# Typologically Diverse QA: How many training examples do you need for a new language anyway?

Stanford CS224N Custom Project

**Caterina Wu**
cwu97@stanford.edu

**Tassica Lim**
tlim98@stanford.edu

**Tatiana Wu**
twu99@stanford.edu

## Abstract

Most current question-answering (QA) systems are catered toward English speakers, despite the fact that there are many non-English-speaking people in the world. There are thousands of different languages out there, and many of these languages use very different approaches to construct meaning. Our project examines just how much training data is needed for a new language in order for a multilingual QA model to perform reasonably well on the new language. Using the TyDi QA dataset provided by Clark et al. [1], we separately held out training examples from Arabic, Russian, and Korean and evaluated model performance. We found for each of the hold-out languages that with only 20% of the training examples, we were able to obtain performance comparable to that when training on all of the data.

## 1 Key Information to include

- Mentor: Matthew Lamm
- External Collaborators: N/A
- Sharing project: N/A

## 2 Introduction

Question-answering (QA) technologies help people in their daily lives — in today's world, users can perform a Google search or ask a voice assistant a question, and expect to receive an answer, when they want to know something. There are English-language QA datasets in the NLP world that reflect the needs of real users, but there are thousands of different languages out there, and many of these languages use very different approaches to construct meaning. Specifically, there are significant typological distinctions between different languages (e.g., word order, plurality systems, spacing convention), underscoring the importance of building models that can accurately represent many languages. By utilizing typologically diverse data, researchers can draw more reliable conclusions about how well their models will generalize across all languages of the world. The goal of our project is to investigate how much training data of a new, previously unseen language is needed, in combination with training data from other languages, for a multilingual QA model to perform reasonably well on the new language.

Essentially, we are exploring how little data is needed to bootstrap a new language. This is important because parallel translation data can be difficult to find or very expensive to acquire, so being able to obtain reasonable performance with smaller amounts of data would increase the chances of being able to generalize to new languages despite the small amounts of available data, especially with lesser known languages.

## 3 Related Work

Previous papers have tackled the problem of multilingual QA. For example, Ture and Boschee [2] propose a translation approach, and Loginova et al. [3] take a deep learning approach to the problem and briefly touch upon the process of collecting parallel data. However, Clark et al. [1] argue for the use of multilingual QA data due to the differences between translated text and purely native text, as well as the limited availability of parallel translation data. In their paper, Clark et al. present TyDi QA, the first public large-scale multilingual corpus of information-seeking question-answer pairs, covering 11 typologically diverse languages, and tackle two tasks:

1. Passage Selection Task: Given a list of the passages in the article, return the index of the passage that answers the question.
2. Minimal Answer Span Task: Given the full text of an article, return (a) the start and end byte indices of the minimal span that completely answers the question; or (b) YES or NO if the question requires a yes/no answer and we can draw a conclusion from the passage.

In addition, they present a (secondary) simpler task called the Gold Passage Task, on which our project will focus. This task is more similar to existing QA tasks, enabling researchers to compare results with prior work and providing compatibility with code for SQuAD [4], XQuAD [5], and MLQA [6]. The main differences between the Gold Passage task and the primary tasks described above are that it only provides the gold answer passage, all questions have an answer, Thai and Japanese are removed, and the model is evaluated with SQuAD 1.1 metrics.

Although Clark et al. present some of their performance results, the most significant contribution is the TyDi QA dataset for future research purposes. Clark et al. offer several directions for future research, including studying the interaction between morphology and question-answer matching; evaluating the effectiveness of transfer learning for languages where parallel data is and is not available; exploring the usefulness of machine translation in QA for data augmentation and as a runtime component, given varying data scenarios and linguistic challenges; and studying zero-shot QA by explicitly not training on a subset of the provided languages. We seek to investigate a version of this last area in our project.

## 4 Approach

Our project investigates just how much training data is needed for a multilingual QA model to perform well on a new language. Rather than acquiring more training data, we take an existing multilingual dataset and instead remove training examples. We choose a "hold-out language" and train our model on a fraction of this hold-out language, as well as the rest of the training data. We incrementally remove training examples of the hold-out language and evaluate model performance on all of the languages separately. Our goal is to use performance on this hold-out language to simulate how a multilingual model might perform on small amounts of a new language, since we don't have easy access to new training data.

### 4.1 Model

For our project, we use the multilingual BERT (mBERT) model[1] used in Clark et al. We modified the code[2] to do the following:

1. add a flag to train the model on a small subset of the training data for debugging purposes;
2. include flags to specify a hold-out language and what fraction of the hold-out language to train on; and
3. save different models for each combination of hold-out language and percent of training examples.

[1] `https://github.com/google-research/bert`

[2] `https://github.com/google-research-datasets/tydiqa/tree/master/gold_passage_baseline`

### 4.2 Baseline

Our baseline is the mBERT model trained on all of the data. Contrary to the traditional baseline model, this baseline serves as our expected upper bound on model performance; as we remove training examples of the hold-out language, we expect model performance to decrease.

Due to memory limitations, we decreased the batch size from 12 to 4, but otherwise kept all hyperparameters the same as in the original model. Nevertheless, we were able to replicate the results from Clark et al., which are summarized in Table 1.

| Language | TyDiQA-GoldP Baseline | Baseline, 2 epochs | Baseline, 3 epochs |
|---|---|---|---|
| English | 67.9 | 67.0 | 66.2 |
| Arabic | 77.2 | 78.2 | 78.3 |
| Bengali | 69.3 | 74.1 | 71.8 |
| Finnish | 71.0 | 71.6 | 71.8 |
| Indonesian | 76.1 | 76.2 | 76.7 |
| Korean | 61.7 | 61.1 | 61.1 |
| Russian | 70.2 | 71.1 | 70.7 |
| Swahili | 79.3 | 78.6 | 78.6 |
| Telugu | 82.4 | 82.1 | 82.9 |
| **OVERALL** | **73.4** | **74.1** | **73.9** |

Table 1: F1 scores for our baselines compared to the GoldP baseline results from Clark et al. Note that the baseline we use for our project is trained on 3 epochs to remain consistent with the rest of our models. For specifics on the hyperparameters, see the Experiments section.

## 5 Experiments

### 5.1 Data

The dataset we are using is the TyDi QA dataset[3] provided by Clark et al., specifically for the Gold Passage (GoldP) Task. TyDiQA-GoldP includes nine languages — English, Arabic, Bengali, Finnish, Indonesian, Korean, Russian, Swahili, and Telugu — and consists of a total of 49,370 usable training examples and 6,174 dev examples. The breakdowns of training and dev examples are provided in Tables 5 and 6 in the Appendix.

We did not preprocess the data and used the same inputs and outputs as in Clark et al.: The input is a question along with the gold answer passage, and the output is the answer to the question.

### 5.2 Evaluation method

In maintaining consistency with the results from Clark et al., as well as other QA models, we evaluate the performance of our models using F1 score. We calculate the F1 score over the dev examples for each of the nine languages separately and compare them to our baseline results. In addition, the overall F1 is calculated by averaging over all the languages (except English).

### 5.3 Experimental details

We trained all of our mBERT models using the same configurations, modifying only the amount of training data. We increased training time to 3 epochs but kept all other hyperparameters the same as our baseline: Specifically, we used a batch size of 4, a learning rate of $3e{-5}$, a maximum sequence length of 384, and a document stride of 128. Our baseline results that we use to compare model performance will therefore be those from training on three epochs rather than two.

Notably, TyDi QA already provides typologically diverse languages, so no two languages in the dataset are all that similar. We chose to separately hold out three different languages: Arabic, Russian,

[3] `https://github.com/google-research-datasets/tydiqa`

| **Language** | **100%** | **80%** | **60%** | **40%** | **20%** | **10%** | **5%** | **1%** |
|---|---|---|---|---|---|---|---|---|
| English | 66.2 | 66.2 | 67.0 | 68.6 | 66.7 | 67.9 | 67.5 | 66.5 |
| Arabic | 78.3 | 77.2 | 76.3 | 76.8 | 76.2 | 74.9 | 74.8 | 73.0 |
| Bengali | 71.8 | 67.4 | 71.1 | 74.7 | 70.8 | 68.9 | 70.1 | 70.5 |
| Finnish | 71.8 | 70.9 | 71.7 | 70.9 | 70.4 | 70.8 | 70.4 | 71.6 |
| Indonesian | 76.7 | 75.7 | 77.1 | 76.0 | 77.3 | 75.4 | 75.0 | 74.1 |
| Korean | 61.1 | 62.5 | 62.1 | 63.9 | 62.1 | 61.3 | 62.3 | 61.7 |
| Russian | 70.7 | 70.4 | 71.9 | 70.5 | 70.4 | 69.5 | 70.1 | 70.5 |
| Swahili | 78.6 | 80.8 | 80.3 | 80.0 | 79.5 | 80.2 | 80.3 | 80.9 |
| Telugu | 82.7 | 83.0 | 83.3 | 82.4 | 81.8 | 82.7 | 82.7 | 82.7 |
| **OVERALL** | **73.9** | **73.5** | **74.2** | **74.4** | **73.6** | **73.0** | **73.2** | **73.1** |

Table 2: F1 for each language as a result of holding out Arabic training examples incrementally.

| **Language** | **100%** | **80%** | **60%** | **40%** | **20%** | **10%** | **5%** | **1%** |
|---|---|---|---|---|---|---|---|---|
| English | 66.2 | 67.9 | 68.2 | 67.5 | 68.4 | 66.7 | 67.8 | 67.9 |
| Arabic | 78.3 | 76.6 | 77.9 | 77.7 | 77.3 | 77.2 | 76.9 | 76.6 |
| Bengali | 71.8 | 73.8 | 71.3 | 74.3 | 71.9 | 73.1 | 71.2 | 72.6 |
| Finnish | 71.8 | 71.4 | 71.5 | 70.2 | 70.9 | 71.1 | 70.2 | 70.8 |
| Indonesian | 76.7 | 77.0 | 76.4 | 75.7 | 76.1 | 76.8 | 74.4 | 75.5 |
| Korean | 61.1 | 62.2 | 63.1 | 60.6 | 62.4 | 61.8 | 60.7 | 59.0 |
| Russian | 70.7 | 70.6 | 70.6 | 69.5 | 70.2 | 70.0 | 70.4 | 70.5 |
| Swahili | 78.6 | 81.8 | 81.8 | 81.2 | 83.0 | 80.9 | 82.3 | 81.3 |
| Telugu | 82.7 | 82.7 | 83.1 | 83.3 | 82.2 | 82.7 | 82.5 | 82.3 |
| **OVERALL** | **73.9** | **74.5** | **74.5** | **74.1** | **74.2** | **74.2** | **73.6** | **73.6** |

Table 3: F1 for each language as a result of holding out Korean training examples incrementally.

| **Language** | **100%** | **80%** | **60%** | **40%** | **20%** | **10%** | **5%** | **1%** |
|---|---|---|---|---|---|---|---|---|
| English | 66.2 | 67.9 | 67.5 | 67.6 | 68.1 | 67.4 | 66.7 | 66.7 |
| Arabic | 78.3 | 77.3 | 76.9 | 77.9 | 77.5 | 76.8 | 76.9 | 76.7 |
| Bengali | 71.8 | 71.4 | 71.2 | 69.9 | 70.0 | 72.4 | 70.7 | 71.7 |
| Finnish | 71.8 | 70.5 | 71.0 | 71.4 | 72.3 | 71.1 | 71.3 | 69.2 |
| Indonesian | 76.7 | 76.8 | 75.9 | 75.9 | 76.5 | 75.2 | 75.6 | 75.4 |
| Korean | 61.1 | 62.9 | 60.0 | 61.7 | 64.1 | 63.3 | 61.9 | 61.9 |
| Russian | 70.7 | 70.6 | 69.3 | 69.5 | 68.8 | 67.2 | 67.0 | 66.5 |
| Swahili | 78.6 | 80.4 | 80.8 | 82.4 | 79.3 | 81.1 | 78.9 | 80.3 |
| Telugu | 82.7 | 82.6 | 83.4 | 82.7 | 83.4 | 82.4 | 82.3 | 82.4 |
| **OVERALL** | **73.9** | **74.1** | **73.6** | **73.9** | **74.0** | **73.7** | **73.1** | **73.0** |

Table 4: F1 for each language as a result of holding out Russian training examples incrementally.

and Korean. Of the languages in the dataset, we found Korean to be the most typologically distinct since it is classified as a language isolate. We also chose to hold out Arabic and Russian because they have the largest numbers of training examples.

For each hold-out language, we incrementally removed 20% of the training examples, trained the model for three epochs, and evaluated the model performance on the entire dev set. Because we saw very small changes in F1 scores between 100% and 20% of the training examples, we also trained on 10%, 5%, and 1% of the training examples.

### 5.4 Results

Our results from holding out Arabic, Korean, and Russian are summarized in Tables 2, 3, and 4, and visualized in Figures 1, 2, and 3, respectively. In general, we found that iteratively removing more Arabic, Korean, and Russian training examples produced a mostly decreasing trend in the F1 score for the respective hold-out language while other languages remained approximately constant (although there are some fluctuations in F1 scores that are likely due to noise in the models).

In general, we expected to see a decreasing trend in model performance as we removed training data. While there is a small decline in F1 scores, we expected a greater drop in performance after removing more than half of the training examples. It appears that there is not a significant drop in performance between having all of the training examples and only having 20% of the training examples, as the F1 score does not drop by more than 2.5 for any of the hold-out languages.

## 6 Analysis

With respect to Arabic and Russian, there is a roughly linear decrease in the F1 score as more of the training examples are held out. This linear trend could be due to the fact that both Arabic and Russian have a large number of training examples (relative to other languages in the training set). In contrast, holding out Korean results in F1 scores that vary somewhat inconsistently as more training examples are held out. This fluctuation could be a result of the unique linguistic structures found in Korean that the multilingual model still has not learned accurately or could be due to the small number of training examples available since Korean has the fewest training examples in the dataset.

In addition, for all of the hold-out languages, there appears to be a noticeable drop-off between having 20% and 10% of the training examples and between having 5% and 1% of the training examples. This could mean that, specific to the TyDi QA dataset, at least 5% of the training examples are needed, and if better performance is necessary, then 20% of the training examples are needed, to show a significant increase in the F1 scores. Since we are using the hold-out language to simulate how our

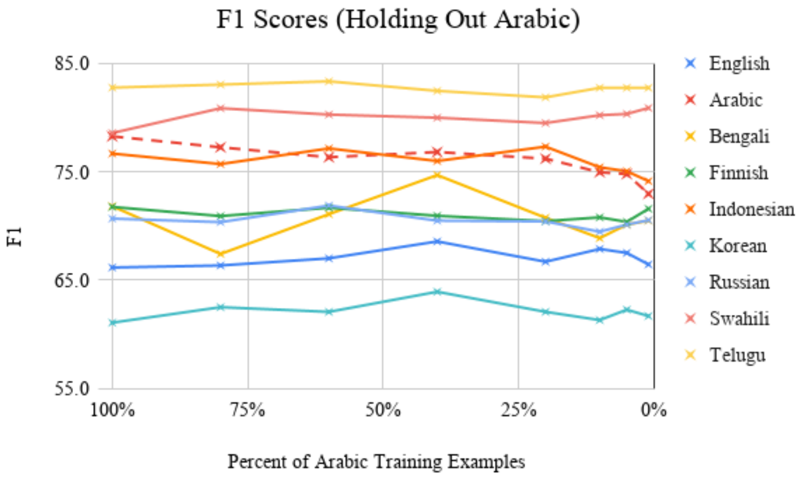

Figure 1: Plot of F1 from holding out Arabic training examples incrementally.

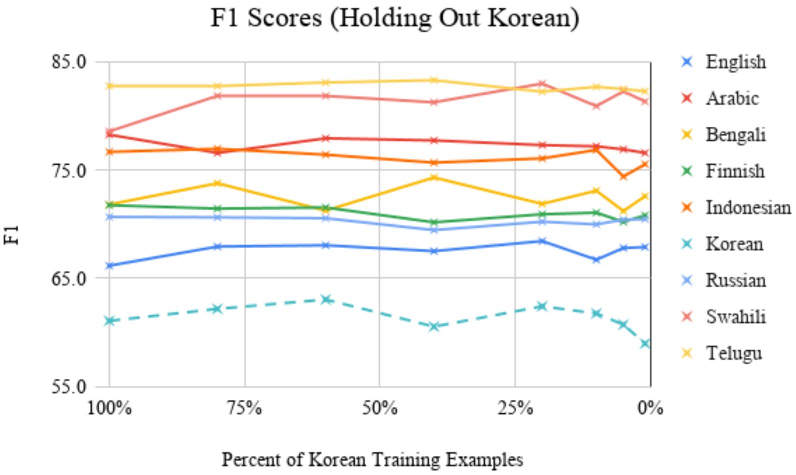

Figure 2: Plot of F1 from holding out Korean training examples incrementally.

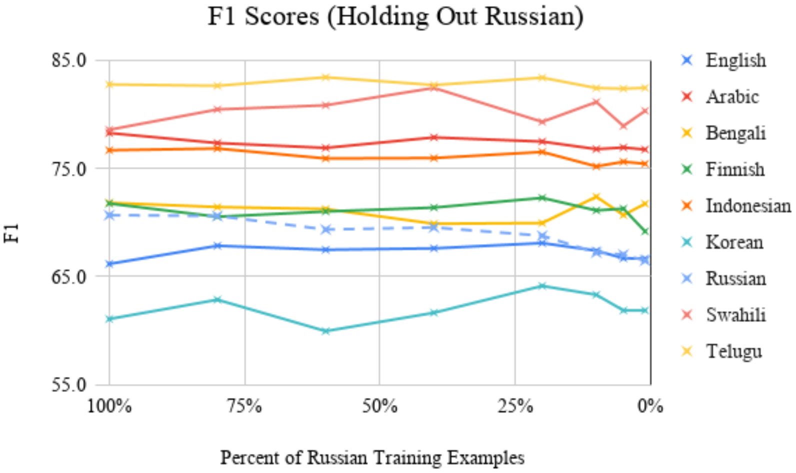

Figure 3: Plot of F1 from holding out Russian training examples incrementally.

model might perform on a new language, our results indicate that even a few hundred QA pairs of a new language may be enough for a multilingual model to generalize to the language and achieve F1 scores in the 60-80 range.

Our models consistently perform better with Arabic examples than Russian, and with Russian than Korean, which may be partly due to the imbalances in the training dataset; there are twice as many Arabic examples as there are Russian, and Korean has the fewest examples in the dataset. However, model performance on only 1% of Arabic examples still far exceeds performance on all of the Korean examples. This indicates that the mBERT model may not generalize very well to certain languages. It is unlikely that the model overfits on the most prevalent language in the dataset, but it may overfit on certain linguistic structures or features. As we noted earlier, Korean is a language isolate, so multilingual models may need more training examples of a language when there are no other typologically similar languages in the dataset.

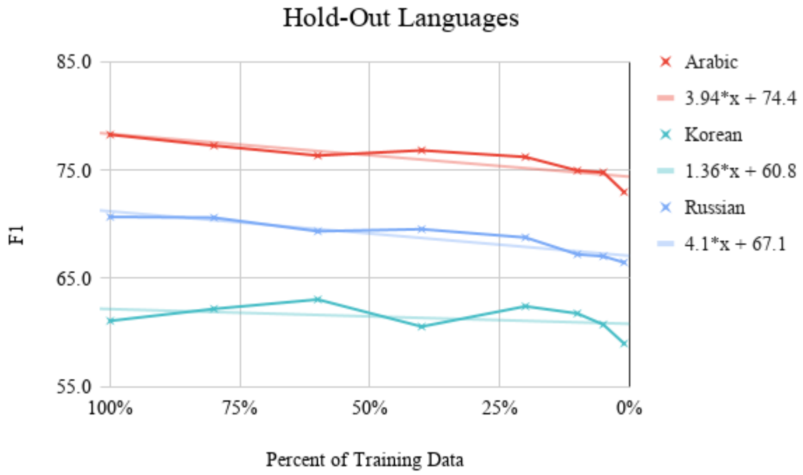

Figure 4: Plot of F1 for each of the hold-out languages (Arabic, Korean, and Russian) as training examples were held out incrementally.

## 7 Conclusion

Overall, we found that models trained on small amounts of training data were still able to obtain comparable performance to our baseline. This is an important result because obtaining training data, especially for more languages, is both difficult and expensive. With the ability to obtain reasonable performance with small amounts of training data for a language, multilingual models such as mBERT hold promise to generalize to new languages even with only a few hundred training examples. In addition, future research in multilingual QA could potentially save significant time and computational resources by training and fine-tuning on fewer examples per language.

### 7.1 Limitations

The main limitation of our work is that we held out training examples on an imbalanced dataset. As we noted previously, it is difficult to compare model performance *between languages* due to these imbalances. It is also difficult to pinpoint exactly how little training data is needed for a reasonable baseline model because there is a lot of noise in our data and due to the subjective definition of a “reasonable baseline.” Our project approached this task using a top-down approach by incrementally removing data, as opposed to a bottom-up approach by incrementally adding data. Rather than attempting to define a reasonable baseline for all future research, our work shows that further research using the TyDi QA dataset may benefit by saving resources from training on fewer examples.

### 7.2 Future Work

Future work expanding upon our results could follow several avenues, including holding out more languages or introducing a small number of training examples in a completely new language and evaluating model performance on the new language. Our main suggestion would be to balance the TyDi QA dataset so that all languages have the same number of training examples before performing hold-out experiments such as the ones in our paper. This would enable comparisons between languages that were not possible in our paper and could lead to insights about relationships between different languages that may allow multilingual models to generalize better to certain families of languages over others.

## References

[1] Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A benchmark for information-seeking question

answering in *Ty*pologically *Di*verse languages. *Transactions of the Association for Computational Linguistics*, 2020.

[2] Ferhan Ture and Elizabeth Boschee. Learning to translate for multilingual question answering. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 573–584, Austin, Texas, November 2016. Association for Computational Linguistics.

[3] Ekaterina Loginova, Stalin Varanasi, and Günter Neumann. Towards multilingual neural question answering. In András Benczúr, Bernhard Thalheim, Tomáš Horváth, Silvia Chiusano, Tania Cerquitelli, Csaba Sidló, and Peter Z. Revesz, editors, *New Trends in Databases and Information Systems*, pages 274–285, Cham, 2018. Springer International Publishing.

[4] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[5] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*, 2019.

[6] Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. MLQA: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*, 2019.

# Appendices

## A TyDiQA-GoldP Dataset Breakdowns by Language

| **Language** | **Number of Training Examples** |
|---|---|
| English | 3,695 |
| Arabic | 14,784 |
| Bengali | 2,390 |
| Finnish | 6,853 |
| Indonesian | 5,701 |
| Korean | 1,624 |
| Russian | 6,018 |
| Swahili | 2,753 |
| Telugu | 5,552 |
| **TOTAL** | **49,370** |

Table 5: Number of training examples for each language.

| **Language** | **Number of Dev Examples** |
|---|---|
| English | 634 |
| Arabic | 1,099 |
| Bengali | 130 |
| Finnish | 983 |
| Indonesian | 711 |
| Korean | 374 |
| Russian | 955 |
| Swahili | 588 |
| Telugu | 700 |
| **TOTAL** | **6,174** |

Table 6: Number of dev examples for each language.