IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 February 2023, accepted 16 March 2023, date of publication 27 March 2023, date of current version 5 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3262308

RESEARCH ARTICLE

Slovak Dataset for Multilingual Question Answering

DANIEL HLÁDEK[1], (Member, IEEE), JÁN STAŠ[1], (Member, IEEE), JOZEF JUHÁR[1], (Member, IEEE), AND TOMÁŠ KOCTÚR[2], (Member, IEEE)

[1]Faculty of Electrical Engineering and Informatics, Technical University of Košice, 04200 Košice, Slovakia
[2]Deutsche Telekom IT & Telecommunications Slovakia, 04001 Košice, Slovakia

Corresponding author: Daniel Hládek (daniel.hladek@tuke.sk)

This work was supported in part by the Deutsche Telekom IT & Telecommunications Slovakia by the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and the Slovak Academy of Sciences under Project VEGA 2/0165/21; and in part by the Cultural and Educational Grant Agency of the Slovak Republic under Project KEGA 055TUKE-4/2023, both funded by the Ministry of Education, Science, Research and Sport of the Slovak Republic and the Slovak Research and Development Agency through the Project of Bilateral Cooperation under Grant APVV-SK-TW-21-0002.

ABSTRACT SK-QuAD is the first manually annotated dataset of questions and answers in Slovak. It consists of more than 91k factual questions and answers from various fields. Each question has an answer marked in the corresponding paragraph. It also contains negative examples in the form of ''unanswered questions'' and ''plausible answers''. The dataset is published free of charge for scientific use. We aim to contribute to the creation of Slovak or multilingual systems for generating an answer to a question in a natural language. The paper provides an overview of the existing datasets for question answering. It describes the annotation process and statistically analyzes the created content. The dataset expands the possibilities of training and evaluation of multilingual language models. Experiments show that the dataset achieves state-of-the-art results for Slovak and improves question answering for other languages in zero-shot learning. We compare the effect of machine-translated data with manually annotated. Additional data improve the modeling for low-resourced languages.

INDEX TERMS Crosslingual dataset, monolingual dataset, multilingual dataset, machine translation, neural language model, question answering, Slovak language.

I. INTRODUCTION

The performance of question answering (QA) systems depends on the amount of available annotated language resources and the quality of statistical models. Common reasons for the lack of language resources are funding and weak research and development infrastructure. If we want to create a new dataset for any language, we must consider whether it is more beneficial to focus on manual annotation or perform machine translation of an existing resource into the target language. Manual annotation is complicated, expensive, and takes a long time. In contrast, machine translation is fast and inexpensive, but its disadvantage is lower accuracy. Even if the conditions and finances for creating a new dataset are available, it is necessary to consider overcoming lengthy organizational difficulties, given the required quality.

The infrastructure for the QA systems works fine in English. There exist multiple language resources and fine-tuned language models. The situation is significantly different for other languages, especially for low-resource languages. That is why we decided to create a new resource – the Slovak question answering dataset. Within this task, we focused on the creation of both types of datasets, a manually annotated, as well as a machine-translated dataset. The new datasets can significantly improve the performance of existing systems for machine reading comprehension (MRC) in Slovak, also multilingual QA systems, and will expand the range of available benchmark corpora for evaluating language

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino.

models. We also tried to get as close as possible to the Stanford Question Answering Dataset, version 2.0 (SQuAD v2.0) [1]. The dataset will be publicly available at.[1]

The article is organized as follows. In Section II, we focus on the problem of question answering in general and briefly summarize the existing monolingual and multilingual resources. Also, we take a closer look at existing Slovak language resources. In Section III, we describe in detail the procedure for creating a manually annotated QA dataset for the Slovak language. Furthermore, Section IV contains a detailed analysis of this dataset. In Section V, we introduce the second created QA dataset based on machine translation of the original English SQuAD v2.0 into the Slovak language. Next, we trained different QA models and compared them in two experiments. The experiments show the usefulness of the dataset for monolingual and multilingual question answering. Finally, Section VII summarizes the contribution of our work and concludes the paper with future directions.

II. STATE OF THE ART

Machine understanding is the ability to formulate meaningful answers to questions in natural language using unstructured text. The formal definition of QA according to [2]:

> Typical machine reading comprehension tasks could be formulated as a supervised learning problem. Given a collection of textual training examples $(p_i, q_i, a_i)_i^n = 1$, where p is a passage of text, and q is a question about the text p. The goal of a typical MRC task is to learn a predictor f that takes a passage of text p and a corresponding question q as inputs and gives the answer a as output, which could be formulated as the following formula:
>
> $$a = f(p, q). \quad (1)$$

The answer is defined as filling in the missing word (clause), choosing one of several options, beginning and ending in the text as a probability distribution between the words of the paragraph, or as a continuous sequence of words [3]. Albilali et al. [4] divide machine understanding tasks accordingly.

Generating an answer to a question is one of the tasks for verifying the performance of neural language models. With the advent of pre-trained language models, there is a need for their accurate evaluation. Although there are already several pre-trained neural language models with support for Slovak, a dataset for their verification has not yet been created, which would be compatible with the standard sets SQuAD [1] and GLUE [5]. For example, Brown et al. [6] tested the GPT3 model on multiple tasks, including QA. The neural language model shows good performance even in few-shot tasks, where not enough data is available.

Chen et al. [2] divide the QA system into two parts: the information retrieval module and the machine understanding module. The retrieval module retrieves the set of passages relevant to the question. The understanding module searches for the answer in the found passage. We focus on searching a span with the answer and ignore the paragraph retrieval process, although this is also a research topic [7].

[1] https://huggingface.co/datasets/TUKE-DeutscheTelekom/skquad

TABLE 1. Content of the English SQuAD datasets [1].

Number of	SQuAD v1.1 Train	SQuAD v1.1 Dev	SQuAD v1.1 Test	SQuAD v2.0 Train	SQuAD v2.0 Dev	SQuAD v2.0 Test
Articles	442	48	46	442	35	28
Questions	87,599	10,570	9,533	130,319	11,873	8,862
Unanswerable	0	0	0	43,498	5,945	4,332

A. ENGLISH QUESTION ANSWERING DATASETS

A majority of question answering and machine reading comprehension research focuses on the English language, which offers a wide selection of benchmark datasets. Since the literature on question answering is broad and extensive, we selected several interesting review and research articles that deal with this topic in more detail [3], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17].

B. THE STANFORD QUESTION ANSWERING DATASET

The most used dataset for validating neural language models is the Stanford Question Answering Dataset (SQuAD).

In SQuAD v1.1 [18], the authors created a basic dataset focused on natural language understanding that contains more than 100k hand-annotated questions and answers in the context. The question refers to a specific paragraph. The answer is the area highlighted in the paragraph. Later, the authors extended the dataset with ''unanswerable'' questions [1]. The SQuAD v2.0 contains additional 50k negative examples of questions and answers to help distinguish the relevant paragraphs. One negative example consists of an ''unanswerable'' question, in words resembling an answerable question, but modified so that the correct answer does not exist or is not found in the current paragraph. Each ''unanswerable'' question is also annotated with a ''plausible'' answer.

Annotated questions and answers serve as training examples to automatically mark the answer, even to an unknown question. Unanswerable questions serve as negative examples for training the distinction between relevant and irrelevant paragraphs.

The statistics on the number of articles, questions, and unanswerable questions for both English SQuAD datasets are summarized in Table 1.

C. NON-ENGLISH QUESTION ANSWERING DATASETS

The options for answering questions in non-English languages are limited. Such datasets usually include only one other language, very rarely a low-resource language.

There are several attempts to clone the original English SQuAD for other languages. Such datasets can be divided

into two groups: created automatically using machine translation, or manually annotated based mostly on crowdsourcing.

Automatically translated datasets are easy to create but have the disadvantage of inaccuracy caused by machine translation. The second disadvantage is that they bring new information because they were created by translating the original articles. Manually created datasets do not have these shortcomings. However, human annotation is significantly more expensive and time-consuming. Therefore, such datasets are significantly smaller and only very rarely contain unanswerable questions.

Chandra et al. [19] provide a comprehensive survey on the recent progress of non-English MRC and open-domain QA datasets. The authors reviewed QA datasets in 14 main languages other than English, as well as several multilingual and cross-lingual QA datasets.

A study by Rogers et al. [20] is the largest survey of current monolingual and multilingual QA/RC resources for languages other than English. This survey focuses exclusively on the typology of existing resources. It brings a systematic review of existing resources with respect to a set of criteria: questions versus statements or extractive, multiple choice, categorical, and freeform answers. The authors consider the conversational features of current resources, their domain coverage, and the available language.

The short overview of SQuAD-like monolingual, machine-translated, and multilingual QA datasets for different languages other than English can also be found in the paper [12].

1) MACHINE-TRANSLATED SQuAD DATASETS

We summarized a total of 17 machine-translated datasets in 14 different languages (see Table 2).

Machine translation was performed mainly using the Google Translate API. In rare cases, another tool was used, such as DeepL [21], the TAR method [22], or LINDAT Translator [23]. Therefore, each paragraph is automatically translated with all relevant questions and marked answers.

Machine-translated datasets approach the original SQuAD v2.0 set in scope. In some cases, only a certain part of the SQuAD v1.1 dataset is translated, or v2.0. However, there is often a problem with the translation itself and the alignment of questions and answers after translation at the word level.

Some of the questions are literally translated, some of the translations may be rephrased, and some translations of the questions may result in multiple translation variants, as summarized in [24]. However, a small part of the questions has to be discarded, usually, because some words cannot be translated correctly or it is not clear from the answer which question it belongs to.

To verify whether the translation between the languages has been done correctly, it is approached either by manual control or by automatic reverse translation into English, and a subsequent comparison of the original with this reverse translation. This also gives us feedback on how well a tool was used to translate between languages. Such a reverse translation was used, for example, for the automatic creation of a machine-translated corpus SQuAD-uk for the Ukrainian language [25], or SQuAD-pl for the Polish language [26]. In both cases, there was no additional manual review of the translated questions and answers.

A problem with alignment occurs when the beginnings and ends of the answers to the relevant questions are not correctly marked in a paragraph or if the links between words in the adjacent context are broken due to translation. Therefore, it is necessary to look for the correctly translated answer in the context of the paragraph. This is usually done by:

- using a special tag inserted during translation (breaks the context);
- using an approximate search for a specially translated answer (requires word vectors and may not be exact);
- alignment of the original and translated contexts.

2) MONOLINGUAL QUESTION ANSWERING DATASETS

By July 2022, we have identified a total of 28 human-annotated datasets in 20 different non-English languages (see Table 3). The size of manually annotated datasets ranges from units of thousands to tens of thousands of triplets (paragraph-question-answer). The range closest to the English SQuAD is only three datasets, namely the FQuAD [35], [36], KorQuAD [37], [38] and partly also SberQuAD [39].

The dataset creation methodology was in all cases very similar to the original English SQuAD. The data source is mostly Wikipedia for the respective language. The paragraphs range from 300 to 500 characters. 3 to 5 questions are manually created for the relevant section. Questions are additionally validated and edited either manually or with the help of various NLP tools, where grammatical errors or correct inflection are checked. Questions that are too general or vague, or that require further justification, are usually discarded. The emphasis is also placed on the length of the marked answer in terms of words. Answers that are too short or too long are discarded. In some cases, the source of data is exam questions and the student's answers assigned to them, official documents containing FAQs, various quizzes, etc. The QA datasets for Bulgarian [40], Portuguese [41], Turkish [42], or Kenyan Swahili [43] were created this way.

The manual creation of the questions and marking of the answers to the relevant paragraphs was usually done by crowd-sourcing by trained workers. To create answers to the questions, custom annotation tools were used, such as AddQA [44], PIAFanno [45], SAJAD [46], or already existing web crowd-sourcing platforms such as Amazon Mechanical Turk [27], Toloka AI [39], or Prolific [47].

The need to create new large-scale datasets for other low-resource languages is still relevant. However, the problem is that the only universal data source is currently Wikipedia, which is too general and fact-oriented and may not sufficiently cover certain domains in which QA systems are presently being created, e.g. from the field of medicine. Factual questions also do not give annotators enough room to

TABLE 2. Overview of machine-translated SQuAD datasets.

Language	Title	Translation	Size	v2.0	Best model	Evaluation metric [%]	Reference
Arabic	Arabic-SQuAD	Google Translate	60k	N	BERT	EM=34.10, F1=48.60	Mozannar et al., 2019 [27]
Bengali	Bengali-SQuAD	Google Cloud API	91k	Y	DistilBERT	EM=50.05, F1=51.18	Mayeesha et al., 2021 [12]
Czech	Czech SQuAD	LINDAT Translator	117k	Y	XLM-R large	EM=75.57, F1=79.19	Macková & Straka, 2020 [23]
Danish	SQuAD-da	TAR Method/Moses	156k	Y	-	-	Carrino et al., 2020 [22]
French	SQuAD-fr	Google Translate	100k	N	BERT base	EM=71.70, F1=86.70	Cattan et al., 2021 [28]
Hindi	Hindi SQuAD	Google Translate	18.5k	N	-	EM=50.11, F1=53.77	Gupta et al., 2019 [29]
Italian	SQuAD-it	DeepL	60k	N	BERT base	EM=75.00, F1=82.20	Croce et al., 2018 [21]
Korean	K-QuAD	Google Translate	77k	N	BiDAF	EM=50.72, F1=71.50	Lee et al., 2018 [24]
Persian	ParSQuAD	Google Translate	95k	Y	mBERT	EM=67.73, F1=70.84	Abadani et al., 2021 [30]
Polish	PolAQ	-	14k	N	mBERT	EM=72.56, F1=80.39	Jurkiewicz, 2020 [31]
Polish	SQuAD-pl	Google Translate	-	Y	-	-	Brodzik, 2022 [26]
Portuguese	SQuAD_v1.1_pt	Google Cloud API	100k	N	BERT	Acc=50	Carvalho, 2019 [32]
Portuguese	SQuAD_v2.0_pt	Google Cloud API	-	Y	-	-	Janiake, 2020 [33]
Spanish	SQuAD-es-v1.1	TAR Method/Moses	87.5k	N	mBERT	EM=48.30, F1=68.10	Carrino et al., 2020 [22]
Spanish	SQuAD-es-v2.0	TAR Method/Moses	46k	Y	mBERT	EM=76.50, F1=86.07	Carrino et al., 2020 [22]
Swedish	SQuAD-v2-sv	Google Translate	125k	Y	BERT base	EM=66.73, F1=70.11	Okazawa, 2021 [34]
Ukrainian	SQuAD-uk	Google Cloud API	30k	N	mBERT	EM=56.10, F1=62.20	Tiutiunnyk & Dyomkin, 2019 [25]

create more complex questions. Some answers to questions lack further justification or clarification of the answer. There may be different opinions on some questions that current fact-oriented datasets do not include. Also, most datasets lack Why? questions. Very few QA datasets are oriented toward solving numerical examples or mathematical problems. These are the challenges scientists face currently when designing and creating new MRC and QA datasets.

3) CROSSLINGUAL AND MULTILINGUAL QUESTION ANSWERING DATASETS

From the above overview, it is clear that for some languages there are not enough resources for the preparation of a natural understanding system. Loginova et al. [61] discuss the current state of the art and the remaining challenges in multilingual QA.

In general, there exist three main approaches to solving multilingual question answering and transferring knowledge from English to other languages:

- zero-shot/few-shot;
- translate-train;
- translate-test.

Zero-shot and few-shot approaches use multilingual language models trained on a corpus of data available in different languages. In the few-shot approach, the multilingual model is subsequently fine-tuned for the task in the target language. In addition, the validation data are from the target language.

On the contrary, the translate-train and translate-test approaches use machine translation. Machine translation can be used at several points in the machine learning chain. But machine translation creates additional inaccuracy and bias. In the case of the translate-train approach, the training set is translated from English to the target language, and then language models are created for the target languages. Test examples are from the target language. The translate-test approach involves the translation of test examples (questions and answers) from the source language into English, and the evaluation takes place with the English model.

Several multilingual datasets have been published in the last five years. Table 4 compares the most significant publicly available multilingual question answering datasets. The table highlights the number of languages and total examples provided for each dataset.

D. SLOVAK LANGUAGE RESOURCES

The Slovak language belongs to the group of west Slavic languages and uses Latin script. Language is characterized by a free-word order in a sentence, a large number of morphological forms, and grammar with a number of exceptions. Words are formed by attaching a prefix or suffix depending on their grammatical or meaning function. The number of speakers (including the diaspora) is less than 10 million.

Slovak has a lack of language resources, but the situation has improved in recent years. Among basic language sources, WordNet [69], Slovak Dependency Treebank [70], (dependency corpus with marked morphological markers and lemmas) and the Aranea Web Corpus [71], are available. Commonly available multilingual language models, such as multilingual BERT [72], also include support for Slovak. Colleagues from Gerulata and the KInIT institute have created a monolingual SlovakBERT model [73], which achieves state-of-the-art results for the Slovak language in both part-of-speech tagging and text categorization tasks. However, Slovak is still one of the languages where few language resources are available.

The existing Slovak language corpora do not yet contain any resources for the creation of QA systems. The only precursor comes from our previous research. We created the Slovak Categorized News Corpus (SCNC) dataset [74] to evaluate the results of the information retrieval. The SCNC dataset contained a set of natural language questions and corresponding newspaper articles that contained the answer.

TABLE 3. Overview of crowdsourced monolingual QA datasets.

Language	Title	Data source	Size	Best model	Evaluation metric [%]	Reference
Arabic	ARCD	Wikipedia	1.4k	BERT	EM=19.60, F1=51.30	Mozannar et al., 2019 [27]
Bulgarian	BG_RC-v1.0	exams, quizzes	2.6k	-	-	Hardalov et al., 2019 [40]
Catalan	ViquiQuAD	Wikipedia	15.2k	-	-	R.-Penagos & A.-Oller, 2021a [48]
Catalan	VilaQuAD	newspapers	6.2k	-	-	R.-Penagos & A.-Oller, 2021b [49]
Czech	Czech SQAD v3.0	Wikipedia	13.5k	-	-	Sabol et al., 2019 [44]
French	FQuAD 1.0	Wikipedia	26k	-	-	d'Hoffschmidt et al., 2020 [35]
French	FQuAD 1.1	Wikipedia	62k	CamemBERT large	EM=82.40, F1=91.80	d'Hoffschmidt et al., 2020 [35]
French	FQuAD 2.0	Wikipedia	80k	CamemBERT large	EM=78.00, F1=83.00	Heinrich et al., 2022 [36]
French	PIAFv1.0	Wikipedia	3.8k	CamemBERT	-	Keraron et al., 2020 [45]
German	GermanQuAD	Wikipedia	13.7k	GELECTRA large	EM=68.60, F1=88,10	Moeller et al., 2021 [50]
Hebrew	ParaShoot	Wikipedia	3k	mBERT	EM=32.00, F1=56,10	Keren et al., 2021 [47]
Chinese	DRCD	Wikipedia	30k	BERT	EM=82.34, F1=89.59	Shao et al., 2018 [51]
Chinese	CMRC	Wikipedia	20k	BERT	EM=67.80, F1=86.00	Cui et al., 2019 [52]
Icelandic	NQiI	Wikipedia	5.6k	IceBERT	EM=51.00, F1=73.00	Snaebjarnnarson, 2021 [53]
Japanese	JaQuAD	Wikipedia	39.7k	BERT	EM=63.38, F1=78.92	So et al., 2022 [54]
Korean	KorQuAD 1.0	Wikipedia	70k	BERT	EM=71.68, F1=89.76	Lim et al., 2019 [37]
Korean	KorQuAD 2.0	Wikipedia	100k	BERT	EM=30.24, F1=45.96	Kim et al., 2020 [38]
Persian	PersianQuAD	Wikipedia	20k	mBERT	EM=78.80, F1=82.97	Kazemi et al., 2022 [46]
Persian	PQuAD	Wikipedia	80k	XLM-RoBERTa	EM=74.80, F1=87.60	Darvishi et al., 2022 [55]
Portuguese	FaQuAD	Wikipedia	0.9k	BiDAF + ELMo	EM=24.50, F1=43.90	Sayama et al., 2019 [41]
Russian	SberQuAD	Wikipedia	50k	BERT	EM=66.60, F1=84.80	Efimov et al., 2020 [39]
Spanish	SQAC	Wikipedia, AnCora	18.8k	RoBERTa large	F1=82.02	G.-Fandino et al., 2022 [56]
Swahili	KenSwQuAD	Kenya lang. corpus	7.5k	-	-	Wanjawa et al., 2022 [43]
Thai	iapp-wiki-qa-squad	Wikipedia	7.2k	XLM-R large	EM=69.51, F1=75.73	Rueangkhajorn & Chan, 2021 [57]
Thai	thaiqa-SQuAD	Wikipedia	4.1k	-	-	PyThaiNLP / NECTEC, 2020 [58]
Turkish	THQuAD	Wikipedia/Ottoman	15.5k	ELECTRA	EM=63.08, F1=81.55	Soygazi et al., 2021 [42]
Vietnamese	UIT-ViQuAD	Wikipedia	23k	XLM-R large	EM=69.18, F1=87.14	Nguyen et al., 2020 [59]
Ukrainian	UA-SQuAD	newspapers	13.9k	-	-	Ivanyuk-Skulsky, et al., 2021 [60]

TABLE 4. Overview of crosslingual and multilingual QA datasets.

Title	#	Languages	Data source	Size	Reference
XQA	9	English, Chinese, French, German, Polish, Portuguese, Russian, Tamil, Ukrainian	Wikipedia	90k	Liu, et al., 2019 [62]
XQuAD	12	English, Arabic, Chinese, German, Greek, Hindi, Romanian, Russian, Spanish, Thai, Turkish, Vietnamese	SQuAD v1.1	13k	Artetxe et al., 2019 [63]
MLQA	7	English, Arabic, German, Hindi, Simplified Chinese, Spanish, Vietnamese	Wikipedia	42k	Lewis et al., 2020 [64]
TyDi QA	11	English, Arabic, Bengali, Finnish, Indonesian, Japanese, Kiswahili, Korean, Russian, Telugu, Thai	Wikipedia	204k	Clark et al., 2020 [65]
XOR-TyDi QA	7	Arabic, Bengali, Finnish, Japanese, Korean, Russian, Telugu	Wikipedia	40k	Asai, et al., 2021 [66]
MKQA	26	English, Arabic, Danish, Dutch, Finnish, French, German, Hebrew, Hungarian, Chinese (Hong Kong, Simplified, Traditional), Italian, Japanese, Khmer, Korean, Malay, Norvegian, Polish, Portuguese, Russian, Spanish, Swedish, Thai, Turkish, Vietnamese	Natural Questions [68]	260k	Longpre et al., 2021 [67]

III. ANNOTATION OF THE DATASET

The goal when designing the dataset was to get to the SQuAD 2.0 data set as closely as possible. Constrained resources did not allow us to reach the size of the original database. The second limitation was the scope of the Slovak section of Wikipedia. The English SQuAD contains questions from the top 500 Wikipedia articles. This approach could not be used for the Slovak because the longest articles do not have enough paragraphs. We had to annotate more articles. Therefore, the SK-QuAD thematically covers almost the entire Slovak Wikipedia.

We used the Prodigy annotation tool [75] to annotate the questions and answers. One annotation task corresponds to one web application deployment and different configurations. We used PostgreSQL to store the results. An additional Flask application helped us to track the progress of the annotations. The proposed crowdsourcing system is depicted in Figure 1.

We divided the prepared text from Wikipedia into several tasks for annotators:

1) annotation of questions and answers;
2) questions and answers validation;
3) annotation of unanswerable questions.

More than 150 volunteers and 9 part-timers annotated the questions and answers. After that, five paid workers checked and corrected them. Some validated items were converted

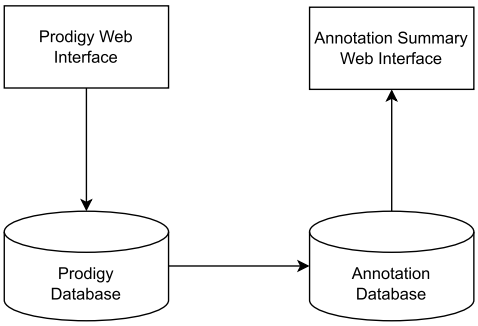

FIGURE 1. Components of the proposed crowdsourcing system.

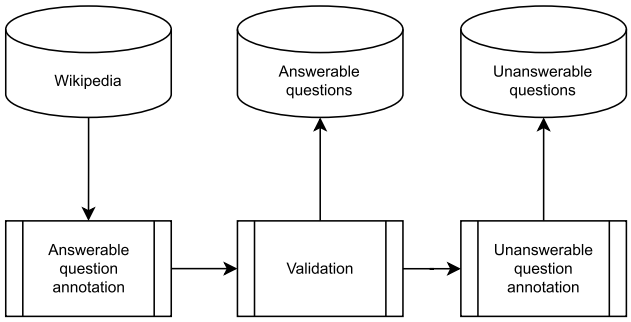

FIGURE 2. Flow of the annotations.

into unanswerable questions and plausible answers by two workers. The flow of the annotation is shown in Figure 2.

A. PREPARATION OF THE TEXT

The input was a compressed dump of the Slovak Wikipedia. The text preprocessing consisted of the following steps:

1) *Parsing the dump*. We parsed the Wikipedia dump and identified individual articles. For each article, we identified its title and text.
2) *Parsing articles*. We have removed the tags so that the resulting text is as clean as possible. We prepared the sections in a form suitable for annotation. We divided each article into paragraphs. We have attached information about the title of the article and its serial number to each paragraph.
3) *Removal of inappropriate articles*. We removed articles with the results of sports events, e.g. World Cups, and articles that do not contain usable text for annotation. Moreover, we removed the following paragraphs from the articles:
   - shorter than 500 characters;
   - containing mathematics;
   - containing bullet points;
   - badly parsed.
4) *Division of the dataset into annotation batches*. During the preparation, we divided the dataset into 100 parts. Each $n-th$ part contains every hundredth article with a shift of n. Each section selected in this way covers the entire Wikipedia evenly. We annotated the selections gradually. Therefore, the SK-QuAD dataset thematically covers almost the entire Slovak Wikipedia.
5) *Transformation of the batch into a form suitable for annotation*. The paragraph was converted into Prodigy format and passed into the question-and-answer annotation.

Table 5 summarizes statistics about the extracted text, suitable for annotation.

TABLE 5. Statistics on Slovak Wikipedia.

Dump date:	1.3.2020
Documents:	208,969
Paragraphs	1,453,878
Paragraphs suitable for annotation	113,658

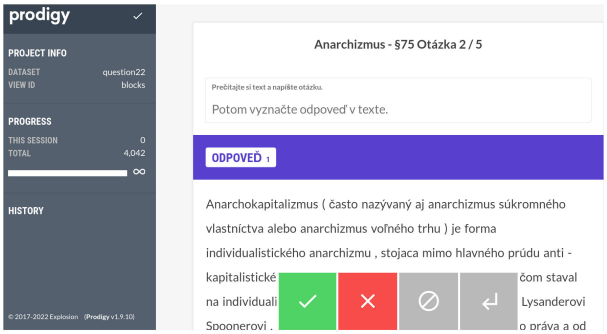

FIGURE 3. Annotation of questions and answers using the Prodigy tool.

B. ANNOTATION OF ANSWERABLE QUESTIONS AND ANSWERS

In the first step, annotators annotated questions and answers in Wikipedia paragraphs longer than 500 characters. The annotator first read the paragraph, wrote the question in the text box, and marked the answer. The same item was shown five times, so the annotator had the opportunity to write a maximum of five questions and answers for each paragraph. The question had to be about the fact that is present in the context. The question had to be short and unambiguous.

The question had to be grammatically correct. This requirement could cause some questions and answers to sound unnatural because the annotators had to respect the morphological form in the context and adapt the question accordingly. This problem does not occur in English, because it has fewer morphological forms, and therefore it is easier to write a grammatically correct question for the highlighted words.

We instructed the annotators to ignore paragraphs that contained too few facts or were poorly formatted. The answer had to exist in the text. The Prodigy interface displays the title of the article and a link to Wikipedia for each paragraph to increase the motivation of the annotators because they could learn something new.

TABLE 6. Correction statistics on SK-QUAD.

	Original count	Corrected
questions	77,808	21,129
question words	585,224	64,287
question letters	3,789,863	293,891
answers	77,808	19,602
answer words	793,664	194,340
answer letters	5,309,878	n/a

C. QUESTIONS AND ANSWERS VALIDATION

After marking the questions and answers to the paragraphs, we automatically validated the highlighted questions and answers. We eliminated any unit where the question or answer was too short or duplicated for that paragraph. In the text, we have marked the words that contain misspellings. We used the Hunspell library with a Slovak dictionary [76].

We put the auto-validated questions and answers into the Prodigy web interface for manual review. Annotators corrected typos, shortened excessively long questions and answers, and edited the questions in the grammatically correct form. If a subject or object was missing from the question, the annotators had to fill it in.

The total number of annotated questions and answers was 82,356 items, but 4,548 were discarded because it was impossible to correct them during validation. The remaining 77,808 questions and answers undergo corrections. Table 6 shows the number of input items and the number of corrected items processed by the annotators. The corrected items are calculated as the Levenshtein distance between the original and corrected versions. However, it was not possible to correct the letters in the answers because the answer is a span in the existing text.

D. ANNOTATION OF UNANSWERABLE QUESTIONS

In the third step, we annotated unanswerable questions. This step results in a question and answer that look correct but are not. The correct answer to the question should not be in the paragraph. These questions serve as negative examples. Thanks to them, the neural network can distinguish the relevance of the proposed span from the answer.

Unanswerable questions were created by annotating the validated answerable questions from the previous step. Annotator was shown a question with the answer highlighted in the paragraph. He modified the correct question into an unanswerable one and had to highlight a plausible answer to the unanswerable question. More details about unanswerable questions and plausible answers are explained in the original SQuAD v2.0 paper [1]. It should be noted that unanswerable questions did not undergo further validation because they were validated in the previous step.

IV. ANALYSIS OF ANNOTATIONS

The manual annotations were exported into the SQuAD v2.0 format. We have ensured that there is exactly one answer to the questions in the training set. However, this may not be true for the test set.

TABLE 7. Statistics on the SK-QuAD dataset.

	SK-QuAD			SQuAD-sk	SQuAD v2.0
Number of	Train	Dev	Total	Train	Train
Documents	8,377	940	9,317	442	442
Paragraphs	22,062	2,568	24,630	18,931	19,035
Questions	81,582	9,583	91,165	120,239	130,319
Answers	65,839	7,822	73,661	79,978	86,821
Unanswerable	15,877	1,784	17,661	40,261	43,498

TABLE 8. Interrogative pronoun word forms, including inflections and prepositions.

Word form	Translation	Occurrence [%]
aký, ktorý	which	32.76
čo, v čom	what	20.84
kedy	when	10.69
kde, kam, odkiaľ	where	8.11
ako	how	7.15
kto	who	6.14
koľko, koľkokrát	how many	5.62
koho, komu	whom	3.52
prečo	why	1.09
je	is	0.15
iné	others	3.86

The size of the SK-QuAD dataset is summarized in Table 7. The table shows that the Slovak dataset has a slightly lower number of questions and covers a much larger number of articles when compared to the original English SQuAD v2.0.

A. PRONOUN FORM ANALYSIS

In the first step, we investigated the most commonly used interrogative pronouns. This analysis should reveal a bias in the morphological forms of the questions.

The interrogative pronoun indicates what will be the object of the question. In Slovak, the interrogative pronoun usually comes first. Therefore, we have found a list of the most frequent words in the first place of the sentence. The interrogative pronoun is often bound to a preposition. The list of the most common prepositions was used to identify the preposition bound to the interrogative pronoun. Our algorithm also joined multiple inflections of the same pronoun. The identified pronouns were grouped to match their English translation.

Table 8 can be compared with the other SQuAD clones.

B. ANALYSIS OF ANSWER TYPES

Next, we tried to approximate Table 2 from the paper [18]. According to the parts of speech and named entities, we classified the question into one of 9 categories:

- AP: adjective phrase;
- NP: noun phrase;
- VP: verb phrase;
- PER: person;
- LOC: location;

TABLE 9. Analysis of answer types.

Type	Ocurrence [%]
NP	29.62
NUM	19.84
LOC	16.77
PER	10.75
ORG	9.77
VP	6.53
UNK	3.38
AP	2.78
DATE	0.52

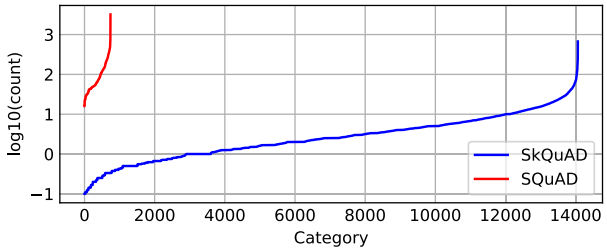

FIGURE 4. Coverage by questions for each Wikipedia category in logarithmic scale.

- ORG: other entity;
- NUM: general quantity;
- DATE: date or time;
- UNK: unrecognized type.

We use a similar algorithm. We separated the numerical (NUM and DATE) and non-numerical answers first. Then, our proposed Slovak spaCy model identified part-of-speech tags and named entities for non-numerical answers. The frequency of parts of speech determined the type of clause (AP – an adjective phrase, NP – noun phrase, VP – verb phrase). If a named entity was present in a noun phrase, it was redefined as PER – person, LOC – location, or ORG – other entity. The rest of the answers were marked UNK. Table 9 summarizes the resulting frequencies of the answer types.

C. TOPIC ANALYSIS

The original English SQuAD dataset annotated around 500 longest articles and covered only a restricted set of topics. We analyze the topics in our database and compare them with the original. The article category is a useful feature because it describes the topic and is available for each article. We obtained the categories for each article in SK-QuAD and SQuAD v2.0 from the Wikipedia API. We calculated co-occurrences of a category of the context and question. One question contributed one point to each category in its context.

The plot in Figure 4 shows the resulting score for each category on a logarithmic scale. The Slovak database contains 14,063 categories, and the English database contains only 741 categories. The plot confirms that our database is thematically broader.

V. SLOVAK TRANSLATION OF THE SQuAD v2.0

We supplemented the human-annotated part with a machine translation of the SQuAD v2.0 dataset. We used the freely available and fast Marian neural MT framework [77] with the Helsinki NLP Opus English-Slovak model [78]. The authors declare that the BLEU score of this model is 36.8.

For machine translation, we used the following procedure:

1) conversion of SQuAD from JSON to paragraph form – the questions and answers have been turned into separate paragraphs;
2) machine translation of paragraphs – for each paragraph, question, and answer, we obtained its translation into Slovak;
3) finding the translated answer in the translated paragraph;
4) validation of the resulting questions and answers – we eliminated answers not present in the paragraph.

Searching for the answer in the translated paragraph can be complicated because the Slovak language is inflectional and has a free-word order. The answer in the context will probably use different word forms than the translated answer. Thus, an approximate search is needed to identify the most semantically similar part in the context. Therefore, we chose the search using word vectors. We tried these steps in sequence:

1) exact search for the original answer;
2) exact search for the translated answer;
3) approximate search for the translated answer.

We searched in context using a floating window of the same size as the translated answer. We calculated the mean word vector A for a floating window and the mean word vector for the translated answer B. Next, we calculated the cosine similarity between the vectors A and B and searched for the window with the maximal similarity to the translated answer. We ignore parts with a similarity lower than a threshold to limit spurious results.

We used the spaCy library [79] to facilitate approximate text searches. We created our word vector model using FastText [80] with Floret vectors [81]. For modeling, a web corpus from our previous research [82], [83] was used. Our Slovak spaCy model is freely available at [84].

A visual inspection of the translated dataset revealed that sometimes not all words were found or the answer does not fit the question grammatically. This type of error results from the differences between the Slovak and English languages. More research is required to identify and eliminate grammatically incorrect answers.

On the other hand, in the vast majority of cases, the result was understandable and grammatically correct. The approximate search fails infrequently. The machine-translated part (SQuAD-sk) is of similar size to the original.

This translation procedure applies to similar languages with a trained word vector and a machine translation model. It is independent of the machine translation method – any other translation model can be used.

VI. EXPERIMENTS

We conducted two experiments. They could help us to evaluate the usefulness of the dataset for monolingual and multilingual question answering.

A. MONOLINGUAL QUESTION ANSWERING EXPERIMENT

The first experiment shows that the model, fine-tuned on the new dataset, achieves a new state-of-the-art for Slovak. The previous methods to train an QA system before our corpus, such as zero-shot classification, translate-train, or translate-test, did not use any annotated data.

We train neural models for QA with multiple sets in the Slovak language. We use the following pre-trained language models with Slovak support:

- monolingual SlovakBERT, base type, 110 million parameters – same as RoBERTa-base;
- multilingual BERT, base type, 110 million parameters.

We fine-tuned both models on the SQuAD v2.0 dataset, the Slovak crowdsourced SK-QuAD dataset, the Slovak translated SQuAD v2.0 dataset (SQuAD-sk), and on the combined dataset, which also contains both manually annotated and machine-translated data together. The purpose of the combined dataset is to verify the impact of additional data on the accuracy. The sizes of the data sets are summarized in Table 7.

During training, we used the Hugging Face Transformers toolkit with the following hyper-parameters:

```
learning_rate = 3e-5, epochs = 3,
max_seq_length = 512, stride = 128,
null_score_diff_threshold = 0.5.
```

We used exact match (EM) and F1 score to evaluate the BERT models. EM measures the percentage of predictions that match any reference answers at the token level. The F1 score is the harmonic mean of precision and recall. Precision measures the ratio of correct tokens in the prediction. Recall rates the ratio of the correct tokens in the prediction to the correct response. If a question has multiple reference answers, the highest F1 score is taken. The average F1 of all predictions is the final F1 of the system.

For verification, we trained the English QA mBERT model on the original SQuAD v2.0 with the same hyperparameters. We achieved EM=74.67 and F1=78.11.

Table 10 summarizes the results of the experiment. We found that, as expected, the manually annotated dataset has the most significant benefit. The accuracy of the Slovak model is comparable to the English one. The performance of the monolingual model is consistently better than the performance of the multilingual model. The machine-translated dataset is not significantly better than zero-shot training on the English training data. However, the performance of the Slovak QA system was quite acceptable, even without manually annotated data.

Different morphological features of the Slovak and English languages cause the lower EM of the system trained on machine-translated SQuAD-sk. It is possible that the grammatical form of the question does not fit the answer because the script translates it together with the context. Automatic elimination of these questions or manual correction could bring improvement.

TABLE 10. Results of monolingual question answering.

Train set	SlovakBERT EM	SlovakBERT F1	mBERT EM	mBERT F1
English SQuAD – zero-shot	28.56	58.28	28.66	52.76
SQuAD-sk – machine-translated	29.03	56.74	28.18	55.35
SK-QuAD – crowdsourced	52.91	74.42	50.23	72.57
SK-QuAD & SQuAD-sk together	52.95	74.49	50.67	72.78

B. MULTILINGUAL QUESTION ANSWERING EXPERIMENT

In the second experiment, we want to show that the language data that we created improve multilingual QA. We use a zero-shot approach that does not require machine translation. We assume that if we use a multilingual model and train it for the QA task, the system will be able to work with a language that is not included in the QA training set to some extent. This feature was confirmed in the previous experiment and in previous research [85].

The multilingual language model is trained on data from Wikipedia. Individual languages are represented in the multilingual BERT (mBERT) training set according to the number of articles on Wikipedia.

For comparison, the number of articles in individual language mutations is as follows: Slovak (SK) – 241,870; English (EN) – 6,096,182; Russian (RU) – 1,633,311; Korean (KO) – 603,748; and Total 59,330,413. The English language is best represented in the language model. We assume that there will be better results for the English language, too.

As a verification task, we selected XQuAD [63], the manual translation of SQuAD v1.1 into 11 other languages. We use KorQuAD [38], SberQuAD [39], and SK-QuAD as a training set. We did not include the English SQuAD, because the test set is its translation and the results would not be relevant. We also used a large concatenated set with data from all three of these sets. That way we will know what effect a large amount of data has on accuracy.

We used the Hugging Face Transformers library and the same mBERT base model as in the previous experiment. The evaluation method was SQuAD v1.0, F1 and EM metrics.

As we can see in Table 11, the average accuracy of F1 is, as expected, the best for the largest "joined" training set. It shows that the new training data improve the accuracy of the zero-shot approach to multilingual QA. The concatenated training set achieves the best results for all languages, except for Spanish and Romanian, where the Russian training set is the best. This means that the new Slovak language resource improves the accuracy of multilingual QA in a zero-shot way.

For separate models, the Russian training set achieves the best results. It is consistent with the fact that the Russian part of the mBERT model training set is larger than the Korean and Slovak parts. The topic of language bias in multilingual models is addressed by [86]. The paper evaluates

TABLE 11. Results of multilingual question answering.

Test/ Train	KorQuAD EM	KorQuAD F1	SberQuAD EM	SberQuAD F1	SK-QuAD EM	SK-QuAD F1	Joined EM	Joined F1
ar	30.50	44.68	30.50	54.52	19.91	45.21	32.01	54.25
de	42.10	55.03	35.96	60.60	25.54	51.91	38.73	63.50
el	34.45	46.13	27.22	55.31	20.92	50.30	28.90	58.44
en	52.10	64.98	43.78	68.42	31.26	56.37	46.21	70.86
es	41.34	57.00	32.01	63.89	25.21	53.51	34.03	63.77
hi	39.66	53.41	29.83	48.77	13.78	35.15	41.42	55.63
ro	39.91	49.97	35.71	63.00	27.64	56.11	31.76	61.27
ru	37.81	51.56	37.89	63.55	28.82	57.33	40.92	66.27
th	30.16	35.64	25.63	38.94	18.48	34.31	32.43	42.67
tr	32.43	50.69	27.47	46.28	12.68	32.18	32.68	51.91
vi	37.47	56.38	35.96	62.18	19.74	45.61	40.16	65.54
zh	47.73	56.70	36.38	53.73	18.73	44.65	46.63	60.30
Average	38.81	51.85	33.20	56.60	21.90	46.89	37.16	59.54

the representation quality of mBERT on 99 languages for named entity recognition and 54 for part-of-speech tagging and dependency parsing. They conclude that while mBERT covers 104 languages, the 30% languages with the least pre-training resources perform worse than using no pre-trained language model at all. The Slovak language is not one of them.

VII. CONCLUSION

SK-QuAD is the first QA dataset for the Slovak language. The manual annotations do not have distortion caused by machine translation. The dataset is thematically diverse; it does not overlap with the original SQuAD, but it brings new knowledge. Each question and the answer were checked by at least two annotators. This dataset will mainly contribute to the creation of new systems for generating an answer to a question in natural language. It enables both training and verification of such a system. The dataset is published free of charge for scientific use.

A new language resource sets the stage for research in natural language processing. This data set will enable a mutual comparison of current and future Slovak and multilingual neuron language models. In addition to the Slovak language, this new dataset is beneficial for natural language processing in general. The fact that the dataset covers other parts of Wikipedia brings new possibilities for exploring multilingualism and crosslinguality.

ACKNOWLEDGMENT

The authors thank all annotators, part-timers, and students of the Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Košice, also the students of the Elementary School Belehradská 21, Košice, under the supervision of Dr. Lenka Macková, Ph.D., and their friends, for their help in contributing to the SK-QuAD dataset. Their special thanks go to Deutsche Telekom IT & Telecommunications Slovakia for fruitful cooperation and personal and financial support.

REFERENCES

[1] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2018, pp. 784–789.

[2] D. Chen. (2018). *Neural Reading Comprehension and Beyond*. [Online]. Available: https://search.proquest.com/openview/702f3833c6d25ddead0d89e3ec8f2299/1?pq-origsite=gscholar&cbl=18750&diss=y and https://stacks.stanford.edu/file/druid:gd576xb1833/thesis-augmented.pdf

[3] C. Zeng, S. Li, Q. Li, J. Hu, and J. Hu, "A survey on machine reading comprehension—Tasks, evaluation metrics and benchmark datasets," *Appl. Sci.*, vol. 10, no. 21, p. 7640, Oct. 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/21/7640

[4] E. Albilali, N. Al-Twairesh, and M. Hosny, "Constructing Arabic reading comprehension datasets: Arabic WikiReading and KaifLematha," *Lang. Resour. Eval.*, vol. 56, pp. 1–36, Mar. 2022, doi: 10.1007/s10579-022-09577-5.

[5] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. EMNLP Workshop BlackboxNLP: Analyzing Interpreting Neural Netw. NLP*, 2018, pp. 353–355.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, and S. Agarwal, "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2020, pp. 1877–1901.

[7] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-T. Yih, "Dense passage retrieval for open-domain question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 6769–6781.

[8] S. Liu, X. Zhang, S. Zhang, H. Wang, and W. Zhang, "Neural machine reading comprehension: Methods and trends," *Appl. Sci.*, vol. 9, no. 18, p. 3698, Sep. 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/18/3698

[9] X. Zhang, A. Yang, S. Li, and Y. Wang, "Machine reading comprehension: A literature review," 2019, *arXiv:1907.01686*.

[10] B. B. Cambazoglu, M. Sanderson, F. Scholer, and B. Croft, "A review of public datasets in question answering research," *ACM SIGIR Forum*, vol. 54, no. 2, pp. 1–23, Aug. 2021, doi: 10.1145/3483382.3483389.

[11] D. Dzendzik, J. Foster, and C. Vogel, "English machine reading comprehension datasets: A survey," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 8784–8804. [Online]. Available: https://aclanthology.org/2021.emnlp-main.693

[12] T. T. Mayeesha, A. Md Sarwar, and R. M. Rahman, "Deep learning based question answering system in Bengali," *J. Inf. Telecommun.*, vol. 5, no. 2, pp. 145–178, 2021.

[13] H. A. Pandya and B. S. Bhatt, "Question answering survey: Directions, challenges, datasets, evaluation matrices," 2021, *arXiv:2112.03572*.

[14] F. Zhu, W. Lei, C. Wang, J. Zheng, S. Poria, and T.-S. Chua, "Retrieving and reading: A comprehensive survey on open-domain question answering," 2021, *arXiv:2101.00774*.

[15] Y. Bai and D. Z. Wang, "More than reading comprehension: A survey on datasets and metrics of textual question answering," 2021, *arXiv:2109.12264*.

[16] R. Baradaran, R. Ghiasi, and H. Amirkhani, "A survey on machine reading comprehension systems," 2020, *arXiv:2001.01582*.

[17] Z. Wang, "Modern question answering datasets and benchmarks: A survey," 2022, *arXiv:2206.15030*.

[18] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," in *Proc. EMNLP*. Austin, TX, USA: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. [Online]. Available: https://aclanthology.org/D16-1264

[19] A. Chandra, A. Fahrizain, Ibrahim, and S. W. Laufried, "A survey on non-English question answering dataset," 2021, *arXiv:2112.13634*.

[20] A. Rogers, M. Gardner, and I. Augenstein, "QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension," 2021, *arXiv:2107.12708*.

[21] D. Croce, A. Zelenanska, and R. Basili, "Neural learning for question answering in Italian," in *Advances in Artificial Intelligence* (Lecture Notes in Computer Science), vol. 11298. New York, NY, USA: Springer, 2018, pp. 389–402, doi: 10.1007/978-3-030-03840-3_29.

[22] C. P. Carrino, M. R. Costa-jussà, and J. A. R. Fonollosa, "Automatic Spanish translation of SQuAD dataset for multi-lingual question answering," in *Proc. 12th Lang. Resour. Eval. Conf.* Marseille, France: European Language Resources Association, May 2020, pp. 5515–5523. [Online]. Available: https://aclanthology.org/2020.lrec-1.677

[23] K. Macková and M. Straka, "Reading comprehension in Czech via machine translation and cross-lingual transfer," in *Text, Speech, Dialogue*, P. Sojka, I. Kopeček, K. Pala, and A. Horák, Eds. Cham, Switzerland: Springer, 2020, pp. 171–179.

[24] K. Lee, K. Yoon, S. Park, and S.-W. Hwang, "Semi-supervised training data generation for multilingual question answering," in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*. Miyazaki, Japan: European Language Resources Association, May 2018, pp. 1–5. [Online]. Available: https://aclanthology.org/L18-1437

[25] S. Tiutiunnyk and V. Dyomkin, "Context-based question-answering system for the Ukrainian language," in *Proc. 1st Masters Symp. Adv. Data Mining, Mach. Learn., Comput. Vis. (MS-AMLV)*, Lviv, Ukraine, 2019, pp. 81—88.

[26] A. Brodzik. (2022). *SQuAD-PL—Polish Google Translated SQuAD v2.0 Dataset*," [Online]. Available: https://github.com/brodzik/SQuAD-PL

[27] H. Mozannar, E. Maamary, K. El Hajal, and H. Hajj, "Neural Arabic question answering," in *Proc. 4th Arabic Natural Lang. Process. Workshop*, 2019, pp. 108–118.

[28] O. Cattan, C. Servan, and S. Rosset, "On the usability of transformers-based models for a French question-answering task," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process. (RANLP)*, 2021, pp. 244–255, [Online]. Available: https://hal.archives-ouvertes.fr/hal-03336060/

[29] D. Gupta, A. Ekbal, and P. Bhattacharyya, "A deep neural network framework for English Hindi question answering," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 19, no. 2, pp. 1–22, Nov. 2019.

[30] N. Abadani, J. Mozafari, A. Fatemi, M. Nematbakhsh, and A. Kazemi, "ParSQuAD: Persian question answering dataset based on machine translation of SQuAD 2.0," *Int. J. Web Res.*, vol. 4, no. 1, pp. 34–46, 2021. [Online]. Available: https://ijwr.usc.ac.ir/article_139661_cbdaea9e77d05933acff25309f4ead9f.pdf

[31] T. Jurkiewicz. (2020). *PolAQ—Polish Answers and Questions*. [Online]. Available: https://github.com/tjur/polaq_master_thesis

[32] N. R. Carvalho. (2019). *Question-Answering Model Fine-Tuned for Portuguese*. [Online]. Available: https://medium.com/nunorc/question-answering-model-fine-tuned-for-portuguese-801bb05b6119

[33] C. Janiake. (2020). *Portuguese Translation of SQuAD 2.0 Dataset*. [Online]. Available: https://github.com/cjaniake/squad_v2.0_pt

[34] S. Okazawa. (2021). *Swedish Version of SQuAD 2.0*. [Online]. Available: https://github.com/susumu2357/SQuAD_v2_sv

[35] M. D'Hoffschmidt, W. Belblidia, T. Brendlé, Q. Heinrich, and M. Vidal, "FQuAD: French question answering dataset," 2020, *arXiv:2002.06071*.

[36] Q. Heinrich, G. Viaud, and W. Belblidia, "FQuAD2.0: French question answering and learning when you don't know," in *Proc. 13th Lang. Resour. Eval. Conf.* Marseille, France: European Language Resources Association, Jun. 2022, pp. 2205–2214. [Online]. Available: https://aclanthology.org/2022.lrec-1.237

[37] S. Lim, M. Kim, and J. Lee, "KorQuAD1.0: Korean QA dataset for machine reading comprehension," 2019, *arXiv:1909.07005*.

[38] Y. Kim, S. Lim, H. Lee, S. Park, and M. Kim, "KorQuAD 2.0: Korean QA dataset for web document machine comprehension," *J. KIISE*, vol. 47, no. 6, pp. 577–586, Jun. 2020. [Online]. Available: http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09353166

[39] P. Efimov, A. Chertok, L. Boytsov, and P. Braslavski, "SberQuAD—Russian reading comprehension dataset: Description and analysis," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, and N. Ferro, Eds. Cham, Switzerland: Springer, 2020, pp. 3–15.

[40] M. Hardalov, I. Koychev, and P. Nakov, "Beyond English-only reading comprehension: Experiments in zero-shot multilingual transfer for bulgarian," 2019, *arXiv:1908.01519*.

[41] H. F. Sayama, A. V. de Araujo, and E. R. Fernandes, "FaQuAD: Reading comprehension dataset in the domain of Brazilian higher education," in *Proc. 8th Brazilian Conf. Intell. Syst. (BRACIS)*, Oct. 2019, pp. 443–448.

[42] F. Soygazi, O. Çiftçi, U. Kök, and S. Cengiz, "THQuAD: Turkish historic question answering dataset for reading comprehension," in *Proc. 6th Int. Conf. Comput. Sci. Eng. (UBMK)*, Sep. 2021, pp. 215–220.

[43] B. W. Wanjawa, L. D. A. Wanzare, F. Indede, O. McOnyango, L. Muchemi, and E. Ombui, "KenSwQuAD—A question answering dataset for swahili low resource language," 2022, *arXiv:2205.02364*.

[44] R. Sabol, M. Medved, and A. Horák, "Czech question answering with extended SQAD v3.0 benchmark dataset," in *Proc. 13th Workshop Recent Adv. Slavonic Natural Lang. Process.*, Dec. 2019, pp. 99–108.

[45] R. Keraron, G. Lancrenon, M. Bras, F. Allary, G. Moyse, T. Scialom, E.-P. Soriano-Morales, and J. Staiano, "Project PIAF: Building a native French question-answering dataset," in *Proc. 12th Lang. Resour. Eval. Conf.* Marseille, France: European Language Resources Association, May 2020, pp. 5481–5490. [Online]. Available: https://aclanthology.org/2020.lrec-1.673

[46] A. Kazemi, J. Mozafari, and M. A. Nematbakhsh, "PersianQuAD: The native question answering dataset for the Persian language," *IEEE Access*, vol. 10, pp. 26045–26057, 2022.

[47] O. Keren and O. Levy, "ParaShoot: A Hebrew question answering dataset," in *Proc. 3rd Workshop Mach. Reading Question Answering*. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 106–112. [Online]. Available: https://aclanthology.org/2021.mrqa-1.11

[48] C. G. Rodriguez-Penagos and C. Armentano-Oller. (2021). *ViquiQuAD: An Extractive QA Dataset From Catalan Wikipedia*. [Online]. Available: https://zenodo.org/record/4562345#.YurD-XbP2bg

[49] C. G. Rodriguez-Penagos and C. Armentano-Oller. (2021). *VilaQuAD: An Extractive QA Dataset From Catalan Newswire*. [Online]. Available: https://zenodo.org/record/4562338#.YurFJnbP2bh

[50] T. Möller, J. Risch, and M. Pietsch, "GermanQuAD and GermanDPR: Improving non-English question answering and passage retrieval," 2021, *arXiv:2104.12741*.

[51] C. C. Shao, T. Liu, Y. Lai, Y. Tseng, and S. Tsai, "DRCD: A Chinese machine reading comprehension dataset," 2018, *arXiv:1806.00920*.

[52] Y. Cui, T. Liu, W. Che, L. Xiao, Z. Chen, W. Ma, S. Wang, and G. Hu, "A span-extraction dataset for Chinese machine reading comprehension," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*. Hong Kong: Association for Computational Linguistics, 2019, pp. 5883–5889. [Online]. Available: https://aclanthology.org/D19-1600

[53] V. Snæbjarnarson, "Automated methods for question-answering in Icelandic," M.S. thesis, Fac. Ind. Eng., Mech. Eng. Comput. Sci., School Eng. Natural Sci., Univ. Iceland, Reykjavik, Iceland, 2021.

[54] B. So, K. Byun, K. Kang, and S. Cho, "JaQuAD: Japanese question answering dataset for machine reading comprehension," 2022, *arXiv:2202.01764*.

[55] K. Darvishi, N. Shahbodaghkhan, Z. Abbasiantaeb, and S. Momtazi, "PQuAD: A Persian question answering dataset," *Comput. Speech Language*, vol. 80, May 2023, Art. no. 101486, doi: 10.1016/j.csl.2023.101486.

[56] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. Pio Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, and M. Villegas, "MarIA: Spanish language models," 2021, *arXiv:2107.07253*.

[57] W. Rueangkhajorn and J. H. Chan. (Dec. 2021). *Question Answering Model in Thai by Using SQuAD Thai Wikipedia Dataset*. [Online]. Available: https://www.techrxiv.org/articles/preprint/Question_Answering_Model_in_Thai_by_using_Squad_Thai_Wikipedia_dataset/17195000
[58] A. Suriyawongkul, K. Chaovavanich, W. Phatthiyaphaibun, and C. Polpanumas. (2020). *Dataset Card for Thaiqa-Squad*. [Online]. Available: https://github.com/pythainlp/thaiqa_squad
[59] K. Van Nguyen, D.-V. Nguyen, A. Gia-Tuan Nguyen, and N. Luu-Thuy Nguyen, "A Vietnamese dataset for evaluating machine reading comprehension," 2020, *arXiv:2009.14725*.
[60] B. Ivanyuk-Skulskiy, A. Zaliznyi, O. Reshetar, O. Protsyk, B. Romanchuk, and V. Shpihanovych. (2021). *UA_Datasets: A Collection of Ukrainian Language Datasets*," [Online]. Available: https://github.com/fido-ai/ua-datasets
[61] E. Loginova, S. Varanasi, and G. Neumann, "Towards end-to-end multilingual question answering," *Inf. Syst. Frontiers*, vol. 23, pp. 227–241, Feb. 2021, doi: 10.1007/s10796-020-09996-1.
[62] J. Liu, Y. Lin, Z. Liu, and M. Sun, "XQA: A cross-lingual open-domain question answering dataset," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 2358–2368. [Online]. Available: https://aclanthology.org/P19-1227
[63] M. Artetxe, S. Ruder, and D. Yogatama, "On the cross-lingual transferability of monolingual representations," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4623–4637. [Online]. Available: https://aclanthology.org/2020.acl-main.421
[64] P. Lewis, B. Oguz, R. Rinott, S. Riedel, and H. Schwenk, "MLQA: Evaluating cross-lingual extractive question answering," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7315–7330. [Online]. Available: https://aclanthology.org/2020.acl-main.653
[65] J. H. Clark, E. Choi, M. Collins, D. Garrette, T. Kwiatkowski, V. Nikolaev, and J. Palomaki, "TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 454–470, Dec. 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.30
[66] A. Asai, J. Kasai, J. Clark, K. Lee, E. Choi, and H. Hajishirzi, "XOR QA: Cross-lingual open-retrieval question answering," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Jun. 2021, pp. 547–564. [Online]. Available: https://aclanthology.org/2021.naacl-main.46
[67] S. Longpre, Y. Lu, and J. Daiber, "MKQA: A linguistically diverse benchmark for multilingual open domain question answering," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 1389–1406, Dec. 2021. [Online]. Available: https://aclanthology.org/2021.tacl-1.82
[68] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural questions: A benchmark for question answering research," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 452–466, Aug. 2019. [Online]. Available: https://aclanthology.org/Q19-1026
[69] O. Dzurjuv, J. Genci, and R. Garabík, "Generating sets of synonyms between languages," *Proc. 6th Int. Conf. Natural Lang. Process., Multilinguality*, 2011, pp. 1–9.
[70] D. Zeman, "Slovak dependency treebank in universal dependencies," *J. Linguistics/Jazykovedný Casopis*, vol. 68, no. 2, pp. 385–395, Dec. 2017. [Online]. Available: http://hdl.handle.net/11234/1-1822
[71] V. Benko, "Two years of Aranea: Increasing counts and tuning the pipeline," in *Proc. 10th Int. Conf. Lang. Resour. Eval. (LREC)*. Portorož, Slovenia: European Language Resources Association, May 2016, pp. 4245–4248. [Online]. Available: https://aclanthology.org/L16-1672
[72] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423
[73] M. Pikuliak, V. S. Grivalský, M. Konôpka, M. Blšták, M. Tamajka, V. Bachratý, M. Simko, P. Balážik, M. Trnka, and F. Uhlárik, "SlovakBERT: Slovak masked language model," in *Findings of the Association for Computational Linguistics: EMNLP*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 7156–7168. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.530
[74] D. Hládek, J. Staš, and J. Juhár, "Evaluation set for Slovak news information retrieval," in *Proc. 10th Int. Conf. Lang. Resour. Eval. (LREC)*. Portorož, Slovenia: European Language Resources Association, May 2016, pp. 1913–1916. [Online]. Available: https://aclanthology.org/L16-1302
[75] Explosion. (2022). *Prodigy—Radically Efficient Machine Teaching. An Annotation Tool Powered by Active Learning*. [Online]. Available: https://prodi.gy/
[76] SK-spell.sk.cx. (2022). *Hunspell-Sk—Slovak Dictionary for HunSpell*. [Online]. Available: https://github.com/sk-spell/hunspell-sk
[77] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, "Marian: Fast neural machine translation in C++," in *Proc. ACL Syst. Demonstrations*. Melbourne, VIC, Australia: Association for Computational Linguistics, Jul. 2018, pp. 116–121. [Online]. Available: https://aclanthology.org/P18-4020
[78] J. Tiedemann and S. Thottingal, "OPUS-MT—Building open translation services for the World," in *Proc. 22nd Annu. Conf. Eur. Assoc. Mach. Transl.* Lisboa, Portugal: European Association for Machine Translation, Nov. 2020, pp. 479–480. [Online]. Available: https://aclanthology.org/2020.eamt-1.61
[79] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength natural language processing in Python," ExplosionAI, Berlin, Germany, Tech. Rep., 2020.
[80] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017. [Online]. Available: https://aclanthology.org/Q17-1010
[81] Explosion. (2022). *Floret: FastText + Bloom Embeddings for Compact, Full-Coverage Vectors With spaCy*. [Online]. Available: https://github.com/explosion/floret
[82] J. Staš, J. Juhár, and D. Hládek, "Classification of heterogeneous text data for robust domain-specific language modeling," *EURASIP J. Audio, Speech, Music Process.*, vol. 2014, no. 1, p. 14, Dec. 2014. [Online]. Available: http://asmp.eurasipjournals.com/content/2014/1/14
[83] D. Hládek and J. Stas, "Text mining and processing for corpora creation in Slovak language," *J. Comput. Sci. Control Syst.*, vol. 3, no. 1, p. 65, 2010.
[84] D. Hládek. (2022). *Slovak spaCy Model*. [Online]. Available: https://github.com/hladek/spacy-skmodel
[85] C.-C. Kuo and K.-Y. Chen, "Toward zero-shot and zero-resource multilingual question answering," *IEEE Access*, vol. 10, pp. 99754–99761, 2022.
[86] S. Wu and M. Dredze, "Are all languages created equal in multilingual BERT?" in *Proc. 5th Workshop Represent. Learn. NLP*, 2020, pp. 120–130. [Online]. Available: https://aclanthology.org/2020.repl4nlp-1.16

DANIEL HLÁDEK (Member, IEEE) was born in Košice, Slovakia, in 1982. He received the M.S. degree in artificial intelligence with the Technical University of Košice, Slovakia, in 2006, and the Ph.D. degree in artificial intelligence, in 2009.

From 2009 to 2015, he was a Research Assistant with the Laboratory of Speech Communication Technologies. Since 2015, he has been an Assistant Professor with the Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Košice. He is the author and coauthor of more than 100 articles. His research interests include natural language processing, natural language understanding, natural language generation, text mining, statistical language modeling, question answering, automatic spelling correction, spontaneous speech recognition, and human–computer interactions. He also investigates questions related to the detection of hate speech and offensive language.

JÁN STAŠ (Member, IEEE) was born in Bardejov, Slovakia, in 1984. He received the M.S. degree in electronics and telecommunications from the Technical University of Košice, Slovakia, in 2007, and the Ph.D. degree in telecommunications, in 2011.

From 2011 to 2015, he was a Research Assistant with the Laboratory of Speech Communication Technologies. Since 2015, he has been an Assistant Professor with the Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Košice. He is the author and coauthor of more than 125 articles. His research interests include natural language processing, statistical language modeling, question answering, text classification, speaker diarization, and spontaneous speech recognition. He also investigates questions related to the detection of hate speech and offensive language, voice stress analysis, and early detection of Alzheimer's disease from speech transcripts.

JOZEF JUHÁR (Member, IEEE) was born in Poproč, Slovakia, in 1956. He received the degree and the Ph.D. degree in radioelectronics from the Technical University of Košice, in 1980 and 1991, respectively.

Currently, he is a Full Professor with the Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Košice, where he is also the Founder of the Laboratory of Speech Communication Technologies. He is the author and coauthor of more than 400 articles. His research interests include digital signal processing, especially speech and audio processing, speech analysis and synthesis, speech acoustics, acoustic event detection, speech enhancement and dereverberation, acoustic modeling, speaker recognition, and the research and development of speech recognition systems, spoken dialogue systems, and human–computer interfaces.

TOMÁŠ KOCTÚR (Member, IEEE) was born in Banská Bystrica, Slovakia, in 1988. He received the M.S. degree in multimedia telecommunications from the Technical University of Košice, Slovakia, in 2014, and the Ph.D. degree in multimedia communication technologies, in 2018.

Since 2018, he has been a Data Scientist with Deutsche Telekom IT & Telecommunications Slovakia. He is focused on NLP topics, such as neural information retrieval, reading comprehension, and text anomaly detection. His corporate social responsibility interests are to help the development of natural language processing in the Slovak language by creating and supporting open-source NLP datasets and language models.

Dr. Koctúr is a member of the Committee on Ethics and Regulation of Artificial Intelligence founded by the Ministry of Investments, Regional Development and Informatization of the Slovak Republic, since 2020.