Pirá: A Bilingual Portuguese-English Dataset for Question-Answering about the Ocean

André F. A. Paschoal*
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo
andre.faleiros.paschoal@usp.br

Valdinei Freire, Karina V. Delgado
Sarajane M. Peres
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo
{valdinei.freire,kvd,sarajane}@usp.br

Paulo Pirozelli*†
Instituto de Estudos Avançados
Universidade de São Paulo
paulo.pirozelli.silva@usp.br

Marcos M. José, Flávio Nakasato
André S. Oliveira, Anarosa A. F. Brandão
Anna H. R. Costa, Fabio G. Cozman
Escola Politécnica
Universidade de São Paulo
{marcos.jose,flavio.cacao,andre.seidel}@usp.br
{anarosa.brandao,anna.reali,fgcozman}@usp.br

ABSTRACT

Current research in natural language processing is highly dependent on carefully produced corpora. Most existing resources focus on English; some resources focus on languages such as Chinese and French; few resources deal with more than one language. This paper presents the *Pirá* dataset, a large set of questions and answers about the ocean and the Brazilian coast both in Portuguese and English. *Pirá* is, to the best of our knowledge, the first QA dataset with supporting texts in Portuguese, and, perhaps more importantly, the first bilingual QA dataset that includes this language. The *Pirá* dataset consists of 2261 properly curated question/answer (QA) sets in both languages. The QA sets were manually created based on two corpora: abstracts related to the Brazilian coast and excerpts of United Nation reports about the ocean. The QA sets were validated in a peer-review process with the dataset contributors. We discuss some of the advantages as well as limitations of *Pirá*, as this new resource can support a set of tasks in NLP such as question-answering, information retrieval, and machine translation.

CCS CONCEPTS

• **Applied computing** → *Document searching*; **Annotation**.

KEYWORDS

Question-answering dataset, Bilingual dataset, Portuguese-English dataset, Ocean dataset

*Both authors contributed equally to this research.
†Corresponding author: paulo.pirozelli.silva@usp.br.

1 INTRODUCTION

The best current solutions to question answering and reading comprehension tasks rely on large scale datasets. That poses a problem for many languages; even though a number of datasets are available in English [6, 31–33], resources in other languages are rather scarce. While a few other languages have received some attention, such as Chinese [8, 18], French [11], and German [27], and while one can find multilingual datasets around [2, 23, 25], many languages, such as Portuguese, still lag behind. Question answering (QA) in non-English languages suffers from an additional difficulty: many, and in some cases most, of the documents used to answer questions are only available in English. Those working with non-English QA must then resort to automated translations without the support of curated bilingual datasets.

In this paper, we describe the creation of the *Pirá* dataset, a high-quality question answering dataset for Portuguese and English that focuses on the ocean and the Brazilian coast. *Pirá* [1] is an openly available bilingual scientific dataset built with the help of 254 volunteer undergraduate and graduate students. To the best of our knowledge, the *Pirá* dataset is the first QA dataset with supporting texts in Portuguese; more importantly, it is the first bilingual QA dataset where Portuguese is one of the languages. *Pirá* is also the first QA dataset in Portuguese with unanswerable questions so as to allow the study of *answer triggering*; finally, it is the first

[1]The word *Pirá* means "fish" in Tupi-Guarani, a family of indigenous languages from South America that heavily influenced Brazilian Portuguese.

QA dataset that deals with scientific knowledge about the ocean, climate change, and marine biodiversity.

Contributions. We offer the following key contributions:

(1) A bilingual (Portuguese-English) QA dataset about ocean data, biodiversity, and climate change, consisting of 4074 texts and 2261 QA sets. In our dataset, a *QA set* consists of four elements: a question in Portuguese and in English, and an answer in Portuguese and in English.

(2) Methods both for enriching datasets through the production of equivalent answers and paraphrased questions, and for manually assessing and describing QA datasets.

Besides, we describe in this paper an open application for producing QA datasets based on supporting documents that can be used for crowdsourcing.

The paper is structured as follows: In Section 2 we outline existing datasets and highlight the *Pirá* dataset main features. In Section 3 we describe the protocol for building the dataset as well as the method for creating and evaluating questions. Section 4 explains the process for augmenting the dataset with the manual creation of answers and questions; it also describes three versions of the *Pirá* dataset that we make available. In Section 5, we present a preliminary analysis of the dataset, describing the main results obtained in the assessment step. Section 6 is dedicated to use cases of the *Pirá* dataset, as well as to a discussion of some of its limitations. We then conclude the paper in Section 7 with plans for future work.

2 BACKGROUND AND MOTIVATION

In this section we first summarize a few facts about the domain of our QA dataset, namely, the ocean and in particular the Brazilian coast. This is a suitable domain not only due to its intrinsic importance and complexity, but also because it naturally lends itself to a dataset in Portuguese and English — one of our main goals is to understand the challenges that bilingual conversational agents may face. We also present in this section a brief survey of existing resources for Natural Language Processing in Portuguese.

2.1 Domain

More than 70 per cent of the surface of the planet is covered by the ocean and 95 per cent of Earth's biosphere lies in it (adopting current view that there is a single connected ocean). Several economic activities directly depend on the ocean, such as fishing, tourism, and extraction of natural resources. Currently, more than 80% of the international trade is made by shipping. The ocean and its ecosystems also provide significant benefits to the global community, including climate regulation, coastal protection, food, employment, recreation, cultural well-being and spiritual bonds.

Recent change in weather patterns caused primarily by human-induced global warming are raising the ocean's temperature and threatening maritime ecosystems. The development of infrastructure in coastal areas, overfishing and garbage dumping are putting many species in danger. Anthropogenic noise is also disturbing maritime life [29, 30].

All these factors demand a greater social awareness of the ocean fundamental importance to human life and the planet. For that reason, the United Nations (UN) established as two of its Sustainable Development Goals "to conserve and sustainably use the oceans,

seas and marine resources" and "take urgent action to combat climate change and its impacts" [28]. The ocean is studied in many fields, such as geology, oceanography, biology, and economics. Despite its importance, up to now no public dataset deals with these topics or, as to our knowledge, any close themes. By filling this gap, we hope that *Pirá* dataset can stimulate more AI researchers to contribute with the advance and diffusion of knowledge on the sustainable use of the ocean.

2.2 Existing resources for Portuguese

Compared to English, and even to other languages such as Chinese or German, resources in Portuguese are rather limited. Most existing resources are geared towards basic syntactic and semantic analysis, such as Mac-Morpho [1, 13] for part-of-speech tagging, or PropBank [12] for Semantic role labeling. Amongst task-oriented resources and associated benchmarks we can cite ASSIN [14] for semantic similarity and textual entailment; SIMPLEX-PB [16, 17] for lexical simplification; and IDPT for irony detection [9, 10, 26].

QA datasets are particularly rare in Portuguese. Large multilingual QA datasets, for example, tend to ignore Portuguese [2, 3, 5, 23, 25], even though it is the sixth most spoken language in the world, with 221 million native speakers.[2] Existing QA datasets in Portuguese usually consist of automatic translations of datasets in English, which are then sampled for manual editing. We mention, for instance, versions of SQuAD[3] and GLUE[4] in Portuguese. The two exceptions are the ENEM-Challeng [36] and MilkQA [7]. The ENEM-Challenge is based on ENEM, the entrance examination valid for almost all universities in Brazil. The dataset contains 1,800 multiple choice questions on Humanities, Languages, Sciences and Mathematics, manually annotated with the types of background knowledge that are required to answer the questions. MilkQA consists of consumer questions from Embrapa's (Brazilian Agricultural Research Corporation) Dairy Cattle unity. The MilkQA dataset contains 2,657 anonymized pairs of questions and answers created directly in Portuguese, in which questions are associated with a pool of 50 candidate answers where only one answer is correct.

3 METHOD

The dataset generation process is depicted in Figure 1. Firstly, we collected two different corpora: abstracts of scientific papers about the Brazilian coast (also known as "Blue Amazon"), and small excerpts of two books about the ocean organized by the United Nations. Secondly, QA sets were manually generated by a set of volunteers. The volunteers were undergraduate and graduate students and researchers at the University of São Paulo. The high educational level of participants let us build a scientific dataset with questions and answers that go beyond mere trivia. In this second phase, volunteers received random texts from both corpora and had to create QA sets based on them. Participants were instructed to produce questions that could be answered with the use of the texts and no other source of information. Third, QA sets went through an extensive manual assessment process. Volunteers received QA sets produced in the previous phase by other individuals, and for each of

---

[2]https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers
[3]https://drive.google.com/file/d/1Q0IaIlv2h2BC468MwUFmUST0EyN7gNkn/view
[4]https://drive.google.com/drive/folders/1WXdarHqxlJqKm30uD2LubwuLfpAGYZX0
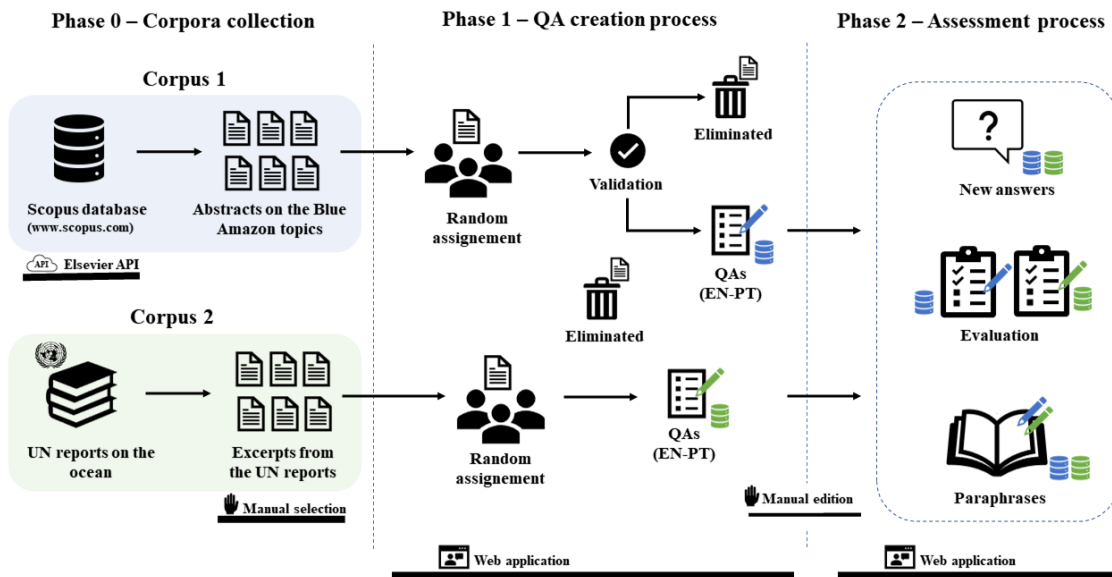
Figure 1: Overview of the *Pirá* dataset generation process

these QA sets, they had to: i) answer the question in both languages without having access to the original answer; ii) assess the whole original QA set (the questions and respective answers) according to a number of aspects; and iii) paraphrase the original question.

A total of 254 volunteers took part in the activity (18 researchers, 169 undergraduate students, 67 graduate students).[5] The remainder of this section describes each phase of the method in detail.

3.1 Phase 0 - Corpora collection

Two sets of texts were used as supporting documents for the QA sets generation. Similarly to the approach taken by PubMedQA [19], our Corpus 1 contains abstracts of scientific papers on the Brazilian coast topics. Abstracts were gathered from Elsevier's Scopus database,[6] an abstract and citation database that covers thousands of scientific journals, conference proceedings and books in different fields of knowledge. The construction of the corpus required us to filter relevant texts from the thousands of documents in the Scopus database. First, we manually analyzed the results for several different keyword sets, in order to find accurate filtering queries and to minimize the number of false positives. Then an expert in the field evaluated the results of our queries and suggested changes to our set of keywords. Finally, we run a final search with the improved query and downloaded the abstracts together with their metadata.

Corpus 2 consists of excerpts of two reports about the ocean organized by the United Nations, the *World Ocean Assessment I* [29] and the *World Ocean Assessment II* [30]. The excerpts were manually selected for the task, following some guidelines: they presented relatively independent contents and dealt with topics that could be understood by readers from the exact sciences with only a generic knowledge of other areas.

Table 1: Summary of the textual corpora

Characteristic	Corpus 1	Corpus 2
Subject	Brazilian coast	Ocean
Type	Abstracts	Text excerpts
Source	Scopus	UN reports
Number of texts	3891	183
Average size (words)	201.72	344.88
Smallest document (words)	15	98
Largest document (words)	1176	1208

The abstracts contained in Corpus 1 are considerably more technical when compared to the text excerpts contained in Corpus 2. Table 1 summarizes the main characteristics of the two corpora.

3.2 Phase 1 - QA creation

After collecting the two corpora, the next step focused on the construction of QA sets by volunteers. A web application was developed for that purpose. Because the corpora were distinct, the method had to be slightly adapted for each case. For this reason, we describe both procedures separately.

3.2.1 Corpus 1 - Scientific abstracts on Brazilian coast topics. Because abstracts were automatically selected, some retrieved abstracts were in fact not related to our domain of interest. Such false positives should not be considered as a basis for constructing the QA dataset. To fix that, we carried out a validation step: after receiving a new abstract, volunteers were asked to confirm its adherence to the topic; in case the document was not related to our domain, participants were instructed to eliminate it. Those abstracts were then marked as false positives in our database and not distributed

---

again in the whole process. If a volunteer thought a document actually dealt with the domain of interest, she was instructed to mark the text as a true positive. To help volunteers in the validation step, the web application provided them with a non-exhaustive list of topics related to the Brazilian coast.

After the validation step, QA sets were generated. Volunteers were asked to produce up to three question/answer pairs per selected abstract. As our dataset is bilingual, volunteers were asked to also produce translations of the questions and answers (recall that each QA set contains four elements: a question in Portuguese, an answer in Portuguese, a question in English, and an answer in English). Volunteers were allowed to use automatic translation tools as long as they checked the translations. They should not use the internet, however, to search for other sources of information about the subject discussed in the abstract under analysis.

Before they started generating QA sets, volunteers were advised as follows: questions should not be generic, but rather they should be contextualized according to the subject discussed in the abstract under analysis; answers should be based on the abstract, but they did not need to be a span, or segment of text, of the abstract; answers should be short, but not repeating the question again; yes or no questions should be avoided (in order to obtain more complex QA sets); and the volunteers should try to diversify the types of questions, varying the use of question words, such as "why", "when", "how", and so on.

We created an optimal scoring scheme (15 points) to encourage participation and reduce early dropout. In fact, the score served as a guide to indicate what we considered to be a minimum expected engagement. Although volunteers were not required to achieve this number of points, the scheme acted as a way to show them whether they were falling short of our expectations and/or contributing beyond them. The application also had a scoring system that rewarded a higher number of questions for abstract: one point for one question, three points for two questions and five points for three questions. The aim was to get more questions per abstract, thus obtaining more complex and diversified questions, instead of easy, and repeatable, patterns of question. Table 2 shows the results of this process on Corpus 1.

3.2.2 Corpus 2 - Text excerpts from UN reports about the ocean. For Corpus 2, the validation step was not necessary as texts came from books that specifically dealt with the subject. In this case, volunteers were asked to make up two QA sets for text excerpt. We lowered the number of questions based on the difficulty volunteers voiced regarding Corpus 1. Again, each QA set comprised four elements: question in Portuguese, answer in Portuguese, question in English, and answer in English. The remaining instructions were the same as for Corpus 1.

We also created a scoring scheme (8 points) to reward volunteers' engagement with the activity. Volunteers had to create two QA sets in order to move to another excerpt; they received one point for QA set submitted. Table 2 shows the results from the QA set generation process regarding Corpus 2.

3.2.3 Manual editing. The volunteers are well educated people, most of them familiar with scientific texts. Nevertheless, the complexity of the topics, the academic style of the texts, and the fact that individuals were not native English speakers led to a number

Table 2: Results of the QA generation process

Characteristics	Corpus 1	Corpus 2
#Abstracts/excerpt texts used	532	149
#Abstracts eliminated	727	- -
#QA sets generated	1675	586
Texts per volunteer	2.96	2.01
QA sets per volunteer	9.31	7.92
QA sets per text	3.15	3.94
Words per question (avg.) (EN | PT)	13.37 | 14.07	12.11 | 13.06
Words per answer (avg.) (EN | PT)	13.34 | 14.56	14.90 | 16.23

of mistakes. To overcome that, the QA sets produced in Phase 1 were manually edited for spelling and grammar.

We also took the opportunity to remove questions that were too broad, such as "Which process does this paper illustrate?". We did not eliminate, however, questions that just did not contextualize it enough, since that aspect should be evaluated in the second phase of the QA dataset generation process. Table 3 presents some examples of the final state of the questions and answers.

3.3 Phase 2 - Assessment process

After Phase 1 was finished, volunteers were invited to evaluate the QA sets in a number of ways. The abstracts and excerpts, with the corresponding QA sets, were randomly assigned to volunteers in order to check their quality. We ensured that volunteers would not evaluate QA sets that they themselves created. QA sets were treated individually in this process; i.e., a volunteer evaluated a single QA set for a given text, even if in the previous phase more than one QA set had been made for that same text. The assessment procedure consisted of three steps described in the remainder of this section.

3.3.1 Answering. In the first step, volunteers received an abstract or excerpt and a question associated with it (both in English and in Portuguese). They then had to answer it in both languages, taking into account only the information available via the abstract/excerpt text. The volunteers could also pass to another abstract/excerpt text in case they were not able to answer the current question.

3.3.2 Assessment. In the second step, volunteers were presented again to the same abstract/excerpt text and question, but now also with the *original* answer and the new answer elaborated in the previous step. They were then asked to evaluate this whole set according to the following aspects:

- answer two yes/no questions: (1) *Is this a generic question? Example: "How big is the oil field?"*; (2) *Can you answer the question only with the information provided by the text?*
- evaluate three statements about the QA set according to a Likert scale (1 - Strongly disagree, 2 - Disagree, 3 - Neither agree nor disagree, 4 - Agree, 5 - Strongly agree): (1) *The question makes sense;* (2) *The question is hard;* (3) *Your answer and the original answer are equivalent.*
- indicate the type of question: *who, what, where, when, why, how, none of them.*

Table 3: Examples of questions and answers

	Context	QA set
Corpus 1	The greenhouse effect and resulting increase in the Earth's temperature may accelerate the mean sea-level rise. The natural response of bays and estuaries to this rise, such as this case study of Santos Bay (Brazil), will include change in shoreline position, land flooding and wetlands impacts. The main impacts of this scenario were studied in a physical model built in the Coastal and Harbour Division of Hydraulic Laboratory, University of São Paulo, and the main conclusions are presented in this paper. [...]	**Question-EN:** What are some consequences of the greenhouse effect in bays and estuaries? **Question-PT:** Quais são algumas consequências do efeito estufa em baías e estuários? **Answer-EN:** Change in shoreline position, land flooding and wetlands impacts. **Answer-PT:** Mudança na posição da linha costeira, inundações de terra e impactos nas zonas húmidas.
Corpus 2	[...] Even in the open ocean, climate warming will increase ocean stratification in some broad areas, reduce primary production and/or result in a shift in productivity to smaller species (from diatoms of 2-200 microns to picoplankton of 0.2-2 microns) of phytoplankton. This has the effect of changing the efficiency of the transfer of energy to other parts of the food web, causing biotic changes over major regions of the open ocean, such as the equatorial Pacific.	**Question-EN:** What is one of the effects of the increased ocean stratification, resulting from the climate warming? **Question-PT:** Qual é um dos efeitos do aumento na estratificação dos oceanos, resultante do aquecimento do clima? **Answer-EN:** Changing the efficiency of energy transfer within the food web in the ocean. **Answer-PT:** Alteração da eficiência da transferência de energia dentro da teia alimentar no oceano.

3.3.3 Rewriting. One of the main shortcomings of QA datasets based on texts is that questions and answers often share words with the texts, a semantic relationship that can be easily grasped by reading comprehension models. For that reason, even relatively simple models based on word matching heuristics (such as counting the number of words that are repeated in the question and candidate sentences) can obtain performances comparable to state-of-the-art baselines [38]. Given this point, some datasets try to fix this dependency *a posteriori*, through a debiasing procedure, excluding QA sets that are easily answerable in this way [35]. Instead, we established an additional step in the assessment phase in which volunteers were asked to paraphrase the question if they considered it acceptable. That procedure resembles the one used in SelQA [20], in which the annotators were instructed to paraphrase the questions to reduce the co-occurrence of words between the texts, and the questions and answers. In our dataset, however, the rewriting process was executed by a different individual, bringing an additional factor of diversity to the generation of QA sets.[7] Some examples of paraphrases are shown in Table 4.

4 DATASET AUGMENTATION AND VERSIONS

With the new questions and answers produced during Phase 2, we were able to expand and enrich the original dataset, obtaining more lexically diverse QA sets. Furthermore, the assessments carried out by the volunteers allowed us to qualify the QA sets. Three versions of the dataset were created: *Pirá-F* (filtered), *Pirá-T* (triggering), and *Pirá-C* (complete).

Pirá-F is the filtered version of the dataset. It includes only QA sets that volunteers indicated as not generic, that could be answered with only the information provided by the text, and moreover that volunteers agreed (4) or strongly agreed (5) that made sense. When

volunteers reported in Phase 2 that their answer did not coincide with the original one (i.e., 1-2 in the Likert scale for "Your answer and the original answer are equivalent"), we excluded the new answer from the dataset. Finally, we also included the paraphrases. The dataset contains up to four possible combinations of QA sets: original question, original answer; original question, new answer; paraphrase, original answer; paraphrase, new answer.

Pirá-T is the answer-triggering version of the dataset. It includes all QA sets generated in Phases 1 and 2 (meaningful and meaningless, contextualized and non-contextualized, answered by the text only or not). As for *Pirí-F*, we also excluded questions that were assessed with one or two points for the statement "Your answer and the original answer are equivalent". In addition, the answers (original e new) for questions that volunteers regarded meaningless (1-2 in the Likert scale for "The question makes sense") were filled with " " (blank value). In datasets that every question possess an answer, models can always guess an answer (and sometimes be right just by luck). Instead, by allowing some questions not to have an answer, models are required to learn when they are able to answer a given question, a task known as "answer triggering" (which is getting more important in QA datasets [20, 32, 39]).

For the sake of completeness, we also make available *Pirá-C*, with all the questions and answers produced in the generation and assessment phases.

5 DATASET ANALYSIS

In this section, we describe the results obtained in Phase 2. In general, volunteers that were assigned to Corpus 1 (Brazilian coast) took more time to finish the activity, as compared to those working with Corpus 2 (Ocean). This result was expected, given the specialized nature of the abstracts, in comparison to the texts extracted from the UN reports, which are intended for a larger audience. In fact, we also selected the text excerpts that seemed more convenient to readers in the composition of Corpus 2. Table 5 shows the

---

[7]In fact, the process of asking volunteers to avoid using the same words in questions as they appear in the texts (or of writing and then rewriting them, as in SelQA [20]) was already among the instructions of the first phase in our method.

Table 4: Examples of paraphrases

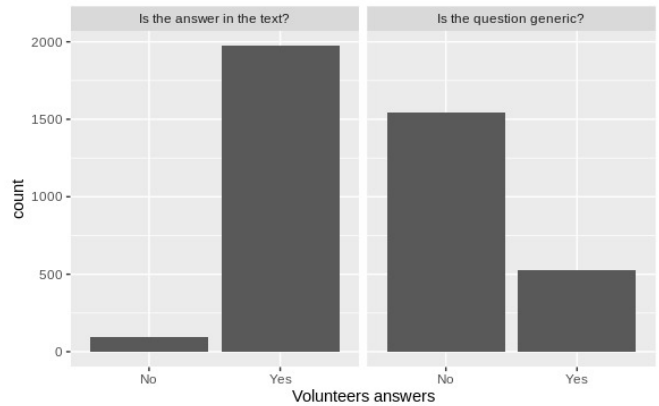|	Old question	New question
Corpus 1	**Question-EN:** What impact did the Guanabara oil spilling pollution caused in foraminiferal taphonomic assemblages? **Question-PT:** Qual o impacto que a poluição do derramamento do óleo da Guanabara causou nas assembléias tafonômicas foraminíferas?	**Question-EN:** What effect did the pollution from the Guanabara oil disaster have on foraminiferal taphonomic assemblages? **Question-PT:** Qual o efeito da poluição do desastre do óleo da Guanabara sobre as assembléias tafonômicas foraminíferas?
Corpus 2	**Question-EN:** What are the consequences on the plankton ecosystem as ocean water warms? **Question-PT:** Quais as consequências no ecossistema de plânctons à medida que a água do oceano aquece?	**Question-EN:** What are the consequences of the rising temperature of ocean water for the plankton ecosystem? **Question-PT:** O que acontece com o ecossistema do plâncton quando a temperatura do oceano sobe?

Table 5: Results of the QA validation process

Characteristics	Corpus 1	Corpus 2
QA sets validated	1506	556
QA sets validated with paraphrase	1342	534
Percentage of QA sets validated	89.91%	94.88%
Words per paraphrase question (avg.) (EN | PT)	13.70 | 14.56	12.34 | 13.10
Words per new answer (avg.) (EN | PT)	12.45 | 13.74	14.10 | 15.45

Figure 2: Answers for the Yes/No questions in Phase 2.

Figure 3: Answers for Likert-scale questions in Phase 2 (cf. Section 3.3.2)

results of the QA validation process on Corpus 1 and 2. Almost the whole dataset was validated, resulting in the generation of 1876 paraphrases.
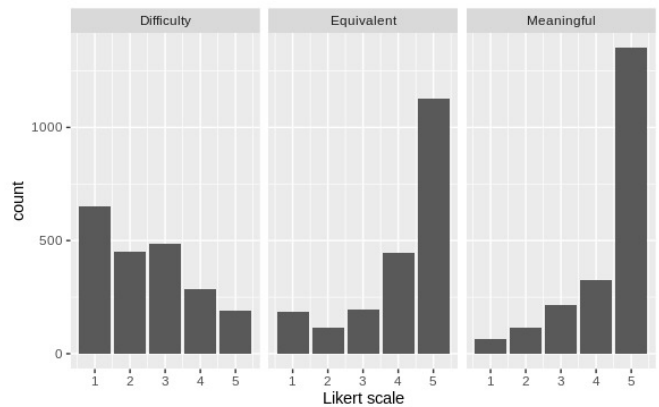
Figure 2 shows the distributions of answers to Yes/No questions for the whole dataset. Almost all questions were considered as being answerable by the content of the text only (95.2%). Approximately three quarters of the questions were properly contextualized, according to volunteers (74.8%). Figure 3 shows the results for the three Likert-scale assessments of QA sets' features. Volunteers generally disagreed that questions were difficult (1 - 31.5%, 2 - 22%). Most of the times, questions were considered meaningful (5 - 65.3%, 4 - 15.8%) and the original and new answers as equivalent (5 - 54.6%, 4 - 21.7%).

Volunteers also classified questions as regards their type. "What" questions made up the majority (62.1%), followed by "How" (12%), "Where" (7.4%), and "Why" (6.9%). "When" questions represented 2.8% of the total and "Who" questions 1.9%. In 4.7% of the cases, questions were not classified in any of the categories. Results can be seen in Figure 4.

In general, assessments were similar for both corpora. A few differences should be stressed, though. For example, answers in Corpus 1 were slightly more generic, according to volunteers (26.8% and 20.6%, respectively). Proportionally, volunteers working with Corpus 2 more often stated they "totally agreed" the questions made sense (70.1%, against 63.6% on Corpus 1). More people responded they "neither agree nor disagree" the question was difficult for Corpus 1 than for Corpus 2 (22% and 26.6%, respectively). As regards the equivalence between the original and the new answer, results were lower for Corpus 1 — 9.7% of the volunteers "totally disagreed"

the answers were equivalent for questions based on texts from Corpus 1, whereas only 5.9% said the same for questions based on texts from Corpus 2. Finally, there was a higher incidence of "where" questions (8.6% for Corpus 1 and 4.1% for Corpus 2) in Corpus 1 and of "how" questions in Corpus 2 (10.7% for Corpus 1 and 15.6% for Corpus 2).
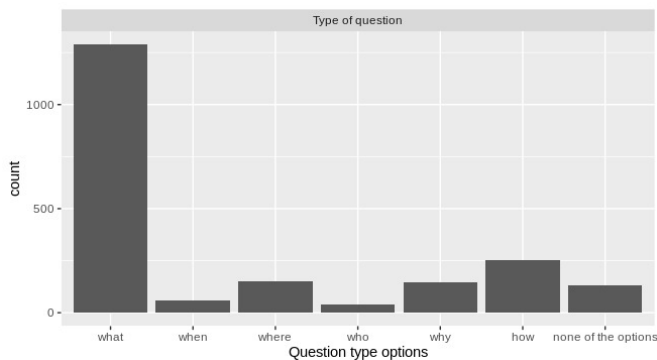
Figure 4: Answers for type of questions in Phase 2.

We then tested the capacity of pre-trained Transformers [37] sentence embeddings to determine equivalences in English sentences, as compared to human judgments. The embeddings were generated by Sentence-BERT [34], the state-of-the-art model for semantic textual similarity assignment.[8] Dissimilarity between answers was measured using Euclidean distance and cosine similarity. We then compared these results with the manual assessments of equivalence generated in Phase 2. The results are reported in Figures 5 and 6. In average, answers to which volunteers attributed a higher degree of equivalence achieved higher similarity scores through Sentence-BERT. Nonetheless, there are some divergences between embedding and human-assigned equivalences, as it can be seen on both figures.

We then carefully examined some of the answers in which manual and automatic assessments diverged the most. To do that, we selected the answers in Corpus 2 which received 5 (cf. Likert scale) in the assessment phase as regards to sentence equivalence but received the highest scores for Euclidean distance (in the Transformer analysis). Three main factors explained the divergence between manual and transformer sentence assessments. First, some volunteers gave short answers while others gave long ones. Second, answers with the same content often had relevant lexical and syntactic differences. Finally, some answers differed in the presence of mathematical and logical reasoning over the text. Table 6 brings examples of each situation.

6 DISCUSSION

In this section, we discuss possible applications, limitations of the *Pirá* dataset, and lessons learned from creating the dataset.

6.1 Impact and applications

The *Pirá* dataset is a valuable resource for reading comprehension and information retrieval. It is particularly useful for researchers working with bilingual tasks, since it mimics a common problem faced by them: that of accessing texts in English to answer questions in their own language. Four characteristics make *Pirá* dataset particularly interesting for that matter: i) it is bilingual; ii) questions

were based on corpora with different complexities; iii) QA sets were manually assessed in a number of ways; and iv) the dataset was enriched by paraphrases.

A first use case for the *Pirá* dataset is in information retrieval. The task, in that case, is finding, for a question, the text which contains the answer to it. Neural retrievers, such as the Dense Passage Retriever, for instance, are trained using QA datasets [21].

The most important use case for the *Pirá* dataset, however, is connected with reading comprehension. Within this complex task, at least two specific paths can be followed, which we refer to as generative [24] and extractive [33] answering. In the generative type, an end-to-end (i.e., question-to-answer) system searches for relevant passages in a corpus and generates answers from them. This task is the one most naturally adapted to the *Pirá* dataset, given the mostly non-span structure of its answers and the high-quality manual assessments made on it.

The *Pirá* dataset can also be employed for extractive answering, in which given a question and a text, the task is that of finding the corresponding span that contains the answer. Although the answers in *Pirá* dataset are not necessarily spans, the dataset can be approximated in this way, as long as one keeps in mind that results tend to be not as good as in a purely extractive dataset.

Some aspects of the *Pirá* dataset make it particularly challenging for these tasks. The first reason is connected to the scientific nature of its corpora, with questions that can be hard even for people to answer. Furthermore, the fact volunteers had to answer and paraphrase each other's questions contributes to a diversified dataset. This is fundamental in reducing biases derived from annotation artifacts, which are frequent in crowdsourcing efforts [15].

Finally, other possible applications for the *Pirá* dataset are classification tasks (taking into account the keywords for Corpus 1 and book sections for Corpus 2 to establish data labels) and for machine translation fine-tuning.

As for the different versions of the dataset, the goal is to test distinct aspects of models. *Pirá-F* is the filtered version of the dataset, composed of easier QA sets. As for *Pirá-T*, the idea is to foster answer triggering tasks, so that models cannot simply "guess" an answer, but must also consider whether a question can actually be answered. Finally, *Pirá-C* is the complete dataset, which can be filtered according to the researchers' interests.

6.2 Limitations

Contextualization. The main limitation of the *Pirá* dataset certainly refers to issues of generality and lack of contextualization. Even seemingly well-defined questions can carry assumptions that go almost unnoticed; especially due to the lack of clear time frames (e.g., "How much will Petrobras invest for gas production in the state of Bahia [in 2010]?"). This is particularly taxing given that some data in the QA sets are regularly updated: for instance, the number of deep-water oil wells on the Brazilian coast.

Translation. Another difficulty is related to typical problems of translation. Many of the expressions in these specialized texts do not possess widely accepted translations. Several alternatives are found in the dataset to deal with these expressions: keeping them in the original form (which sometimes uses words that have already been incorporated into Portuguese, such as "design", and other
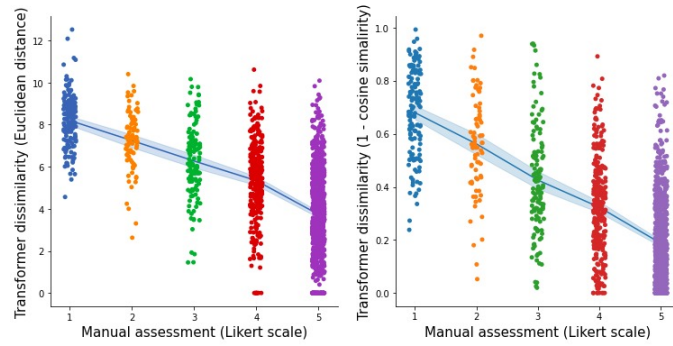
---

[8]We applied "paraphrase-distilroberta-base-v1" checkpoint from Hugging Face, a flavor of Sentence-BERT on top of DistilBERT model, fine-tuned on several paraphrase-related tasks.

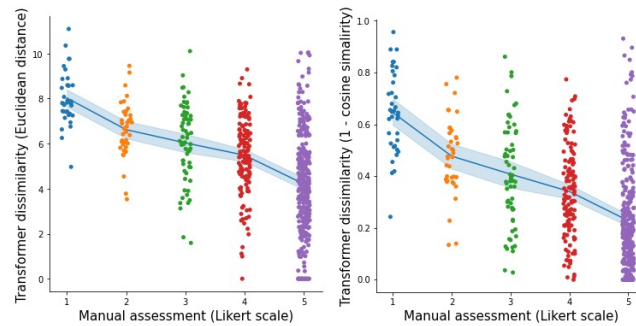**Figure 5: Manual *vs* transformer assessment of sentence equivalence (Corpus 1)**

Figure 6: Manual *vs* transformer assessment of sentence equivalence (Corpus 2)

Table 6: Human *vs* Transformers embeddings semantic similarity

Divergence	Question	Answers
Short/long answer	What is the proportion of women among the seafarers?	**Original answer:** Worldwide approximately two per cent of the seafarers are women. **New answer:** About two per cent.
Lexical/syntactic diversity	What are the risks that nanoplastics and microplastics in the oceans impose to the human health?	**Original answer:** For now, according to the European Food Safety Authority, it is not possible to evaluate the risks. **New answer:** There's not enough scientific evidence to back the claim that micro and nano plastics pose a risk to human health.
Reasoning	How much was the increase in per capita consumption of fish from 2013 to 2016?	**Original answer:** 0.8 kg. **New answer:** Per capita food fish consumption was estimated at 20.3 kg in 2016, compared with 19.5 kg in 2013.

less popular ones), translating them literally (sometimes through adaptations that do not adjust perfectly with Portuguese, such as "stressors" → "*estressores*"), and even translating only part of the expression ("blocky effect" → "*efeito* blocky" or " offshore fields" → "*campos* offshore"). Similarly, stylistic difficulties are encountered: for example, in which cases should the name of a site, a technique or a project be translated? Ideally, the solution should depend on a convention. In the dataset, some volunteers left expressions in English, others preferred to translate them and still others placed the original and its corresponding translation together. The same

happened with Brazilian companies and placenames: should "Marlin Sul" be kept unchanged or be translated to "South Marlin"? In addition, English has much more flexibility to create compounds than Portuguese, and constructions that sound natural in the former may become clumsy in the latter: "How are visual feeding species affected by low-oxygen water areas?" → "*Como as áreas de água com baixo oxigênio afetam as espécies que usam a visão para se alimentar?*".

Linguistic idiosyncrasies. Other difficulties are related to essential differences between Portuguese and English. A very common example derives from the fact that Portuguese has genders

(male and female). Although for most words there is a unique translation, in many cases the common Portuguese speaker finds it difficult to decide which gender to use. Particularly challenging are the nominal phrases that involve elliptical constructions. "Why has Cengroup Petroleum signed a contract with the Azerbaijani government?" → "*Por que A Cengroup Petroleum assinou contrato com o governo do Azerbaijão*?", due to the word "*empresa*" [company – female in Portuguese] or "*Por que O Cengroup Petroleum assinou contrato com o governo do Azerbaijão*?", due to the word "*grupo*" [group – male in Portuguese] or "*petróleo*" [petroleum – male in Portuguese]?

Portuguese-bias. Because QA sets were produced by Portuguese-speaking people, there is a tendency for questions in the dataset to be better formulated in Portuguese than in English. This "Portuguese-bias", although not the ideal scenario, may perhaps counterbalance the preponderance of English datasets or English-adapted datasets.

Structure. Even if correct, several questions are hard to grasp. For instance: "Are Petrobras semisubmersible platforms SS-20 used in connection with aut-leg and conventional mooring systems operating at which water depth?". One reason for this kind of question may be a difficulty in handling complex nominal phrases, leading volunteers to begin questions with these expressions. In addition, many abstracts in Corpus 1 have been written by non-English speakers in academic language, making it even more difficult to understand their specialized content.

Method subdetermination. Despite giving detailed instruction and explaining the activity in online meetings, rules were not sufficient to completely standardize results. For example, although volunteers were supposed to write short answers, the results varied a great deal. Moreover, despite asking volunteers to avoid "yes/no questions", some of them can be found in the dataset. Other questions were simply not addressed by our instructions: for instance, some volunteers wrote questions with several sentences, a possibility that we did not explictly forbid.

6.3 Further steps

Although preliminary tasks have been tried with *Pirá*, we intend to conduct more extensive and systematic explorations. A first step would be to produce robust baselines for three tasks: information retrieval, extractive answering, and generative answering. Those tasks can be then tested with different input-output language combinations as follows:

- Information retrieval: Portuguese (question only); English (question only);
- Extractive answering: Portuguese → English; English → English; Any language → English;
- Generative answering: Portuguese → Portuguese; English → English; Any language → Same language as the input.

We also plan to expand *Pirá* by regularly asking for volunteers, thus including other topics related to the ocean and the Brazilian coast. We would like also to expand the size of the dataset by filtering natural questions related to the ocean topics from the datasets Natural Questions [22] and MS-Marco [31], and to perform experiments with dialogues about the texts of Corpus 2 [4].

Another line of investigation is the comparison of the results in the three tasks mentioned above by considering both languages,

Portuguese and English, so as to establish a baseline for other studies in these languages, and possibly for bilingual tasks in general. We would like to conduct further analyzes on the variations arising from human behavior during the generation and assessment of QAs sets (e.g., sentence length; syntactic complexity, as measured by dependency tree; and topics).

Finally, we hope to examine more in-depth the results on equivalence assessments, in order to improve estimates of semantic similarity for open answers.

7 FINAL REMARKS

In this paper, we presented *Pirá*, a bilingual QA dataset containing 2261 manually generated QA sets about the Brazilian coast and the ocean. Three versions of the dataset are available: *Pirá-F*, containing meaningful and contextualized QA sets; *Pirá-T*, including questions without answers; and *Pirá-C*, which brings the complete dataset. We hope these datasets provide valuable training data for bilingual QA systems; careful manual assessments for questions and answers; and data enrichment based on paraphrases. Also, we presented a replicable method for dataset generation (with freely available code for an open source application). Possible applications of the dataset are: reading comprehension, information retrieval, and machine translation. Finally, we hope that *Pirá* can contribute to AI research as applied to the ocean sciences, climate change, and marine biodiversity.

Availability. All versions of the dataset and the corresponding code are freely available for download and use under the *Creative Commons Attribution 4.0 International license. Pirá* is published as an online resource at Github.[9] Instructions and code for the web application are also available at Github.

Ethical issues. This project fits into Art. 1, VI, of the Resolution Nº 510 of the Brazilian Ministry of Health, which dispenses the necessity of submitting it to an ethical committee. Before engaging in the activity, volunteers signed a consent form stating they were voluntarily participating of the activity and also giving up the rights on the produced material. Furthermore, no information of personal level is included in the dataset.

---

[9]https://github.com/C4AI/Pira

REFERENCES

[1] Sandra Aluísio, Jorge Pelizzoni, Ana Raquel Marchi, Lucélia de Oliveira, Regiana Manenti, and Vanessa Marquiafável. 2003. An account of the challenge of tagging a reference corpus for Brazilian Portuguese. In *International Workshop on Computational Processing of the Portuguese Language.* Springer, 110–117.

[2] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR* abs/1910.11856 (2019). arXiv:1910.11856

[3] Akari Asai, Jungo Kasai, Jonathan H Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2020. XOR QA: Cross-lingual open-retrieval question answering. *arXiv preprint arXiv:2010.11856* (2020).

[4] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036* (2018).

[5] Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Ty pologically Di verse Languages. *Transactions of the Association for Computational Linguistics* 8 (2020), 454–470.

[6] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457* (2018).

[7] Marcelo Criscuolo, Erick Rocha Fonseca, Sandra Maria Aluísio, and Ana Carolina Sperança-Criscuolo. 2017. MilkQA: a Dataset of Consumer Questions for the Task of Answer Selection. In *Proceedings of the 6th Brazilian Conference on Intelligent Systems (BRACIS)* (2-5), Vol. 1. IEEE, Uberlândia, Brazil, 354–359. https://doi.org/10.1109/BRACIS.2017.12

[8] Yiming Cui, Ting Liu, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2018. Dataset for the First Evaluation on Chinese Machine Reading Comprehension. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (Eds.). European Language Resources Association (ELRA), Miyazaki, Japan.

[9] Fábio Ricardo Araújo Da Silva. 2018. Detecção de ironia e sarcasmo em língua portuguesa: Uma abordagem utilizando deep learning. (2018).

[10] Larissa A de Freitas, Aline A Vanin, Denise N Hogetop, Marco N Bochernitsan, and Renata Vieira. 2014. Pathways for irony detection in tweets. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing.* 628–633.

[11] Martin d'Hoffschmidt, Maxime Vidal, Wacim Belblidia, and Tom Brendlé. 2020. FQuAD: French Question Answering Dataset. *CoRR* abs/2002.06071 (2020). arXiv:2002.06071 https://arxiv.org/abs/2002.06071

[12] Magali Sanches Duran and Sandra Maria Aluísio. 2012. Propbank-Br: a Brazilian Treebank annotated with semantic role labels.. In *LREC.* 1862–1867.

[13] Erick Fonseca and João Luís G Rosa. 2013. Mac-morpho revisited: Towards robust part-of-speech tagging. In *Proceedings of the 9th Brazilian symposium in information and human language technology.*

[14] E Fonseca, L Santos, Marcelo Criscuolo, and S Aluisio. 2016. ASSIN: Avaliacao de similaridade semantica e inferencia textual. In *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal.* 13–15.

[15] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324* (2018).

[16] Nathan Hartmann, Gustavo Paetzold, and Sandra Aluísio. 2020. SIMPLEX-PB 2.0: A Reliable Dataset for Lexical Simplification in Brazilian Portuguese. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop.* 18–22.

[17] Nathan S Hartmann, Gustavo H Paetzold, and Sandra M Aluísio. 2018. SIMPLEX-PB: A lexical simplification database and benchmark for Portuguese. In *International Conference on Computational Processing of the Portuguese Language.* Springer, 272–283.

[18] Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. 2017. Dureader: a chinese machine reading comprehension dataset from real-world applications. *arXiv preprint arXiv:1711.05073* (2017).

[19] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. arXiv:1909.06146 [cs.CL]

[20] Tomasz Jurczyk, Michael Zhai, and Jinho D Choi. 2016. Selqa: A new benchmark for selection-based question answering. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI).* IEEE, 820–827.

[21] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906* (2020).

[22] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466.

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

[24] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401 [cs.CL]

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

[26] Gabriel Schubert Marten and Larissa Astrogildo de Freitas. 2021. The Construction of a Corpus for Detecting Irony and Sarcasm in Portuguese. *Brazilian Journal of Development* 7, 5 (2021), 47973–47984.

[27] Timo Möller, Julian Risch, and Malte Pietsch. 2021. GermanQuAD and GermanDPR: Improving Non-English Question Answering and Passage Retrieval. *arXiv preprint arXiv:2104.12741* (2021).

[28] United Nations. 2015. Transforming our world: the 2030 Agenda for Sustainable Development. (2015). https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E

[29] United Nations. 2017. *World Ocean Assessment I.* United Nations publication. https://www.un.org/regularprocess/content/first-world-ocean-assessment

[30] United Nations. 2021. *World Ocean Assessment II.* United Nations publication. https://www.un.org/regularprocess/woa2launch

[31] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@ NIPS.*

[32] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822* (2018).

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

[34] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[35] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8732–8740.

[36] Igor Cataneo Silveira and Denis Deratani Mauá. 2017. University Entrance Exam as a Guiding Test for Artificial Intelligence. In *Proceedings of the 6th Brazilian Conference on Intelligent Systems (BRACIS).* 426–431.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[38] Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Making neural qa as simple as possible but not simpler. *arXiv preprint arXiv:1703.04816* (2017).

[39] Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, Lisbon, Portugal, 2013–2018. https://doi.org/10.18653/v1/D15-1237