Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets

Julia Kreutzer[1,2], **Isaac Caswell**[3], **Lisa Wang**[3,4], **Ahsan Wahab**[5,47], **Daan van Esch**[6], **Nasanbayar Ulzii-Orshikh**[7], **Allahsera Tapo**[8,9], **Nishant Subramani**[10,11], **Artem Sokolov**[4], **Claytone Sikasote**[12,13], **Monang Setyawan**[14], **Supheakmungkol Sarin**[14], **Sokhar Samb**[15,16], **Benoît Sagot**[17], **Clara Rivera**[18], **Annette Rios**[19], **Isabel Papadimitriou**[20], **Salomey Osei**[21,22], **Pedro Ortiz Suarez**[17,23], **Iroro Orife**[10,24], **Kelechi Ogueji**[2,25], **Andre Niyongabo Rubungo**[26,27], **Toan Q. Nguyen**[28], **Mathias Müller**[19], **André Müller**[19], **Shamsuddeen Hassan Muhammad**[29,30], **Nanda Muhammad**[30], **Ayanda Mnyakeni**[31], **Jamshidbek Mirzakhalov**[5,32], **Tapiwanashe Matangira**[33], **Colin Leong**[10], **Nze Lawson**[14], **Sneha Kudugunta**[3], **Yacine Jernite**[10,34], **Mathias Jenny**[19], **Orhan Firat**[3,5], **Bonaventure F. P. Dossou**[35,36], **Sakhile Dlamini**[14], **Nisansa de Silva**[37], **Sakine Çabuk Ballı**[19], **Stella Biderman**[38], **Alessia Battisti**[19], **Ahmed Baruwa**[10,39], **Ankur Bapna**[3], **Pallavi Baljekar**[1], **Israel Abebe Azime**[40,41], **Ayodele Awokoya**[29,42], **Duygu Ataman**[19,43], **Orevaoghene Ahia**[10,44], **Oghenefego Ahia**[14], **Sweta Agrawal**[45], **Mofetoluwa Adeyemi**[29,46]

[1]Google Research, Canada, [2]Masakhane NLP, USA, [3]Google Research, USA, [4]Google Research, Germany, [5]Turkic Interlingua, [6]Google Research, The Netherlands, [7]Haverford College, USA, [8]Masakhane NLP, Mali, [9]RobotsMali, Mali, [10]Masakhane NLP, USA, [11]Allen Institute for Artificial Intelligence, USA, [12]Masakhane NLP, Zambia, [13]University of Zambia, Zambia, [14]Google, USA, [15]Masakhane NLP, Senegal, [16]AIMS-AMMI, Senegal, [17]Inria, France, [18]Google Research, UK, [19]University of Zurich, Switzerland, [20]Stanford University, USA, [21]Masakhane NLP, Ghana, [22]Kwame Nkrumah University of Science and Technology, Ghana, [23]Sorbonne Université, France, [24]Niger-Volta LTI, USA, [25]University of Waterloo, Canada, [26]Masakhane NLP, Spain, [27]Universitat Politècnica de Catalunya, Spain, [28]University of Notre Dame, USA, [29]Masakhane NLP, Nigeria, [30]Bayero University Kano, Nigeria, [31]Google, South Africa, [32]University of South Florida, USA, [33]Google, Canada, [34]Hugging Face, USA, [35]Masakhane NLP, Germany, [36]Jacobs University Bremen, Germany, [37]University of Moratuwa, Sri Lanka, [38]EleutherAI, USA, [39]Obafemi Awolowo University, Nigeria, [40]Masakhane NLP, Ethiopia, [41]AIMS-AMMI, Ethiopia, [42]University of Ibadan, Nigeria, [43]Turkic Interlingua, Switzerland, [44]Instadeep, Nigeria, [45]University of Maryland, USA, [46]Defence Space Administration Abuja, Nigeria, [47]University of South Florida, USA

Abstract

With the success of large-scale pre-training and multilingual modeling in Natural Language Processing (NLP), recent years have seen a proliferation of large, Web-mined text datasets covering hundreds of languages. We manually audit the quality of 205 language-specific corpora released with five major public datasets (CCAligned, ParaCrawl, WikiMatrix, OSCAR, mC4). Lower-resource corpora have systematic issues: At least 15 corpora have no usable text, and a significant fraction contains less than 50% sentences of acceptable quality. In addition, many are mislabeled or use nonstandard/ambiguous language codes. We demonstrate that these issues are easy to detect even for non-proficient speakers, and supplement the human audit with automatic analyses. Finally, we recommend techniques to evaluate and improve multilingual corpora and discuss potential risks that come with low-quality data releases.

1 Introduction

Access to multilingual datasets for NLP research has vastly improved over the past years. A variety of Web-derived collections for hundreds of languages is available for anyone to download, such as ParaCrawl (Esplà et al., 2019; Bañón et al., 2020), WikiMatrix (Schwenk et al., 2021),

Transactions of the Association for Computational Linguistics, vol. 10, pp. 50–72, 2022. https://doi.org/10.1162/tacl_a_00447
Action Editor: Sebastian Padó. Submission batch: 6/2021; Revision batch: 9/2021; Published 1/2022.

CCAligned (El-Kishky et al., 2020), OSCAR (Ortiz Suárez et al., 2019; Ortiz Suárez et al., 2020), and several others. These have in turn enabled a variety of highly multilingual models, like mT5 (Xue et al., 2021), M2M-100 (Fan et al., 2020), and M4 (Arivazhagan et al., 2019).

Curating such datasets relies on the Web sites giving clues about the language of their contents (e.g., a language identifier in the URL) and on automatic language classification (LangID). It is commonly known that these automatically crawled and filtered datasets tend to have overall lower quality than hand-curated collections (Koehn et al., 2020), but their quality is rarely measured directly, and is rather judged through the improvements they bring to downstream applications (Schwenk et al., 2021).

Building NLP technologies with automatically crawled datasets is promising. This is especially true for low-resource languages, because data scarcity is one of the major bottlenecks for deep learning approaches. However, there is a problem: There exists very little research on evaluating both data collections and automatic crawling and filtering tools for low-resource languages. As a result, although many low-resource languages are covered by the latest multilingual crawl data releases, their quality and thus usability is unknown.

To shed light on the quality of data crawls for the lowest resource languages, we perform a manual data audit for 230 per-language subsets of five major crawled multilingual datasets:[1] CCAligned (El-Kishky et al., 2020), ParaCrawl (Esplà et al., 2019; Bañón et al., 2020), WikiMatrix (Schwenk et al., 2021), OSCAR (Ortiz Suárez et al., 2019; Ortiz Suárez et al., 2020), and mC4 (Xue et al., 2021). We propose solutions for effective, low-effort data auditing (Section 4), including an error taxonomy. Our quantitative analysis reveals surprisingly low amounts of valid in-language data, and identifies systematic issues across datasets and languages. In addition, we find that a large number of datasets is labeled with nontransparent or incorrect language codes (Section 5). This leads us to reflect on the potential harm of low-quality data releases for low-resource languages (Section 6), and provide a set of recommendations for future multilingual data releases (Section 7).

[1]Annotations are available for download (last accessed: 12 Oct 2021).

2 Related Work

Corpora collected by web crawlers are known to be noisy (Junczys-Dowmunt, 2019; Luccioni and Viviano, 2021). In highly multilingual settings, past work found that web-crawls of lower-resource languages have serious issues, especially with segment-level LangID (Caswell et al., 2020). Cleaning and filtering web-crawls can boost general language modeling (Gao et al., 2020; Brown et al., 2020; Raffel et al., 2020) and downstream task performance (Moore and Lewis, 2010; Rarrick et al., 2011; Xu and Koehn, 2017; Khayrallah and Koehn 2018; Brown et al., 2020).

As the scale of ML research grows, it becomes increasingly difficult to validate automatically collected and curated datasets (Biderman and Scheirer, 2020; Birhane and Prabhu, 2021; Bender et al., 2021). Several works have focused on advancing methodologies and best practices to address these challenges. Bender and Friedman (2018) introduced data statements, a documentary framework for NLP datasets that seeks to provide a universal minimum bar for dataset description. Similar work has been done on systematizing documentation in other areas in data science and machine learning, including work focusing on online news (Kevin et al., 2018), data ethics (Sun et al., 2019), and data exploration (Holland et al., 2018), as well as generalist work such as Gebru et al. (2018). Data quality is also implicitly documented by successes of filtering methods. There is a large literature on filtering data for various NLP tasks, for example, Axelrod et al., 2011, Moore and Lewis (2010), Rarrick et al., 2011, Wang et al. (2018), Kamholz et al. (2014), Junczys-Dowmunt (2018), and Caswell et al., 2020.

Closest to our work is the analysis of a highly multilingual (non-publicly available) web-crawl and LangID-related quality issues by (Caswell et al., 2020). They perform a brief analysis of the quality of OSCAR with the focus only on the presence of in-language content. Dodge et al. (2021) automatically documented and analyzed the contents and sources of C4 (Raffel et al., 2020), the English counterpart of mC4, which surfaced

	Parallel			Monolingual	
	CCAligned	**ParaCrawl v7.1**	**WikiMatrix**	**OSCAR**	**mC4**
#languages	137	41	85	166	101
Source	CC 2013–2020	selected Web sites	Wikipedia	CC 11/2018	CC all
Filtering level	document	sentence	sentence	document	document
Langid	FastText	CLD2	FastText	FastText	CLD3
Alignment	LASER	Vec/Hun/BLEU-Align	LASER	–	–
Evaluation	TED-6	WMT-5	TED-45	POS/DEP-5	XTREME

Table 1: Comparison of parallel and monolingual corpora extracted from web documents, including their downstream evaluation tasks. All parallel corpora are evaluated for machine translation (BLEU). TED-6: da, cr, sl, sk, lt, et; TED-45: 45-language subset of (Qi et al., 2018); WMT-5: cs, de, fi, lv, ro. POS/DEP-5: part-of-speech labeling and dependency parsing for bg, ca, da, fi, id.

the presence of machine-translated contents and NLP benchmark data.

3 Multilingual Corpora

Table 1 provides an overview of the corpora of interest in this work. We selected the corpora for their multilinguality and the inclusion of understudied languages in NLP. With the exception of WikiMatrix and ParaCrawl, all corpora are derived from CommonCrawl (CC).[2]

CCAligned (El-Kishky et al., 2020) is a parallel dataset built off 68 CC snapshots. Documents are aligned if they are in the same language according to FastText LangID (Joulin et al., 2016, 2017), and have the same URL but for a differing language code. These alignments are refined with cross-lingual LASER embeddings (Artetxe and Schwenk, 2019). For sentence-level data, they split on newlines and align with LASER, but perform no further filtering. Human annotators evaluated the quality of document alignments for six languages (de, zh, ar, ro, et, my) selected for their different scripts and amount of retrieved documents, reporting precision of over 90%. The quality of the extracted parallel sentences was evaluated in a machine translation (MT) task on six European (da, cr, sl, sk, lt, et) languages of the TED corpus (Qi et al., 2018), where it compared favorably to systems built on crawled sentences from WikiMatrix and ParaCrawl v6.

Multilingual C4 (mC4) (Xue et al., 2021) is a document-level dataset used for training the mT5 language model. It consists of monolingual text in 101 languages and is generated from 71 CC snapshots. It filters out pages that contain less than three lines of at least 200 characters and pages that contain bad words.[3] Since this is a document-level dataset, we split it by sentence and deduplicate it before rating. For language identification, it uses CLD3 (Botha et al., 2017),[4] a small feed-forward neural network that was trained to detect 107 languages. The mT5 model pre-trained on mC4 is evaluated on 6 tasks of the XTREME benchmark (Hu et al., 2020) covering a variety of languages and outperforms other multilingual pre-trained language models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020).

OSCAR (Ortiz Suárez et al., 2019; Ortiz Suárez et al., 2020) is a set of monolingual corpora extracted from CC snapshots, specifically from the plain text *WET* format distributed by CC which removes all the HTML tags and converts the text to UTF-8. It is deduplicated and follows the approach by Grave et al. (2018) of using FastText LangID (Joulin et al., 2016, 2017) on a line-level.[5] No other filtering was applied. For five languages (bg, ca, da, fi, id), OSCAR was used by its original authors to train language models which were then evaluated on parsing and POS tagging (Ortiz Suárez et al., 2020). OSCAR has also been used in independent studies to train monolingual or multilingual language models (ar, as, bn, de, el, fr, gu, he, hi, kn, ml, mr, nl, or, pa, ro, ta, te) and subsequently evaluate them on various downstream tasks (Antoun et al., 2021;

[2] http://commoncrawl.org/.

[3] https://github.com/LDNOOBW/.

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

[5] https://fasttext.cc/docs/en/language-identification.html.

Kakwani et al., 2020; Wilie et al., 2020; Chan et al., 2020; Koutsikakis et al., 2020; Martin et al., 2020; Chriqui and Yahav, 2021; Seker et al., 2021; Delobelle et al., 2020; Dumitrescu et al., 2020; Masala et al., 2020).

ParaCrawl v7.1 is a parallel dataset with 41 language pairs primarily aligned with English (39 out of 41) and mined using the parallel-data-crawling tool Bitextor (Esplà et al., 2019; Bañón et al., 2020) which includes downloading documents, preprocessing and normalization, aligning documents and segments, and filtering noisy data via Bicleaner.[6] ParaCrawl focuses on European languages, but also includes 9 lower-resource, non-European language pairs in v7.1. Sentence alignment and sentence pair filtering choices were optimized for five languages (`mt`, `et`, `hu`, `cs`, `de`) by training and evaluating MT models on the resulting parallel sentences. An earlier version (v5) was shown to improve translation quality on WMT benchmarks for `cs`, `de`, `fi`, `lv`, `ro`.

WikiMatrix (Schwenk et al., 2021) is a public dataset containing 135M parallel sentences in 1620 language pairs (85 languages) mined from Wikipedia. Out of the 135M parallel sentences, 34M are aligned with English. The text is extracted from Wikipedia pages, split into sentences, and duplicate sentences are removed. FastText LangID is used before identifying bitext with LASER's distance-based mining approach. The margin threshold is optimized by training and evaluating downstream MT models on four WMT benchmarks (`de-en`, `de-fr`, `cs-de`, `cs-fr`). The final dataset is used to train translation models that are then evaluated by automatically measuring the quality of their translations against human translations of TED talks in 45 languages, with highest quality for translations between English and, for example, `pt`, `es`, `da`, and lowest for `sr`, `ja`, `mr`, `zh_TW`. In the audit we focus on language pairs with English on one side.

4 Auditing Data Quality

None of the above datasets has been evaluated for quality on the sentence level (exception: several languages in ParaCrawl v3), and downstream evaluations are centered around a small fraction of higher-resource languages. This is insufficient for drawing conclusions about the quality of individual or aligned sentences, and about the entirety of languages. In addition, there might be a publication bias preventing negative results with any of the above corpora with lower quality being published.

To close this gap, we conduct a human data quality audit focused on the lowest-resource and most under-evaluated languages, but also covering mid- and high-resource languages for comparison.

4.1 Auditing Process

Participants We recruited 51 volunteers from the NLP community, covering about 70 languages with proficient language skills.[7] Each sentence is annotated by one rater. To verify our hypothesis that those annotations can largely done by non-native speakers, we repeat a set of language expert annotations by a non-expert, and measure the accuracy of the non-expert.

Sample Selection For each language in each dataset, we took a random sample of 100 lines, which may be anywhere from single words to short paragraphs depending on segmentation. We manually annotated them according to the error taxonomy described below. For WikiMatrix and CCAligned, we selected those languages that are paired with English, and for ParaCrawl, we also included those paired with Spanish (''total'' counts in Table 3). We did not annotate all languages, but focused on the ones with the least number of sentences in each dataset (at least the smallest 10) and languages for which we found proficient speakers. Since we annotate the same maximum number of sentences[8] across all chosen languages regardless of their total number of sentences, the annotated samples are not an unbiased sample from the whole dataset.

Non-expert Labeling Strategies Although many of the volunteers were familiar with the languages in question or spoke related languages, in cases where no speaker of a relevant language could be found, volunteers used dictionaries and Internet search to form educated guesses. We discuss this deeper in Appendix C to highlight how much of this low-resource focused evaluation

[6] `https://github.com/bitextor/bicleaner`.

[7] This surprisingly high number comes in part because there are many closely related languages, e.g., one person may be proficient enough to rate many different Slavic or Turkic languages even if only one is their native language.

[8] Some languages had fewer than 100 sentences.

Correct Codes	
C: *Correct translation, any*	Combined label for CC, CB, CS
CC: *Correct translation, natural sentence*	
en The Constitution of South Africa	nso Molaotheo wa Rephabliki ya Afrika Borwa
en Transforming your swimming pool into a pond	de Umbau Ihres Swimmingpools zum Teich
CB: *Correct translation, Boilerplate or low quality*	
en Reference number: 13634	ln Motango ya référence: 13634
en Latest Smell Stop Articles	fil Pinakabagong mga Artikulo Smell Stop
CS: *Correct translation, Short*	
en movies, dad	it cinema, papà
en Halloween - without me	ay Hallowen –janiw nayampejj
Error Codes	
X: *Incorrect translation, but both correct languages*	
en A map of the arrondissements of Paris	kg Paris kele mbanza ya kimfumu ya Fwalansa.
en Ask a question	tr Soru sor Kullanıma göre seçim
WL: *Source OR target wrong language, but both still linguistic content*	
en The ISO3 language code is zho	zza Táim eadra bracach mar bhionns na frogannaidhe.
en Der Werwolf—sprach der gute Mann,	de des Weswolfs, Genitiv sodann,
NL: *Not a language: at least one of source and target are not linguistic content*	
en EntryScan 4 _	tn TSA PM704 _
en organic peanut butter	ckb ◆◆◆◆◆◆◆

Table 2: Annotation codes for parallel data with sentence pair examples. The language code before each sentence indicates the language it is supposed to be in.

can actually be done by non-proficient speakers with relatively low effort. In general, we aim to find an upper bound on quality, so we encouraged annotators to be forgiving of translation mistakes when the overall meaning of the sentence or large parts thereof are conveyed, or when most of the sentence is in the correct language.

Effort The individual effort was dependent on the quality and complexity of the data, and on the annotator's knowledge of the language(s), for example, it took from less than two minutes for an English native speaker to pass through 100 well-formed English sentences (or similarly to annotate languages with 0% in-language sentences), to two hours of ''detective work'' for well-formed content in languages for an annotator without familiarity.

Taxonomy In order to quantify errors, we developed a simple error taxonomy. Sentences and sentence pairs were annotated according to a simple rubric with error classes of Incorrect Translation (X, excluded for monolingual data), Wrong Language (WL), and Non-Linguistic Content (NL). Of correct sentences (C), we further mark single words or phrases (CS) and boilerplate contents (CB). In addition, we asked annotators to flag offensive or pornographic content. Table 2 provides examples for parallel data, and Appendix B contains detailed annotation instructions.

4.2 Human Audit Results

Interpretation of Results For each language, we compute the percentage of each label within the 100 audited sentences. Then, we either aggregate the labels across languages with equal weights (macro-average), or weight them according to their presence in the overall dataset (micro-average). Results are shown in Table 3. The statistics for the correct codes (CC, CB, CS) are combined as C. The number of languages, the numbers of sentences per language, and the choice of languages differ across datasets, both

		Parallel			Monolingual	
		CCAligned	**ParaCrawl v7.1**	**WikiMatrix**	**OSCAR**	**mC4**
	#langs audited / total	65 / 119	21 / 38	20 / 78	51 / 166	48 / 106
	%langs audited	54.62%	55.26%	25.64%	30.72%	44.44%
	#sents audited / total	8037 / 907M	2214 / 521M	1997 / 95M	3517 / 8.4B	5314 / 8.5B
	%sents audited	0.00089%	0.00043%	0.00211%	0.00004%	0.00006%
macro	C	29.25%	76.14%	23.74%	87.21%	72.40%
	X	29.46%	19.17%	68.18%	–	–
	WL	9.44%	3.43%	6.08%	6.26%	15.98%
	NL	31.42%	1.13%	1.60%	6.54%	11.40%
	offensive	0.01%	0.00%	0.00%	0.14%	0.06%
	porn	5.30%	0.63%	0.00%	0.48%	0.36%
micro	C	53.52%	83.00%	50.58%	98.72%	92.66%
	X	32.25%	15.27%	47.10%	–	–
	WL	3.60%	1.04%	1.35%	0.52%	2.33%
	NL	10.53%	0.69%	0.94%	0.75%	5.01%
	offensive	0.00%	0.00%	0.00%	0.18%	0.03%
	porn	2.86%	0.33%	0.00%	1.63%	0.08%
	#langs =0% C	7	0	1	7	0
	#langs <50% C	44	4	19	11	9
	#langs >50% NL	13	0	0	7	1
	#langs >50% WL	1	0	0	3	4

Table 3: Averages of sentence-level annotations across datasets and selected languages. Macro-avg: Each language is weighted equally in the aggregation, regardless of its size. Micro-avg: Each label is weighted by the fraction of sentences for that language in the overall annotated corpus, i.e., the annotations for higher-represented languages are upweighted, and annotations for lower-represented languages are downweighted. The bottom rows contain the number of languages that have 0% labeled C, etc. Note that these are not true expectations since the languages audited were not randomly sampled.

in the original release and in the selection for our audit, so the comparison of numbers across datasets has to be taken with a grain of salt. Since the numbers are based on a small sample of sentences that were partially annotated by non-experts, the error statistics are only rough estimates. Our audit captures a decent ratio of languages (25–55%, 2nd row in Table 3), but only a tiny fraction of the overall number of sentences (0.00004–0.002%). When we speak of ‘‘low-’’ and ‘‘high’’-resource languages, we mean languages with smaller or larger representation in the datasets at hand. When reporting language-specific results we use the original language identifiers of the datasets.

Which Datasets Have Quality Issues? The macro-averaged results show that the ratio of correct samples (C) ranges from 24% to 87%, with a large variance across the five audited datasets. Particularly severe problems were found in CCAligned and WikiMatrix, with 44 of the 65 languages that we audited for CCAligned containing under 50% correct sentences, and 19 of the 20 in WikiMatrix. In total, 15 of the 205 language-specific samples (7.3%) contained not a single correct sentence. For the parallel datasets we are also interested in the quantity of misaligned/mistranslated sentences (X). For WikiMatrix, two-thirds of the audited samples were on average misaligned. We noticed that sentences were often similar in structure, but described different facts (see Table 6). This might originate from the nature of the underlying Wikipedia articles, since they are often comparable rather than parallel (Schwenk et al., 2021).

Figure 1 illustrates per-corpus correctness more completely, showing for each dataset what percent of audited corpora are under each possible threshold of correctness.

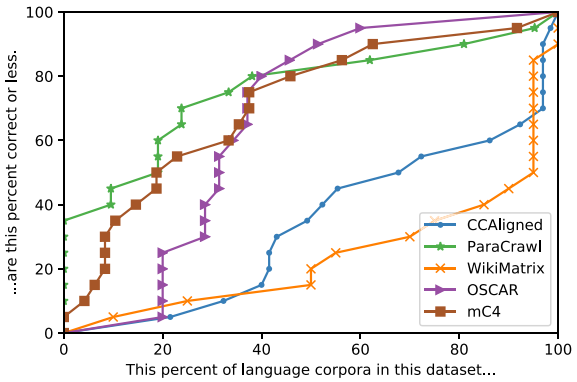


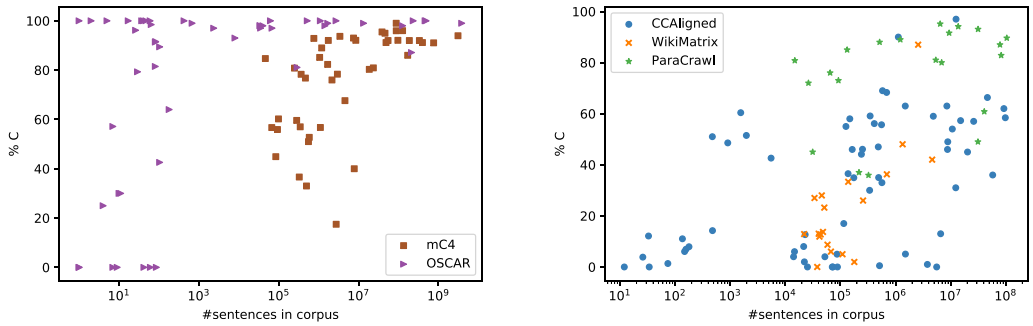
Figure 1: Fraction of languages in each dataset below a given quality threshold (percent correct).

Why Haven't These Problems Been Reported Before? The findings above are averaged on a per-language basis (i.e., macro-average), and therefore give low and high-resource languages equal weight. If we instead estimate the quality on a per-sentence basis (i.e., down-weight lower-resource languages in the computation of the average), the numbers paint a more optimistic picture (''micro'' block in Table 3). This is especially relevant for the monolingual datasets because they contain audits for English, which makes up for 43% of all sentences in OSCAR and 36% in mC4. To illustrate the effect of this imbalance: A random sample from the entire mC4 dataset with over 63% chance will be from one of the 8 largest languages (`en`, `ru`, `es`, `de`, `fr`, `it`, `pt`, `pl`, >100M sentences each), of which all have near perfect quality. Analogously, evaluation and tuning of web mining pipelines and resulting corpora in downstream applications focused largely on higher-resource languages (Section 3), so the low quality of underrepresented languages might go unnoticed if there is no dedicated evaluation, or no proficient speakers are involved in the curation (Nekoto et al., 2020).

How Much Content is Nonlinguistic or in the Wrong Language? Nonlinguistic content is a more common problem than wrong-language content. Among the parallel datasets, CCAligned contains the highest percentage of nonlinguistic content, at 31.42% on average across all rated corpora, and also the highest percent of wrong-language content, at 9.44%. Among the monolingual datasets, mC4 contains the highest ratio both of sentences in incorrect languages (15.98% average) and nonlinguistic content (11.40% average), with 4 of the 48 audited languages having more than 50% contents in other languages. The low amount of wrong language in ParaCrawl shows the benefits of selecting domains by the amount in-language text, but the dataset also covers the smallest amount of languages. The low ratio of wrong language samples in OSCAR may reflect the success of line-level LangID filtering. These numbers provide evidence that more research in LangID could improve the overall quality, especially with respect to nonlinguistic content.

Which Languages Got Confused? The languages that were confused were frequently related higher-resource languages. However, there were also a significant number of ''out-of-model cousin'' cases, where languages not supported by the LangID model ended up in a similar-seeming language. For instance in mC4, much of the Shona (`sn`, Bantu language spoken in Zimbabwe and Mozambique) corpus is actually Kinyarwanda (`rw`, Bantu language spoken in mostly in Rwanda and Uganda)—and, peculiarly, much of the Hawaiian (`haw`, Polynesian language spoken in Hawaii) is actually Twi (`tw`/`ak`, Central Tano language spoken mostly in Ghana).

Do Low-resource Languages Have Lower Quality? Low-resource datasets tend to have lower human-judged quality. The Spearman rank correlation between quality (%`C`) and size is positive in all cases. The trend is strongest for mC4 ($r = 0.66$), and gradually declines for CCAligned ($r = 0.53$), WikiMatrix ($r = 0.49$), ParaCrawl ($r = 0.43$), and OSCAR ($r = 0.37$). Figure 2 compares the number of sentences for each language against the proportion of correct sentences: Not all higher-resource languages ($> 10^6$ sentences) have high quality, in particular for CCAligned (e.g., Javanese (`en-jv_ID`) with 5%`C`, or Tagalog (`en-tl_XX`) with 13%`C`). For mid-resource languages (10^4–10^6 sentences) the picture is inconclusive, with some languages having high quality, and others having extremely low quality, even within the same datasets (e.g., Urdu in CCAligned `en-ur_PK` has 100%`C` vs. its romanized counterpart `en-ur_PK_rom` 0.5% `C`). For individual error codes trends are less clear (not depicted).



(a) Monolingual corpora

(b) Parallel corpora

Figure 2: Percentage of sentences labeled as correct vs. log N sentences for all audited languages.

	es_XX	bm_ML	yo_NG	tr_TR	ku_TR	zh_CN	af_ZA	jv_ID	zh_TW	it_IT	**mean**
Acc-6	0.58	0.73	0.41	0.45	0.43	0.55	0.65	0.55	0.46	0.55	0.66
Acc-4	0.77	0.73	0.60	0.55	0.56	0.72	0.72	0.57	0.58	0.66	0.72
Acc-2	0.91	0.96	0.72	0.64	0.71	0.79	0.77	0.92	0.81	0.69	0.79

Table 4: Rater evaluation for a subset of audits from **CCAligned** (translated from English) measured by the accuracy (Acc-n) of annotations by non-proficient speaker against annotations by proficient speakers.

Which Languages Have the Lowest Quality? Across datasets we observe that the quality is particularly poor for languages that are included in romanized script (_rom/_latn), but are more commonly written in other scripts (e.g., Urdu (ur), Japanese (ja), Arabic (ar)). These are not transliterations of other scripts, but mostly contain non-linguistic material or wrong languages (e.g., the romanized Japanese corpus in mC4 (ja_latn) contains Spanish, French, English, Portuguese, among others). In terms of geography, the poorest quality is found for African languages (Bambara (bm), Fula (ff), Kikongo (kg), Luganda (lg), Lingala (ln), Norther Sotho (nso), Oromo (om), Shona (sn), Somali (so), Tswana (tn), Wolof (wo)), minority languages in Europe and the Middle East that are closely related to higher-resource languages (Azerbaijani (az-IR), North Frisian (frr), Neapolitan (nap), Silesian (szl), Zaza (zza)), lesser spoken Chinese languages sharing a script with Mandarin (Yue (yue), Wu (wuu)), four major Austronesian (Central Bikol (bcl), Chavacano (cbk), Javanese (jv), Sundanese (su)), and some South-Asian languages, in particular Sinhala (si). Appendix D contains the detailed per-language statistics for all corpora.

What Is the Incidence of Offensive and Pornographic Content? Overall, the sampled sentences did not contain a large amount of offensive content. However, there were notable amounts of pornographic content ($> 10\%$) found in CCAligned for 11 languages.

Annotation Quality For a subset of audited languages from CCAligned and OSCAR we measure the accuracy (Acc) of the labels assigned by non-proficient speakers against the labels assigned by proficient speakers for all audited sentences. This can be understood as a directed measure of annotator agreement for the special case where one rater is an expert and the other is not. Results for varying label granularity are reported in Tables 4 and 5. For $n = 6$ all classes of the taxonomy were distinguished, for $n = 4$ the C subclasses were combined, and for $n = 2$ it is binary decision between C and the rest of the error classes. With the full 6-class taxonomy (Acc-6) we find a mean accuracy of 0.66 for CCAligned audits, and 0.98 for OSCAR audits. With a binary taxonomy (Acc-2) distinguishing C from the rest, the accuracy further increases to 0.79 for CCAligned. This provides strong evidence that good quality

	tyv	rm	bar	eml	zh	la	**mean**
Acc-6	1.0	0.98	1.0	1.0	0.86	1.0	0.98
Acc-4	1.0	1.0	1.0	1.0	0.87	1.0	0.98
Acc-2	1.0	1.0	1.0	1.0	0.87	1.0	0.98

Table 5: Rater evaluation for a subset of audits from **OSCAR** measured by the accuracy (Acc-n) of annotations by non-proficient speaker against annotations by proficient speakers.

annotations are not limited to those proficient in a language.

However, the significant drop of accuracy for finer-grained labels hints at that our taxonomy can be further improved, especially for parallel sentences. The error taxonomy lacks at least one category of error, namely, ''correct/in-language but unnatural''. Similarly, the definition of ''correct-short'' and ''correct-boilerplate'' were not understood equally by all annotators and the concept of ''correct-short'' has potential issues for agglutinative languages like Turkish. Finally, it was unclear what to do with related dialects, for example, when a sentence is ''almost correct but wrong dialect'' or when it is unclear which dialect a sentence belongs to. We recommend including these categories for future audits.

4.3 Automatic Filtering

Given the frequency of `WL` and `NL` annotations, it might be tempting to use open-source LangID models to post-filter data on a per-sentence(-pair) level, as OSCAR does. Unfortunately, this turns out to have its own issues.

Sentence-level *n*-gram LangID Filtering We classify all sentence pairs of CCAligned with CLD3, an *n*-gram based LangID model. By comparing its predictions to the audit labels, we evaluate its quality on the subset of annotated samples: The classifier should detect both correct languages when the pair is annotated as `C` and `X`, and should detect incorrect languages in the pair when `WL` and `NL`. On this task, the CLD3 classifier achieves an average precision of only 40.6%.

Sentence-level Transformer LangID Filtering *n*-gram LangID models like CLD3 have known problems. However, Caswell et al. (2020) demonstrate that semi-supervised Transformer-based LangID models strongly out-perform them. We train a comparable Transformer-based LangID model and apply it to our annotated CCAligned data. We find that filtering noisy corpora ($<$ 50% correct) on LangID for both source and target leads to gains in median precision, rising from 13.8% pre-filter to 43.9% post-filter. However, this comes at a steep cost of 77.5% loss in recall. The biggest winners were Lingala, whose precision climbs from 8% to 80%, and Oromo, which soars from 2% to 33% in-language. Both of these, however, come at the cost of losing 50% of the correct in-language sentences, being reduced from 22k sentences to 3k and 1k sentences, respectively, which would likely be too small for building downstream models. The moral is that, at least at the current stage, there is no one-size-fits-all approach for sentence-level LangID filtering.

5 Dataset Mis-labeling

Standardized and unambiguous representations of language codes are important for practical data use and exchange. The standard used by most academic and industry applications is BCP-47 (Phillips and Davis, 2005), which builds off the two-letter ISO639-2 codes and three-letter ISO639-3 codes, but also allows for adding subtags for scripts (e.g., Hindi in Latin script: `hi-Latn`) or regional varieties (e.g., French spoken in Canada: `fr-CA`). It would enhance transparency and interoperability if adopted consistently, especially with growing language diversity in NLP.

We find a variety of errors and inconsistencies in language code usage, ranging from serious mislabelings to small transgressions against standard conventions. For this analysis, we also include the JW300 (Agić and Vulić, 2019) dataset, a multilingual dataset crawled from `jw.org`. In summary, we find 8 nonstandard codes in CCAligned, 3 in OSCAR, 1 in mC4, 1 in WikiMatrix, and 70 in JW300, for 83 in total. This does not include the 59 codes affected by superset issues. Full details are given in Appendix A.

Inconsistent Language Codes One common issue is simply using nonstandard or invented codes. For example, CCAligned uses only two-letter codes, so when the BCP-47 code for a language is three letters it is either shortened (e.g., `zza` → `zz`) or invented (`shn` → `qa`). Similarly, OSCAR contains data labeled as `als` (BCP-47 for Tosk

Albanian) that is actually in gsw (Allemannic).[9] Twenty-two additional language codes in JW300 have similar issues, including 12 codes that start with jw_ but are not Javanese.

False Sign Languages Twelve percent (48/417) of JW300 carry language codes for sign languages. Instead of sign language transcripts they are texts in another high-resource language, mostly English or Spanish—for example, the en-zsl (Zambian sign language) data is actually English-English parallel data (copies), details in Appendix A. This was likely caused by videos with sign language interpretation embedded on the crawled Web sites.[10]

Mysterious Supersets When datasets contain language codes that are supersets of other language codes, it is difficult to determine which particular language the text contains. WikiMatrix has Serbian (sr), Croatian (hr), Bosnian (bs), and Serbo-Croatian (sh)—their superset.[11] The issue of codes that are supersets of others is common enough to include a small table dedicated to it (Appendix Table 7). In some cases this may not be an issue, as with Arabic, where ar conventionally refers to Modern Standard Arabic, even though the code technically encompasses all dialects. In many cases, the nature of the data in the superset code remains a mystery.

Deprecated Codes Finally, there are several deprecated codes that are used: sh in WikiMatrix, iw in mC4, sh and eml in Oscar, and daf in JW300.

6 Risks of Low-Quality Data

Low Quality in Downstream Applications Text corpora today are building blocks for many downstream NLP applications like question answering and text summarization—for instance, a common approach is to first train translation models on such data and then automatically translate training data for downstream models (Conneau et al., 2018). If the data used for the original systems is flawed, derived technology may fail for those languages far down the line without knowing the causes. This risk of undesired downstream effects calls for future studies with a careful treatment of intertwined effects such as data size and domain, language-specific phenomena, evaluation data and metric biases. To give the reader a brief glimpse of the impact of data quality for the example of translation, we compare the C% metric from our audit with the translation quality (sentencepiece-BLEU, spBLEU) of the multilingual translation model M2M124 for 124 languages (Goyal et al., 2021). It was trained on WikiMatrix and CCAligned, and similar data collected with the same tools, which we expect to show similar biases. Translation quality is evaluated on the trusted, human-translated FloReS benchmark (Goyal et al., 2021). For the 21 languages present in both the audit and the FloReS benchmark, we found a positive correlation (Spearman) between the data quality scores and spBLEU of $\rho = 0.44$ $(p = 0.041)$. This is not as large as the correlation with data size ($\rho = 0.66$, $p = 0.00078$), but it nonetheless helps to explain translation quality—the correlation between the product of C% and data size (in other words, the expected total number of good sentences in the dataset), is the highest yet, with a value of $\rho = 0.73$ $(p = 0.00013)$.[12]

Representation Washing Since there are datasets that contain many low-resource languages, the community may feel a sense of progress and growing equity, despite the actual quality of the resources for these languages. Similarly, if low-quality datasets are used as benchmarks they may exaggerate model performance, making low-resource NLP appear more solved than it is—or conversely, if models perform poorly when trained with such data, it may be wrongly assumed that the task of learning models for these languages is harder than it actually is or infeasible given current resources. These effects could result in productive effort being redirected away from these tasks and languages.

Trust in Incorrect ''Facts'' We found many instances of parallel-looking sentences that are structurally and semantically similar, but not factually correct translations (Table 6). They can cause models to produce plausible ''translations'' that are factually wrong, but users may still trust them (*algorithmic trust*) without verifying the

[9]This is a result of the language code used by the Alemannic Wikipedia and affects any corpus or tool that uses Wikipedia data without correcting for this, like FastText.

[10]Kudos to Rebecca Knowles for this explanation.

[11]https://iso639-3.sil.org/code/hbs.

[12]For the translation from English, BLEU scores are less comparable but the trend holds nonetheless, with values of ($\rho = 0.32$, $p = 0.14$), ($\rho = 0.74$, $p = 0.000078$), and ($\rho = 0.80$, $p = 0.0000087$), respectively.

en	The prime minister of the **UK** is **Boris Johnson**.
nl	De minister-president van **Nederland** is **Mark Rutte**.
	en: The prime minister of the Netherlands is Mark Rutte.
en	**24 March** 2018
pt	**14 Novembro** 2018
	en: 14 November 2018
en	The current local time in **Sarasota** is **89** minutes.
nn	Den lokale tiden i **Miami** er **86** minutt.
	en: The local time in Miami is 86 minutes.
en	In **1932** the highway was extended **north to LA**.
bar	**1938** is de Autobahn bei **Inglstod** fertig gstellt.
	en: The highway near Inglstod was completed in 1938.

Table 6: Examples of ‘‘parallel’’ data where the translation has a different meaning than the source, but the form looks the same. (We added translations of the non-English side.) Such data may encourage hallucinations of fake ‘‘facts’’.

information. Similarly, *automation bias* (Skitka et al., 1999), referring to humans favoring decisions made by automated systems over decisions made by humans, might amplify the issues of inaccurate translations caused by misalignments.

7 Future Work and Recommendations

Of the five multilingual corpora evaluated, we consistently found severe issues with quality, especially in the lower-resource languages. We rated samples of 205 languages, and found that 87 of them had under 50% usable data, with a full 15 languages at 0% in-language. We furthermore found consistent issues with mislabeled data and nonstandard language codes, particularly in the JW300 dataset, and identified 83 affected corpora, at least 48 of which were entirely spurious (Section 5). While there might have been anecdotal evidence of insufficient quality for some of the datasets, the majority of these quality issues had not been reported, nor been investigated in depth. These issues might go unnoticed for languages that are not represented in the evaluation of the crawling methods, and cause harm in downstream applications (Khayrallah and Koehn, 2018).

There are a variety of ways to improve both the ease and accuracy of human evaluation, as well a few classes of issues we ignored in this paper, like close dialects. Ideally we would like to build a standard suite of automatic metrics for datasets, but more research is necessary to determine what the appropriate metrics would be. One important area missing from our analyses, however, is the estimated portion of a dataset which has been generated by MT (Rarrick et al., 2011), LM systems, or bots/templates, as for example in the analysis of C4 (Dodge et al., 2021). The information captured in machine-generated content might still be useful for modeling, but might falsely overrepresent typical generation patterns and introduce linguistic errors or unnatural artifacts.

We therefore strongly recommend looking at samples of any dataset before using it or releasing it to the public. As we have shown, one does not need to be proficient in a language to see when there are serious quality issues, and a quick scan of 100 sentences can be sufficient to detect major problems. Moreover, going through and annotating a small sample of data can bring actionable insights about new ways to filter or use it.

If data quality issues are found, a wide variety of techniques can be explored, like filtering on length-ratio, LangID, TF-IDF wordlists (Caswell et al., 2020), or dictionaries (Kamholz et al., 2014); to neural approaches like LM scoring (Axelrod et al., 2011; Moore and Lewis, 2010; Wang et al., 2018). Unfortunately, none of these provides a quick and easy fix, especially for low-resource languages—data cleaning is no trivial task!

Noisy datasets are by no means useless, at least if they contain some desirable content. Therefore an alternative to filtering can be documentation (Bender et al., 2021). This can take the form of a per-language quality score and notes about known issues, a datasheet (Gebru et al., 2018) or nutrition label (Holland et al., 2018). However, we suggest researchers not release corpora with near-zero in-language content, as this may give the mistaken impression of usable resources.

Finally, we encourage the community to continue conducting evaluations and audits of public datasets—similar to system comparison papers.

Acknowledgments

We would like to thank the TACL editors and reviewers, and AfricaNLP and Google reviewers who have helped us shape this paper. Furthermore, we are grateful for Ahmed El-Kishky’s support and help with CCAligned and WikiMatrix size statistics.

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1310

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraELECTRA: Pre-training text discriminators for Arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

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. https://doi.org/10.1162/tacl_a_00288

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

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. https://doi.org/10.18653/v1/2020.acl-main.417

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604. https://doi.org/10.1162/tacl_a_00041

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Stella Biderman and Walter J. Scheirer. 2020. Pitfalls in machine learning research: Reexamining the development cycle. *arXiv preprint arXiv:2011.02832*.

Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546. https://doi.org/10.1109/WACV48630.2021.00158

Jan A. Botha, Emily Pitler, Ji Ma, Anton Bakalov, Alex Salcianu, David Weiss, Ryan McDonald, and Slav Petrov. 2017. Natural language processing with small feed-forward networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2879–2885, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1309

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever,

and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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. https://doi.org/10.18653/v1/2020.coling-main.579

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.598

Avihay Chriqui and Inbal Yahav. 2021. HeBERT & HebEMO: A Hebrew BERT Model and a Tool for Polarity Analysis and Emotion Recognition. *arXiv preprint arXiv:2102.01909*.

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. https://doi.org/10.18653/v1/2020.acl-main.747

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1269

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.292

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, and Matt Gardner. 2021. Documenting the english colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.

Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.387

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. https://doi.org/10.18653/v1/2020.emnlp-main.480

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal,

Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-centric multilingual machine translation. *arXiv preprint arXiv:2010.11125*.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online. https://doi.org/10.18653/v1/2020.findings-emnlp.195

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv preprint arXiv:2106.03193*.

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

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomás Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

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. https://doi.org/10.18653/v1/E17-2068

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. https://doi.org/10.18653/v1/W18-6478

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task*

Papers, Day 1), pages 225–233, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5321

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N. C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.findings-emnlp.445

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC' 14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

Vincentius Kevin, Birte Högden, Claudia Schwenger, Ali Şahan, Neelu Madan, Piush Aggarwal, Anusha Bangaru, Farid Muradov, and Ahmet Aker. 2018. Information nutrition labels: A plugin for online news evaluation. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 28–33, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5505

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. https://doi.org/10.18653/v1/W18-2709

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.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. Greek-bert: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, pages 110–117, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3411408.3411440

Alexandra Sasha Luccioni and Joseph D. Viviano. 2021. What's in the box? an analysis of undesirable content in the common crawl corpus. *arXiv preprint arXiv:2105.02732*. https://doi.org/10.18653/v1/2021.acl-short.24

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: A tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.645

Mihai Masala, Stefan Ruseti, and Mihai Dascalu. 2020. RoBERT—a Romanian BERT model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6626–6637, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.581

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.156

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low

resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9–16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Addison Phillips and Mark Davis. 2005. Tags for Identifying Languages. Internet Engineering Task Force. Work in Progress.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2084

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. *Journal of Machine Learning Research*, 21:1–67.

Spencer Rarrick, Chris Quirk, and Will Lewis. 2011. MT detection in Web-scraped parallel corpora. In *Proceedings of MT Summit XIII*. Asia-Pacific Association for Machine Translation.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. 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. https://doi.org/10.18653/v1/2021.eacl-main.115

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Shaked Greenfeld, and Reut Tsarfaty. 2021. AlephBERT:A Hebrew large pre-trained language model to start-off your Hebrew NLP application with. *arXiv preprint arXiv:2104.04052*.

Linda J. Skitka, Kathleen L. Mosier, and Mark Burdick. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006. https://doi.org/10.1006/ijhc.1999.0252

Chenkai Sun, Abolfazl Asudeh, H. V. Jagadish, Bill Howe, and Julia Stoyanovich. 2019. Mithralabel: Flexible dataset nutritional labels for responsible data science. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2893–2896, New York, NY, USA. Association for Computing Machinery.

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018. Denoising neural machine translation training with trusted data and online data selection. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 133–143, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6314

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Hainan Xu and Philipp Koehn. 2017. Zipporah: A fast and scalable data cleaning system for noisy Web-crawled parallel corpora. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2945–2950, Copenhagen, Denmark. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Dataset	Supercode	Subcode(s)
JW300	kg	kwy
JW300	mg	tdx
JW300	qu	que, qug, qus, quw, quy, quz, qvi, qvz
JW300	sw	swc
OSCAR	ar	arz
OSCAR	az	azb
OSCAR	sh	bs, hr, sr
OSCAR	ku	ckb
OSCAR	ms	id, min
OSCAR	no	nn
OSCAR	sq	als*
OSCAR	zh	yue, wuu
WikiMatrix	ar	arz
WikiMatrix	sh	bs, hr, sr
WikiMatrix	zh	wuu

Table 7: Situations where two language codes are represented, but one is a superset of another by the ISO standard, leading to unclarity about the data in the supercode dataset. *The als dataset is actually in gsw.

Actual language	Code in JW300
cs	cse
de	gsg
el	gss
en	ase, asf, bfi, ins, psp, sfs, zib, zsl
es	aed, bvl, csf, csg, csn, csr, ecs, esn, gsm, hds, lsp, mfs, ncs, prl, pys, ssp, vsl
fi	fse
fr	fcs,fsl
hu	hsh
id	inl
it	ise
ja	jsl
ko	kvk
pl	pso
pt	bzs, mzy, psr, sgn_AO
ro	rms
ru	rsl
sk	svk
sq	sql
st	jw_ssa
zh	csl, tss

Table 8: There are 48 languages in the JW300 corpus with language codes that correspond to sign languages, but in reality are unrelated high-resource languages (usually the most spoken language in the country of origin of the sign language). This table shows the actual language of the data corresponding to each sign language code.

A Details on Language Code Issues

Table 7 provides a complete lists of the corpora where one code is defined as a superset of the other by the ISO standard, and in Table 8 we provide a complete list of the language codes in JW300 which purport to be sign language but are actually unrelated high-resource languages.

Special attention needs to be given to the JW300 dataset, which, in addition to the sign languages and superset code issues, has a variety of other peculiarities. These problems seem to originate in the codes used by jw.org,[13] which were apparently not checked in the creation of the JW300 dataset. An overview is provided in Table 9, and the following paragraphs give specifics.

Twelve languages in JW300 have codes starting in jw_, suggesting they are varieties of Javanese (ISO639-1 jw), but are instead attempts to represent language dialects for which there are no BCP-47 codes. These codes seem to have been updated in jw.org to appropriate BCP-47 private-use extensions in the form <supercode>_x_<tag>, which are provided in Table 9. Twelve languages have codes starting in jw_, suggesting they are varieties of Javanese, but are instead mis-parsed private-use extensions. Three codes appear in addition to equivalent ISO codes, making it unclear which languages they are. One language uses a deprecated ISO code. Four languages use the ISO639-3 code instead of the ISO639-2 code, and therefore are not BCP-47.

In addition to the jw_ tags, there are two other mis-used private subtags: hy_arevmda, which in addition to lacking the mandatory _x_ appears to represent standard Western Armenian (hyw); and

[13]The jw.org Web site seems to use correct BCP-47 extensions now, however, and entering a code such as ''jw_dmr'' redirects to ''naq_x_dmr''.

Code in JW300	BCP-47 code	Actual Language Name
Incorrect private-use extensions		
hy_arevmda	hyw	Western Armenian
jw_dgr	os_x_dgr	Digor Ossetian
jw_dmr	naq_x_dmr	Damara Khoekhoe
jw_ibi	yom_x_ibi	Ibinda Kongo
jw_paa	pap_x_paa	Papiamento (Aruba)
jw_qcs	qxl	Salasaca Highland Kichwa
jw_rmg	rmn_x_rmg	Greek Romani (South)
jw_rmv	rmy_x_rmv	Vlax Romani, Russia
jw_spl	nso_x_spl	Sepulana
jw_ssa	st_ZA	Sesotho (South Africa)
jw_tpo	pt_PT	Portuguese (Portugal)
jw_vlc	ca_x_vlc	Catalan (Valencia)
jw_vz	skg_x_vz	Vezo Malagasy
rmy_AR	rmy_x_?	Kalderash
Equivalent codes used in place of extensions		
kmr_latn	kmr_x_rdu	Kurmanji (Caucasus)
nya	ny_x_?	Chinyanja (Zambia)
que	qu_x_?	Quechua (Ancash)
Deprecated codes		
daf	dnj/lda	Dan
ISO-693-3 used in place of ISO-693-2		
cat	ca	Catalan
gug	gn	Guarani
run	rn	Kirundi
tso_MZ	ts_MZ	Changana (Mozambique)

Table 9: Language code issues in the JW300 datasets for 22 language varieties not covered by Tables 7 and 8. Private use extensions are given as they appear in `jw.org`, and specified as '?' if they are absent from `jw.org`.

`rmy_AR`, which, rather than being Romany from Argentina, is Kalderash Romany.

There are also a few anomalies where private use extensions should have been used but other methods were found to convey the distinctions. Three codes appear in addition to equivalent ISO codes, making it unclear which languages they are. Two of these are equivalencies between ISO639-2 and ISO639-3 (`nya` and `ny` are both Chichewa, `qu` and `que` are both Quechua), and one is a script equivalency (`kmr` and `kmr_latn` are both in Latin script). In these three cases the two codes do represent different languages—so a private use extension would have been appropriate.

Finally, there is the more minor issue that three languages use the ISO639-3 code instead of the ISO639-2 code, and therefore are not BCP-47.

Dataset	Code in Corpus	Correct Code
CCAligned	`zz`	`zza`
CCAligned	`sz`	`szl`
CCAligned	`ns`	`nso`
CCAligned	`cb`	`ckb`
CCAligned	`tz`	`ber`
CCAligned	`qa`	`shn`
CCAligned	`qd`	`kac`
CCAligned	`cx`	`ceb`
mC4	`iw`	`he`
OSCAR	`eml`	`egl`
OSCAR	`als`	`gsw`
OSCAR	`sh`	`hbs`
WikiMatrix	`sh`	`hbs`

Table 10: Miscellaneous errors in language codes.

In addition to the JW300-specific errors, Table 10 summarizes miscellaneous errors in CCAligned and OSCAR that were detailed in Section 5.

B Complete Error Taxonomy and Instructions

In addition to the examples given in Table 2, raters were provided with the following verbal notes on the error codes:

- **`CC`: Correct translation, natural sentence:** It's OK if it's a sentence fragment instead of a whole sentence, as long as it is not too short (about 5 words or greater). The translation does not have to be perfect.
- **`CS`: Correct translation, but single word or short phrase:** Also includes highly repeated short phrases, like ''the cat the cat the cat the cat the cat ...''
- **`CB`: Correct translation, but boilerplate:** This can be auto-generated or formulaic content, or content that one deems ''technically correct but generally not very useful to NLP models''. Unfortunately, it's often not clear what should be counted as boilerplate...do your best.
- **`X`: Incorrect translation** [for parallel sentences] both source and target are in the correct language, but they are not adequate translations.

- **`WL`: Wrong language** For short sentences, especially with proper nouns, there is often a fine line between ‘‘Wrong language’’ and ‘‘Not language’’. Do your best.
- **`NL`: Not language** At least one of source and target are not linguistic content. Any sentence consisting only of a proper noun (e.g. ‘‘Tyrone Ping’’) should be marked as `NL`.
- **`U`: Unknown** for sentences that need verification by a native speaker. This is an auxiliary label that is resolved in most cases.

C Methodological Notes

A surprising amount of work can be done without being an expert in the languages involved. The easiest approach is simply to search the internet for the sentence, which usually results in finding the exact page the sentence came from, which in turn frequently contains clues like language codes in the URL, or a headline like *News in X language*, sometimes with references to a translated version of the same page. However, for the cases where this is insufficient, here are a few tips, tricks, and observations.

No Skills Required: Things that do not require knowledge of the language(s) in question.

1. ‘‘Not language’’ can usually be identified by anyone who can read the script, though there are tricky cases with proper nouns.
2. Frequently, ‘‘parallel’’ sentences contain different numbers in the source and target (especially autogenerated content), and are easy to disqualify.
3. Errors tend to repeat. If a word is mistranslated once, it will often be mistranslated many more times throughout a corpus, making it easy to spot.

Basic Research Required: Things that do not require knowledge of the language(s) in question but can be done with basic research.

1. If it’s written in the wrong script it’s considered wrong language. (Sometimes the writing system is indicated in the published corpus, e.g., `bg-Latn`, but usually the language has a ‘‘default’’ script defined by ISO.)
2. Some types of texts come with inherent labels or markers, such as enumerators or verse numbers.
3. When all else fails, search the internet for the whole sentence or n-grams thereof! If the whole sentence can be found, frequently the language is betrayed by the web page (the language’s autonym is useful in this case).

D Complete Audit Results

Tables 11, 12, 13, 14, and 15 give the complete annotation percentages for CCAligned, WikiMatrix, ParaCrawl, mC4 and OSCAR, respectively. For each annotation label, we report the ratio of the annotated sentences (of max 100 sentences) that were assigned that label by the primary annotator. Repeated annotations done for agreement measurement are not included. The `C` column aggregates all correct sub-codes (`CC`, `CS`, `CB`). We also report the total number of sentences that each dataset contains for each language and the average sentence length for the audited sentences to illustrate differences across languages. The original language codes as they are published with the datasets are maintained for the sake of consistency (but should be handled with care in future work, see Section 5), and those with less than 20% correct sentences are highlighted.

	C	CC	CS	CB	X	WL	NL	porn	#sentences	avg target length
en-sz_PL	0.00%	0.00%	0.00%	0.00%	0.00%	8.33%	91.67%	0.00%	12	71.42
en-mt_MT	3.85%	0.00%	3.85%	0.00%	50.00%	26.92%	19.23%	0.00%	26	12.58
en-tz_MA	12.12%	6.06%	6.06%	0.00%	45.45%	36.36%	6.06%	0.00%	33	57.33
en-zz_TR	0.00%	0.00%	0.00%	0.00%	8.82%	61.76%	29.41%	0.00%	34	46.53
en-kg_AO	1.35%	0.00%	1.35%	0.00%	14.86%	2.70%	81.08%	0.00%	74	29.20
en-qa_MM	11.03%	5.88%	3.68%	1.47%	72.06%	3.68%	13.24%	0.00%	136	55.28
en-bm_ML	6.04%	4.03%	2.01%	0.00%	26.85%	6.71%	60.40%	0.00%	149	32.19
en-az_IR	6.93%	6.93%	0.00%	0.00%	20.79%	13.86%	58.42%	0.00%	158	115.85
en-qd_MM	7.92%	4.95%	1.98%	0.99%	81.19%	3.96%	6.93%	0.00%	179	60.34
en-ay_BO	51.00%	33.00%	18.00%	0.00%	29.00%	3.00%	17.00%	0.00%	475	92.19
en-ak_GH	14.23%	13.60%	0.63%	0.00%	46.86%	19.25%	19.67%	0.00%	478	45.85
en-st_ZA	48.57%	42.14%	0.00%	6.43%	40.71%	1.43%	9.29%	0.00%	904	111.83
en-ve_ZA	60.40%	29.70%	21.78%	8.91%	28.71%	3.96%	6.93%	0.00%	1555	82.99
en-ts_ZA	51.49%	34.65%	11.88%	4.95%	40.59%	2.97%	4.95%	0.00%	1967	73.93
en-or_IN	42.61%	6.09%	24.35%	12.17%	38.26%	9.57%	9.57%	0.00%	5526	71.39
en-ns_ZA	4.00%	2.00%	0.00%	2.00%	23.00%	15.00%	58.00%	4.00%	14138	33.52
en-lg_UG	6.00%	0.00%	6.00%	0.00%	68.00%	17.00%	9.00%	2.00%	14701	15.83
en-ln_CD	8.00%	4.00%	3.00%	1.00%	14.00%	4.00%	74.00%	4.00%	21562	28.80
en-om_KE	2.00%	2.00%	0.00%	0.00%	31.00%	38.00%	29.00%	24.00%	22206	23.83
en-ss_SZ	12.65%	9.04%	3.61%	0.00%	13.25%	24.10%	50.00%	13.86%	22960	25.30
en-te_IN_rom	0.00%	0.00%	0.00%	0.00%	25.00%	8.00%	67.00%	5.00%	25272	24.21
en-cb_IQ	4.00%	1.00%	3.00%	0.00%	30.00%	18.00%	48.00%	11.00%	52297	30.04
en-tn_BW	0.00%	0.00%	0.00%	0.00%	6.90%	8.97%	63.45%	10.34%	71253	16.80
en-ff_NG	0.00%	0.00%	0.00%	0.00%	0.00%	8.00%	92.00%	2.00%	73022	33.59
en-sn_ZW	5.00%	1.00%	3.00%	1.00%	81.00%	14.00%	0.00%	0.00%	86868	102.59
en-wo_SN	0.00%	0.00%	0.00%	0.00%	1.71%	3.31%	94.98%	18.46%	88441	27.25
en-br_FR	17.00%	3.00%	1.00%	13.00%	37.00%	14.00%	32.00%	1.00%	115128	41.68
en-zu_ZA	55.00%	39.00%	3.00%	13.00%	30.00%	7.00%	8.00%	3.00%	126101	79.32
en-ku_TR	36.52%	12.17%	13.04%	11.30%	33.04%	28.70%	1.74%	1.74%	137874	90.51
en-ig_NG	58.00%	49.00%	3.00%	6.00%	29.00%	12.00%	1.00%	0.00%	148146	83.42
en-kn_IN	46.00%	9.00%	6.00%	31.00%	46.00%	2.00%	5.00%	4.00%	163921	70.20
en-yo_NG	34.93%	6.16%	10.96%	17.81%	34.93%	12.33%	17.81%	0.00%	175192	75.01
en-ky_KG	44.12%	24.51%	17.65%	1.96%	33.33%	22.55%	0.00%	0.98%	240657	69.56
en-tg_TJ	46.08%	18.63%	24.51%	2.94%	32.35%	20.59%	0.98%	4.90%	251865	75.31
en-ha_NG	30.00%	25.00%	3.00%	2.00%	49.00%	9.00%	12.00%	1.00%	339176	60.78
en-am_ET	59.11%	35.47%	2.46%	21.18%	37.44%	2.96%	0.49%	0.00%	346517	58.29
en-km_KH	56.12%	12.24%	33.67%	10.20%	42.86%	1.02%	0.00%	0.00%	412381	71.35
en-ne_NP	47.00%	10.00%	13.00%	24.00%	15.00%	8.00%	30.00%	14.00%	487155	79.14
en-su_ID	35.00%	15.00%	15.00%	5.00%	13.00%	13.00%	39.00%	0.00%	494142	57.08
en-ur_PK_rom	0.50%	0.00%	0.50%	0.00%	18.91%	27.36%	53.23%	5.47%	513123	18.41
en-ht_HT	55.67%	8.25%	10.31%	37.11%	35.05%	6.19%	3.09%	1.03%	558167	101.95
en-mn_MN	33.00%	8.00%	14.00%	11.00%	42.00%	7.00%	18.00%	12.00%	566885	44.43
en-te_IN	69.00%	42.00%	11.00%	16.00%	27.00%	1.00%	3.00%	1.00%	581651	97.95
en-kk_KZ	68.32%	40.59%	18.81%	8.91%	18.81%	8.91%	3.96%	1.98%	689651	72.36
en-be_BY	90.00%	57.00%	13.00%	20.00%	10.00%	0.00%	0.00%	2.00%	1125772	118.45
en-af_ZA	63.00%	40.00%	23.00%	0.00%	31.00%	2.00%	4.00%	12.00%	1504061	105.45
en-jv_ID	5.05%	1.01%	1.01%	3.03%	25.25%	10.10%	59.60%	8.08%	1513974	18.34
en-hi_IN_rom	1.00%	0.00%	0.00%	1.00%	39.00%	21.00%	39.00%	8.00%	3789571	18.13
en-lv_LV	59.00%	37.00%	9.00%	13.00%	31.00%	7.00%	3.00%	14.00%	4850957	83.67
en-ar_AR_rom	0.00%	0.00%	0.00%	0.00%	0.00%	4.00%	96.00%	4.00%	5584724	16.69
en-tl_XX	13.00%	6.00%	3.00%	4.00%	24.00%	26.00%	37.00%	5.00%	6593250	37.03
en-uk_UA	63.00%	42.00%	8.00%	13.00%	35.00%	1.00%	1.00%	5.00%	8547348	67.88
en-zh_TW	46.00%	11.00%	31.00%	4.00%	47.00%	6.00%	1.00%	1.00%	8778971	24.89
en-el_GR	49.00%	15.00%	5.00%	29.00%	38.00%	3.00%	10.00%	8.00%	8878492	54.90
en-nl_NL	46.00%	27.00%	19.00%	0.00%	49.00%	2.00%	3.00%	0.00%	36324231	85.95
en-da_DK	54.00%	31.00%	18.00%	5.00%	29.00%	5.00%	12.00%	7.00%	10738582	73.99
en-vi_VN	31.00%	18.00%	0.00%	13.00%	54.00%	1.00%	14.00%	6.00%	12394379	74.19
en-sv_SE	97.00%	91.00%	3.00%	3.00%	0.00%	3.00%	0.00%	0.00%	12544075	103.91
en-zh_CN	57.29%	22.92%	12.50%	21.88%	31.25%	1.04%	10.42%	1.04%	15181410	33.55
en-tr_TR	45.00%	14.50%	14.00%	16.50%	44.50%	5.00%	5.50%	4.00%	20282339	83.80
en-ja_XX	57.00%	35.00%	21.00%	1.00%	34.00%	6.00%	0.00%	0.00%	26201214	34.44
en-pt_XX	66.34%	36.63%	10.89%	18.81%	20.79%	3.96%	8.91%	0.00%	46525410	87.20
en-it_IT	36.00%	14.00%	18.00%	4.00%	60.00%	1.00%	3.00%	0.00%	58022366	97.44
en-de_DE	62.00%	29.00%	14.00%	19.00%	28.00%	2.00%	8.00%	2.00%	92597196	78.08
en-es_XX	58.42%	16.83%	25.74%	15.84%	22.77%	2.97%	15.84%	4.95%	98351611	72.18

Table 11: Audit results for a sample of 100 sentences from **CCAligned** for each language pair, compared to the number of sentences available in the dataset. If fewer than 100 sentences were available, all sentences were audited. Language codes are as originally published. The length is measured in number of characters and averaged across the audited portion of each corpus. Languages with less than 20% correct sentences are boldfaced.

	C	CC	CS	CB	X	WL	NL	porn	# sentences	avg target length
en-ug	12.87%	8.91%	1.98%	1.98%	72.28%	9.90%	1.98%	0.00%	22012	95.55
en-mwl	27.00%	26.00%	0.00%	1.00%	73.00%	0.00%	0.00%	0.00%	33899	135.26
en-tg	0.00%	0.00%	0.00%	0.00%	95.10%	3.92%	0.98%	0.00%	37975	88.87
en-ne	13.00%	7.00%	6.00%	0.00%	60.00%	23.00%	4.00%	0.00%	40549	69.26
en-ka	11.88%	2.97%	2.97%	5.94%	73.27%	10.89%	2.97%	0.00%	41638	144.74
en-lmo	12.75%	11.76%	0.00%	0.98%	81.37%	4.90%	0.98%	0.00%	43790	89.38
en-io	28.00%	27.00%	0.00%	1.00%	69.00%	2.00%	1.00%	0.00%	45999	83.26
en-jv	13.73%	9.80%	0.00%	3.92%	70.59%	12.75%	2.94%	0.00%	48301	91.87
en-wuu	23.23%	14.14%	7.07%	2.02%	65.66%	7.07%	4.04%	0.00%	51024	34.77
br-en	8.70%	7.61%	1.09%	0.00%	82.61%	4.35%	0.00%	0.00%	58400	90.68
bar-en	6.00%	6.00%	0.00%	0.00%	75.00%	16.00%	3.00%	0.00%	67394	103.51
en-kk	5.00%	2.00%	2.00%	1.00%	81.00%	14.00%	0.00%	0.00%	109074	56.03
en-sw	33.33%	27.27%	4.04%	2.02%	64.65%	2.02%	0.00%	0.00%	138590	111.61
en-nds	1.96%	1.96%	0.00%	0.00%	95.10%	1.96%	0.98%	0.00%	178533	91.95
be-en	26.00%	24.00%	2.00%	0.00%	73.00%	1.00%	0.00%	0.00%	257946	121.22
en-hi	36.27%	32.35%	0.98%	2.94%	59.80%	0.98%	2.94%	0.00%	696125	96.77
en-ko	48.04%	33.33%	2.94%	11.76%	48.04%	2.94%	0.98%	0.00%	1345630	55.18
en-uk	87.00%	84.00%	2.00%	1.00%	10.00%	1.00%	2.00%	0.00%	2576425	104.39
en-it	42.00%	42.00%	0.00%	0.00%	58.00%	0.00%	0.00%	0.00%	4626048	140.27
en-simple	37.62%	24.75%	0.00%	12.87%	56.44%	2.97%	2.97%	0.00%	N/A	77.53

Table 12: Audit results for a sample of 100 sentences from **WikiMatrix** for each language pair, compared to the number of sentences available in the dataset. Language codes are as originally published. The length is measured in number of characters and averaged across the audited portion of each corpus. Languages with less than 20% correct sentences are boldfaced.

	C	CC	CS	CB	X	WL	NL	porn	# sentences	avg target length
en-so	80.81%	61.62%	1.01%	18.18%	14.14%	5.05%	0.00%	0.00%	14879	189.83
en-ps	72.00%	53.00%	9.00%	10.00%	17.00%	10.00%	0.00%	0.00%	26321	141.01
en-my	45.00%	9.00%	16.00%	20.00%	32.00%	9.00%	14.00%	0.00%	31374	147.07
en-km	76.00%	51.00%	13.00%	12.00%	18.00%	6.00%	0.00%	0.00%	65113	121.20
en-ne	73.00%	48.00%	1.00%	24.00%	23.00%	2.00%	0.00%	0.00%	92084	153.42
en-sw	85.00%	60.00%	15.00%	10.00%	11.00%	2.00%	2.00%	0.00%	132517	167.34
en-si	37.00%	31.00%	6.00%	0.00%	62.00%	0.00%	1.00%	0.00%	217407	123.06
en-nn	35.92%	24.27%	8.74%	2.91%	49.51%	13.59%	0.97%	0.00%	323519	56.24
es-eu	88.00%	66.00%	15.00%	7.00%	10.00%	1.00%	1.00%	0.00%	514610	121.31
es-gl	89.00%	46.00%	6.00%	37.00%	4.00%	7.00%	0.00%	0.00%	1222837	107.88
en-ru	81.00%	73.00%	6.00%	2.00%	19.00%	0.00%	0.00%	6.00%	5377911	101.28
en-bg	95.15%	85.44%	0.97%	8.74%	4.85%	0.00%	0.00%	0.97%	6470710	112.29
es-ca	80.00%	54.00%	19.00%	7.00%	11.00%	9.00%	0.00%	5.00%	6870183	107.21
en-el	91.59%	68.22%	0.93%	22.43%	7.48%	0.93%	0.00%	0.00%	9402646	135.66
en-pl	94.12%	76.47%	0.98%	16.67%	3.92%	1.96%	0.00%	0.98%	13744860	95.95
en-nl	49.00%	32.00%	17.00%	0.00%	46.00%	3.00%	2.00%	0.00%	31295016	95.05
en-pt	93.07%	92.08%	0.00%	0.99%	4.95%	1.98%	0.00%	0.00%	31486963	108.68
en-it	60.82%	36.08%	16.49%	8.25%	38.14%	0.00%	1.03%	0.00%	40798278	127.55
en-es	87.00%	54.00%	20.00%	13.00%	12.00%	0.00%	1.00%	0.50%	78662122	119.72
en-de	82.83%	64.65%	13.13%	5.05%	13.13%	3.03%	1.01%	0.00%	82638202	111.43
en-fr	89.62%	82.08%	4.72%	2.83%	10.38%	0.00%	0.00%	0.00%	104351522	144.20

Table 13: Audit results for a sample of 100 sentences from **ParaCrawl** for each language pair, compared to the number of sentences available in the dataset. Language codes are as originally published. The length is measured in number of characters and averaged across the audited portion of each corpus.

	C	CC	CS	CB	WL	NL	porn	# sentences	avg length
yo	84.69%	71.43%	2.04%	11.22%	14.29%	1.02%	0.00%	46214	117.71
st	56.70%	42.27%	14.43%	0.00%	35.05%	8.25%	0.00%	66837	132.13
haw	44.90%	34.69%	1.02%	9.18%	33.67%	21.43%	1.02%	84312	129.99
ig	55.91%	41.73%	10.24%	3.94%	0.00%	44.09%	0.79%	92909	98.03
sm	60.20%	58.16%	2.04%	0.00%	27.55%	12.24%	0.00%	98467	126.42
ha	80.81%	79.80%	1.01%	0.00%	14.14%	5.05%	2.02%	247479	155.76
su	59.60%	58.59%	1.01%	0.00%	25.25%	15.15%	2.02%	280719	107.10
sn	36.63%	32.67%	2.97%	0.99%	58.42%	4.95%	0.00%	326392	145.59
mg	57.00%	57.00%	0.00%	0.00%	18.00%	25.00%	0.00%	345040	116.23
pa	78.30%	68.87%	3.77%	5.66%	4.72%	10.38%	0.00%	363399	134.43
ga	76.77%	58.59%	6.06%	12.12%	10.10%	13.13%	0.00%	465670	147.35
co	33.00%	29.00%	2.00%	2.00%	48.00%	19.00%	0.00%	494913	195.30
zu	51.00%	48.00%	2.00%	1.00%	30.00%	19.00%	0.00%	555458	137.81
jv	52.73%	19.09%	19.09%	14.55%	40.00%	7.27%	1.82%	581528	97.96
km	92.86%	92.86%	0.00%	0.00%	7.14%	0.00%	0.00%	756612	162.57
kn	85.15%	73.27%	3.96%	7.92%	2.97%	9.90%	0.00%	1056849	105.39
fy	56.73%	50.00%	3.85%	2.88%	39.42%	3.85%	0.00%	1104359	234.25
te	89.00%	76.00%	9.00%	4.00%	3.00%	8.00%	0.00%	1188243	108.49
la	82.31%	65.38%	6.15%	10.77%	10.00%	7.69%	0.00%	674463	67.25
be	92.04%	86.73%	2.65%	2.65%	4.42%	3.54%	0.00%	1742030	110.86
af	76.00%	76.00%	0.00%	0.00%	15.00%	9.00%	0.00%	2152243	99.52
lb	17.48%	17.48%	0.00%	0.00%	7.77%	74.76%	0.00%	2740336	481.68
ne	78.35%	77.32%	1.03%	0.00%	21.65%	0.00%	0.00%	2942785	102.88
sr	93.69%	85.59%	7.21%	0.90%	5.41%	0.00%	0.00%	3398483	131.72
gl	67.62%	57.14%	10.48%	0.00%	13.33%	17.14%	0.00%	4549465	151.45
bn	93.00%	86.00%	1.00%	6.00%	3.00%	4.00%	0.00%	7444098	92.60
mr	40.00%	35.24%	2.86%	1.90%	49.52%	10.48%	0.00%	7774331	281.94
sl	92.08%	82.18%	4.95%	4.95%	2.97%	4.95%	0.00%	8499456	149.45
hi	80.30%	76.77%	1.01%	2.53%	19.70%	0.00%	2.53%	18507273	105.54
bg	80.90%	75.88%	2.51%	2.51%	2.01%	17.09%	0.00%	23409799	93.86
uk	95.48%	81.41%	7.54%	6.53%	2.01%	2.51%	0.00%	38556465	116.79
ro	94.95%	78.79%	12.12%	4.04%	3.03%	2.02%	0.00%	45738857	130.08
sv	91.18%	84.31%	2.94%	3.92%	4.90%	3.92%	1.96%	8570979	114.45
zh	92.00%	87.00%	1.00%	4.00%	1.00%	7.00%	0.00%	54542308	94.77
ja	99.00%	89.00%	6.00%	4.00%	0.00%	1.00%	1.00%	87337884	59.94
tr	95.96%	88.89%	0.00%	7.07%	3.54%	0.51%	0.00%	87595290	152.75
nl	92.08%	85.15%	6.93%	0.00%	1.98%	5.94%	0.00%	96210458	103.67
pl	96.00%	82.00%	7.00%	7.00%	2.00%	2.00%	0.00%	126164277	170.70
pt	86.00%	79.00%	4.00%	3.00%	2.00%	12.00%	1.00%	169239084	133.51
it	92.00%	79.00%	9.00%	4.00%	1.00%	7.00%	0.00%	186404508	180.26
fr	92.00%	82.00%	7.00%	3.00%	1.00%	7.00%	0.00%	332674575	143.69
de	91.18%	77.45%	7.84%	5.88%	6.86%	1.96%	0.00%	397006993	107.71
ru	91.06%	69.11%	11.38%	10.57%	4.07%	4.88%	0.00%	755585265	109.28
en	93.94%	83.84%	8.08%	2.02%	1.01%	5.05%	0.00%	3079081989	130.97
bg_latn	9.09%	9.09%	0.00%	0.00%	51.52%	39.39%	1.01%	N/A	139.92
ja_latn	13.00%	7.00%	4.00%	2.00%	60.00%	27.00%	0.00%	N/A	218.92
ru_latn	36.45%	25.23%	10.28%	0.93%	34.58%	28.97%	0.93%	N/A	123.14
zh_latn	5.00%	4.00%	1.00%	0.00%	64.00%	31.00%	0.00%	N/A	186.84

Table 14: Audit results for a sample of 100 sentences from **mC4** for each language, compared to the number of sentences available in the dataset. Language codes are as originally published. The length is measured in number of characters and averaged across the audited portion of each corpus. Languages with less than 20% correct sentences are boldfaced.

	C	CC	CS	CB	WL	NL	porn	# sentences	avg length
diq	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1	131.00
bcl	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	1	623.00
cbk	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	1	519.00
pam	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	2	139.00
bar	25.00%	25.00%	0.00%	0.00%	0.00%	75.00%	0.00%	4	53.50
myv	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	5	127.00
yue	0.00%	0.00%	0.00%	0.00%	57.14%	42.86%	0.00%	7	177.00
mwl	57.14%	57.14%	0.00%	0.00%	42.86%	0.00%	0.00%	7	141.00
frr	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	9	231.56
ht	30.00%	30.00%	0.00%	0.00%	0.00%	70.00%	0.00%	10	329.10
ie	30.00%	30.00%	0.00%	0.00%	30.00%	40.00%	0.00%	11	121.70
scn	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	17	155.59
tyv	96.15%	96.15%	0.00%	0.00%	0.00%	3.85%	0.00%	26	167.96
mai	79.31%	75.86%	0.00%	3.45%	20.69%	0.00%	0.00%	29	141.17
bxr	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	37	160.76
dsb	100.00%	97.56%	0.00%	2.44%	0.00%	0.00%	0.00%	41	155.15
so	0.00%	0.00%	0.00%	0.00%	28.57%	71.43%	0.00%	42	208.24
rm	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	47	137.66
nah	100.00%	96.67%	0.00%	3.33%	0.00%	0.00%	0.00%	60	164.53
nap	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	61	152.11
yo	98.46%	96.92%	0.00%	1.54%	1.54%	0.00%	0.00%	64	281.57
gn	81.48%	81.48%	0.00%	0.00%	2.47%	16.05%	0.00%	81	234.95
vec	91.36%	91.36%	0.00%	0.00%	0.00%	8.64%	0.00%	81	184.90
kw	91.57%	90.36%	0.00%	1.20%	3.61%	4.82%	0.00%	83	162.75
wuu	0.00%	0.00%	0.00%	0.00%	98.84%	1.16%	0.00%	86	157.15
eml	42.57%	42.57%	0.00%	0.00%	0.00%	57.43%	0.00%	104	177.88
bh	89.42%	21.15%	0.00%	68.27%	1.92%	8.65%	0.00%	104	137.17
min	64.00%	6.00%	0.00%	58.00%	27.00%	9.00%	0.00%	180	649.85
qu	100.00%	98.97%	0.00%	1.03%	0.00%	0.00%	0.00%	425	167.27
su	99.00%	99.00%	0.00%	0.00%	0.00%	1.00%	0.00%	676	221.00
jv	97.00%	86.00%	0.00%	11.00%	1.00%	2.00%	0.00%	2350	203.08
als	93.00%	93.00%	0.00%	0.00%	6.00%	1.00%	0.00%	7997	375.44
la	98.00%	98.00%	0.00%	0.00%	2.00%	0.00%	0.00%	33838	224.11
uz	98.00%	98.00%	0.00%	0.00%	2.00%	0.00%	0.00%	34244	369.99
nds	97.03%	95.05%	0.00%	1.98%	2.97%	0.00%	0.00%	35032	344.74
sw	98.00%	98.00%	0.00%	0.00%	0.00%	2.00%	0.00%	40066	196.70
br	100.00%	96.00%	0.00%	4.00%	0.00%	0.00%	0.00%	61941	239.56
fy	97.00%	97.00%	0.00%	0.00%	2.00%	1.00%	0.00%	67762	340.23
am	81.09%	79.10%	0.00%	1.99%	18.91%	0.00%	0.00%	287142	267.43
af	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	517353	339.18
eu	100.00%	98.00%	0.00%	2.00%	0.00%	0.00%	0.00%	1099498	330.93
mn	98.00%	94.00%	0.00%	4.00%	2.00%	0.00%	0.00%	1430527	309.94
te	98.99%	93.94%	1.01%	4.04%	0.00%	1.01%	1.01%	1685185	412.31
kk	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	2719851	318.93
ca	99.00%	91.00%	0.00%	8.00%	1.00%	0.00%	0.00%	13292843	333.38
nl	98.00%	94.00%	2.00%	2.00%	2.00%	0.00%	4.00%	126067610	305.01
it	87.13%	71.29%	1.98%	13.86%	11.88%	0.99%	1.98%	210348435	393.66
zh	100.00%	97.00%	0.00%	3.00%	0.00%	0.00%	1.00%	232673578	195.60
fr	100.00%	93.00%	0.00%	7.00%	0.00%	0.00%	5.00%	461349575	306.62
es	100.00%	94.00%	0.00%	6.00%	0.00%	0.00%	3.00%	488616724	268.07
en	99.00%	96.00%	0.00%	3.00%	0.00%	1.00%	1.00%	3809525119	364.65

Table 15: Audit results for a sample of 100 sentences from **OSCAR** for each language, compared to the number of sentences available in the dataset. If fewer than 100 sentences were available, all sentences were audited language codes are as originally published. Length is measured in number of characters. Languages with less than 20% correct sentences are boldfaced.