*Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 4488–4496
Marseille, 20-25 June 2022

# Natural Questions in Icelandic

**Vésteinn Snæbjarnarson**[1,2], **Hafsteinn Einarsson**[2]
[1]Miðeind ehf, [2]University of Iceland
vesteinn@mideind.is, hafsteinne@hi.is

### Abstract

We present the first extractive question answering (QA) dataset for Icelandic, Natural Questions in Icelandic (NQiI). Developing such datasets is important for the development and evaluation of Icelandic QA systems. It also aids in the development of QA methods that need to work for a wide range of morphologically and grammatically different languages in a multilingual setting. The dataset was created by asking contributors to come up with questions they would like to know the answer to. Later, they were tasked with finding answers to each others questions following a previously published methodology. The questions are Natural in the sense that they are real questions posed out of interest in knowing the answer. The complete dataset contains 18 thousand labeled entries of which 5,568 are directly suitable for training an extractive QA system for Icelandic. The dataset is a valuable resource for Icelandic which we demonstrate by creating and evaluating a system capable of extractive QA in Icelandic.

**Keywords:** QA, question answering, Icelandic

## 1. Introduction

Most of us use QA systems daily through the use of search engines. Generally, the user asking a question wants to know, in their formulation, the answer to the question and its location within a source document. To build, or at the very least, validate such systems it is currently necessary to have a properly labeled dataset in the underlying language. Such datasets contain questions and underlying context, along with labelled answer spans within the context.

Several datasets exist for this task in English, such as WikiQA (Yang et al., 2015), and the Natural Questions (NQ) dataset (Kwiatkowski et al., 2019). The need to create similar datasets in other languages was recognized recently (Clark et al., 2020). While English is spoken by many, there exist thousands of languages and most of the world's population does not speak English. Furthermore, languages differ widely in typology such as grammar and morphology. All these differences might affect a QA system's performance. To study the effects of language differences on QA systems, such a QA dataset on 11 typologically diverse languages was released by Clark et al. (2020). Other such datasets have also been created for a variety of languages (Rogers et al., 2021) as they are a prerequisite for accurate evaluation of QA systems for a specific language.

In this paper, we describe and release Natural Questions in Icelandic (NQiI), the first dataset for boolean (questions with yes-no answers) and extractive QA in Icelandic that has not been scraped from the web. The dataset contains around 13.7k questions and 18k question-passage pairs with roughly 5k containing answers. We used the methodology presented by Clark et al. (2020) to create questions that reflect information-seeking behavior. We use the terminology *Natural Questions* following (Kwiatkowski et al., 2019) to emphasizes the fact that while the creation of the dataset was controlled, the questions are real questions posed by real users. We also train QA-models using the subset of the data with labelled answer spans. To train the models we use transfer learning from an Icelandic language model, IceBERT (Snæbjarnarson et al., 2022), and compare the performance to a model on a multilingual model XLM-Roberta (Conneau et al., 2020).

## 2. Related Work

QA systems can be broadly split into two categories: *extractive* and *abstractive* (or *generative*) QA systems (Fan et al., 2019). The abstractive systems may generate an answer which can not be directly found in any underlying text. The extractive systems, which are the focus of this work, locate and return a segment of text (within a given document or corpus) that hopefully contains an answer to the question posed. Note that abstractive systems can also be trained to search within a document library but their output is generated, i.e. not copied directly from an existing document. The methods used to build such systems have been referred to as *generative QA based on machine reading*. Such methods can be used to enrich answers with external knowledge that is not found within the document the answer is based on (Bi et al., 2019). Such an approach can, in theory, improve QA systems since background knowledge or commonsense might be necessary to derive an answer if the context from a document is insufficient. Furthermore, there also exist multi-hop systems (Ho et al., 2020) and datasets (Ponti et al., 2020) where multiple segments are used to infer an answer to a question.

QA systems are also categorized into *open domain* (sometimes referred to as simply *open* (Herzig et al., 2021) or open-book systems) and *closed* QA systems. The open systems target many documents at once, use large databases or apply neural networks with embedded information such as GPT (Radford et al., 2019). The *closed* methods target a single document at once and resemble traditional *reading comprehension* (Rajpurkar et al., 2016) tasks. For reading comprehension systems, some specific part of the text is targeted, e.g. a paragraph or article which is provided as input along with the question. The system is then tasked with finding the answer or possibly indicating that no such answer can be found within the context provided. The data presented here can be used to train both open and closed systems.

The type of QA system determines the structure of the training data that needs to be available to train the system. For generative systems that are not required to find answers in a corpus, question-answer pairs can suffice. For systems that search in a corpus, a question needs to be paired with a passage containing the answer, and possibly also with passages that do not contain the answer as negative training examples.

By now many QA datasets have been created that reflect the diversity of QA systems. Datasets such as SQuAD (Rajpurkar et al., 2016) and NewsQA (Trischler et al., 2017) are extractive datasets. Natural Questions (Kwiatkowski et al., 2019) and TyDi-QA (Clark et al., 2020) are also extractive but more suitable for open questions answering than closed domain systems since the questions are in some sense "natural", i.e., they better reflect information-seeking behaviour. Work on datasets like TyDi-QA has led to work on similar datasets in other languages such as Indian languages (Kakwani et al., 2020; Tahsin Mayeesha et al., 2021), Russian (Efimov et al., 2020; Korablinov and Braslavski, 2020), Portuguese (Paschoal et al., 2021), Hebrew (Keren and Levy, 2021), and Persian (Khashabi et al., 2021). Multilingual datasets (Longpre et al., 2021) have also been released that include the nordic languages Danish, Norwegian and Swedish except for Icelandic. We note though that Icelandic is a part of the MFAQ dataset[1] a multilingual dataset extracted from Common Crawl (De Bruyn et al., 2021). The vast amount of new QA datasets has even been documented in recent surveys, some which refer to a QA dataset explosion (Rogers et al., 2021; Cambazoglu et al., 2021; Chandra et al., 2021).

Work on multilingual QA datasets has been incorporated into general benchmarks for multilingual models (Hu et al., 2020; Ruder et al., 2021). Such work is highly valuable for lower resource languages when deciding on what language model to use where extensive experimentation can be time consuming and expensive. Fortunately, evaluation on diverse languages is by now encouraged to justify performance claims and a broad applicability of multilingual models (Artetxe et al., 2020). Evaluation on multilingual benchmark datasets have revealed a wide spread of results across languages. Furthermore, although human level performance is reached for English it is not yet achieved for other languages using cross-lingually transferred models (Hu et al., 2020). These observations emphasize the need for good benchmark datasets in many languages.

Some work has been done in QA for Icelandic that relies on a dataset restricted to question-answer pairs, i.e., without context. QA datasets without answer contexts are valuable resources, even if they can not be used for training in the same way as those with labeled answer spans, they can be used to evaluate performance of open domain QA systems. Two such resources exist for Icelandic. Geirsson (2013) uses a set of trivia style questions, the system they built leverages term frequencies and implements modules based on three question types: persons, locations and years. The dataset contains 4,569 questions with answers. Another such set, a collection of community collected trivia-style questions in Icelandic is available online[2], it contains 11,610 questions with answers.

## 3. Methods

The creation of the dataset is based on the methods presented in Clark et al. (2020) on typologically diverse languages that we recap here for the sake of completeness. We chose this method as it is a proven method that leads to questions that are information-seeking. In methods to build older datasets the annotators that created the questions knew the answer at the time the question was written. That approach can introduce a bias in the dataset where the questions do not reflect those posed in actual usage of information retrieval systems such as web search engines. We summarize the method below.

**Question elicitation:** Human annotators received the first 100 characters from an Icelandic Wikipedia article as a prompt. Based on the prompt, the annotator should write a question that they want to know the answer to and that the prompt does not answer. The prompt serves as an inspiration, and the questions do not need to have a strong connection to the prompt.

To create the prompts, we used a database dump of the Icelandic Wikipedia from the 20th of May 2020. We only selected articles with at least 250 characters, and we presented the prompts ordered by the length of the corresponding Wikipedia article in descending order. The choice of using Wikipedia follows the approach in (Clark et al., 2020). For the sake of efficiency, it is important that at least some fraction of the questions can be answered using text from a reference corpus like Wikipedia. For that reason, we use prompts from Wikipedia pages to improve the chances of answers being found. However, using prompts from other sources

[1] Despite the impressive size of this dataset, we note that it has not been thoroughly cleaned and contains multiple questions in English labelled as being in Icelandic as well as machine-translated text. Additionally, almost all of the questions (94%) concern hotel bookings on two popular booking platforms.

[2] The data is available at `https://github.com/sveinn-steinarsson/is-trivia-questions`.

and alternative reference corpora is certainly also feasible. Using the prompts, the annotators worked on creating questions in separate spreadsheet documents where the prompts were in one column and the questions in another column next to it. This turned out to be a managable workflow with five annotators but it might not scale well with a larger group.

**Article retrieval:** We programatically perform a Google search for each question and select the top-ranked Icelandic Wikipedia page as a candidate that could contain the answer. We refer to these articles as passages. Note that this process is not guaranteed to be successful. For 34% of our questions, the search did not return any Icelandic Wikipedia page. This approach is taken following the original TyDi methodology to encourage question writers to come up with questions they are genuinely interested in knowing the answer to. Simply using the same passages as were used as inspiration for the questions would not suit this purpose well. Consider, for instance, the Wikipedia page about the fruit apple; reading the section might have one think about the company Apple and then wonder when the company was founded.

**Answer labeling:** In a separate task, the annotators received question-passage pairs. The annotators could either label the passage as not containing an answer or select a paragraph containing the answer. If a paragraph was chosen, the annotator could decide if it was a yes/no question or provide a short minimal answer by selecting a span defined by start and end coordinates within the text.

**Software:** We received permission to use and modify the software interface used in (Asai et al., 2020) to collect the data and annotate the question-passage pairs.

## 4. The Dataset

Five undergraduate students, all native speakers of Icelandic, were employed to create the dataset and their contribution summed up to nine months of work. The students were from different departments within the university and of mixed genders but of similar age (20-25). While we do not investigate this further, it is not unlikely that there is some bias in the questions posed that links them to university students. The dataset contains 18k labeled question-answer pairs where the answers come from 1,400 unique Wikipedia articles. Summary statistics can be found in Table 1. In total, 13,740 questions were written. However, for 4,680 questions (34%), no article was found in the article retrieval step. A few examples of questions and answers are included in Table 6 in the Appendix.

In Table 2, we compare question types between NQiI and the English development sets of TyDi QA and SQuAD. The question types are somewhat more evenly distributed than in the TyDi QA dataset. For each question type, we list the Icelandic wh-word at the beginning of the sentence. The *Other* category is composed of questions that did not start with any of the words listed and account for 3% of the total. Generally, these start with a verb (a typological difference, compared to English), and most of them are yes/no questions.

| | |
|---|---|
| **No. of questions written:** | 13,740 |
| With an associated passage: | 9,060 |
| **No. of labeled pairs:** | 18,378 |
| With answer found: | 5,405 |
| With no answer found: | 12,973 |
| 1-way annotated: | 3,153 |
| 2-way annotated (54.1% agr.): | 2,721 |
| 3-way annotated (36.1% agr.): | 2,817 |
| 4-way annotated (37% agr.): | 333 |

Table 1: Summary statistics for the dataset. The agreement (agr.) numbers are measured over the cases where answers spans were labeled. *N-way* annotated refers to the questions having been seen by *N* annotators during the second phase of annotation.

| **Question words** | NQiI | TyDi QA | SQuAD |
|---|---|---|---|
| **What** | 27% | 30% | 51% |
| Hvað | 27% | | |
| **How** | 11% | 19% | 12% |
| Hvernig | 5% | | |
| Hversu | 6% | | |
| **When** | 10% | 14% | 8% |
| Hvenær | 10% | | |
| **Where** | 7% | 14% | 5% |
| Hvar | 3% | | |
| Hvert | 3% | | |
| Hvaðan | 1% | | |
| **(Yes/No)** | 7% | 10% | <1% |
| Er | 5% | | |
| Eru | 1% | | |
| Var | 1% | | |
| **Who** | 20% | 9% | 11% |
| Hver | 18% | | |
| Hverjir | 1% | | |
| Hverjar | 1% | | |
| **Which** – Hvaða | 12% | 3% | 5% |
| **Why** | 3% | 1% | 2% |
| Af hverju | 2% | | |
| Hvers vegna | 1% | | |
| **Other** | 3% | | |

Table 2: Distribution of question words compared to the English portion of TyDi QA and SQuAD.

During the project, we realized that the small size of the Icelandic Wikipedia has an effect on the number of answers that can be labeled. The archive we used from the 20th of May 2020 contained over 102k pages, but only 3,730 of them had more than 250 characters.

### 4.1. Comparison to TyDi

The original TyDi dataset that our work is based on contains data in 11 typologically diverse languages: English, Arabic, Bengali, Finnish, Indonesian, Japanese, Kiswahili, Korean, Russian, Telugu, and Thai. These 11 languages were chosen as they represent a wide range of the 192 typological features that have been categorized in over 2,600 languages (Haspelmath et al., 2005; Dryer and Haspelmath, 2013).

Icelandic is a west Scandinavian language in the Germanic branch of the Indo-European language family. The language is most closely related to Faroese, extinct Norn and western Norwegian dialects. Icelandic has a rich morphology, where nouns, adjectives and verbs are inflected. An Icelandic noun typically has 16 forms and verbs are conjugated for tense, mood, person, number, and voice with three voices and up to ten tenses. The basic word order is subject-verb-object but every combination is allowed (Árnason et al., 2005), which means that questions can, for example, start with a verb. Icelandic has changed relatively little throughout its history, to some extent due to purist language attitudes, when compared to other Germanic languages which have greatly reduced levels of inflection (Hilmarsson-Dunn and Kristinsson, 2010).

For the sake of comparison, we list the features below that were used to categorize the languages in the TyDi dataset. The (+) denotes the presence of a feature and the (-) its absence. Many + symbols denote the degree of that feature.

- **Latin script** (+). Icelandic is written in latin script with a few characters not used in English (á, é, í, ó, ú, ý, þ, æ, ö). Note that Icelandic does not use c and z. Letters with diacrits are treated as separate letters.
- **White space tokens** (+). Tokens are separated by a white space, similar to English.
- **Sentence boundaries** (+). Sentences are separated by a period, similar to English.
- **Word formation** (+++). Icelandic is highly inflected and compounding is used actively for constructing new words.
- **Gender** (+). Icelandic has three grammatical genders, masculine, feminine and neuter.
- **Prodrop** (-). Icelandic does not have prodrop. However, null subjects were historically a part of Icelandic until the 20th century (Kinn et al., 2016).

Based on these features, Icelandic is not the same as any of the languages in the TyDi dataset but has most of the listed features in common with Finnish. If we consider old Icelandic, as can be studied in the Icelandic Parsed Historical Corpus (Wallenberg et al., 2011), then it has all the features listed above, that is, it includes prodrop.

### 4.2. Adapting the dataset for use in training

While the dataset contains yes-no answers and questions with no associated paragraphs, a portion of the data is compatible with the commonly used QA-dataset SQuAD (Rajpurkar et al., 2016) which is suitable for training of extractive QA models. We release a split of this data, with 80% taken aside for training, 10% for validation at training time and 10% for testing as shown in Table 3. The dataset has been made available on Huggingface[3]

| **Subset** | **Questions** | **With answer** |
|---|---:|---:|
| Train | 4,552 | 2,234 |
| Development | 513 | 259 |
| Test | 503 | 244 |
| Total | 5,568 | 2,737 |

Table 3: NQiI split for training

## 5. Results

We have adapted a Transformer model for Icelandic, IceBERT[4] (Snæbjarnarson et al., 2022), on the NQiI. The model is adapted for transfer learning by adding a classification head on top of the model that outputs two integers, a start and end location that mark an answer span. At training and inference time the model is then provided with a paired context and question.

The model is trained for 5 epochs using the Transformers library (Wolf et al., 2020) using a batch size of 12, learning rate of $3e{-}5$, max sequence length 384 and document stride 128. When there are multiple annotations, a single answer variant is used in training. The result is a model that performs with an F1 score of 76.0 and exact match score of 58.4 over the test set as shown in Table 4. The scores for the respective question types using the IceBERT model are shown in Table 5. The why questions score lowest with a 43.8% e.m., while the *other, is* and *how* score above 50%. Note that the E.m. column in the table stands for "exact match" and denotes the fraction of examples where the answer span was chosen correctly. To calculate the exact match and F1 metrics we use the official scoring script for SQuAD.

We compare the performance of a fine-tuned IceBERT model with a fine-tuned XLMR-base (Conneau et al., 2020) model where the fine-tuning is done over NQiI and TyDi. We chose English and Finnish from the TyDi dataset as they most resemble Icelandic on the feature level as described earlier. This comparison might be beneficial for low-resource languages where a language model pre-trained on a monolingual corpus is not available.

[3] `https://huggingface.co/datasets/vesteinn/icelandic-qa-NQiI`.

[4] Available on HuggingFace at `https://huggingface.co/mideind/`.

Not surprisingly, IceBERT performs better than the XLMR model, its vocabulary is tailored to Icelandic and the model is pre-trained on more Icelandic text than XLMR. However, the multilingual XLMR-base model is not a bad choice as it reaches F1-scores close to 70 for Icelandic, English and Finnish. We hypothize that the large differences between the F1-scores and E.m. are due to the inter annotator disagreement shown in Table 1.

| Model | Dataset | F1 | E.m. |
|---|---|---|---|
| IceBERT | NQiI | 76.0 | 58.4 |
| XLMR-base | NQiI | 72.1 | 56.1 |
| XLMR-base | TyDi English | 67.7 | 56.6 |
| XLMR-base | TyDi Finnish | 70.3 | 44.4 |

Table 4: Accuracy for models adapted from IceBERT and the original XLMR-base model on NQiI and TyDi training data. We show both F1 scores and the exact match score (e.m.) for the corresponding test sets.

| Question type | Entries | F1 | E.m. |
|---|---|---|---|
| What | 182 | 65.4 | 37.4 |
| How | 25 | 78.8 | 60.0 |
| When | 60 | 64.0 | 41.7 |
| Where | 40 | 70.6 | 50.0 |
| Is | 18 | 79.6 | 72.2 |
| Who | 110 | 66.3 | 47.3 |
| Which | 43 | 62.0 | 46.5 |
| Why | 32 | 56.6 | 43.8 |
| Other | 16 | 97.9 | 87.5 |

Table 5: F1 scores and e.m. values by question type as measured using the IceBERT model fine tuned on NQiI.

At this time we do not conduct experiments with questions-passage pairs containing yes/no answers as there are only 245 such pairs in the dataset whereof only 10 were checked by more than one annotator with an agreement rate of 56%.

## 6. Discussion

When building the NQiI dataset we applied the same methodology as was used to build the TyDi dataset. Since a crowdsourcing platform like Amazon mechanical turk with Icelandic speaking contributors was not available at the time we built the dataset we employed five university students for the task. This limited number of annotators can introduce some form of annotator bias in the dataset that is probably less prominent in other datasets with a larger set of annotators. However, we do believe that by meeting the annotators, by reviewing their work, and by giving them feedback, they became better equipped to make high-quality questions and annotations than anonymous workers that are incentivized for quantity over quality. It is not unlikely that this homogenous background, and age, may have caused a bias to some topics or phrasing of the questions posed.

We also note that each of the 11 languages in the original TyDi dataset has millions of speakers and all of them have more pages in the Wikipedia of their language than Icelandic has as of this writing[5]. Nine of them (with the exception of Telugu and Kiswahili) have over a hundred thousand articles in their corresponding Wikipedia. If Wikipedia is used as a knowledge source to build a QA dataset with the methods we used then its size must be taken into account since it could hamper the diversity of the resulting dataset.

In our case, we managed to exhaust the Icelandic Wikipedia just barely and we do encourage others building QA datasets for languages with Wikipedias of a similar size or smaller to consider other sources of information, ideally sources of high-quality text that are likely to contain answers to information-seeking questions. As an example for Icelandic, we have the Icelandic *Web of Science*[6], a collection of detailed answers, written by faculty at the University of Iceland, to questions submitted by the public. Importantly, using alternative reference corpora can improve the efficiency of retrieving article candidates using Google search. In our setting, 34% of questions could not be used because no Wikipedia article was found. This may be due to the fact that the Icelandic Wikipedia is small with only around 3,730 pages containing more than 250 characters. The small size of the Icelandic Wikipedia also introduces a bias in our question elicitation process since the questions are most likely to be on the set of topics determined by the few Wikipedia pages and its contributors.

In future work, we aim to improve the dataset and develop further benchmarks for QA in Icelandic[7]. Ideally, we want to attract more annotators to ensure the question creators are of a more varied background, both with respect to age and academic background, which could reduce annotator and cultural bias and increase diversity. We also aim to introduce quality checks, e.g., to filter out erroneous question-passage pairs. The inter annotator agreement can also be improved by providing clearer guidelines such as encouraging the selection of a minimal span containing the answer. Examples of annotator disagreement are shown in Table 7 in the Appendix. To leverage the human labor more efficiently for quality checks, we plan to guide the annotator by ranking paragraphs using a model as has been done

[5] For an overview, see for example `https://meta.wikimedia.org/wiki/Wikipedia_article_depth`.

[6] `https://visindavefur.is`

[7] After we finished building NQiI, a crowd-sourcing initiative was started for building another QA dataset in Icelandic, see spurningar.is.

by Asai et al. (2020) or by selecting question-passages pairs by uncertainty and by having annotators review labelled spans in an active learning fashion.

Our aim is to eventually build more capable QA systems such as multi-hop systems (Ho et al., 2020) for Icelandic and the required datasets such as XCOPA (Ponti et al., 2020) or a conversational question answering dataset like CoQA (Reddy et al., 2019). Given the size of the Icelandic language community, we are concerned about the human labor required to build these datasets. It will be challenging for a small community to keep up with the pace of high-resource languages such as English. Methods to build datasets efficiently, perhaps through better translation methods or more efficient use of human labor, would be beneficial. We also note that with progress in cross-lingual transfer, large datasets for training might not be necessary, but datasets for evaluation of cross-lingual transfer results will still be required.

## 7. Conclusion

In this paper, we describe work on building the first public extractive question-answering dataset for Icelandic along with benchmark evaluation results. Datasets such as this one are necessary to accurately evaluate QA models for native speakers and important when developing QA systems in general to ensure they fit a variety of different languages, both morphologically and grammatically.

## 8. Access and licensing

The full dataset, as well as a SQuAD-style variant have been made available in the Icelandic CLARIN repository (Snæbjarnarson et al., 2021) with an open, permissive license, CC BY 4.0.

## 9. Acknowledgements

We would like to thank Akari Asai for granting us access to the annotation software we adapted for this project.

We would also like to thank the student annotators: Bergur Tareq Tamimi, Ingibjörg Iða Auðunardóttir, Unnar Ingi Sæmundsson, Hildur Bjarnadóttir and Helgi Valur Gunnarsson. They were supported by a grant from the Icelandic student innovation fund.

Finally, we thank the anonymous reviewers for their helpful comments and questions.

## Appendix: Examples from the dataset

Example questions, answers and passages from the NQiI dataset are shown in Table 6. Translations to English are included for the questions and answers. We note that the full dataset includes multiple answers to some questions and thus multiple location spans.

## Appendix: Examples of disagreeing annotations

As inter annotator disagreement is quite high in the dataset, we include some examples of disagreeing annotations in Table 7.

## 10. Bibliographical References

Árnason, K., Pind, J., Kvaran, G., and Þráinsson, H. (2005). *Íslensk tunga*, volume 1. Almenna bókafélagið.

Artetxe, M., Ruder, S., Yogatama, D., Labaka, G., and Agirre, E. (2020). *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388. 00028 arXiv: 2004.14958.

Asai, A., Kasai, J., Clark, J. H., Lee, K., Choi, E., and Hajishirzi, H. (2020). XOR QA: Cross-lingual Open-Retrieval Question Answering. *arXiv:2010.11856 [cs]*, October. 00003 arXiv: 2010.11856.

Bi, B., Wu, C., Yan, M., Wang, W., Xia, J., and Li, C. (2019). Incorporating External Knowledge into Machine Reading for Generative Question Answering. *arXiv:1909.02745 [cs]*, September. 00020 arXiv: 1909.02745.

Cambazoglu, B. B., Sanderson, M., Scholer, F., and Croft, B. (2021). A review of public datasets in question answering research. *ACM SIGIR Forum*, 54(2):5:1–5:23, August. 00006.

Chandra, A., Fahrizain, A., Ibrahim, and Laufried, S. W. (2021). A Survey on non-English Question Answering Dataset. *arXiv:2112.13634 [cs]*, December. 00000 arXiv: 2112.13634.

Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., and Palomaki, J. (2020). Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

De Bruyn, M., Lotfi, E., Buhmann, J., and Daelemans, W. (2021). MFAQ: a Multilingual FAQ Dataset. *arXiv:2109.12870 [cs]*, October. 00000 arXiv: 2109.12870.

Matthew S. Dryer et al., editors. (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Efimov, P., Chertok, A., Boytsov, L., and Braslavski, P. (2020). SberQuAD – Russian Reading Comprehension Dataset: Description and Analysis. In Avi

| Question | Passage |
|---|---|
| Frá hvaða landi er Mozart? (e. *From what country is Mozart?*) | Austurríki: Margir af þekktustu tónskáldum og tónlistarmönnum klassíska tímans komu frá **Austurríki**. Þeirra helstu eru Wolfgang Amadeus Mozart, Joseph Haydn, Franz Schubert, Anton Bruckner og Franz Liszt, sem var Ungverji sem fæddist í Burgenland. Auk þeirra fluttu ýmsir aðrir til Vínar til að starfa þar, svo sem Ludwig van Beethoven (fæddist í Bonn). **Answer: Austurríki (e. *Austria*)** |
| Hvenær var Búdapest orðin að einni borg? (e. *When had Budapest become a single city?*)' | Berlín: Berlín er stærsta borg og höfuðborg Þýskalands með tæpar 3,44 milljónir íbúa (2014) en flestir hafa íbúarnir verið 4,4 milljónir (2011), fyrir síðari heimsstyrjöld. Borgin er einnig sú næstfjölmennasta innan Evrópusambandsins á eftir London ef miðað er við opinber borgarmörk. Berlín stendur við árnar Spree og Havel í norðaustanverðu Þýskalandi og er umlukt sambandslandinu Brandenborg, en borgin sjálf er sjálfstætt sambandsland. **Answer: None** |
| Er England í Evrópusambandinu? (e. *Is England in the European Union?*) | Bretland: Bretland var eitt af tólf löndum sem stofnuðu Evrópusambandið árið 1992 þegar Maastrichtsáttmálinn var undirritaður. Fyrir stofnun ESB var Bretland aðildarríki Evrópubandalagsins frá 1973. Árið 2016 ákvað Bretland hins vegar að segja sig úr sambandinu með þjóðaratkvæðagreiðslu. Bretar yfirgáfu sambandið í lok janúar 2020. **Answer: No** |

Table 6: Examples from the NQiI dataset, note that the dataset includes information about answer span locations and sometimes more than one answer and location.

| Question | Annotations |
|---|---|
| Hver er fólksfjöldinn í Bandaríkjunum? | **A1:** yfir 324 milljónir íbúa (árið 2017)<br>**A2:** yfir 324 milljónir íbúa |
| Hvenær var Rammstein með tónleika á Íslandi? | **A1:** 15. júní 2001<br>**A2:** Rammstein hélt tvenna tónleika á Íslandi, þá fyrstu 15. júní 2001 |
| Hvenær var Háskóli Ísland stofnaður? | **A1:** árið 1911<br>**A2:** 1911 |
| Eftir hvern er óperan Hollendingurinn fljúgandi? | **A1:** Richard Wagner<br>**A2:** Wagner |
| Hvar lést Robert Fischer? | **A1:** í Reykjavík<br>**A2:** dáinn í Reykjavík 17. janúar 2008 |

Table 7: Examples of disagreeing answer annotations from the NQiI dataset.

Arampatzis, et al., editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, pages 3–15, Cham. Springer International Publishing. 00026.

Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., and Auli, M. (2019). ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy, July. Association for Computational Linguistics.

Geirsson, P. (2013). Iceqa: Developing a question answering system for icelandic.

Haspelmath, M., Dryer, M. S., Gil, D., and Comrie, B. (2005). *The world atlas of language structures*. OUP Oxford.

Herzig, J., Müller, T., Krichene, S., and Eisenschlos, J. (2021). Open domain question answering over tables via dense retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 512–519, Online, June. Association for Computational Linguistics.

Hilmarsson-Dunn, A. and Kristinsson, A. P. (2010). The language situation in Iceland. *Current Issues in Language Planning*, 11(3):207–276, August. 00000 Publisher: Routledge _eprint: https://doi.org/10.1080/14664208.2010.538008.

Ho, X., Duong Nguyen, A.-K., Sugawara, S., and Aizawa, A. (2020). Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Con-*

*ference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4411–4421. PMLR, November. 00263 ISSN: 2640-3498.

Kakwani, D., Kunchukuttan, A., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M. M., and Kumar, P. (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, November. Association for Computational Linguistics. 00003.

Keren, O. and Levy, O. (2021). ParaShoot: A Hebrew Question Answering Dataset. *arXiv:2109.11314 [cs]*, September. 00001 arXiv: 2109.11314.

Khashabi, D., Cohan, A., Shakeri, S., Hosseini, P., Pezeshkpour, P., Alikhani, M., Aminnaseri, M., Bitaab, M., Brahman, F., Ghazarian, S., Gheini, M., Kabiri, A., Mahabadi, R. K., Memarrast, O., Mosallanezhad, A., Noury, E., Raji, S., Rasooli, M. S., Sadeghi, S., Azer, E. S., Samghabadi, N. S., Shafaei, M., Sheybani, S., Tazarv, A., and Yaghoobzadeh, Y. (2021). ParsiNLU: A Suite of Language Understanding Challenges for Persian. *arXiv:2012.06154 [cs]*, July. 00003 arXiv: 2012.06154.

Kinn, K., Rusten, K. A., and Walkden, G. (2016). Null subjects in early icelandic. *Journal of Germanic Linguistics*, 28(1):31–78.

Korablinov, V. and Braslavski, P. (2020). Rubq: a russian dataset for question answering over wikidata. In *International Semantic Web Conference*, pages 97–110. Springer.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. (2019). Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Longpre, S., Lu, Y., and Daiber, J. (2021). MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. *arXiv:2007.15207 [cs]*, August. 00021 arXiv: 2007.15207.

Paschoal, A. F. A., Pirozelli, P., Freire, V., Delgado, K. V., Peres, S. M., José, M. M., Nakasato, F., Oliveira, A. S., Brandão, A. A. F., Costa, A. H. R., and Cozman, F. G. (2021). Pirá: A Bilingual Portuguese-English Dataset for Question-Answering about the Ocean. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4544–4553. Association for Computing Machinery, New York, NY, USA, October. 00000.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.

Rogers, A., Gardner, M., and Augenstein, I. (2021). QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. *arXiv:2107.12708 [cs]*, July. 00008 arXiv: 2107.12708.

Ruder, S., Constant, N., Botha, J., Siddhant, A., Firat, O., Fu, J., Liu, P., Hu, J., Garrette, D., Neubig, G., and Johnson, M. (2021). XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation. *arXiv:2104.07412 [cs]*, October. 00013 arXiv: 2104.07412.

Snæbjarnarson, V., Símonarson, H. B., Ragnarsson, P. O., Ingólfsdóttir, S., Jónsson, H. P., Þorsteinsson, V., and Einarsson, H. (2022). A Warm Start and a Clean Crawled Corpus – A Recipe for Good Language Models. *arXiv:2201.05601 [cs]*, January. 00000 arXiv: 2201.05601.

Tahsin Mayeesha, T., Md Sarwar, A., and Rahman, R. M. (2021). Deep learning based question answering system in Bengali. *Journal of Information and Telecommunication*, 5(2):145–178, April. 00002 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/24751839.2020.1833136.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., and Suleman, K. (2017). NewsQA: A Machine Comprehension Dataset. *arXiv:1611.09830 [cs]*, February. 00520 arXiv: 1611.09830.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Yang, Y., Yih, W.-t., and Meek, C. (2015). WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal, September. Association for Computational Linguistics. 00515.

## 11. Language Resource References

Ponti, Edoardo Maria and Glavaš, Goran and Majewska, Olga and Liu, Qianchu and Vulić, Ivan and Korhonen, Anna. (2020). *XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning*.

Reddy, Siva and Chen, Danqi and Manning, Christopher D. (2019). *CoQA: A Conversational Question Answering Challenge*.

Snæbjarnarson, Vésteinn and Einarsson, Bergur Tareq Tamimi and Auðunardóttir, Ingibjörg Iða and Sæmundsson, Unnar Ingi and Bjarnadóttir, Hildur and Gunnarsson, Helgi Valur and Einarsson, Hafsteinn. (2021). *NQiI - Natural Questions In Icelandic - v1.0*.

Wallenberg, Joel C and Ingason, Anton Karl and Sigurðsson, Einar Freyr and Rúnarsson, Kristján and Rögnvaldsson, Eiríkur. (2011). *The Icelandic Parsed Historical Corpus (IcePaHC)*. University of Iceland.