Beyond Counting Datasets: A Survey of Multilingual Dataset Construction and Necessary Resources

Xinyan Velocity Yu[◇*] **Akari Asai**[◇*] **Trina Chatterjee**[♣]
Junjie Hu[♡] **Eunsol Choi**[♣]
[◇]University of Washington, [♡]The University of Wisconsin-Madison, [♣]The University of Texas at Austin
{xyu530,akari}@cs.washington.edu,
junjie.hu@wisc.edu, {tchatter,eunsol}@utexas.edu

Abstract

While the NLP community is generally aware of resource disparities among languages, we lack research that quantifies the extent and types of such disparity. Prior surveys estimating the availability of resources based on the number of datasets can be misleading as dataset quality varies: many datasets are automatically induced or translated from English data. To provide a more comprehensive picture of language resources, we examine the characteristics of 156 publicly available NLP datasets. We manually annotate *how* they are created, including input text and label sources and tools used to build them, and *what* they study, tasks they address and motivations for their creation. After quantifying the qualitative NLP resource gap across languages, we discuss how to improve data collection in low-resource languages. We survey language-proficient NLP researchers and crowd workers per language, finding that their estimated availability correlates with dataset availability. Through crowdsourcing experiments, we identify strategies for collecting high-quality multilingual data on the Mechanical Turk platform. We conclude by making macro and micro-level suggestions to the NLP community and individual researchers for future multilingual data development.

1 Introduction

Datasets play fundamental roles in advancing language technologies (Paullada et al., 2021). However, large disparities exist among languages in terms of the scale of existing datasets (Kreutzer et al., 2022; Joshi et al., 2020) and the resulting task performance (Blasi et al., 2022). Multilingual resources also harbor unique annotation artifacts, such as translationese (Clark et al., 2020; Artetxe et al., 2020a). Understanding dataset construction processes can help explain the true landscape of multilingual NLP datasets.

[*]Equal contributions.

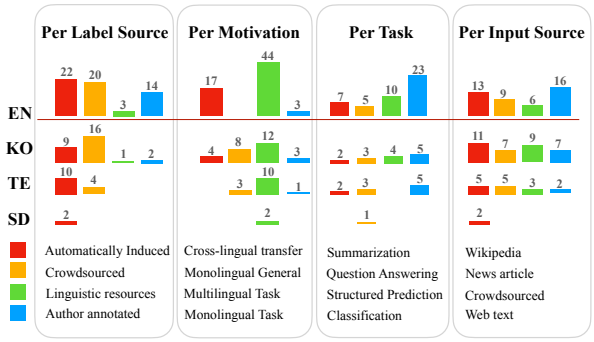


Figure 1: Fine-grained, qualitative comparison of dataset availability in four languages: English (EN), Korean (KO), Telugu (TE) and Sindhi (SD). Low-resource languages often only feature automatically induced labels and lack diverse input sources (frequently limited to Wikipedia text) as a part of parallel multilingual evaluation datasets.

Despite low-level awareness among multilingual researchers about disparities in linguistic resources, scant research to date has examined the *quality* of labeled data in non-English languages. Existing surveys such as Joshi et al. (2020) have studied the scale of available resources per language, without exploring how such datasets are developed, and often simply count number of datasets.

We provide the first large-scale survey of multilingual dataset characteristics, examining in depth how and why these datasets are constructed and quantifying their availability. We study 156 datasets, each covering at least one non-English language, from the ACL anthology and HuggingFace (Lhoest et al., 2021). We propose a new annotation scheme with 13 attributes, including the task it addresses, its source of input text, its label collection process, its stated motivation, and tools used to create it, such as translation services (Section 3).

Our annotation yields rich information about the status of multilingual datasets (Section 4). As shown in Figure 1, language coverage varies significantly across tasks, and label collection methods differ across languages and tasks. 222 languages

are covered by the 156 datasets, and, on average, 5.6 datasets per language are available. However, we posit that this number is misleading when interpreting the landscape of multilingual NLP datasets: we observe the prevalence of automatically induced labels particularly in low resource languages; one-third of the datasets we surveyed use automatically induced labels, and 68% of the languages have no manually annotated data. While Wikipedia and news texts are available for a wide range of languages, texts specifically written for the NLP task are not available for most languages. Furthermore, about 20% of surveyed datasets involved translation in their creation, and these were used at a much higher rate in highly cited resources to study cross-lingual transfer. These types of quantitative and qualitative disparities persist in multilingual supervised datasets, motivating us to identify underlying bottlenecks to create multilingual datasets.

We investigate whether the resources required for NLP data collection affect the prevalence of multilingual datasets. These include the availability of crowdworkers and NLP researchers with sufficient language proficiency as well as raw input text (Section 5). These data collection resources all correlate with the number of datasets in each language, suggesting paths for a more equitable data landscape. To assess the paths for crowdsourcing in non-English languages on a popular crowdsourcing platform, we design controlled data collection experiments in six languages, quantifying the challenges even in relatively high-resource languages.

We conclude our survey by relating research-derived suggestions to the NLP community and to individual researchers for multilingual data construction. We also provide concrete suggestions for crowdsourcing platforms and crowdsourcing quality control tips, recommended translation services, and publication venues. We host our survey at `https://multilingual-dataset-survey.github.io`, permitting readers to review multilingual resources and submit their datasets following our schema.

2 Related Work

Kreutzer et al. (2022) study the quality of raw text data available for researchers, such as ParaCrawl (Esplà et al., 2019), Wikimatrix (Schwenk et al., 2021), and mC4 (Xue et al., 2021), by manually evaluating 100 sample texts for each language. The data quality for low-resource languages is poor, and parallel sentences are often misaligned (Koehn et al., 2019). Blasi et al. (2022) focus on the models' performance on diverse tasks, suggesting that economic powers of the languages' users drives NLP technology development, while we focus on analyzing the quality of existing multilingual datasets. A recent position paper (Hershcovich et al., 2022) discusses the representation of various cultures in NLP datasets. Bender and Friedman (2018) argue that providing details about data can help alleviate issues related to exclusion and bias in language technology. Our schema includes additional information such as text sources. Most closely related to our work, Joshi et al. (2020) study the availability of Linguistic Data Consortium (LDC) Catalog datasets and their relationship to certain factors (e.g., the number of Wikipedia articles). To our knowledge, our work is the first large-scale study (on 156 datasets beyond LDC catalog) of the curation process of labeled, open-sourced multilingual datasets and relevant external factors (e.g., the availability of crowdsourcing and language-proficient researchers). Several work study the quality and availability of crowdsourced workers to gather translation data on MTurk (Callison-Burch, 2009; Bloodgood and Callison-Burch, 2010; Pavlick et al., 2014). Our MTurk experiment examines qualification control methods for multilingual crowdsourcing.

3 Survey Scope, Scheme and Process

In the following section, we first identify the 156 open-sourced NLP datasets published in NLP venues that contain at least one non-English language. We then describe our proposed annotation scheme with 13 attributes, seven of which are analyzed for the first time (Table 1).

Scope. The scope of our work includes *labeled* datasets, where a system is expected to generate a label (output) y given input text x. The output label is not limited to a single categorical label and includes generated text. We filter out (1) unlabeled data, including parallel corpora (studied in Kreutzer et al. 2022); (2) language identification datasets; (3) machine translation (MT) datasets (inherently cross-lingual); and (4) multi-modal datasets (other modalities can be language agnostic)

We compiled dataset lists in December 2021 sourced from: (1) the ACL paper anthology identified through a keyword search, and (2) the open-source Hugging Face Datasets after filtering ex-

Aspect	Descriptions	Categories
Language♦	Target languages	{ISO 639-1 Language Code, not mentioned}
Task Type♦	Ten coarse-grained NLP task type that the dataset addresses	{classification (sentiment analyis), classification (sentence pair), classification (other), QA (w/ retrieval), QA (machine reading), structured prediction, sequence tagging, generation (summarization), generation (other), other}
Dataset size	Avg. # of data in each language	{ < 100, $100 \sim 1000$, $1000 \sim 10k$, $> 10k$ }
Creator♦	Who led dataset creation	{industry, individual researchers, university}
Pub. Venue	Venue of paper publication	{*CL, LREC, *ACL Workshop, Findings, NeurIPS Datasets and Benchmarks Track, arXiv, N/A}
Pub. Year	Year of paper publication	Year of publication between 2008 - 2021
Motivation	Motivation for dataset creation	{cross-lingual transfer, single task (multilingual) w/ ML training, single task (single lang), multi-task (single lang)}
Input text (x) Source♦	Where input text is from	{annotated (authors, linguists), commercial sources, crowdsourced, curated linguistic resources (wordnet, etc.), curated source (exams, scientific papers, etc.), media, template-based, web, Wikipedia}
x Language	Languages where x is collected	{English, its own language, both, other language, not mentioned}
Label (y) Collection♦	How label is collected	{annotated (authors, linguists), automatically induced, crowdsourced, curated linguistic resources, not mentioned}
y Language	Languages where y is collected	{English, its own language, not mentioned}
Reuse	Reusing released datasets?	{Yes-Eng, Yes-other-lang, Yes-Eng&other-lang, No}
Translation♦	Type of translation used during data collection	{automatic translation, human (author), human (non-author), no translation, unclear}

Table 1: Annotation scheme. Each row represents an **aspect** that we label with **categories**. ♦ marks aspects on which a dataset can be assigned more than one categories. The highlight marks an aspect where no prior survey has analyzed. Input text (x) refers to the source of the text of the dataset, and label (y) refers to the target that the model should generate. For example, for a news summarization dataset, the label is the news summary and the input text source can be news articles from media.

cluded tasks and English-only datasets. Our preliminary study shows that the combination of HuggingFace and ACL anthology keyword search gives reasonable coverage of existing multilingual NLP dataset papers: Most NLP dataset papers have been published in ACL-related conferences (e.g., LREC, *CL workshops, *CL) and listed in ACL anthology. HuggingFace datasets[*] are widely used in the NLP community and many recently-published papers make their datasets available there. We searched the ACL anthology using case-insensitive keyword matches by first selecting papers that includes specific keywords in their titles, and then filtering out the papers mentioning excluded tasks in their titles and abstracts:

- **Included keywords:** multilingual, cross-lingual, dataset, annotation, labelled, benchmark.
- **Excluded keywords:** machine translation, language identification, vision, topic model, induction, speech, multimodal.

We selected these keywords to focus our study on labeled multilingual NLP *text-only* datasets. The preceding steps helped us select 151 papers of the 73,384 papers in the ACL anthology, and we subsequently filtered out papers due to unavailability of datasets. For multi-task datasets for a specific language (such as KLUE; Park et al. 2021), we decomposed the unified benchmark into each sub-task and included each sub-task as a unique dataset. For example, in the KLUE benchmark, there are eight tasks in total, and we annotate each sub-task dataset independently. Standardized multi-task test suites that directly reuse existing resources such as XTREME (Hu et al., 2020), or dataset / sub-task dataset that directly reuses a previously published dataset with no significant modification to any of the schemes were not included, and we only include the original dataset. Our final survey includes 156 datasets from 112 papers.

Annotation Scheme. Table 1 describes the annotation scheme, consisting of **aspects** (e.g., task type) and their **categories** (e.g., summaries, sequence tagging). We cover three topics related to each dataset: (1) *coverage*, i.e., what languages and tasks does it cover? (2) *metadata*, i.e., why and who created the dataset? when was it created? (3) *source*, i.e., how were the input and labels collected? We started with a set of categories for each aspect and updated it periodically. Four (motiva-

[*] https://huggingface.co/docs/datasets/index

Category	Lang	Task	Train	Main Goal (an example dataset)
cross-lingual transfer	multi	single		Evaluating across languages w/o training (e.g., XNLI; Conneau et al. 2018)
multilingual task	multi	single	✓	Improving a task across languages (e.g., TyDi QA; Clark et al. 2020)
monolingual task	single	single	✓	Improving a task in a single language (e.g., KQuAD; Lim et al. 2019)
monolingual general	single	multi	✓	Improving multiple tasks in a single language (e.g., CLUE; Xu et al. 2020)

Table 2: Motivation category explanations. The *Train* column indicates presence/absence of a data split. We label datasets with training data as *cross-lingual transfer* when the original papers explicitly mention that as their primary goal rather than developing a system for the target task.

tion, use/type of translation, input language, output language) out of our 13 annotation attributes focus on multilingual dataset collection. The rest of the scheme can also be useful to analyze the distribution of English datasets for future work.

Many aspects are self-explanatory (e.g., language of the dataset), but the *motivation* aspect requires elaboration. While datasets are often repurposed after their introduction, dataset creators aim to address specific research questions at the creation time. We identified four types of motivation after manually reading the papers, including cross-lingual transfer, multilingual task, monolingual task, and monolingual general, which are summarized and explained in Table 2.

Annotation Process. Three of the authors of this paper, each of whom have more than one year of NLP research experience and speaks more than one language, have manually annotated the datasets from December 2021 until March 2022. When a dataset fell within the boundaries of the pre-defined category set, each annotator tagged each borderline case, and all authors resolved the final category by discussion. For multi-category aspects noted by ♦ in Table 1, we incremented *the count of each*.

4 Survey Results

We summarize our findings, focusing on novel aspects. For each category of each aspect, we report the number of datasets belonging to that category and the unique languages they cover. We also highlight noteworthy correlations between multiple aspects, e.g., how label collection methods correlated with the type of tasks addressed.

4.1 Coverage

Task Types. We classify NLP tasks into five major categories: structured prediction, sequence tagging, generation, question answering (QA) and classification. For the last three, we provide subcategories. Table 3 shows per-task resources availability. Sequence tagging, summarization, and information retrieval have a wide coverage of languages. However, our analysis reveals that many datasets of these tasks were constructed with distant supervision. Classification task is common, while fewer resources are available for complex tasks such as structured prediction.* Limited resources exist for text generation, potentially because of the difficulty of quality control for generated texts. As an exception, summarization is covered broadly as they are often scraped from encyclopedic or news websites (Hasan et al., 2021), although recent work shows such automatically created summarization datasets can be noisy (Goyal et al., 2022).

Task type		# Data	# Langs
Classification	sentiment	20	29
	sentence pair	17	25
	other	33	108
Sequence tagging		16	194
Generation	summarization	10	95
	other	9	12
QA	machine reading	20	47
	w/ retrieval	10	142
Structured prediction		19	33
Other		3	13

Table 3: The statistics of datasets by task types.

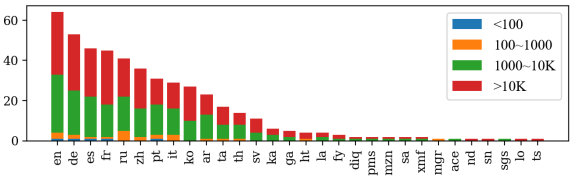


Figure 2: The number of datasets per data size bucket (the number of examples in the dataset) per language.

Dataset Size. Each bar in Figure 2 presents the number of datasets bucketed by data size for a specific language for the top 10 languages and 20 randomly sampled languages. Low-resource lan-

*We interpret "structured prediction" coarsely, referring to tasks with a complex output space, such as parsing, coreference resolution, dialogue state tracking, and discourse analysis.

Source	Category	# Data	# Langs
generated by	crowdworkers	16	33
	authors, linguists	2	9
	template	2	11
collected from	web	35	114
	social media/commerce	24	33
	Wikipedia	32	184
	media (news)	40	103
curated source	linguistics	15	32
	others (exams, etc)	25	59

Table 4: Statistics on the source of input texts.

guages do not necessarily have smaller datasets, but they have fewer manually annotated datasets. Automatically induced datasets, where label y is determined without human supervision, also tend to be larger: 35 out of 81 datasets that have more than 10K examples are automatically induced.

4.2 Input and Label Collection

Input Source: Collection Process for Input Texts (x). We classify input text source into three high-level categories (i.e., *generated by* human, *collected from* websites, and extracted from *curated sources*) and break them down into nine fine-grained categories. Table 4 shows the categories and annotation results. While we cover 222 languages, only 40 of them (18%) have input text specifically written by humans for the task.[*] The news is the most common source, used by 40 of 156 datasets, often in summarization or classification tasks, followed by web corpora[*] and Wikipedia. Many languages have datasets derived only from Wikipedia text. Figure 6 in appendix shows per-task input source distributions.

Overall, we observe that high-resource languages entertain a variety of input sources, while low-resource ones rely on fewer resources such as Wikipedia and news.

Label Source: Collection Process for Labels (y). Table 5 presents the statistics on how the output labels were collected, split into five categories: annotated by authors or linguists, crowdsourced, automatically induced, derived from linguistic resources, and not mentioned. Label collection methods affects dataset quality. While manually annotated datasets can exhibit artifacts (Gururangan et al., 2018; Poliak et al., 2018), they are often validated via inter-annotator agreement. In contrast, automatically induced datasets are often introduced without such kind of validation phases and tend to be noisy. For example, bullet points from news article are often considered to be the summary of the article, which can contain missing background information in the rest of the article (Kang and Hashimoto, 2020; Goyal et al., 2022). Fifty-three datasets had automatically induced labels, most commonly seen for summarization and classification tasks, and 95 used manual annotation. This further breakdowns to 27 datasets *solely* annotated by domain experts and 56 datasets *solely* annotated by crowdworkers. We investigated annotator pools for non-English languages in Section 5.

Label Source and Task Types. Figure 3 presents label collection methods per task type. QA with retrieval (e.g., XQA; Liu et al. 2019) and generation tasks show a high proportion of automatically induced datasets. In contrast, structured prediction datasets were rarely automatically induced; they were more often annotated by authors or linguists. Crowdsourcing is commonly used to construct reading comprehension and classification datasets.

Label Source and Language Diversity. Figure 4 shows the distributions of the label data collection methods for the top 10 languages and for 20 sampled languages. In high-resource languages, a large number of datasets are labeled manually, where in low-resource languages, the percentage of automatically induced datasets increases, with 135 languages have only automatically induced datasets. On (macro-)average, the 10 highest resource languages show 43.4% of their datatsets with only automatically induced labels; however, for all languages, 84.9% of the datasets use only automatically induced labels. Prior work often uses the total number of datasets in a target language as a proxy for resource availability of the language, which our analysis suggests is limited.

4.3 Translations

Many datasets are created by translating existing high-resource language datasets into target languages, which allows the creation of parallel data across many languages and removes reliance on the limited number of language-proficient annotators. Table 6 reports the number of datasets and language covered by different translation methods. Thirty-three datasets used some translation during the creation process, compared to 123 that

[*]Due to overlap of language covered among datasets, there are 40 unique languages instead of 53 (from Table 4).

[*]The "web" fine-grained category refers to the collection of sentences scraped or sampled at large-scale from the web.

Label source	Description	# Data	# Langs
annotated by authors or linguists	manual annotation by domain experts.	37	43
crowdsourced	manual annotation by crowdworkers.	63	56
automatically induced	automatically aligned or deduced from labeled or unlabeled data.	53	210
linguistic data	derived from curated linguistic resources (e.g., WordNet).	5	24
not mentioned	No details provided or inadequate documentation.	9	18

Table 5: Label collection method statistics. If dataset creation involved multiple methods (e.g., automatically induced and then manually verified by authors), they are counted for each dataset.

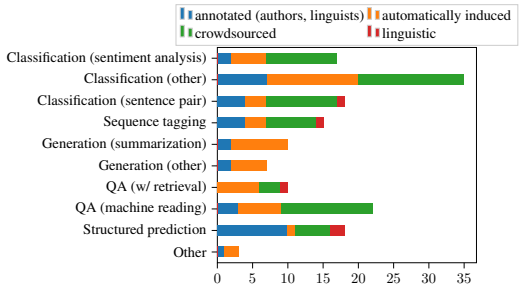


Figure 3: The distribution over label collection methods per task type. The size of bar for each collection method represents the number of datasets of that task type.

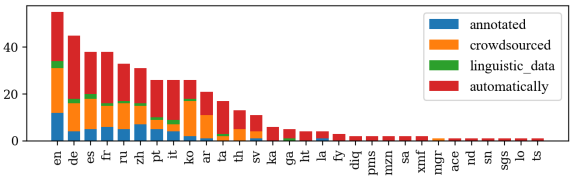


Figure 4: Label source per language for the top 10 and the 20 sampled languages below top 10.

Translation involved		# Data	# Langs
Yes	automatic	12	178
	human (author)	2	3
	human (non-author)	15	37
	unclear	7	8
No		123	185

Table 6: Statistics on translation involved.

did not. Quality issues can arise when using automatic translation for dataset creation; these include quality degradation for long sentences and translation artifacts (Lembersky et al., 2011; Eetemadi and Toutanova, 2014; Koehn and Knowles, 2017; Artetxe et al., 2020a); Clark et al. (2020) suggests that besides translation artifacts, translation-based approaches can result in data that does not reflect native speakers' interests. Despite these problems, automatic translation was still used for 12 datasets.

Input and Label Derivation via Translation. Translating English data into the target languages are used in 33 datasets, but most datasets collected data in its original language. Yet, many recent and highly cited datasets for cross-lingual transfer evaluation (Artetxe et al., 2020b; Conneau et al., 2018) are created with translation-based approaches, which we discuss in detail below.

4.4 Motivation for Dataset Creation

Table 7 summarizes statistics on the motivation aspect, with a breakdown for the number of datasets for each motivation with and without translation. The most frequent driver for dataset creation was to cover multiple languages for a single task (62 datasets, covering 217 languages), often for downstream tasks with high economic demands, such as QA or summarization (Blasi et al., 2022). There were 16 languages (e.g., Chinese, Arabic) that had their own benchmark suites labeled as the monolingual general model category, which seemed to align with the availability of language-proficient NLP researchers. We discuss the relationship between dataset availability and the number of language-proficient researchers in Section 5.

Motivation and Translation Used. The datasets studying cross-lingual transfer used translation at a much higher rate (61.9%) than datasets with other motivations. Monolingual task datasets rarely used translation (8.5%).

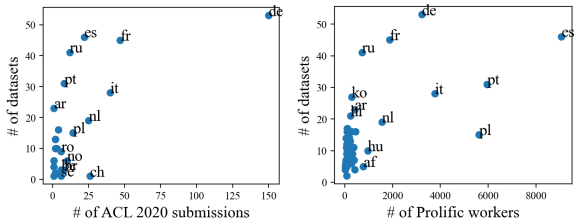
Motivation and Affiliation. We also find that (1) industry researchers focused on cross-lingual transfer (12 of 20 papers), while academic researchers focused more on single task-oriented (either multi- or monolingual task) benchmarks (67 of 80 papers), and (2) MRC and QA had a higher proportion of task-oriented datasets than other tasks, potentially because they are close to downstream products.

5 What's Needed to Develop NLP Datasets for Global Languages?

We study the building blocks to create multilingual datasets: (1) NLP researchers who speak the target language, and (2) ways to collect labels in the target language, including hiring crowdsourced workers.

Motivation	Translation No	Translation Yes	# Data	# Lang
cross-lingual transfer	8	13	21	48
multilingual task	52	11	62	217
monolingual task	32	3	35	25
monolingual general	31	6	37	16

Table 7: Motivations for dataset creation and their reliance on external translations.



(a) The # of the ACL 2020 submissions from the countries where the languages are spoken.

(b) The number of active crowdworkers on Prolific.

Figure 5: The number of non-English datasets and the factors that correlates. Each dot represents a language.

We computed the Pearson correlation coefficient (ρ) between the number of surveyed datasets and the proxy for each building block.[*]

Availability of NLP Researchers. Collecting data in a language without the availability of someone who understands the task and language to patrol data collection is challenging. We approximate the number of NLP researchers with language proficiency by the number of submissions to ACL 2020 from released general conference statistics.[*] We heuristically map country names to a set of languages commonly spoken in those countries. We use the number of submissions as a proxy to reflect research activities. Figure 5a shows a scatter plot for all surveyed languages except two highly dominant languages (i.e., English and Chinese). Its x-axis is the number of ACL 2020 submissions, and its y-axis is the number of labeled datasets, which are correlated with $\rho = 0.57$. The correlation between the number of manually annotated datasets and the number of researchers was even higher ($\rho = 0.71$). As the availability of NLP researchers affects datasets' availability, a linguistically diverse group of NLP researchers is required for equitable dataset development.

Availability of Crowdworkers. We estimate the demographics on the crowdsourcing platform using worker's demographic statistics per their first language listed on the academic research crowdsourcing Prolific,[*] a total of 136,884 workers are available, 70% of whom speak English as their first language, and the remaining 30% cover 60 additional languages. Figure 5b shows the relationship between the number of annotators and the number of datasets in our survey. We observe a weak correlation of $\rho = 0.58$ (and $\rho = 0.59$ when considering only crowdsourced datasets), possibly because Prolific is not yet widely used in the NLP community except for a handful of datasets (Liu et al., 2021), and our proxy might miss crowdworkers with proficiency in other languages besides their native one. However, no other platforms, including Amazon Mechanical Turk (MTurk)[*], the most widely used crowdsourcing platform according to our survey, provide statistics about annotators. To investigate the potential to gather high-quality multilingual data on English-centric crowdsourcing platforms, we conducted the following pilot study about multilingual worker availability on MTurk.

6 Pilot Study on Crowdsourcing Multilingual Data on MTurk

How easy is it to collect multilingual dataset on popular annotation platform (MTurk) at the moment? Crowdsourcing enables large-scale, cost efficient data collection; however, for many languages, the number of language-proficient crowdworkers is limited (Garcia et al., 2021). We quantify the availability of MTurk workers with proficiency in non-English languages.

We formulated a four-way sentiment analysis task using the Multilingual Amazon Review Corpus (Keung et al., 2020) and analyzed the annotation quality, cost, and time to finish tasks in English, Spanish, German, French, Japanese and Chinese of crowdworkers.

In all settings, we asked annotators to translate the same English sentence to assess their actual (rather than professed) language proficiency. We found that many production-level MT systems fail to translate his sentence due to its compositionality. Further, we investigated the newly introduced "language qualification" in MTurk, which available for only four aforementioned languages and Brazilian

[*]Appendix B.1 studies the availability of unlabeled text data in a target language and the number of surveyed datasets, finding a positive correlation, as in prior work (Joshi et al., 2020) which surveyed LDC datasets.

[*]https://acl2020.org/blog/general-conference-statistics/

[*]https://prolific.co

[*]https://www.mturk.com/

Portuguese as of 2022.

For our sentiment analysis task, without the language qualification, the accuracy of human binary classification performance in all non-English languages (55.2%) was significantly worse than that of English (77%). Past recommendations, such as constraining location and HIT acceptance rate, are insufficient (as of 2022) to collect high quality data even for languages considered "easier" to crowdsource in prior work (Pavlick et al., 2014). Language qualification improved performance by up to 40% and reduced the prevalence of cheating across all languages. However, with the language qualification, the data collection process usually took more time and cost ($1 per assignment). More pilot study details are in Appendix C.

Quality Control Using Translation Task. We investigate whether crowdworkers relied on automatic machine translation, despite our instruction saying **not** to use them. We ask native speakers to compare the crowdsourced translation with the translation results from three major translation platforms: Google Translate,[*] Microsoft Bing Translator,[*] and DeepL Translator.[*] Without language qualification, we identified 33% of crowdworkers copy-and-pasted automatic translation outputs (with qualification, 7%). This is significantly higher than what Pavlick et al. (2014) report (10%), suggesting more crowdworkers have started to use MT services.

We found that we could potentially use the translation task to identify good submissions: if we take only the submissions whose translation (1) do not match translation from MT and (2) valid translated judged by native speakers (labeled as either correct or partially correct),[*] binary task accuracy rises to 81.0% from 63.5%, matching the binary accuracy of English.

Our pilot study suggests that translation quality can reflect the target task performance if workers who copy from MT systems are filtered, and can be a good proxy for the languages without aforementioned language qualifications.

7 Discussion and Suggestions

This work provides the first large-scale meta survey on public multilingual NLP datasets, focusing on the novel aspects under-explored in prior work. We found that many languages lack a diverse set of **manually annotated** datasets and coverage of tasks and input&label sources. Particularly, except for summarization, generation tasks for non-English languages show limited language coverage. We conclude this work by presenting concrete suggestions to both the NLP community (Section 7.1) and to individual researchers aspiring to create new multilingual NLP datasets (Section 7.2).

7.1 Suggestions for the NLP Community

To Foster Language-proficient Researchers and Community Efforts. Our analysis shows that the availability of NLP researchers who are fluent in languages highly correlates with the availability of datasets. Moreover, monolingual test suites cover only 16 languages, such as Chinese (Xu et al., 2020), Indic Languages (Kakwani et al., 2020), Polish (Rybak et al., 2020), Persian (Khashabi et al., 2021), Russian (Shavrina et al., 2020) or Arabic (Seelawi et al., 2021), where efforts are driven by language-proficient NLP researchers. Organizing these large-scale, inter-organization efforts can be challenging but have profound effects. Recent community efforts such as Masakhane[*] spur research for under-resourced languages, resulting in new valuable resources for underrepresented languages (e.g., MasakhaNER; Adelani et al. 2021). Developing a directory of language-proficient NLP researchers interested in global collaboration could foster more cooperation. In the long run, globalized NLP education like AFIRM[*] will be necessary. A directory of potential funding sources to support multlingual data collection can also be helpful.

On Inclusive Venues. The academic publication/conference reviewing system should also reward efforts to develop language-specific resources, without perceiving this as a niche, low-impact effort (Rogers et al., 2022). As a community, we should encourage efforts to create and provide region-specific (e.g., Nordic Conference on Computational Linguistics, Pacific Asia Conference on Language, Information and Computation), language-oriented (e.g., Deep Learning for Low-Resource NLP, AfricaNLP, Workshop on Indian Language Data: Resources and Evaluation), and data-oriented (e.g., NeurIPS dataset and benchmark track) venues for introducing multilingual datasets.

[*] https://translate.google.com
[*] https://www.bing.com/translator
[*] https://www.deepl.com
[*] The details of translation quality study can be found in the appendix.

[*] https://www.masakhane.io/
[*] https://sigir.org/afirm2020/

On Multilingual Shared Tasks. Several recent shared tasks have driven dataset creation for both low-resource languages and novel tasks. For example, the WMT 2022 General MT task added four new languages pairs (e.g., Ukrainian), and MIA 2022 Workshop released the first annotated open-domain QA data in Tagalog and Tamil (Asai et al., 2022). Similarly, the WMT 2022 Large-scale Machine Translation Evaluation for African Language track[*] presents a data collection track for African languages. Large-scale multilingual NLP shared tasks have often focused on major, particularly European languages (Callison-Burch et al., 2010; Hajič, 2009), leaving many world languages behind. Adapting existing systems to new and low-resource languages poses a challenging and intriguing task as well as substantial research inquiries. The community should continue supporting such efforts and expand evaluation data for diverse target languages.

7.2 Suggestions for Individual Researchers

For Crowdsourcing. Our pilot study reveals both the difficulty of crowdsourcing for non-English languages and the high reliance on MT systems on English-centric platform. To conduct crowdsourcing on MTurk, one can either (1) adding language qualification newly introduced on MTurk for the 5 languages available, (2) introducing translation qualification and pruning workers based on their translation quality, and (3) translate original input into English and then crowdsource in English (Asai et al., 2021). We also recommend using language-specific crowdsourcing platforms, when available.[*] Alternative crowdsourcing platforms like Prolific or freelance platforms, such as CrowdFlower[*] or Upwork,[*] can be explored, though they tend to be more expensive.

For Translating English Datasets. Another option to create a multilingual dataset is to translate datasets in high-resource languages into target languages (Conneau et al., 2018; Lewis et al., 2020). Fortunately, there are many crowdsourced translation services that offer semi-professional translation at cheaper costs and better availabilities than translation services provided by trained professional translators. In our survey, Gengo[*] and One Hour Translation,[*] are the most highly used platforms for translation-based multilingual dataset creation. However, translation artifacts in these datasets remain unclear. Future studies can further quantify the quality of each translation method in existing translation-based datasets, the distribution of translation artifacts and mistakes, and the impact of such artifacts on final downstream task performances using our meta-annotations.

On Funding Sources. As discussed previously, multilingual dataset creation is often more expensive than English dataset creation. We summarize funding sources from the paper we surveyed that had over 50 citations. For general multilingual research, funding sources mostly consist of national funding agencies, such as the Spanish Ministry of Education and Science, the Catalan Secretary of Linguistic Policy, the Science Foundation Ireland, the Irish Research Council, the National Natural Science Foundation of China, the Department of Defense (e.g., DARPA, ARL, ARO), the US National Science Foundation, and the National Centre for Human Language Technology in the South African Department of Arts and Culture. For computational supports, researchers could apply to Google's Tensorflow Research Cloud and NVIDIA's academic hardware grant.

8 Conclusion

We present the first large-scale comprehensive survey on characteristics of multilingual datasets, exposing that the disparity among languages is not only quantitative (i.e., the number of the datasets) but also qualitative (e.g., *how* and *why* those datasets are created). We also discuss building blocks for constructing data resources, from language proficient researchers to crowdworkers. Our MTurk experiments show the challenges of quality and costs of annotating multilingual datasets on MTurk and suggest several approaches to tackle those challenges. We conclude our survey with a list of concrete suggestions for researchers interested in constructing language resources.

[*] https://statmt.org/wmt22/large-scale-multilingual-translation-task.html

[*] Toloka (https://toloka.yandex.com) is widely used by Russian language researchers. SelectStar (https://selectstar.ai) and DeepNatural (https://deepnatural.ai) are South Korea-based crowdsourcing platforms.

[*] https://visit.figure-eight.com/People-Powered-Data-Enrichment_T

[*] https://www.upwork.com

[*] https://gengo.com

[*] https://onehourtranslation.com

Limitations

On Dataset Documentation. Throughout our survey process, we found inadequate dataset documentation, limiting the coverage of our survey. We suggest that individual researchers provide the input data source and the labeling methodology; if people were involved in dataset creation, their demographic information should be provided, as well. Such information can help researchers analyze potential bias embedded in the dataset (Bender and Friedman, 2018).

Surveyed Dataset Collection Process. Despite our best efforts, we do not claim to cover all relevant datasets. Our collection process overlooks datasets that are published at non-ACL venues and not in Hugging Face as well as papers that do not match our search keywords. For instance, we missed multilingual machine reading comprehension datasets (Gupta and Khade, 2020; Asai et al., 2018) and morphology datasets (McCarthy et al., 2020). We also found a very low presence of indigenous language datasets. None of 10 indigenous American languages from a recent study (Ebrahimi et al., 2022) was represented in our survey. That said, we host `http://multilingual-dataset-survey.github.io` where researchers can submit their dataset information and periodically update our analysis. Furthermore, we constantly encountered poorly written documentation or unavailable datasets during our annotation processes. During annotation, whenever this paper's dataset annotators encountered unclear documentation, they made their best guess to put datasets into predefined categories. If no evidence could be found for the inference, they put "not mentioned" as a result. All unclear decision were adjudicated by at least three annotators.

Using Country Names as a Proxy for Languages Spoken. In Section 5, we attempted to approximate the number of NLP researchers with language proficiency in different languages. To do this, we mapped the names of ACL submission country to the most commonly spoken languages in those countries. We acknowledge that (1) the country of origin of researchers might be different from the country of submission, (2) researchers native language might not be listed in the commonly spoken languages and (3) the mapping might be incomprehensive.

MTurk Pilot Study. Due to our limited data points, although our MTurk study showed that data quality could be improved if the language qualification were applied in the collection process on MTurk, and our previous recommendations do not currently apply, we acknowledge that more research at scale should be done to statistically confirm the conclusion. Furthermore, languages other than the supported 5 languages might still be unsuitable for gathering multilingual data on MTurk

Acknowledgements

We thank Melanie Sclar, Thibault Sellam, and Tobias Rohde for evaluating our mTurk experiment translations. We thank Alisa Liu, Peter West, Zhaofeng Wu, Li Du, Kyle Mahowald and the members in the UW NLP group for their helpful feedback on this work.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020a. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020b. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 4623–4637, Online. Association for Computational Linguistics.

Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. *arXiv preprint arXiv:1809.03275*.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.

Akari Asai, Shayne Longpre, Jungo Kasai, Chia-Hsuan Lee, Rui Zhang, Junjie Hu, Ikuya Yamada, Jonathan H. Clark, and Eunsol Choi. 2022. MIA 2022 shared task: Evaluating cross-lingual open-retrieval question answering for 16 diverse languages. In *Proceedings of the Workshop on Multilingual Information Access (MIA)*, pages 108–120, Seattle, USA. Association for Computational Linguistics.

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.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Michael Bloodgood and Chris Callison-Burch. 2010. Using Mechanical Turk to build machine translation evaluation sets. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 208–211, Los Angeles. Association for Computational Linguistics.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Singapore. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Sauleh Eetemadi and Kristina Toutanova. 2014. Asymmetric features of human generated translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 159–164, Doha, Qatar. Association for Computational Linguistics.

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.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. 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 2730–2741, Online. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Somil Gupta and Nilesh Khade. 2020. Bert based multilingual machine comprehension in english and hindi. *arXiv preprint arXiv:2006.01432*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. 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 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Jan Hajič, editor. 2009. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*. Association for Computational Linguistics, Boulder, Colorado.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

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 *International Conference on Machine Learning*, pages 4411–4421. PMLR.

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.

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, Online. Association for Computational Linguistics.

Daniel Kang and Tatsunori B. Hashimoto. 2020. Improved natural language generation via loss truncation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 718–731, Online. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabagdi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, Erfan Sadeqi Azer, Niloofar Safi Samghabadi, Mahsa Shafaei, Saber Sheybani, Ali Tazarv, and Yadollah Yaghoobzadeh. 2021. ParsiNLU: A suite of language understanding challenges for Persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162.

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 and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. 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.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2011. Language models for machine translation: Original vs. translated texts. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 363–374, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. Korquad1.0: Korean qa dataset for machine reading comprehension. *ArXiv*, abs/1909.07005.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019. XQA: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368, Florence, Italy. Association for Computational Linguistics.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2022. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. KLEJ: Comprehensive benchmark for Polish language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1191–1201, Online. Association for Computational Linguistics.

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.

Haitham Seelawi, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. 2021. ALUE: Arabic language understanding evaluation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 173–184, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on 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.

Appendix

A Additional Data Analysis

A.1 Task types and input sources.

We anticipate correlations between the task type and the source of input text. We visualize it in Figure 6, and highlight a few findings. First, QA (with and without retrieval) datasets are often created using either Wikipedia articles such as d'Hoffschmidt et al. (2020) or curated sources such as exams and scientific papers (Vilares and Gómez-Rodríguez, 2019), while social media such as Twitter or commercial websites like Amazon.com are mainly used to construct sentiment analysis datasets. Secondly, summarization dataset mostly derived from news media, MRC and sentence pair classification tasks often involved multiple input sources. For instance, evidence passage can come from existing Wikipedia passage, but the question is crowdsourced (Lewis et al., 2020).

B Details of the Meta Analysis

B.1 Availability of Unlabeled Text

Unlabeled text can be used for pre-training (Blasi et al., 2022) or as input sources of the new labeled datasets. Joshi et al. (2020) reports a correlation between the amount of unlabeled data such as Wikipedia articles and the number of datasets on the LDC catalog.* We study the correlation between unlabeled corpora and our surveyed datasets, most of which are not included in licensed LDC. As we identify that many datasets use texts beyond Wikipedia (see Table 4), we instead use the mC4 corpora (Xue et al., 2021), a much larger collection of texts in 101 languages drawn from the public Common Crawl web scrape and used for training the mT5 model. Specifically, we use the number of tokens in mC4 to estimate the amount of unlabeled data. Figure 7a shows a scatter plot where the x-axis represents the number of tokens in mC4 and y-axis represents the number of labeled datasets available in the languages. The availability of unlabeled text corpora and the number of labeled datasets show a high correlation ($\rho = 0.794$). We also analyze the relationship between the number of labeled datasets and the number of Wikipedia articles in Figure 7b. Again we observe a high correlation of $\rho = 0.767$.

*https://catalog.ldc.upenn.edu/

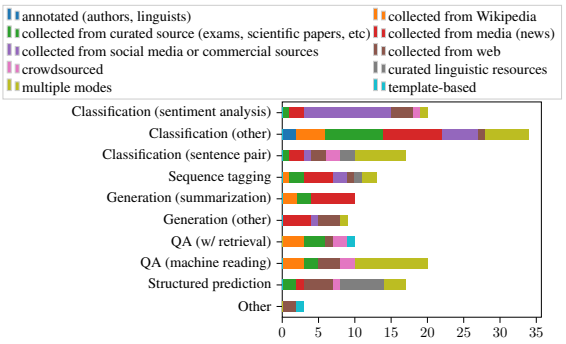
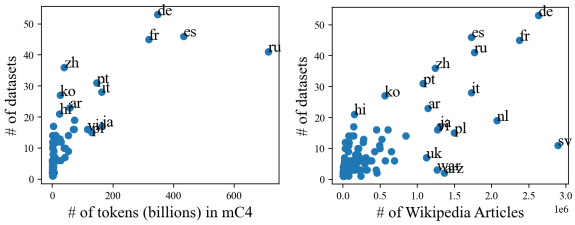


Figure 6: The distribution over input source per task type. The size of bar represents the # of datasets of that task type and input source.



(a) The number of tokens (billions) in mC4 (Xue et al., 2021).

(b) The number of Wikipedia articles.

Figure 7: Relationships between the number of datasets and number of tokens in mC4 and Wikipedia articles. English is removed.

C Pilot Study: Investigating the Viability of Crowdsourcing for Six Languages

Previous work (Callison-Burch, 2009; Bloodgood and Callison-Burch, 2010) studied the feasibility of using crowdsourcing platform to evaluate machine translation systems. Pavlick et al. (2014) expanded the study to translating 100 languages and recommended several "best" languages (high quality results with fast completion speed) to target on MTurk platform. We re-visit the worker availability eight years later, comparing our findings with the previous findings.

C.1 Task Design

We design a sentiment analysis task that is trivial for a native speaker but is challenging for someone who does not have native proficiency, along with a translation task to evaluate workers' true proficiency in the language of interest. Our sample interface layout is shown in 8.

Source data and languages. We use the Multilingual Amazon Review Corpus (MARC; Keung et al. 2020), which contains 5-way sentiment labels for reviews from Amazon in English, German, French, Spanish, Japanese and Chinese (Mandarin). While all of them are relatively high-resource lan-

Language	Afrikaans	Albanian	Amharic	Arabic	Armenian	Basque	Bengali
# Workers	774	70	25	417	39	27	192
Language	Bulgarian	Cantonese	Catalan	**Chinese**	Croatian	Czech	Danish
# Workers	168	80	149	**1241**	101	324	156
Language	**Dutch**	**English**	Estonian	Farsi	Finnish	**French**	**German**
# Workers	**1551**	**93056**	375	127	219	**1870**	**3220**
Language	**Greek**	Gujarati	Hebrew	Hindi	Hungarian	Icelandic	Indonesian
# Workers	**1391**	74	387	245	955	25	105
Language	**Italian**	Japanese	Khmer	Korean	Latvian	Lithuanian	Macedonian
# Workers	**3762**	95	27	287	273	133	30
Language	Malay	Malayalam	Mandarin	Nepali	Norwegian	**Polish**	**Portuguese**
# Workers	41	64	34	88	129	**5609**	**5948**
Language	Punjabi	Romanian	Russian	Serbian	Slovak	Slovenian	**Spanish**
# Workers	96	328	715	76	62	377	**9060**
Language	Swahili	Swedish	Tagalog-Filipino	Tamil	Telugu	Thai	Turkish
# Workers	89	377	422	108	67	43	288
Language	Twi	Ukrainian	Urdu	Vietnamese	Welsh		
# Worker	31	52	276	445	89		

Table 8: The number of active crowd-workers in the last 90 days available on Prolific (as of January 6th, 2022) in terms of first language. Note that the number of active workers that has a first language of either Belarusian, Burmese, Dari, Dzongkha, Esperanto, Faroese, Gaelic, Galician, Georgian, Hakka, Inuktitut (Eskimo), Kurdish, Laotian, Lappish, Maltese, Papiamento, Pashto, Scots, Somali, Tajik, Tibetan, Tigrinya, Tongan, Turkmen, or Uzbek is less than 25, and therefore we omit them in the table.

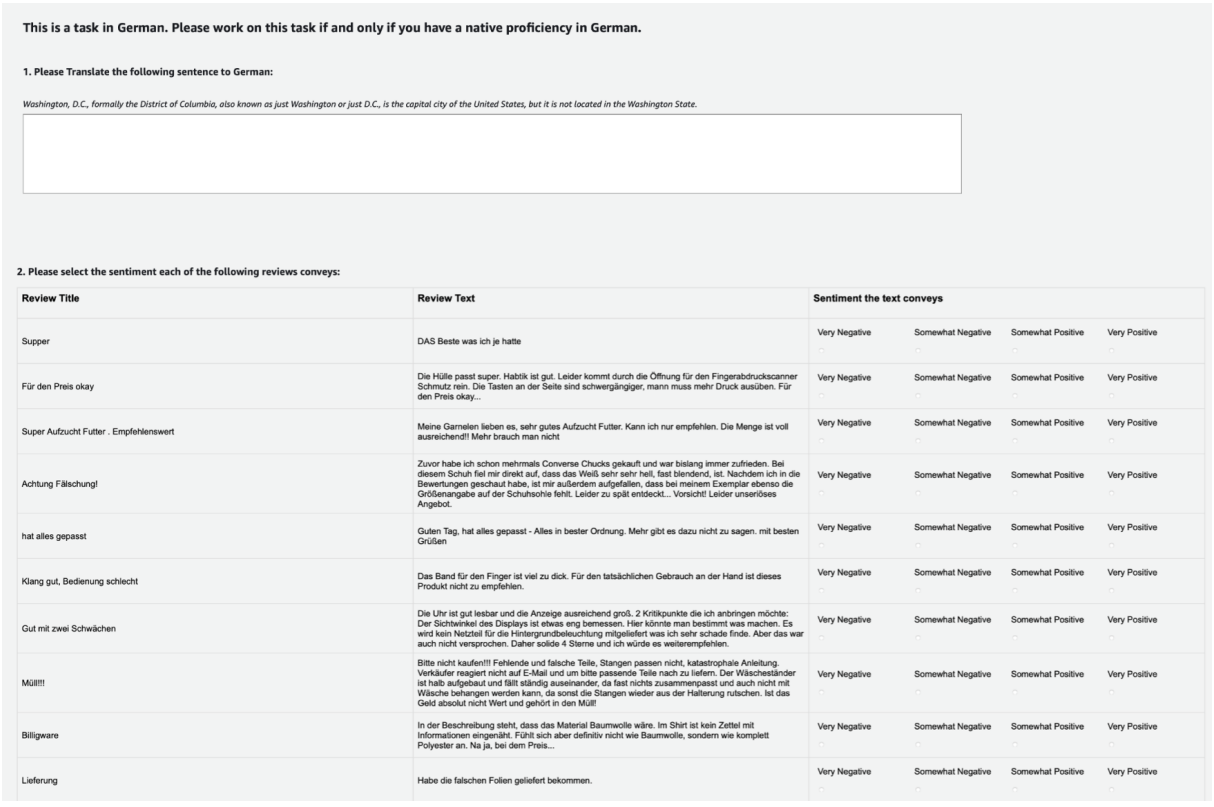


Figure 8: Layout of the tasks on Amazon Mechanical Turk for German

guages, their numbers of the available crowdworkers varies significantly (e.g., English has 93K annotators while Japanese only has 95 on Prolific).

Sentiment analysis. We extract the reviews with 1 (very negative), 2 (somewhat negative), 4 (somewhat positive) and 5 (very positive) stars, omitting reviews with 3 stars to keep annotation task less ambiguous. We evenly sample 5 reviews of each of four ratings for each respective language. We compute four-way classification accuracy, and binary

classification accuracy by merging 1, 2-star ratings as *negative* and 4, 5-star ratings as *positive*.

Translation. For non-English tasks, as an additional check (for worker's understanding ability and cheating detection), we require crowdworkers to translate a simple yet compositional English sentence from an Wikipedia article,[*] *"Washington, D.C., formally the District of Columbia, also known as just Washington or just D.C., is the capital city of the United States, but it is not located in the Washington State."* Some widely-used online translation tools give sub-par translations on this sentence because of its compositional structure. We forbid workers to use online translation platforms in the task prompt.

We collect the gold translation from native speakers as references. In addition, we ask them to rate the collected translations as correct (3), partially correct (2), or incorrect (1).[*] We also compute the BLEU score against the reference answer.

We present some sample translation results as well as the gold translation provided by human native speaker annotators in Table 9. We found that the most easily made mistake is the misreading of the word "formally" to "formerly", which also exists from the google translated German sentence, and it also happened when we initially ask native speakers to translate the sentence.

C.2 Task Setting

Worker qualification. We use following qualifications provided by MTurk: (i) workers must have at least a 95% acceptance rate, (ii) workers must be in the US or Top 5 countries with the largest population speaking for each language from WorldData.info,[*] and (iii) for Spanish, German, French and Chinese with "Language Fluency (Basic)" premium qualifications available, we collect labels on the same data with and without this qualification. Note that as of 2022, it's only available for the four aforementioned languages and Brazilian Portuguese with an additional $1 per assignment.

Task statistics. We collect 20 sentiment annotations per languages along with one translation example, each of which accepts up to 10 unique MTurkers' annotations, resulting in 200 annotations for all six languages (**without language qualification**). In Spanish, Germany, French and Chinese, we release the same HITs with the language proficiency requirements, resulting in additional 200 annotations (**with language qualification**). We aimed at an hourly pay of $12 for this task.

C.3 Evaluation and Analysis

Table 10 shows the time elapsed, the number of annotations collected, the four-way and binary accuracy for the sentiment analysis task, the average human evaluation score and sacreBLEU (Post, 2018) score for translation task for each of the language experiments with and without the qualification.

Time elapsed to collect annotations. Without language qualification, all annotations finished in a single day, with English being the fastest, and Chinese being the slowest (0.8 v.s. 6.9 hours). When language proficiency criteria is added, the tasks expired after four days without gathering all annotations, with Spanish being the most available and Chinese being the least available (140 v.s. 30 annotations). Conversely, their language qualification may overly shrinks the worker pool; according to a web forum among crowdworkers,[*] paths to acquiring this language qualification is unclear. We assume that many workers with proficiency in the target language might not obtain the qualification.

Annotation quality on sentiment analysis task. We evaluate the binary and 4-way classification accuracy. Random baseline yield 50% and 25% accuracy, respectively. Although Pavlick et al. (2014) recommends French, German, and Spanish to be some of the best target language on MTurk, our results were unsatisfactory on these languages; without language qualification, the classification performance for all languages is significantly worse than English (77%), indicating that only by constraining location and HIT acceptance is insufficient as of 2022. Language qualification improves performance across languages (e.g., 46.5% and 30.6% 4-way accuracy improvements in Germany and Spanish, respectively).

Annotation quality on translation task. Table 10 shows that the average sacreBLEU score without language qualification is significantly lower than the one with qualification for all the languages. As shown in Table 10, human rating

[*] https://en.wikipedia.org/wiki/Washington,_D.C.

[*] A partially correct translation has some minor grammatical errors or lose details, but overall conveys the information.

[*] https://www.worlddata.info

[*] https://turkerview.com/qualifeye/

lang	Gold tranlsation	w/qual translation samples	w/o qual translation samples
de	Washington, D.C., formal der District of Columbia, auch bekannt als Washington oder nur D.C., ist die Hauptstadt der Vereinigten Staaten, liegt aber nicht im Staat Washington.	Washington, D.C., früher District of Columbia, auch nur Washington oder nur D.C. genannt, ist die Hauptstadt der Vereinigten Staaten, befindet sich jedoch nicht im Bundesstaat Washington. Washington, D.C., früher District of Columbia, auch nur Washington oder nur D.C. genannt, ist die Hauptstadt der Vereinigten Staaten, befindet sich jedoch nicht im Bundesstaat Washington	Washington D.C., formal "District of Columbia" genannt, auch bekannt als "Washington" oder nur "D.C." ist die Hauptstadt der USA, liegt aber nicht im Staat Washington Washington D.C., formal der "District of Columbia", auch bekannt als "Washington" oder nur "D.C.", ist die Hauptstadt der USA, liegt aber nicht im Staat Washington.
es	Washington, D.C., formalmente el Distrito de Columbia, también conocido simplemente como Washington o como D.C., es la ciudad capital de los Estados Unidos, pero no se encuentra en el Estado de Washington.	Washington, D.C., formalmente el Distrito de Columbia, también conocido simplemente como Washington o simplemente D.C., es la ciudad capital de los Estados Unidos, pero no está ubicada en el estado de Washington. Washington,D.C.,formalmente el Distrito de Columbia, también conocido simplemente como Washington o simplemente D.C.,es la ciudad capital de los Estados Unidos,pero no se encuentra enel estado de Washington.	Washingto, D.c.,formalmente el Distrito de Columbia, tambein conocido simplemente como Washington o simplemente D.C., es la ciudad capital de los estados unidos, pero no se encuentra en el estado de washington. Washington, D.C., oficialmente el Distrito de Columbia, también conocido tan solo como Washington, o únicamente D.C., es la ciudad capital de los Estados Unidos, pero no se encuentra en el estado de Washington.
fr	Washington, D.C., officiellement nommée District of Columbia, aussi connue sous le simple nom de Washington ou juste D.C., est la capitale des Etats-Unis, mais elle n'est pas située dans l'Etat de Washington.	Washington, D.C., anciennement le District de Columbia, également connu sous le nom de Washington ou simplement D.C., est la capitale des États-Unis, mais elle n'est pas située dans l'État de Washington. Washington, D.C., anciennement le District de Columbia, également connu sous le nom de Washington ou simplement D.C., est la capitale des États-Unis, mais elle n'est pas située dans l'État de Washington.	Washington, D.C, officiellement le District de Columbia, aussi connue juste comme Washington ou juste D.C, est la capitale des Etats-Unis, mais n'est pas située dans l'etat de Washington. Washington D.C, officiellement The District of Columbia, également connue sous le nom de Washington, ou simplement D.C, est la capitale des États-Unis, mais n'est pas située dans l'état de Washington.
ja	ワシントンDCは公式にはコロンビア特別区、もしくは単にワシントンおよびDCと呼ばれ、アメリカ合衆国の首都ではあるがワシントン州に位置しているわけではない。	ワシントンD.C.、正式にはコロンビア特別区、別名ワシントンD.C.は米国の首都ですが、ワシントン州にはありません。	—
zh-cn	华盛顿特区，正式名称为哥伦比亚特区，也被称为华盛顿或D.C.，是美国的首都，但它并不位于华盛顿州内。	国会图书馆，华盛顿特区。新的联邦领土被命名为哥伦比亚特区，以纪念探险家克里斯托弗·哥伦布，新的联邦城市以乔治·华盛顿的名字命名	华盛顿DC，理论上叫哥伦比亚特区，也被称作华盛顿或者DC，是美国的首都，但是不位于华盛顿州
zh-tw	華盛頓特區，正式名稱為哥倫比亞特區，也被稱為華盛頓或D.C.，是美國的首都，但它並不位於華盛頓州內。	華盛頓特區，正式名稱為哥倫比亞特區，也稱為華盛頓或華盛頓特區，是美國的首都，但並不位於華盛頓州。	—

Table 9: Sample translation results, We provide gold translation and select samples for both simplified Chinese and traditional Chinese to display.

with language qualification is higher (e.g., 2.5 v.s. 1.7 in German). The human evaluation correlates with sacreBLEU metric ($\rho = 0.79$). Without language qualification, it is non trivial to collect high-quality translation data from workers proficient in the target language.

language	# annotations	time (hour)	# matched	4-way acc(%)	binary acc(%)	human score	sacreBLEU
English	200 / –	0.83 / –	–	42.0 / –	**77.0** / –	–	–
Spanish	200 / 140	1.43 / 96	6 / 0	28.0 / 58.6	51.5 / 95.0	**1.80** / 2.36	39.09 / **63.09**
German	200 / 40	1.30 / 96	11 / 0	23.5 / **70.0**	*48.5* / 95.0	1.70 / **2.50**	**39.13** / 57.33
French	200 / 60	1.25 / 96	8 / 2	*23.0* / 50.0	57.0 / **100**	1.55 / 2.33	33.30 / 49.33
Japanese	200 / –	3.78 / –	10 / –	**44.0** / –	65.5 / –	1.74 / –	10.20 / –
Chinese	200 / 30	6.98 / 96	5 / 0	29.5 / 56.7	53.5 / 80.0	1.60 / 2.33	32.03 / 42.63

Table 10: Results from MTurk pilot study for data collection on six languages. Each cell reports the results from without / with language qualification. We report the number collected annotations, elapsed time to finish (max 96 hours which is when the HIT expires), the number of annotators whose translation matched the output from MT systems, average binary classification accuracy and 4-way classification accuracy of sentiment analysis task, average human evaluation score and sacreBLEU score for translation task.