Toward More Meaningful Resources for Lower-resourced Languages

Anonymous ACL submission

Abstract

In this paper, we describe our perspective on how meaningful resources for lower-resourced languages can be developed in connection with the speakers of those languages. We examine two massively multilingual resources in detail. We explore the contents of the names stored in Wikidata for a few lower-resourced languages and find that many of them are not in fact in the languages they claim to be and require non-trivial effort to correct. We discuss quality issues present in WikiAnn and evaluate whether it is a useful supplement to hand annotated data. We then discuss the importance of creating annotation for lower-resourced languages in a thoughtful and ethical way that includes the languages' speakers as part of the development process. We conclude with recommended guidelines for resource development.

1 Introduction

Recent years have seen increased interest from the ACL community in developing both models and datasets for what may be termed "lower-resourced" languages. Advances in transfer learning and the increased availability of data and benchmarks in these languages have made it straightforward to create what appear to be high-performing models for these languages with little or no annotated data.

Yet despite the popularity and apparent effectiveness of these systems, the unique challenges and best practices for developing datasets and models for lower-resourced languages are rarely discussed alongside the system themselves. In this paper, through commentary on our experiences working with existing datasets and a discussion of current trends, we explore what resources can be most useful for the development of meaningful language technology for lower-resourced languages, advancing our perspective that *open* data and models and a *participatory* approach to research are the keys to progress.

Before we can elaborate on our perspective, we must address a terminological issue which creates a stumbling block when attempting to discuss resources and language technology for the the languages of interest to us. Throughout this paper, we will use the term *lower-resourced language* to refer to languages that have received fewer resources—as measured in any numbers of dimensions such as models, datasets, papers, funding, etc.—than the most popularly-studied languages in the field of natural language processing. We explicitly use the comparative *lower* rather than *low* to emphasize the continuum that exists across languages regarding the resources available for developing language technology.

We also acknowledge that whether a language is lower-resourced in a specific context may depend on what is available for the task at hand. Due to singular efforts, a language that may have otherwise been underserved by the research community may have a rich set of resources for a single task like machine translation, but might not have annotation for other tasks. For example, consider the Inuktitut language, which has a substantial amount of government-domain machine translation data (Joanis et al., 2020) that enabled a shared task (Barrault et al., 2020) creating many machine translation models, but has few labeled datasets for other tasks and is not included in any large multilingual language models we are aware of.

We take an open and intersectional perspective to what might be called a lower-resourced language, acknowledging that this designation is both imperfect and often the result of many contributing factors. For example, many languages referred to in this way may be less-widely spoken, underserved by the research community and funding agencies, or used by marginalized or minoritized populations. In summary, our use of the term *lower-resourced* is intended to reflect a continuous, not categorical, status and one that is multi-

dimensional, depending on task and context.

This paper's contributions come at two levels. At the more concrete level, we discuss particular issues related to using Wikidata and WikiAnn as sources of information about names, and highlight how automatic processes to take advantage of this information can go wrong when no human expertise is involved. At a higher level, we discuss the problematic nature of developing language technology datasets and models without no or limited interaction with the population of speakers of the languages involved.

The goal of this paper is argue for our perspective regarding what are the most effective ways to construct and use resources for building language technology for lower-resourced languages. While anonymity poses a challenge to being transparent about the experiences that have helped shape our views, we will provide this relevant background to our perspective: we are researchers who are deeply committed to building impactful language technology and we work with the speakers of lower-resourced languages to develop datasets and models that benefit their language communities.

The structure of the paper is as follows. We first begin a review of two popular resources used in NLP for lower-resourced languages and the impact that they have had on the field. While we will discuss shortcomings of these resources, sometimes demonstrating them with experiments, the goal of this section is not to publish a critique of these resources, but rather to make other researchers aware of their shortcomings and limitations. By acknowledging and understanding their limitations, we can better understand how to use them most appropriately and develop future resources that do not share the same limitations.

We then turn to the importance of annotation and dataset creation processes that meaningfully involve speakers of the languages under study. We discuss open challenges for NLP for lower-resourced languages, and conclude with suggested guidelines for researchers performing research in this area.

2 Wikidata: A source for name labels

Wikidata[1] is an open and collaboratively edited knowledge graph, hosted by the Wikimedia Foundation. The Wikidata graph consists of entity nodes connected by labeled edges that represent relations. Each entity and relation is identified by a unique Wikidata identifier, e.g. Q4346375 (*Association for Computational Linguistics*) and P361 (`part-of` relation).

While English labels are typically used for the page titles on the Wikidata website, most entities have labels available in several languages, with the most well-edited entries having labels in hundreds of languages. This makes Wikidata an appealing data source for constructing multilingual NLP resources related to entity names, as parallel names can in theory be trivially extracted from each entity. For example, Wikidata could be used to harvest name lists for a named entity recognition (NER) system, or as a source of parallel names for translation or transliteration models. In this section, we show Wikidata's promise for extracting names in lower-resourced languages as well as the data issues that arise in attempting to use it for this purpose.

2.1 Name quality in lower-resourced languages

Given the limited name-related annotation available for many lower-resourced languages, Wikidata is a promising source of information regarding entity names. For example, previous work has used it as a source of data for multilingual name transliteration (Benites et al., 2020; Irvine et al., 2010). Specifically for lower-resourced languages, many approaches to NER and linking for the LORELEI program (Strassel and Tracey, 2016) used Wikidata, Wikipedia, DBpedia, GeoNames, and similar resources to provide name lists relevant to the languages and regions for which systems were developed.

However, there are many caveats hidden in the data present in Wikidata and using the contents without scrutiny can be problematic. One such caveat is the mixing of languages and scripts occurring within the entity labels of a single language, especially lower-resourced ones.

Tigrinya, a Semitic language spoken in Africa by over 9 million speakers, is a particularly good example to explore, as it is written using the Ge'ez script and has only 539 entity labels in Wikidata. However, only 269 out of 539 labels are actually written in the Ge'ez script, with the rest being in Latin script.[2] This problem is particularly

[1] https://www.wikidata.org

[2] We have confirmed with a native speaker of the language that this does not represent meaningful variation where some names may be borrowed in Latin script; they believed only

pronounced among entities referring to persons, where only 36 entities are written in Ge'ez, and 245 in Latin script. Out of all Tigrinya entity labels, nearly 50% are identical to the English label.

While we have not done an exhaustive analysis, we believe that many other lower-resourced languages are affected by issues of this type. Another example is Inuktitut, an Indigenous and historically minoritized language spoken in the Canadian Arctic by approximately 35,000–40,000 people, contains 25,222 entity labels in Wikidata, yet only 429 are written in the official Inuktitut syllabics. Out of all Inuktitut labels, 97% are identical to the corresponding English label, which suggests the entities are not only in the wrong script, but also the wrong language.

This vast number of Latin script labels relative to the ones written in Ge'ez script and Inuktitut syllabics suggests that labels are potentially being copied over from other languages, presumably as a result of bot activity. This script pollution may have adverse effects, particularly when the amount of data in the desired script is very small, not only on training models on the raw data but also on any heuristic filtering methods that try to, for example, filter out all entity labels not in the most common script for the language (which may end up being the incorrect script).

2.2 Name copying

Another minority language heavily impacted by likely copying is Asturian, a language spoken by 100,000–450,000 speakers in Spain. Wikidata contains over 5 million entity labels for Asturian, with 97% of them identical to the English label. For comparison, 93.5% out of 5.8 million Spanish entity labels are identical to English.

Overlap with English labels is not necessarily indicative of the labels being incorrect, as both Asturian and Spanish use the Latin alphabet and many named entities, particularly persons and organizations, can be written in the same way across languages. However, the vast number of labels relative to the number of Asturian speakers (and proportionately, active Wikidata editors), and the extra-high level of English-matching suggests that labels are being copied from other languages. This automated copying is widespread, so much so that Asturian ranks as the fourth largest language in Wikidata as measured by number of entity labels, following English, Dutch, and Spanish.

All in all, these examples show that extra care must be taken when harvesting multilingual data for lower-resourced languages in cases where it is possible that data may have been copied from a higher-resourced language. This problem is most visible in cases where a language uses a non-Latin script, but is likely to exist for many other languages. Failing to exercise caution may result in creating low-quality derived datasets that may do more harm than good. For example, if a multilingual dataset contains a large number of incorrect copies of English in other languages, it may make tasks appear easier than they are because of trivial transfer from English.

2.3 Summary

In spite of the shortcomings in data quality, we still believe Wikidata may be a valuable resource for language technology development, provided that enough effort is invested in data cleaning and validation. This process can take many forms depending on the application and could include, among other things, automated identification and filtering of languages and scripts, analysis of label copying from higher-resourced languages, and even analysis of who is making edits to Wikidata (for example, to identify automated edits).

Instead of shunning Wikidata, we encourage researchers to contribute to making it better for the global research community, and especially for lower-resourced languages for which Wikidata may be one of the only resources available. We also encourage researchers to use Wikidata collaboratively with the speakers of lower-resourced languages who can provide guidance on the quality of resources derived from it and help with the process of removing incorrect information.

3 WikiAnn

We now turn our attention to a different Wikipedia-related resource, one derived from it. WikiAnn (Pan et al., 2017) is a dataset originally created for named entity tagging and linking of 282 languages present in Wikipedia. Pan et al. generate "silver-standard" named entity annotations "by transferring annotations from English to other languages through cross-lingual links and KB properties, refining annotations through self-training and topic selection, deriving language-specific morphology features from anchor links, and mining word trans-

Ge'ez script names should count as valid data.

lation pairs from cross-lingual links."

3.1 A "silver" standard for multilingual NER

Since the WikiAnn dataset was created it has been used as a multilingual NER benchmark and it is included as part of the XTREME (Hu et al., 2020) massively multilingual multi-task benchmark. It is not uncommon for WikiAnn to be mentioned as a multilingual NER benchmark with sometimes no mention of the fact that it is system output and not, in fact, annotation (despite the *Ann* in the name).

We question the appropriateness of treating WikiAnn as a multilingual benchmark for NER. Even in a higher resourced language, the practice of evaluating a task on automatically derived data is sub-standard, hence the original authors referring to it as a "silver standard." This kind of evaluation can only show how close one model comes to replicating the automated data collection process and does not reflect human performance.

Just examining the English data of WikiAnn reveals a number of entity names that would otherwise never be marked as names by a human. Strings of text such as *Independently released*, *If I were a boy*, and *List of books written by teenagers* are annotated as organizations. *I was glad*, *the latter's studio*, and *were promoted* are annotated as locations, and *range has expanded*, *a twelve-year-old passenger was found alive* and *Artavasdes II, who served as* are tagged as person. While these examples are a small sample, we identified hundreds like these with either span issues, Wikipedia-specific entities like lists, incorrect entity types or entities that were simply not names. One might argue that this represents less than 1% of the English WikiAnn data and is therefore noise. However, human annotation does not typically have such mistakes and also has the benefit of being able to report inter annotator agreement so that researchers can better understand the difficulty of the task.

3.2 Subsampled splits: Unnecessarily discarded data

There can be some confusion as to which dataset is referred to by "WikiAnn." The original WikiAnn paper (Pan et al., 2017) contains data in 282 languages, but there is also another version derived from the original with only 176 languages (Rahimi et al., 2019) with fixed data splits. This derived version is the one that can be found in Hugging Face's Datasets.[3] The original (unsplit) WikiAnn datasets can be found on Google Drive.[4]

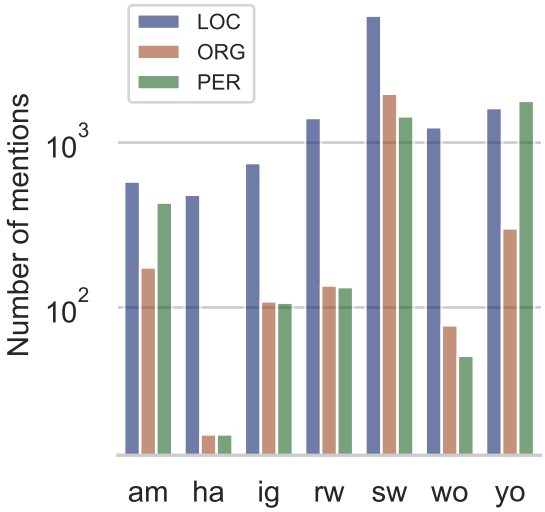

Figure 1: Counts of mentions in the full WikiAnn datasets in seven African languages that will be of interest in experiments in Section 3.5. A logarithmic scale is used so that all languages can be visualized.

We describe the differences between the two versions of the dataset in detail because many may use the currently popular Hugging Face Datasets library (Lhoest et al., 2021) and never realize they have access to less data than is in the original. Additionally, we highlight how the subsampled data no longer reflects a natural distribution of entity types and discards substantial amounts of data.

The original WikiAnn datasets are much larger than what was kept in the splits used by Rahimi et al. (2019). As can be seen in Figure 1, in the original data the languages display an uneven distribution of mention types, with most languages having far more mentions of LOC than ORG or PER.[5] This likely reflects a mix of the true distribution of named entities in the data and the fact that recall is typically highest for LOC entities.

Rahimi et al. (2019) used stratified sampling to select sentences for inclusion in their splits. The process is as follows: first, the sentences in the dataset are categorized into three groups (LOC,

[3] https://HuggingFace.co/datasets/wikiann

[4] https://drive.google.com/drive/folders/1bkK6ly_awxe9IgAKL16VVvCtjcYcDSw8

[5] In all figures showing WikiAnn data, we use Wikimedia language codes rather than ISO 639-3 codes for consistency with the original data.

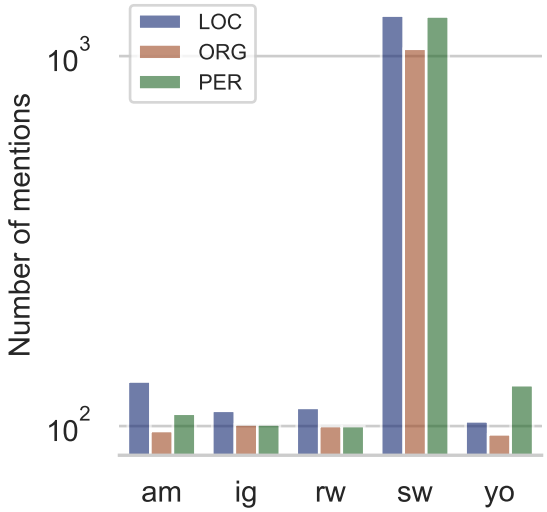

Figure 2: Counts of mentions in the stratified data selected by (Rahimi et al., 2019). A logarithmic scale is used so that all languages can be visualized.

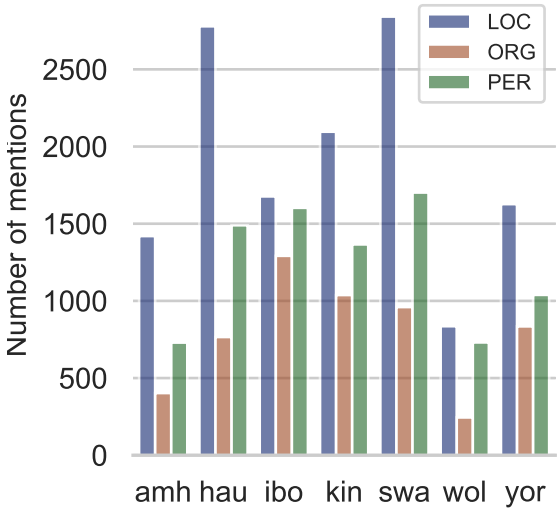

Figure 3: Counts of mentions in the MasakhaNER data for languages present in WikiAnn. The DATE tag was excluded for consistency with WikiAnn.

ORG, PER) based on the entity type of the last mention in the sentence. The size of the smallest of these groups is defined to be the minimum count, and this number of sentences is taken from each group and added to a new list of sentences. This list is shuffled, and if it is large enough, will be used to make 10,000/10,000/10,000 splits. If not, the splits will be 1,000/1,000/1,000, and if there are not enough sentences for that, then the next step is 100/100/100. If there is not enough data for 100/100/100 splits, then the language will be skipped. However, this was not the process used to create splits for the 41 languages whose performance was examined in (Rahimi et al., 2019), though the authors provide information on those splits in the appendix.

As can be seen in Figure 2, stratified sampling was largely successful in balancing the mention types across the splits. However, a large amount of data is discarded by this process, and Hausa and Wolof are entirely thrown out because the entity type with the minimum count had too few mentions for this splitting method to be used. Furthermore, the distribution of entity types is artificial, and does not match up with what might be the natural distribution of entity types. For comparison see the distribution of entity types from the MasakhaNER data (Adelani et al., 2021), a human annotated NER dataset for 10 African languages, as shown in Figure 3.

3.3 Is WikiAnn useful for languages with human annotation?

For languages where there is no annotated NER data, WikiAnn is likely better than nothing. However, when annotated data is available for evaluation, can WikiAnn still be a useful resource? We conducted experiments by comparing models fine-tuned only on the MasakhaNER training data to those fine-tuned on the concatenation of WikiAnn with the MasakhaNER training data to see whether adding WikiAnn could improve performance by providing additional in-language training data.[6] For comparison, we also experimented with fine-tuning using the concatenation of the training data across all languages in MasakhaNER.

Qualitatively, the WikiAnn data differs greatly from any typical NER dataset annotated on news, such as MasakhaNER. As can be seen in Table 1, the WikiAnn data is generally very dense in mentions, contains many "sentences" not ending in periods (which are likely not actually sentences at all), has a high number of "sentences" that consist of only a single mention.

For all experiments, training was done for 50 epochs. Training was done using the `train_ner.py` script from the MasakhaNER

[6]The MasakhaNER dataset makes use of the DATE tag in addition to LOC, ORG, and PER which appear in both datasets. For our experiments all of the DATE mentions were removed and annotated as O.

Lang.	Sentences	Sentences ending in a period	Sentences consisting of a single mention	Mentions	Tokens	Tokens inside mentions (%)	Average tokens per mention
amh	1,032	101	381	1,189	6,477	38.74	2.11
hau	489	176	223	517	3,650	16.58	1.17
ibo	937	284	568	968	6,387	27.59	1.82
kin	1,517	176	1,163	1,680	6,496	42.90	1.66
swa	7,589	2,353	3,113	9,315	43,085	57.36	2.65
wol	1,196	337	624	1,370	10,800	17.05	1.34
yor	3,438	396	2,285	3,716	18,319	62.44	3.08

Table 1: Token, sentence, and mention statistics for all data in seven African languages contained in WikiAnn.

Lang.	MasakhaNER	+WikiAnn	Δ	*p*-value
amh	**70.63** ±1.16	69.02 ±1.72	1.61	**0.0191**
hau	**90.54** ±0.56	90.03 ±0.58	0.51	0.1041
ibo	**86.37** ±0.83	85.44 ±0.58	0.93	**0.0211**
kin	**73.89** ±2.12	72.14 ±1.79	1.75	0.0821
swa	87.90 ±0.64	**88.16** ±0.85	0.26	0.4963
wol	68.12 ±1.38	**68.18** ±1.22	0.06	0.8798
yor	78.23 ±0.99	**79.25** ±0.98	1.02	0.0539

Table 2: Comparison of F1 scores between XLM-R fine-tuned using only MasakhaNER data and fine-tuned using MasakhaNER data and all available WikiAnn data in each language.

Lang.	Single lang.	All langs.	Δ	*p*-value
amh	71.19 ±1.20	**71.70** ±1.01	0.51	0.3847
hau	89.78 ±0.41	**90.85** ±0.48	1.07	**0.0005**
ibo	84.18 ±0.94	**85.72** ±0.60	1.54	**0.0024**
kin	73.29 ±1.40	**74.67** ±0.79	1.38	**0.0283**
lug	80.02 ±0.91	**80.88** ±0.73	0.86	**0.0413**
luo	74.43 ±1.60	**77.19** ±1.17	2.76	**0.0015**
pcm	87.89 ±0.72	**89.14** ±0.49	1.25	**0.0002**
swa	**87.43** ±0.56	87.19 ±0.42	0.24	0.1988
wol	64.74 ±1.82	**65.33** ±1.42	0.59	0.4963
yor	77.63 ±1.17	**80.75** ±0.52	3.12	**0.0002**

Table 3: Comparison of F1 scores between XLM-R fine-tuned on MasakhaNER data and fine-tuned on the concatenation of the MasakhaNER train splits for all languages.

GitHub.[7] Unless otherwise specified, each experiment was run with 10 different seeds.[8]

For evaluation, the SeqScore[9] toolkit (Palen-Michel et al., 2021) was used, with the `conlleval` method of repairing invalid label sequences unless otherwise specified. Reported scores are the average of all training runs.[10]

3.4 Fine-tuning with WikiAnn and MasakhaNER

We experimented with fine-tuning XLM-R on the concatenation of the MasakhaNER training data with all available WikiAnn data in the corresponding language. The results are shown in Table 2. On average, F1 decreased by 0.49 when adding WikiAnn to the training data. Three languages (Swahili, Wolof, and Yoruba) show increases in performance, and while none are statistically significant, the improvement in Yoruba is marginally significant.

3.5 Fine-tuning with all MasakhaNER languages

The inclusion of WikiAnn data in the training data offers mixed results at best. Another option is to pool the training data across languages, creating a multilingual NER model trained on just the MasakhaNER data that is evaluated on each language's test set individually.[11] Adelani et al. (2021) previously performed this experiment, but we replicate it here using our methodology, which includes statistical significance testing, a larger number of random seeds, and includes Amharic in this experiment even though it uses a different script than the other languages.

The results, shown in Table 3, are similar to those reported by (Adelani et al., 2021), with all languages seeing improved performance except for Swahili. Many of the improvements are statistically significant, showing that simply using

[7] https://github.com/masakhane-io/masakhane-ner

[8] The learning rate was 5*e*-5. The optimizer used was AdamW, with an epsilon value of 1*e*-8, and the scheduler was the script's default scheduler, called `get_linear_schedule_with_warmup`. The maximum sequence length was set to 164, and the training batch size was set to 32. Prediction was done with a maximum sequence length of 512, because smaller values led to a handful of tokens not receiving any predicted labels.

[9] https://github.com/bltlab/seqscore

[10] Throughout this paper, the Wilcoxon rank-sum test is used to evaluate statistical significance.

[11] For this experiment, the DATE tag was left in the data, as consistency with WikiAnn data was not necessary.

more higher quality in-domain human-annotated data improves performance, while WikiAnn does not appear to help.

3.6 Summary

Many languages have no annotated data of the type provided by WikiAnn, and for those languages WikiAnn may prove useful as long as users are aware of its shortcomings. Being an automatically created dataset, it contains noisy data, and it was constructed without input from speakers of almost all the languages contained in it. Given that the core types (LOC, ORG, and PER) are intended to have the same definition across both WikiAnn and the MasakhaNER datasets, we predicted that augmenting MasakhaNER data with WikiAnn would improve performance.

But even when all available WikiAnn data is used, it does not improve the performance of models for the MasakhaNER data, and the simpler approach of simply pooling the MasakhaNER data across all languages produces better results. This suggests that the noise level of the WikiAnn "silver" standard is very high, raising into doubt the validity of benchmarks which treat it as gold standard data.

4 Human annotation: Still essential

We have spent much of our paper describing many of the complexities of working with large-scale, Wiki-derived datasets, demonstrating that while they have some utility, they must be used with caution. This caution stems from the fact that though the data contained in them originated from human contributions, in its final form in a resource it has been removed from its original quality checks. For example, in Wikidata, names may be copied from one language to another *en masse*, and in WikiAnn, the NER "annotation" is system output trained on relatively distant human supervision.

We have spent so much of our paper describing these shortcomings to demonstrate that there is no "free lunch" when it comes to avoiding human annotation or quality checking of datasets. We believe that human annotation processes that are ultimately participatory—involving speakers of the languages as stakeholders and collaborators, not mere annotators for hire—like that of the MasakhaNER project which we have featured through the previous section and related projects (Nekoto et al., 2020; Orife et al., 2020) are the most important direction for developing language technology for lower-resourced languages.

A discussion of efforts to annotate lower-resourced languages would not be complete without a discussion of the resources developed as part of the DARPA[12] LORELEI (Low Resource Languages for Emergent Incidents) program. The LORELEI program began in fall 2015, and a major thrust of the program was producing annotation for many lower-resourced languages (Strassel and Tracey, 2016). The Linguistic Data Consortium (LDC) is in the process of releasing the 31 language packs developed as a part of the program, which have been available to the primarily U.S.-based groups funded by the program for years. Tracey and Strassel (2020) stated that they planned to release 1-2 packs per month in 2020. However, as of November 2021, even though the research efforts of the program have largely concluded, only 7 language packs have been released to the general public: Akan, Amharic, Oromo, Somali, Tigrinya, Ukranian, and Vietnamese. Each of these language packs is available for $200 USD to non-members of the LDC. While that is less expensive than typical LDC datasets, that cost can be prohibitively expensive for people who are speakers of the languages in the packs, who may live in countries with substantially lower wages and may not have the backing of a well-funded research lab.

While it is unfortunate that so little data has been released to date and that the data is not freely available, the main contrast we would like to draw between LORELEI and efforts such as MasakhaNER is the involvement of the speakers of lower-resourced languages. Speakers of the languages included in the LORELEI datasets did not have any significant involvement in the construction of the datasets beyond their role as annotators. This is in no way unique to the LORELEI program; it is the status quo for annotation projects.

Returning to the issues we raise in our introduction, we want to highlight that a confluence of factors come together to make a language lower-resourced, among them often a marginalization and/or minoritization of its speakers. We should consider whether it is ethical to have a paradigm in which the marginalized have no say in research that involves them, and in this case may not even

[12] Defense Advanced Research Projects Agency, a research agency that is part of the United States Department of Defense and funds a large proportion of US-based computer science research.

be able to access the result of their work years after it is performed. In the United States, institutional review boards which approve research give extra consideration to the harms that can be caused by the exploitation of marginalized populations. Due to this marginalization, developing resources for languages that have had fewer resources created for them to date poses a unique set of ethical challenges that differs from higher-resourced language work, and engaging language speaker in a participatory fashion can help mitigate the risk of harm.

5 Challenges

Before we conclude, we wish to highlight challenges that we believe should be addressed as a part of continuing to develop resources and models for lower-resourced languages.

Reducing reliance on religious text. Due to the large amount of translation of religious texts into lower-resourced languages by religious organizations in attempts to spread their message, religious materials are a common place to look for parallel or monolingual data.[13] While religious texts can be a convenient source of data due to their broad coverage of languages, it is important to be aware of potential biases, especially when the religion of the text is not the predominant religion of the speakers of the language—and thus may not match their norms—or when the target task could be affected by bias from the religious data.

JW300 (Agić and Vulić, 2019) is a source of parallel data for over 300 languages with roughly 100,000 parallel sentences per language pair on average. The data was scraped from jw.org, which is the website of the Jehovah's Witnesses. Despite being sourced from a religious organization, it contains articles on a variety of topics translated into many languages.[14] Inclusion of articles on a variety of topics does not fully prevent the potential for religious bias. As Azunre et al. (2021) demonstrate with a few masked sentence completion examples, a model trained on JW300 frequently produces completions with biblical names. Although these types of completions are not grammatically incorrect, they are suggestive of a low level of generalization beyond religious data.

Quality control for text resources. A popular way to gather multilingual data is through web crawling. These datasets include CCAligned, Multilingual C4 (mC4), OSCAR, ParaCrawl, WikiMatrix, and the aforementioned JW300. However, as detailed in (Caswell et al., 2021), there can be fairly serious quality issues when web-crawled data collection is not done carefully. Currently, hen working with lower-resourced languages in large multilingual datasets, it is not a certainty (and sometimes, not even likely) that data is actually in the language that it claims to be.

6 Conclusion

In closing, we want to refer to the situation highlighted by Bird (2020), where it can be the case that researchers are preoccupied with a data-centric view to the point of completely removing the need to involve speakers of the language in any part of the process. Through our discussion of the shortcomings of Wikidata and WikiAnn for the specific purposes that we have evaluated them, we demonstrated the gaps that are created when the dataset creation process is divorced from the speakers of the language. Our perspective on the process of dataset and model creation can be summarized through these guidelines we propose for future work on lower-resourced languages:

1. Maximize interaction with and listening to the native speakers of languages included in resources you are developing.
2. When feasible, engage with speakers of included languages for quality control.
3. Consider the potential negative consequences of releasing datasets known to be of low-quality, as regardless of how you intend the resources to be used, they will likely be used for evaluation purposes.
4. Prefer human annotation by speakers of the language to automatic processes, and release all human annotator decisions as discussed in (Davani et al., 2021; Prabhakaran et al., 2021).

[13] From the perspective of decolonizing language technology (Bird, 2020), these sources may be especially problematic as many of them were used as tools of colonization.

[14] At the time of submission, this resource is not available. The site distributing the dataset claimed that the dataset was freely available for non-commercial use, referred readers to jw.org's copyright at https://www.jw.org/en/terms-of-use/, and stated that "for all practical purposes their custom terms of use are very closely aligned with the more well-known CC-BY-NC-SA license." However, recently the dataset has been taken down due to a copyright complaint until formal permission can be obtained.

References

David Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Azime, Shamsuddeen Muhammad, Chris Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Yimam, Tajuddeen Gwadabe, Ignatius Ezeani, 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 Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane Mboup, Dibora Haile, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure Dossou, Kelechi Ogueji, Thierno 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(0):1116–1131.

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.

Paul Azunre, Salomey Osei, Salomey Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac Kojo Essel Ampomah, Joseph Otoo, Reindorf Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James Ben Hayfron-Acquah. 2021. Contextual text embeddings for Twi. *arXiv preprint arXiv:2103.15963*.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Fernando Benites, Gilbert François Duivesteijn, Pius von Däniken, and Mark Cieliebak. 2020. TRANSLIT: A large-scale name transliteration resource. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3265–3271, Marseille, France. European Language Resources Association.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Auguste Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara E. Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Fred nm Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan 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 Ball, 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. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*. In press.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2021. Dealing with disagreements: Looking beyond the majority vote in subjective annotations.

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.

Ann Irvine, Chris Callison-Burch, and Alexandre Klementiev. 2010. Transliterating from all languages. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 100–110.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France. European Language Resources Association.

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.

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*, pages 2144–2160, Online. Association for Computational Linguistics.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Z. Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan Van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. 2020. Masakhane - machine translation for africa. *arXiv preprint arXiv:2003.11529*.

Chester Palen-Michel, Nolan Holley, and Constantine Lignos. 2021. SeqScore: Addressing barriers to reproducible named entity recognition evaluation. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 40–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

Jennifer Tracey and Stephanie Strassel. 2020. Basic language resources for 31 languages (plus English): The LORELEI representative and incident language packs. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 277–284, Marseille, France. European Language Resources association.