Quality Beyond A Glance: Revealing Large Quality Differences Between Web-Crawled Parallel Corpora

Rik van Noord[♦], **Miquel Esplà-Gomis**[★], **Malina Chichirau**[♦]
Gema Ramírez-Sánchez[†] **and Antonio Toral**[♦]
[♦]University of Groningen, [★]Universitat d'Alacant, [†]Prompsit
rikvannoord@gmail.com

Abstract

Parallel corpora play a vital role in advanced multilingual natural language processing tasks, notably in machine translation (MT). The recent emergence of numerous large parallel corpora, often extracted from multilingual documents on the Internet, has expanded the available resources. Nevertheless, the quality of these corpora remains largely unexplored, while there are large differences in how the corpora are constructed. Moreover, how the potential differences affect the performance of neural MT (NMT) systems has received only limited attention. This study addresses this gap by manually and automatically evaluating four well-known publicly available parallel corpora across eleven language pairs.

Our findings are quite concerning: all corpora contain a substantial amount of noisy sentence pairs, with CCMatrix and CCAligned having well below of 50% reasonably clean pairs. MaCoCu and ParaCrawl generally have higher quality texts, though around a third of the texts still have clear issues.

While corpus size impacts NMT models' performance, our study highlights the critical role of quality: higher-quality corpora consistently yield better-performing NMT models when controlling for size.

1 Introduction

Parallel data, which comprises collections of texts in one language aligned with their corresponding translations in another language, are crucial for numerous natural language processing tasks in cross-lingual scenarios (Conneau et al., 2020; Reid and Artetxe, 2022). Its significance is particularly pronounced in the field of neural machine translation (NMT), where parallel data is a crucial component of the current state-of-the-art approaches (Vaswani et al., 2017; Junczys-Dowmunt et al., 2018; Klein et al., 2018; Huang et al., 2023; Maillard et al., 2023; Edman et al., 2024; Wu et al., 2024).

The need for parallel data is fuelling a growing interest in building larger and higher quality parallel corpora. Many of these efforts use the Web as a source for parallel data, either by crawling (Bañón et al., 2023) or by using existing general purpose large crawls, such as Common Crawl[1] or the Internet Archive (El-Kishky et al., 2020; Bañón et al., 2020; Schwenk et al., 2021b).[2]

Building parallel corpora from web-crawled content has proven to be a successful strategy. However, this approach is susceptible to producing noisier data due to the inherent heterogeneity of sources (Kreutzer et al., 2022). In this work, we compare several parallel corpora harvested from the Internet in order to measure their quality following two methods: we first run an intrinsic manual evaluation, by hiring language experts to assess the quality of a sample of each corpus; then we run an extrinsic automatic evaluation by using the corpora to train a variety of NMT systems. We include four corpora in our evaluation: CCAligned (El-Kishky et al., 2020), CCMatrix (Schwenk et al., 2021b), ParaCrawl (Bañón et al., 2020) and MaCoCu (Bañón et al., 2023). The case of the MaCoCu corpus is especially interesting for the aims of this work: two versions of this corpus are available, with the second one being substantially smaller than the first one, as authors carried out an additional effort in cleaning their data. We aim to answer two main research questions:[3]

- **RQ1**: How noisy are web-crawled parallel corpora and what issues do they have?
- **RQ2**: How do these differences affect the performance of NMT systems trained on the corpora?

[1] https://commoncrawl.org/
[2] https://archive.org
[3] All data and code available here: https://github.com/RikVN/MaCoCu_Parallel

Proceedings of the 31st International Conference on Computational Linguistics, pages 1824–1838
January 19–24, 2025. ©2025 Association for Computational Linguistics

Contributions To answer **RQ1**, we create a novel language-independent annotation scheme that is specifically tailored towards evaluating web-crawled parallel corpora. We find that there are indeed large differences in (perceived) quality between the four corpora, with MaCoCu being of the highest quality, while CCAligned and CCMatrix being the noisiest corpora. We find the results quite concerning: measured across 11 languages, the best corpus (MaCoCu) only has 64% of acceptable translations, while this drops to an alarming 31% for CCMatrix. Moreover, these differences in quality do generally impact the performance of NMT systems (**RQ2**), though the differences are not as pronounced as in the manual evaluation.

2 Related work

Noise & Annotation Khayrallah and Koehn (2018) annotate a small subset of the ParaCrawl corpus for different types of noise and find that the broad category of misaligned sentences is the biggest source of noise. Herold et al. (2022) extend this work by using more refined noise categories and show that automatically recognizing such noisy sentence pairs is still challenging. Ramírez-Sánchez et al. (2022) extend these two previous works by running a new type of human evaluation based on measuring post-editing effort in fixing segment pairs from the corpus, and by running an automatic evaluation consisting of training NMT models on this corpus. The conclusions of all these papers are clearly aligned as all of them work on the ParaCrawl corpus.

Quality at a glance Our work mostly takes inspiration from the milestone work of Kreutzer et al. (2022), who manually annotated a large subset of languages present in the parallel corpora of ParaCrawl v7.1, CCAligned and WikiMatrix (Schwenk et al., 2021a). They find that such corpora have systematic issues: many corpora have less than 50% usable sentence pairs, while there are also large issues regarding language identification. We built on their work, but with a focus on corpora comparison. We achieve this by looking into languages that are present in several of the corpora under comparison, while Kreutzer et al. (2022) rather look into languages representative of each corpus that are not necessarily covered by the other corpora. Moreover, we use a more fine-grained annotation scheme that also identifies in what way sentence-pairs are wrongly aligned or identified. Finally, and also different from Kreutzer et al. (2022), we also perform an automatic evaluation, by training NMT systems on different corpora and evaluating translations of the resulting systems, thereby analyzing how the quality of the corpora affects the quality of MT. Other related work includes Caswell et al. (2020), who manually evaluate their language identification system across a large number of languages, and Dodge et al. (2021), who analyse and document the English C4 corpus (Raffel et al., 2020).

Parallel corpus cleaning Our second research question is closely related to the task of parallel corpus cleaning or bitext filtering, which is the task of filtering possibly detrimental sentence pairs from a parallel corpus. There is a large body of related work that shows that such sentences indeed hurt the performance of NMT systems (Carpuat et al., 2017; Khayrallah and Koehn, 2018; Junczys-Dowmunt, 2018; Chaudhary et al., 2019; Briakou and Carpuat, 2021; Steingrímsson et al., 2023), which was driven by a number of shared tasks on the topic (Barbu et al., 2016; Koehn et al., 2018, 2019, 2020; Sloto et al., 2023). Bansal et al. (2022) on the other hand show that some filtering advantages disappear when using larger data sets. Popular publicly available filtering tools include LASER (Artetxe and Schwenk, 2019), BiCleaner (Ramírez-Sánchez et al., 2020) and OpusFilter (Aulamo et al., 2020).

Aim of this paper However, we want to emphasize that we do not investigate which filtering method is preferable. Our aim is to evaluate the parallel corpora *as they are released*, since this reflects how most practitioners and researchers would use them. The parallel corpora under investigation here were released after already applying different filtering processes, making this a confounding factor that we cannot control for. However, since the corpora were actually filtered already, we believe it is still fair to compare them in this manner, without applying any extra preprocessing or filtering.

3 Data

We include four parallel corpora in our evaluation: CCAligned (El-Kishky et al., 2020), ParaCrawl (Bañón et al., 2020), CCMatrix (Schwenk et al., 2021b), and MaCoCu (Bañón et al., 2023). This section briefly describes the main features of each corpus and characterizes them from a quantitative point of view.

	MaCoCu-V1		**MaCoCu-V2**		**CCAligned**		**CCMatrix**		**ParaCrawl**	
	Sents	**Tokens**	**Sents**	**Tokens**	**Sents**	**Tokens**	**Sents**	**Tokens**	**Sents**	**Tokens**
Albanian	—	—	494	23,662	1,328	47,089	9,513	264,737	—	—
Bosnian	—	—	464	20,367	154	10,510	—	—	—	—
Bulgarian	2,156	89,314	1,676	68,789	5,829	207,661	24,214	731,854	10,235	342,468
Croatian	1,920	84,646	2,147	94,817	4,793	159,913	8,327	243,097	2,597	120,112
Icelandic	291	11,793	258	10,258	848	31,503	3,399	69,779	2,103	60,636
Macedonian	402	19,654	359	17,447	1,168	41,624	4,263	129,650	—	—
Maltese	979	54,309	867	50,762	—	—	—	—	988	38,079
Montenegrin	—	—	204	10,502	—	—	—	—	—	—
Serbian	—	—	1,664	78,816	1,590	328,207	13,428	328,207	—	—
Slovenian	2,158	93,632	1,788	81,244	2,647	92,627	14,574	384,688	7,047	219,324
Turkish	3,801	205,000	1,533	83,669	8,541	271,859	31,016	702,223	—	—

Table 1: Data set sizes (thousands of sentence pairs and thousands of tokens) per corpus and per language pair. Not each language is present in each corpus. Sizes are after normalization and near-deduplication.

CCAligned This corpus was created through URL-based document alignment on a collection of 68 Common Crawl Snapshots. Document language is identified by FastText (Joulin et al., 2017) and they ensure that the document URL contains the corresponding language code. Then pairs of URLs are compared by trying to strip any language identification elements from them; if two URLs are exactly the same after removing these elements, they are considered to correspond to parallel documents. Aligned documents are split into sentences based on punctuation, and documents are aligned at the level of segments by using LASER (Artetxe and Schwenk, 2019). Finally, deduplication is performed at the level of segment pairs. The resulting corpus contains 392 million document pairs covering 138 languages.

CCMatrix This corpus is similar to CCAligned as it was created through sentence alignment on a collection of documents from 10 Common Crawl Snapshots. FastText is used for language identification and additional filtering is carried out at document level by discarding documents with a low perplexity according to language models trained on Wikipedia content. LASER is used to mine parallel segments from the collection of multilingual data. The main difference between this corpus and CCAligned is that in CCMatrix parallel segments are found across the whole dataset, while in CCAligned there is a previous step of document alignment. The resulting corpus contains 4.5 billion segment pairs for 28 language pairs.

ParaCrawl Several incremental versions of this corpus exist, with early ParaCrawl versions obtaining data by automatically crawling the Web, while later versions build on data extracted from the Internet Archive. CLD2[4] is used to identify the language of documents, after which the non-English documents are machine translated to English. Then, a TF-IDF-based metric is used to identify the most similar documents. Pairs of documents are then aligned at the segment level by using BLEU-align (Sennrich and Volk, 2011). Additional filtering is finally applied with the set of tools Bifixer, Bicleaner-hardrules and Bicleaner AI (Ramírez-Sánchez et al., 2020; Zaragoza-Bernabeu et al., 2022). The resulting corpus covers 42 languages, and includes 46 language pairs. Corpora size ranges from 278 million segment pairs for the highest-resourced language pair to 14 thousand for the lowest-resourced one.

MaCoCu The MaCoCu corpus builds on data crawled from the web for eleven low to medium-resource European languages. Crawling was conducted automatically on top-level domains identified as relevant for the targeted languages. Two versions of the parallel corpora exist for some of the languages covered by the corpus. The same collection of crawled documents was used to build both versions, but an improved pipeline was used for the second one, which lead to smaller and higher quality corpora. We will evaluate both versions in this paper. The main differences in the pipelines used for both versions are: (a) a custom trigram model was used for language identification in the first version, while the tool CLD2 was used for the second one; and (b) an improved and more precise version of the document alignment tool was used for the second version. For both document/segment

[4] https://github.com/CLD2Owners/cld2

alignment and parallel data cleaning, updated versions of the same tools mentioned for ParaCrawl were used. The resulting corpora contain about 10 million pairs of segments for the most represented language pair to about half a million pairs of segments for the least represented language.

Languages MaCoCu is an interesting corpus to include in the analysis, since (i) it uses a different crawling method compared to the other corpora, and (ii) it released two different versions built on the same raw data, but with different levels of cleaning applied. However, this does mean that our evaluation is limited to the eleven languages covered by the MaCoCu corpus.[5] Nevertheless, this is still quite a diverse set of languages, including Albanic, Germanic, Semitic, Slavic and Turkic languages. Moreover, none of these languages are considered highly resourced in the taxonomy of Joshi et al. (2020), which divides languages from class 0 (minimal resources available) to class 5 (highly resourced): Albanian and Macedonian belong to class 1, Maltese and Icelandic to class 2, Bulgarian, Bosnian and Slovenian to class 3, and Turkish, Serbian and Croatian to class 4.[6]

Statistics Each corpus we use is pre-processed by the script used for training models in the Tatoeba Translation challenge (Tiedemann, 2020), after which duplicate and near-duplicate sentence pairs are removed. Table 1 describes the amount of data available in each of these corpora for each language. CCMatrix is, by far, the corpus with the most data. Only four out of the eleven language pairs included in our evaluation are covered by all the corpora compared, and there is one language pair (English–Montenegrin) that is only covered by the MaCoCu-V2 corpus.

Corpus overlap Table 2 shows the percentage of overlapping segment pairs for each pair of corpora. We define overlapping instances as those having either a source or target near-duplicate.[7] The highest overlap is, unsurprisingly, observed between the two versions of MaCoCu. CCAligned seems to be the corpus with the lowest amount of overlapping segment pairs with the rest of the corpora. In general, while the creation of the corpora follows a similar process, the overlap is relatively small.

	CCAligned	**CCMatrix**	**MaCoCuV1**	**MaCoCuV2**	**ParaCrawl**
CCAligned	100	4	1	1	6
CCMatrix	1	100	2	2	5
MaCoCuV1	3	13	100	65	18
MaCoCuV2	3	14	74	100	18
ParaCrawl	3	11	5	5	100

Table 2: Instance-level overlap percentages between the corpora. The rows are leading, i.e. the first row describes the percentage of CCAligned that is also present in the other corpora. For a fair comparison, we average the percentages over the four languages for which all four corpora have data for: Bulgarian, Croatian, Icelandic and Slovenian.

4 Manual Evaluation

This section aims to answer **RQ1** by presenting the results of the human evaluation run on the four publicly available web-crawled parallel corpora that are described in Section 3. Therefore, in this section we are interested in the *quality* of the sentence pairs in the corpora, regardless of the size of each corpus.

Motivation Evaluating web-crawled parallel corpora for the purpose of building downstream applications (e.g. training NMT systems) is quite different from evaluating translations. We are not necessarily interested in fine-grained error analysis of the translations. A reasonable translation, though not fully accurate or fully fluent, is still highly likely to be useful for training NMT systems. Web-crawled corpora are created by automatically aligning potential sentence pairs, and this alignment is known to be noisy (Khayrallah and Koehn, 2018). Therefore, we aim to identify 1) the number of times this process is indeed imperfect and 2) the exact issue the wrong sentence pairs suffer from. In other words, we aim to detect where the automatic tools to derive parallel corpora made mistakes (as this can be improved), rather than detecting somewhat flawed translations (as we cannot reasonably expect automatic tools to be able to capture or fix this). To this end, we have created a novel annotation scheme based on this philosophy, inspired by Kreutzer et al. (2022), which is shown in Table 3.

[5] The MaCoCu project is EU-funded, which explains their focus on European languages.

[6] Montenegrin is not included in the taxonomy.

[7] Near-duplicates are sentence pairs that are the exact same after removing whitespace and non-alphabetic characters.

Level 1	**E**	**Examples**
Wrong Language (WL): The content of one of the two sentences is not in the expected language.	✖	**S1:** The meeting takes place on Friday the 28th of May. **S2:** Fundurinn fer fram föstudaginn 28. maí.
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.	✖	**S1:** The meeting takes place on Friday the 28th of May. **S2:** The meeting takes place on föstudaginn 28. maí.
Correct Languages (CL). The content of both sentences is in the expected languages.	✔	
Level 2	**E**	**Examples**
Missing Content (MC): The content in one sentence is missing a substantial part of the content from the other sentence.	✖	**S1:** The meeting takes place on Friday the 28th of May. **S2:** The meeting takes place on Friday.
Replaced Content (RC): The second sentence looks like a reasonable translation of the first, but one or more content words are replaced by a wrong word or phrase. Common examples are different dates, proper nouns and numbers.	✖ ✖ ✖	**S1:** The meeting takes place on Friday the 28th of May. **S2:** The meeting takes place on Friday the 12th of May. **S2:**The meeting takes place on Friday the 27th of April. **S2:**The meeting takes place on Monday the 28th of May.
Complete Misalignment (MA): The content of both sentences is completely different.	✖	**S1:** The meeting takes place on Friday the 28th of May. **S2:** We had a great party last week.
Same Content (SC). The content is (roughly) the same.	✔	
Level 3	**E**	**Examples**
Correct, but boilerplate (CB): 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.	✖	**S1 or S2:** 850 Acres of Land Stock 355 Parcels **S1 or S2:** Click here to go back to Home. **S1 or S2:** By accepting all cookies, you agree to our use of cookies to deliver our services **S1 or S2:** <header> Abstract </header>
Low Quality Translation (LQ): The content of both sentences is roughly the same but there are serious translation errors.	✖	**S1:** The meeting takes place on Friday the 28th of May. **S2:** Meeting take place Friday 28 May.
Reasonable Translation (RT): The content of both sentences is roughly the same and the translation is at least reasonable.	✔ ✔	**S1:** The meeting takes place on Friday the 28th of May. **S2:** Our meeting is on Friday 28-05. **S2:** On Friday the 28th of May we have a meeting.

Table 3: Annotation hierarchy that was given to the annotators, including simplified example sentence pairs. See Appendix A for specific annotation instructions.

4.1 Annotation scheme

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 different alignment issues. The first is that a translation of a single sentence is split into two sentences on the target side, but the automatic tools aligned 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: e.g. a sentence containing the number 17 is aligned to a sentence containing the number 28, or *Thursday* is aligned to *Wednesday*. This should be annotated as *Replaced Content*. Again, annotation stops if one of these issues occurs.

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, which are tried to be filtered out from corpora. Examples are shown in Appendix A. If none of the previous options apply, annotators automatically have to pick *Reasonable Translation*. As stated previously, translations do not have to be perfect to be useful for training MT systems: a reasonable translation is considered good enough by 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 (PR)?
- Is the source or target not running text (NR)? 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.

4.2 Annotation results

We hired professional annotators for the 11 languages and 4 corpora under consideration. 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 completes 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 different annotations, we pick one of the annotations at random for the analysis. The KEOPS online annotation tool was used to perform this task.[8] The inter-annotator agreement in terms of exact annotation overlap and Cohen's kappa coefficient are shown in Table 5. The annotators agree to a reasonable extent (κ between 0.3 and 0.71), except for Maltese and Montenegrin, for which the results are somewhat concerning (κ values of 0.22 and 0.23).

Main results The detailed results of the annotation process across the 4 corpora and 11 languages are shown in Table 4. The main conclusion that stands out is that the 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, confirming the claims made by the authors of these corpora that the refined processing steps for the second release had a positive effect on data quality. After MaCoCu, ParaCrawl is generally the corpus with the most *Reasonable Translations*, followed by CCMatrix and then CCAligned. Even for MaCoCu, though, it is clear that this corpus is far from perfect. Albanian, for instance, seems to have a lot of boilerplate or machine generated content. After clarification from annotators, we found that they also included potentially machine-translated texts in this category that nevertheless looked mostly good.

Averaging across languages To get a more clear and general 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 evaluated corpora, we also show the results for each corpus when we average over all languages

Bulgarian	**WL**	**ML**	**MC**	**RC**	**MA**	**LQ**	**CB**	**RT**	**NR**	**PR**
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**	**LQ**	**CB**	**RT**	**NR**	**PR**
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**	**LQ**	**CB**	**RT**	**NR**	**PR**
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**	**LQ**	**CB**	**RT**	**NR**	**PR**
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**	**LQ**	**CB**	**RT**	**NR**	**PR**
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**	**LQ**	**CB**	**RT**	**NR**	**PR**
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**	**LQ**	**CB**	**RT**	**NR**	**PR**
MaCoCuV2	14	10	24	14	0	19	8	111	0	0
Slovenian	**WL**	**ML**	**MC**	**RC**	**MA**	**LQ**	**CB**	**RT**	**NR**	**PR**
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**	**LQ**	**CB**	**RT**	**NR**	**PR**
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**	**LQ**	**CB**	**RT**	**NR**	**PR**
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**	**LQ**	**CB**	**RT**	**NR**	**PR**
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 4: Detailed statistics for the human evaluation of the corpora. **RT** is short for *Reasonable Translation*, which we consider the most important indicator of quality. Please see Table 3 for an overview of all abbreviations.

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

	%	κ
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 5: Inter-annotator agreement between the two annotators for each language pair. 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.

available for this corpus. These results are shown in Table 6. Again, it is clear that there is quite a difference between MaCoCu-V2 and the other corpora. At the same time, though, we observe that there are still serious issues with all the evaluated 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 is even worse: only around a third of the sentence pairs are free from major issues. It is clear that the large size of both CCAligned and CCMatrix (see Table 1) comes at a cost of including lower quality sentence pairs.

Analysis 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 (around 5% of the instances), 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 such texts. Alignment issues represented by MC, RC and MA categories are more acute in CCAligned, CCMatrix and MaCoCu-V1 than in ParaCrawl or MaCoCu-V2.

5 Automatic Evaluation

In this section we aim to answer **RQ2** by running an extrinsic automatic evaluation for the corpora compared by training and evaluating NMT systems.

5.1 Training details

We build NMT systems from English into 10 of the languages targeted in this paper.[9] We train the models from scratch, using a Transformer (Vaswani et al., 2017) model implemented in Marian (Junczys-Dowmunt et al., 2018). We train a Transformer-based 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 (Sennrich et al., 2016; Kudo and Richardson, 2018). 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 NVIDIA V100 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 across all our experiments.

Evaluation To evaluate performance, we use the well-established MT metrics COMET (Rei et al., 2020) and BLEU (Papineni et al., 2002).[10] We evaluate performance on the FLoRes dev and devtest data sets (Goyal et al., 2022), showing only the devtest scores for brevity. All scores will be made publicly available, though our conclusions were similar across metrics and test sets.

5.2 Results

We train NMT systems (i) on the corpora as they are released (referred to as full size in the first half of Table 7) and (ii) on subsets of each corpus of equal size to the smallest one (MaCoCu-V2), shown in the second half of Table 7.

Full size The results show that data set size still matters a lot: CCMatrix (the largest corpus by far, see Table 1) obtains the best performance for all language pairs it has data for (except for English–Croatian), despite being clearly of lower quality than MaCoCu and ParaCrawl (see Section 4). That

[9] Montenegrin lacked enough data to train NMT systems.

[10] For brevity, we only show COMET scores, though BLEU scores are available in Appendix B.

	Langs	WL	ML	MC	RC	MA	LQ	CB	RT	NR	PR
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 6: Percentage of annotations for each of the annotation categories, averaged over corpus across either the four languages that had all corpora available, or all available languages.

	en-bg	en-bs	en-hr	en-is	en-mk	en-mt	en-sl	en-sq	en-sr	en-tr
Full size:										
MaCoCu-V2	86.5	76.9	87.3	—	73.0	71.2	84.3	82.3	85.2	85.2
MaCoCu-V1	86.2	—	85.9	—	77.3	71.5	83.9	—	—	83.2
CCAligned	86.2	38.0	84.2	71.7	76.8	—	81.1	80.4	81.6	84.3
CCMatrix	90.0	—	86.8	79.7	87.3	—	86.9	87.4	86.4	88.6
ParaCrawl	86.7	—	89.0	—	—	70.6	82.4	—	—	—
Equal size to MaCoCu-V2										
MaCoCu-V2	86.5	—	87.3	—	73.0	71.2	84.3	82.3	85.2	85.2
MaCoCu-V1	**-1.0**	—	**-1.4**	—	**+2.1**	**+0.1**	**-0.7**	—	—	**-6.5**
CCAligned	**-4.2**	—	**-4.6**	—	**+2.1**	—	**-3.9**	**-8.0**	**-3.6**	**-5.7**
CCMatrix	**+0.9**	—	**-0.5**	—	**-17.5**	—	**+0.0**	**-5.3**	**-1.0**	**+0.0**
ParaCrawl	**+0.2**	—	**-0.3**	—	—	**-0.6**	**-1.9**	—	—	—

Table 7: COMET scores on the FLoRes devtest set for our machine translation systems trained on either the full corpus (first half of the table) or on a randomly selected subset of each corpus that is equal to MaCoCu-V2 (same size) in terms of number of sentence-pairs (second half).

said, the results also show that the MaCoCu corpora offer quite competitive performance to the other web-crawled corpora, despite often being of smaller size.[11] Here, we can already compare the performance of the first and second versions of MaCoCu, for the 6 language pairs shared. In this case, data set size clearly does not tell the whole story: MaCoCu-V2 improves on 4 out of the 6 language pairs, despite actually having less data for three of these pairs. Especially the case of English–Turkish is striking: the parallel corpus goes from 3.8M sentence pairs in MaCoCu-V1 to 1.5M sentence pairs in MaCoCu-V2, but the model trained on the second version actually obtains a 2-point increase in COMET score.

Controlling for size Our main aim in this work is to evaluate the corpora *as they are released*. The amount of filtering performed is a design decision, with less filtering leading to a bigger corpus that would contain lower-quality sentence pairs, which nevertheless could potentially be beneficial for training NMT systems. We have indeed seen this in the first half of Table 7, where the biggest corpus leads to the best scores even if it is not the highest-quality one. However, we are also interested in investigating whether the perceived quality of the corpora (see Table 4 and Table 6) influences NMT performance at all. To this end, we perform a controlled experiment in which we limit the size of each corpus to be the same as that of the smallest corpus: MaCoCu-V2. Its results are shown in the second half of Table 7.

[11] The English-Icelandic MaCoCu data was too small for training a stable NMT system from scratch.

We find that MaCoCu-V2 does rather well in this controlled setting, outperforming all the other corpora in most cases: 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 indeed MaCoCu-V2 has the highest-quality sentence pairs for the purpose of training NMT systems, though the differences are often small.

Takeaway Our takeaway of Section 5 is that data quality (or corpus selection) clearly matters for training NMT systems (as shown in the second half of Table 7), even if a lack of quality can be mitigated by an increased data set size (as shown in the first half of Table 7). We believe these findings are of particular interest to practitioners who have limited computational resources, i.e. they might prefer a small drop in performance if it comes with a large increase in efficiency during training.

6 Conclusion

In this study, we compared four web-crawled parallel corpora across 11 language pairs. We first conduct an intrinsic evaluation, wherein professional translators annotated samples from each corpus. The results here are quite concerning: all corpora contain a substantial amount of noise in the form of segments that either are not in the expected language, are simply not running text, or are not a correct translation of their aligned counterpart. The amount ranges from about 30-40% for the less noisy corpora (MaCoCu and ParaCrawl), to about 60-70% for the noisier corpora (CCMatrix and CCAligned). It is clear that the larger size of the latter two corpora comes at a cost of quality. In follow-up automatic extrinsic evaluation, in which we train NMT systems on the corpora, we show that it is hard to determine which corpus is *better*. The more noisy corpora clearly benefit from their larger size (presumably due to less strict filtering), but when size is controlled for, the cleaner corpora are superior. The complex relationship between the amount of cleaning and the total size of the corpus is an important direction for future research.

7 Limitations and Impact

This section discusses the most relevant limitations of the evaluation described in this work, as well as some relevant aspects regarding the impact of our research related to ethical considerations.

Sample size The manual annotation process, as detailed in Section 4, involved the annotation of samples comprising 200 pairs of segments for each corpus and language pair. This translated to a total annotation of 1,600 to 2,200 samples per corpus. While a larger sample size would undoubtedly improve the robustness of our evaluation, we were constrained by budgetary considerations. However, despite the limitations imposed by budget constraints, the consistency of the evaluation results across all languages is evident, particularly in terms of the quality ranking among the corpora.

Number of corpora While other web-crawled parallel corpora exist that could have been included in this evaluation, we aimed at including some of the most used and larger parallel corpora publicly available. We were especially interested in the MaCoCu corpus which, while not being as popular as the other three corpora, allowed us to compare two versions of the same corpus built on the same data, but one being cleaned more aggressively than the other one.

Languages included The main drawback of including the MaCoCu corpus is that it covers a lower amount of languages than the other three, which constrained the set of languages covered in our experiments. However, including more languages in our evaluation would have been difficult anyway, given the budget limitations for human evaluation. Nevertheless, we encourage the community to extend this research to other corpora and language pairs.

Fine-tuning In the extrinsic evaluation described in Section 5, we train NMT models from scratch. Another option could have been to fine-tune a pre-trained multilingual NMT model on the different corpora. However, we could not find a high-quality pre-trained NMT model publicly available for which none of the four corpora were included in the pre-training process. For this reason, we discarded this option for the evaluation. We do not consider this a problem as our aim is not to train the best-performing models possible, but to run a fair comparison of the corpora.

Human annotators In order to run our human evaluation, we contacted several language service providers (LSPs), and among those capable to provide the required service, we chose the one that provided the best value for cost. Professional translators were hired through the LSP following the corresponding local legislation of their countries of residence.

Random subset In Section 5.2 we evaluate the corpora by taking a randomly selected subset. One could argue that this is not fair: we could also apply some sophisticated data selection method that aims to select the higher quality sentence pairs, giving a potential advantage to the larger corpora. However, our goal in this paper is to evaluate the corpora *as they are released*, as the corpora already went through an (often unclear) cleaning and filtering process we have no control over. We therefore believe that, given this goal, taking a random subset is not unfair. In fact, we would argue that it is actually the only fair way to compare the corpora. Any other method does not evaluate what is released, but a filtered subset of what is released.

Acknowledgments

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. 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 helpful discussions. We thank Raoul Buurke for his feedback on the annotation scheme and for participating in a pilot annotation run. Finally, we thank the reviewers for their comments.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A configurable parallel corpus filtering toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156, Online. Association for Computational Linguistics.

Marta Bañón, Mălina Chichirău, Miquel Esplà-Gomis, Mikel Forcada, Aarón Galiano-Jiménez, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vit Suchomel, Antonio Toral, and Jaume Zaragoza-Bernabeu. 2023. MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 505–506, Tampere, Finland. European Association for Machine Translation.

Yamini Bansal, Behrooz Ghorbani, Ankush Garg, Biao Zhang, Colin Cherry, Behnam Neyshabur, and Orhan Firat. 2022. Data scaling laws in nmt: The effect of noise and architecture. In *International Conference on Machine Learning*, pages 1466–1482. PMLR.

Eduard Barbu, Carla Parra Escartín, Luisa Bentivogli, Matteo Negri, Marco Turchi, Constantin Orasan, and Marcello Federico. 2016. The first automatic translation memory cleaning shared task. *Machine Translation*, 30:145–166.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Eleftheria Briakou and Marine Carpuat. 2021. Beyond noise: Mitigating the impact of fine-grained semantic divergences on neural machine translation. 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 7236–7249, Online. Association for Computational Linguistics.

Marine Carpuat, Yogarshi Vyas, and Xing Niu. 2017. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 69–79, Vancouver. Association for Computational Linguistics.

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.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 261–266, Florence, Italy. 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.

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.

Lukas Edman, Gabriele Sarti, Antonio Toral, Gertjan van Noord, and Arianna Bisazza. 2024. Are character-level translations worth the wait? comparing ByT5 and mT5 for machine translation. *Transactions of the Association for Computational Linguistics*, 11:392–410.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Christian Herold, Jan Rosendahl, Joris Vanvinckenroye, and Hermann Ney. 2022. Detecting various types of noise for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2542–2551, Dublin, Ireland. Association for Computational Linguistics.

Kaiyu Huang, Peng Li, Jin Ma, Ting Yao, and Yang Liu. 2023. Knowledge transfer in incremental learning for multilingual neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15286–15304, Toronto, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, Melbourne, Australia.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander Rush. 2018. OpenNMT: Neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 177–184, Boston, MA. Association for Machine Translation in the Americas.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. 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.

Taku Kudo and John Richardson. 2018. 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*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Gema Ramírez-Sánchez, Marta Bañón, Jaume Zaragoza-Bernabeu, and Sergio Ortiz Rojas. 2022. Human evaluation of web-crawled parallel corpora for machine translation. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 32–41, Dublin, Ireland. Association for Computational Linguistics.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Machel Reid and Mikel Artetxe. 2022. PARADISE: Exploiting parallel data for multilingual sequence-to-sequence pretraining. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 800–810, Seattle, United States. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. 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)*, pages 6490–6500, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based sentence alignment of parallel texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 175–182, Riga, Latvia. Northern European Association for Language Technology (NEALT).

Steve Sloto, Brian Thompson, Huda Khayrallah, Tobias Domhan, Thamme Gowda, and Philipp Koehn. 2023. Findings of the WMT 2023 shared task on parallel data curation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 95–102, Singapore. Association for Computational Linguistics.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023. Filtering matters: Experiments in filtering training sets for machine translation. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 588–600, Tórshavn, Faroe Islands. University of Tartu Library.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, Long Beach, USA.

Minghao Wu, Yufei Wang, George Foster, Lizhen Qu, and Gholamreza Haffari. 2024. Importance-aware data augmentation for document-level neural machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 740–752, St. Julian's, Malta. Association for Computational Linguistics.

Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner goes neural. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 824–831, Marseille, France. European Language Resources Association.

A Annotation instructions

The example parallel texts that were actually given to annotators are shown in the table below (see next page). The first and second sentence are shown both in English, so we could use the same instructions across languages, but the annotators were shown the second sentence in the corresponding language. Annotators were shown both the hierarchy and this overview prior to annotation and could always easily refer back to both instructions.

B BLEU results

Table 8 shows the exact same evaluation as was performed in Table 7, but now with BLEU scores instead of COMET. The conclusions are largely the same, except that MaCoCu-V1 now outperforms MaCoCu-V2 for training a full-size Turkish model from scratch (25.0 and 24.4 BLEU versus 83.2 and 85.2 COMET).

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 (CB):

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 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 (LQ):

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 (PR):

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
Sent 1 or 2: 1500mmx3000mm hot sale and good price fiber laser cutting machine
Sent 1 or 2: Photo White-spotted Puffer (Arothron hispidus), Spotted, Aquarium Fish

	en-bg	**en-bs**	**en-hr**	**en-is**	**en-mk**	**en-mt**	**en-sl**	**en-sq**	**en-sr**	**en-tr**
Full size:										
MaCoCu-V2	35.3	21.4	28.1	—	21.3	35.0	25.8	25.3	30.1	24.4
MaCoCu-V1	34.7	—	26.3	—	23.6	36.0	25.4	—	—	25.0
CCAligned	37.5	2.3	26.4	18.1	23.5	—	24.3	23.9	28.2	27.1
CCMatrix	42.8	—	28.2	24.3	33.9	—	29.9	31.2	31.5	31.6
ParaCrawl	36.4	—	31.2	—	—	34.2	25.0	—	—	—
Equal size to MaCoCu-V2										
MaCoCu-V2	35.3	—	28.1	—	21.3	35.0	25.8	25.3	30.1	24.4
MaCoCu-V1	**-0.4**	—	**-1.8**	—	**+0.7**	**+0.2**	**-0.6**	—	—	**-5.1**
CCAligned	**-2.7**	—	**-2.9**	—	**+0.9**	—	**-1.7**	**-5.8**	**-1.9**	**-3.1**
CCMatrix	**+2.7**	—	**+0.1**	—	**-10.5**	—	**+1.3**	**-4.6**	**-0.1**	**-1.3**
ParaCrawl	**+1.1**	—	**-0.9**	—	—	**-0.8**	**-0.8**	—	—	—

Table 8: **BLEU** scores on the FLoRes devtest set for our machine translation systems trained on either the full corpus (first half of the table) or on a randomly selected subset of each corpus that is equal to MaCoCu-V2 (same size) in terms of number of sentence-pairs (second half).