	**NR**	**PRN**
CCAligned	9	9	11	15	41	40	5	**70**	63	10
CCMatrix	0	2	30	24	22	33	3	86	10	0
MaCoCuV1	0	1	29	30	48	16	1	75	11	0
MaCoCuV2	0	3	30	14	9	29	4	**111**	17	0
ParaCrawl	2	0	30	23	11	27	6	101	27	0
Bosnian	**WL**	**ML**	**MC**	**RC**	**MA**	**LQT**	**CBT**	**RT**	**NR**	**PRN**
CCAligned	2	19	19	21	39	7	11	**82**	1	1
MaCoCuV2	0	10	15	9	3	10	11	**142**	0	0
Croatian	**WL**	**ML**	**MC**	**RC**	**MA**	**LQT**	**CBT**	**RT**	**NR**	**PRN**
CCAligned	25	14	13	14	56	27	20	**31**	79	17
CCMatrix	5	1	21	38	55	14	2	64	4	0
MaCoCuV1	6	3	37	22	43	13	4	72	6	0
MaCoCuV2	6	1	23	15	5	13	7	**130**	6	0
ParaCrawl	1	4	30	20	13	27	13	92	13	2
Icelandic	**WL**	**ML**	**MC**	**RC**	**MA**	**LQT**	**CBT**	**RT**	**NR**	**PRN**
CCAligned	2	52	7	29	16	34	1	**59**	20	2
CCMatrix	0	2	12	89	20	15	2	60	0	3
MaCoCuV1	1	1	20	36	9	12	3	118	1	0
MaCoCuV2	0	0	18	19	0	11	0	**152**	0	0
ParaCrawl	3	16	13	49	2	33	0	84	3	0
Macedonian	**WL**	**ML**	**MC**	**RC**	**MA**	**LQT**	**CBT**	**RT**	**NR**	**PRN**
CCAligned	3	11	9	21	38	36	10	**72**	15	2
CCMatrix	1	1	13	34	32	32	4	83	4	0
MaCoCuV1	0	2	13	27	7	27	2	122	2	0
MaCoCuV2	0	2	10	15	2	29	1	**141**	0	0
Maltese	**WL**	**ML**	**MC**	**RC**	**MA**	**LQT**	**CBT**	**RT**	**NR**	**PRN**
MaCoCuV1	0	0	10	23	3	21	16	127	16	0
MaCoCuV2	0	1	3	7	2	20	12	**155**	19	0
ParaCrawl	0	8	7	18	4	52	14	**97**	46	0
Montenegrin	**WL**	**ML**	**MC**	**RC**	**MA**	**LQT**	**CBT**	**RT**	**NR**	**PRN**
MaCoCuV2	14	10	24	14	0	19	8	111	0	0
Slovenian	**WL**	**ML**	**MC**	**RC**	**MA**	**LQT**	**CBT**	**RT**	**NR**	**PRN**
CCAligned	10	7	11	26	37	13	3	**93**	9	24
CCMatrix	1	1	14	32	17	6	22	107	1	0
MaCoCuV1	1	0	28	20	28	2	5	116	0	0
MaCoCuV2	0	1	16	5	4	4	5	**165**	0	0
ParaCrawl	0	4	16	26	11	1	12	130	3	2
Albanian	**WL**	**ML**	**MC**	**RC**	**MA**	**LQT**	**CBT**	**RT**	**NR**	**PRN**
CCAligned	18	29	69	5	0	24	34	**21**	0	0
CCMatrix	0	2	86	1	2	20	58	31	0	0
MaCoCuV2	0	0	26	0	0	24	84	**66**	0	0
Serbian	**WL**	**ML**	**MC**	**RC**	**MA**	**LQT**	**CBT**	**RT**	**NR**	**PRN**
CCAligned	0	3	6	12	18	82	13	**66**	0	0
CCMatrix	0	0	20	55	6	21	4	94	0	0
MaCoCuV2	6	1	23	18	0	7	2	**143**	0	0
Turkish	**WL**	**ML**	**MC**	**RC**	**MA**	**LQT**	**CBT**	**RT**	**NR**	**PRN**
CCAligned	8	12	12	15	33	29	22	69	81	6
CCMatrix	0	5	14	13	20	41	28	**79**	29	0
MaCoCuV1	1	3	21	55	86	5	7	22	18	0
MaCoCuV2	0	5	25	32	25	6	16	**91**	25	0

Table 3.4: Detailed statistics for the human evaluation of parallel data.

The main conclusion that stands out is that MaCoCu-V2 corpora are the best valued by annotators. For all 10 languages where a comparison is possible, MaCoCu-V2 has the highest number of "Reasonable Translations", which we consider the main indicator of quality. MaCoCu-v2 is also clearly better than MaCoCu-V1, indicating that the refined processing steps for the second release clearly had a positive effect on data quality (**RQ1**). After MaCoCu, ParaCrawl is generally the corpus with the most "Reasonable Translations", followed by CCMatrix and then CCAligned. Even for MaCoCu, though, it's clear that our tools are far from perfect. Albanian seems to have a lot of boilerplate or machine generated content. After clarification from annotators, we found that they included in this category potentially machine translated texts that looked mostly good.

To get a more clear picture, we also average over all languages, to get more reliable scores per corpus. Since there are only four languages (Bulgarian, Croatian, Icelandic and Slovene) included in all five evaluated corpora, we also show the results for each corpus when we average over all languages available for this corpus. These results are shown in Table 3.5. Again, it's clear that there is quite a difference between MaCoCu-V2 and the other corpora (**RQ2**). We observe, though, that there are still serious issues with web-crawled parallel corpora. For MaCoCu and ParaCrawl, only around half the sentence-pairs (a bit more for MaCoCu-V2) can be considered a "Reasonable Translation". For CCAligned and CCMatrix this worse: only around a third of the sentence pairs are free from major issues.

	Langs	WL	ML	MC	RC	MA	LQT	CBT	RT	NR	PRN
Only shared languages:											
CCAligned	4	5.8%	10.1%	4.9%	10.6%	19.0%	14.4%	3.6%	**31.6%**	21.4%	6.2%
CCMatrix	4	0.8%	0.8%	9.0%	23.0%	14.6%	8.8%	3.8%	39.4%	2.0%	0.4%
MaCoCuV1	4	1.0%	0.6%	14.5%	13.9%	15.8%	5.2%	1.5%	47.5%	2.4%	0.0%
MaCoCuV2	4	0.8%	0.6%	10.6%	6.5%	2.0%	7.2%	1.9%	**70.4%**	2.8%	0.0%
ParaCrawl	4	0.8%	3.1%	11.6%	14.5%	4.5%	11.0%	3.9%	50.6%	5.8%	0.8%
All possible languages:											
CCAligned	9	4.3%	8.7%	8.7%	8.8%	15.4%	16.2%	6.6%	**31.3%**	14.9%	3.4%
CCMatrix	8	0.4%	0.9%	13.1%	17.9%	10.9%	11.4%	7.7%	37.8%	3.0%	0.2%
MaCoCuV1	7	1.8%	0.8%	11.2%	14.8%	16.1%	6.7%	2.1%	46.4%	4.3%	0.0%
MaCoCuV2	11	2.1%	1.5%	9.7%	6.7%	2.2%	7.5%	6.5%	**63.7%**	3.3%	0.0%
ParaCrawl	5	4.2%	2.7%	9.6%	13.4%	4.1%	11.3%	3.6%	51.1%	8.7%	0.4%

Table 3.5: Percentage of annotations for each of the annotation categories, averaged over corpus across either the four languages that had all five corpora available, or all available languages.

CCAligned especially seems to suffer from texts that are not actually running text, though the other corpora also struggle with this. Variability among languages is observed: Serbian and Albanian never have any not-running text, while Bulgarian, Maltese and Turkish have this quite often. This is surprising to us and might be related to the preferences of individual annotators. CCAligned is also the corpus that most often misidentified one of the languages, though this is never a huge issue for any of the corpora. Similarly, CCAligned is virtually the only corpus with offensive or pornographic sentence pairs, meaning that the other corpora successfully filtered them. Alignment issues represented by MC, RC and MA categories are more acute in CCAlign, CCMatrix and MaCoCuV1 than in ParaCrawl or MaCoCu-v2. Again, improvements to the MaCoCu pipeline from version 1 to version 2 show their impact in the final quality of the data.

4. Conclusion

In this report we evaluated the monolingual and parallel corpora included in the second data release of the MaCoCu action. In both settings, we compared the MaCoCu corpora to similar other large, web-crawled and publicly available corpora. We automatically evaluated the quality of the data sets by training either language models (from monolingual data) or neural machine translation (NMT) systems (from parallel data). Moreover, we hired professional linguists to directly annotate randomly selected fragments of text from each of the corpora, to get a fair comparison between them. In the automatic evaluation for monolingual data, we found that the MaCoCu corpora work well, though not better than other similar corpora. In the manual evaluation, though, we find that the MaCoCu and OSCAR corpora are generally considered to be the corpora with the highest quality texts. Interestingly, this difference is not reflected in the automatic evaluation.

For the parallel data, the findings are more clear. We find that, when controlling for data set size, the MaCoCu-V2 corpus is generally the best corpus for training NMT systems. Moreover, in the human evaluation, we found that the MaCoCu-V2 corpora have the most reasonable translations, and the least amount of other issues, such as misaligned sentences. There is still room for improvement, though: even for the MaCoCu corpus, only 63.7% of sentence pairs were judged to be at least a reasonable translation. This seems to have a moderate impact on the quality of models evaluated through automatic metrics.

A. Monolingual annotation scale examples

Wrong language / Not language

1: STRAND 240 242 ECO:0000244—PDB:1FMK.
2: box-shadow: 1px 1px 3px 2px #121e03;
3: このテキストは完全に間違った言語で書かれています。

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.

B. Parallel annotation scheme examples

Wrong Language (WL):
Sent 1: The meeting takes place on Thursday the 28th of March and will be about our finances.
Sent 2: Fundurinn fer fram fimmtudaginn 28. mars og mun fjalla um fjármálin okkar.

Mixed Languages (ML):
Sent 1: The meeting takes place on Thursday the 28th of March and will be about our finances.
Sent 2: The meeting takes place on Thursday the 28th og mun fjalla um fjármálin okkar.

Note that usage of specific terms or toponyms is not considered mixed language:
Sent 1 or 2: I got accepted at *Háskólinn Reykjavík* for the next academic year.
Sent 1 or 2: From the house you get to a large terrace with dining table and lounge, where you can relax, or have your *siesta* in a hammock.

Missing Content (MC):
Sent 1: The meeting takes place on Thursday the 28th of March and will be about our finances.
Sent 2: The meeting takes place on Thursday the 28th of March.
Sent 1: The meeting takes place on Thursday the 28th of March.
Sent 2: The meeting takes place on Thursday the 28th of March and will be about our finances.

Replaced Content (RC):
Sent 1: The meeting takes place on Thursday the 28th of March and will be about our finances.
Sent 2: The meeting takes place on Wednesday the 27th of March and will be about our merger.
Sent 1: Turkey is a beautiful country to visit in the summer.
Sent 2: Greece is a beautiful country to visit in the summer.
Sent 1: Book your tickets for only 500 euro here!
Sent 2: You can book tickets for only 400 euro here.

Complete Misalignment (MA):
Sent 1: The meeting takes place on Thursday the 28th of March and will be about our finances.
Sent 2: John and Mary went to the zoo and had a great time.
Sent 1: The meeting takes place on Thursday the 28th of March and will be about our finances.
Sent 2: In our previous meeting, which took place on April 2nd, we discussed our current situation and any plans we had for the future.

Correct, but boilerplate (CBT):
Sent 1 or 2: 850 Acres of Land Stock 355 Parcels
Sent 1 or 2: By accepting all cookies, you agree to our use of cookies to deliver and maintain our services.
Sent 1 or 2: Click here to go back to Home.
Sent 1 or 2: Premier Apartment with Sea Front View

Low Quality Translation (LQT):
Sent 1: The meeting takes place on Thursday the 28th of March and will be about our finances.
Sent 2: Meeting take place thursday 28 march about money.

Reasonable Translation (RT):
Sent 1: The meeting takes place on Thursday the 28th of March and will be about our finances.
Sent 2: Our meeting about our final situation takes place on Thursday the 28th of March.
Sent 2: On 28-03 we will meet about our finances.
Sent 2: Next week Thursday 28-03 the meeting about the budget will take place.

Offensive or pornographic content (PRN):
Sent 1 or 2: What the fuck is wrong with you dumb idiot
Sent 1 or 2: Amateur Teen Sex Porn Now Order Here

Not running text (NR):
Sent 1 or 2: <start=”204.771” dur=”1.868”>Well, you guys,
Sent 1 or 2: Vacation Holiday Turkey Slovenia Ankara Book Now
Sent 1 or 2: TO007 Stone Granite Display Cabinet
Sent 1 or 2: Home >Products >Circuit Protection >Electrical, Specialty Fuses >004176029
Sent 1 or 2: 1500mmx3000mm hot sale and good price fiber laser cutting machine with 500w,700w
Sent 1 or 2: Photo White-spotted Puffer (Arothron hispidus), Spotted, Aquarium Fish

Bibliography

[1] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. [Online]. Available: https://aclanthology.org/2020.acl-main.747

[2] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan, "The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 522–538, 05 2022. [Online]. Available: https://doi.org/10.1162/tacl_a_00474

[3] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for mt evaluation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 2685–2702.

[4] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[5] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[6] P. J. Ortiz Su'arez, L. Romary, and B. Sagot, "A monolingual approach to contextualized word embeddings for mid-resource languages," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 1703–1714. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.156

[7] A. El-Kishky, V. Chaudhary, F. Guzmán, and P. Koehn, "CCAligned: A massive collection of cross-lingual web-document pairs," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*. Online: Association for Computational Linguistics, November 2020, pp. 5960–5969. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.480

[8] H. Schwenk, G. Wenzek, S. Edunov, E. Grave, A. Joulin, and A. Fan, "CCMatrix: Mining billions of high-quality parallel sentences on the web," 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)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 6490–6500. [Online]. Available: https://aclanthology.org/2021.acl-long.507

[9] M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. L. Forcada, A. Kamran, F. Kirefu, P. Koehn, S. Ortiz Rojas, L. Pla Sempere, G. Ramírez-Sánchez, E. Sarrías, M. Strelec, B. Thompson, W. Waites, D. Wiggins, and J. Zaragoza, "ParaCrawl: Web-scale acquisition of parallel corpora," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 4555–4567. [Online]. Available: https://aclanthology.org/2020.acl-main.417

[10] M. Roemmele, C. A. Bejan, and A. S. Gordon, "Choice of plausible alternatives: An evaluation of commonsense causal reasoning." in *AAAI spring symposium: logical formalizations of commonsense reasoning*, 2011, pp. 90–95.

[11] M.-C. De Marneffe, M. Simons, and J. Tonhauser, "The commitmentbank: Investigating projection in naturally occurring discourse," in *proceedings of Sinn und Bedeutung*, vol. 23, no. 2, 2019, pp. 107–124.

[12] S. L. Ingólfsdóttir, Á. A. Gudhjónsson, and H. Loftsson, "MIM-GOLD-NER – named entity recognition corpus (21.09)," 2020, CLARIN-IS. [Online]. Available: http://hdl.handle.net/20.500.12537/140

[13] J. Kreutzer, I. Caswell, L. Wang, A. Wahab, D. van Esch, N. Ulzii-Orshikh, A. Tapo, N. Subramani, A. Sokolov, C. Sikasote, M. Setyawan, S. Sarin, S. Samb, B. Sagot, C. Rivera, A. Rios, I. Papadimitriou, S. Osei, P. O. Suarez, I. Orife, K. Ogueji, A. N. Rubungo, T. Q. Nguyen, M. Müller, A. Müller, S. H. Muhammad, N. Muhammad, A. Mnyakeni, J. Mirzakhalov, T. Matangira, C. Leong, N. Lawson, S. Kudugunta, Y. Jernite, M. Jenny, O. Firat, B. F. P. Dossou, S. Dlamini, N. de Silva, S. Çabuk Ballı, S. Biderman, A. Battisti, A. Baruwa, A. Bapna, P. Baljekar, I. A. Azime, A. Awokoya, D. Ataman, O. Ahia, O. Ahia, S. Agrawal, and M. Adeyemi, "Quality at a glance: An audit of web-crawled multilingual datasets," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 50–72, 2022. [Online]. Available: https://aclanthology.org/2022.tacl-1.4

[14] M. Artetxe, I. Aldabe, R. Agerri, O. Perez-de Viñaspre, and A. Soroa, "Does corpus quality really matter for low-resource languages?" in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 7383–7390. [Online]. Available: https://aclanthology.org/2022.emnlp-main.499

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, "Marian: Fast neural machine translation in C++," in *Proceedings of ACL 2018, System Demonstrations*, Melbourne, Australia, 2018. [Online]. Available: https://arxiv.org/abs/1804.00344

[17] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. [Online]. Available: https://aclanthology.org/P16-1162

[18] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71. [Online]. Available: https://aclanthology.org/D18-2012

[19] M. Popović, "chrF: character n-gram F-score for automatic MT evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 392–395. [Online]. Available: https://aclanthology.org/W15-3049

[20] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2019.

[21] T. Sellam, D. Das, and A. Parikh, "BLEURT: Learning robust metrics for text generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7881–7892. [Online]. Available: https://aclanthology.org/2020.acl-main.704

[22] P. Bozovic, T. Erjavec, J. Tiedemann, N. Ljubesic, and V. Gorjanc, "Opus-montenegrinsubs 1.0: First electronic corpus of the montenegrin language," in *Proceedings of the conference on Language Technologies & Digital Humanities 2018*. Ljubljana University Press, 2018.

[23] J. Tiedemann, "The tatoeba translation challenge – realistic data sets for low resource and multilingual MT," in *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1174–1182. [Online]. Available: https://aclanthology.org/2020.wmt-1.139

[24] G. Ramírez-Sánchez, M. Bañón, J. Zaragoza-Bernabeu, and S. Ortiz Rojas, "Human evaluation of web-crawled parallel corpora for machine translation," in *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 32–41. [Online]. Available: https://aclanthology.org/2022.humeval-1.4