Proceedings of the 10th Workshop on the Representation and Processing of Sign Languages (sign-lang@LREC 2022), pages 154–158
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC 4.0

Towards Large Vocabulary Kazakh-Russian Sign Language Dataset: KRSL-OnlineSchool

Medet Mukushev[1], **Aigerim Kydyrbekova**[1], **Vadim Kimmelman**[2], **Anara Sandygulova**[1]
[1]Department of Robotics and Mechatronics, School of Engineering and Digital Sciences
Nazarbayev University, Nur-Sultan, Kazakhstan
[2]Department of Linguistic, Literary, and Aesthetic Studies
University of Bergen, Bergen, Norway
{mmukushev, aigerim.kydyrbekova, anara.sandygulova}@nu.edu.kz
vadim.kimmelman@uib.no

Abstract

This paper presents a new dataset for Kazakh-Russian Sign Language (KRSL) created for the purposes of Sign Language Processing. In 2020, Kazakhstan's schools were quickly switched to online mode due to COVID-19 pandemic. Every working day, the El-arna TV channel was broadcasting video lessons for grades from 1 to 11 with sign language translation. This opportunity allowed us to record a corpus with a large vocabulary and spontaneous SL interpretation. To this end, this corpus contains video recordings of Kazakhstan's online school translated to Kazakh-Russian sign language by 7 interpreters. At the moment we collected and cleaned 890 hours of video material. A custom annotation tool was created to make the process of data annotation simple and easy-to-use by Deaf community. To date, around 325 hours of videos have been annotated with glosses and 4,009 lessons out of 4,547 were transcribed with automatic speech-to-text software. KRSL-OnlineSchool dataset will be made publicly available at https://krslproject.github.io/online-school/

Keywords: sign language dataset, kazakh-russian sign language, sign language processing

1. Introduction

Sign Language Processing (SLP) combines three related research and development directions, such as automated Sign Language recognition, generation, and translation, with the goal of developing technological solutions that will help break down communication barriers for the Deaf community and sign language users (Bragg et al., 2019). To date, more than half of published vision-based research for SLP utilizes isolated sign language data with a limited vocabulary size (Koller, 2020). However, the real-world value of SLP solutions demands continuous signing recognition, which is significantly harder than individual sign recognition due to co-articulation (the end of one sign affecting the beginning of the next), depiction (visually representing or enacting content), and generalization (Bragg et al., 2019). There are considerable limitations in publicly available sign language datasets that restrict the strength and applicability of recognition systems trained on them. Limitations of datasets include the size of the vocabulary, which is mostly related to expensive annotation methodologies, or datasets that only include isolated signs, which are insufficient for most real-world use cases involving continuous signing (Bragg et al., 2019). As a result, in order to progress SLP, realistic, generalizable, and extensive datasets are required.

This paper proposes a new large-scale KRSL dataset created for the needs of SLP. The objective of KRSL-OnlineSchool is to address shortcomings of commonly used datasets such as lack of continuous signing and small vocabulary size. KRSL-OnlineSchool's main advantage is in its large vocabulary size, extensive gloss annotation, and high number of recorded videos.

In 2020, classes in Kazakhstan's schools were quickly switched to online mode due to COVID-19 pandemic. Every working day, the El-arna TV channel was broadcasting video lessons for grades from 1 to 11 with sign language translation. This opportunity was used to create a large corpus consisting of video recordings of online school's translations to sign language performed by 7 interpreters. We collected more than 1,000 hours of raw video recordings which were later pre-processed and divided into categories by subject and grade. At the end we obtained 890 hours of cleaned videos with sign language. Additionally, web-based annotation tool was created to make the process of data annotation simple and easy-to-use by deaf annotators. To date, around 325 hours of videos were annotated with glosses and 4,009 lessons out of 4,547 were transcribed with automatic speech-to-text software. Thus, this paper makes the following contributions:

- we release the first large-scale Kazakh-Russian Sign Language dataset consisting of 4547 video lessons (890 hours), translated by 7 signers, and divided into categories by subject and grade;
- we release transcripts for 4009 lessons collected with automatic speech-to-text software (a total of 1 million sentences);
- we release more than 39,000 gloss annotations of 30-seconds video segments (a total of 325 hours).

Datasets	Language	Signers	Vocabulary	Samples	Duration (h)
RWTH-BOSTON-400 (Dreuw et al., 2008)	ASL	4	483	843 sentences	-
SIGNUM (Agris and Kraiss, 2010)	DGS	25	450	780 sentences	55,3
RWTH-PHOENIX 2014T (Camgoz et al., 2018)	DGS	9	2887	8,257 sentences	10,96
Video-Based CSL (Huang et al., 2018)	CSL	50	178	25,000 videos	100
BSL-1K (Albanie et al., 2020)	BSL	40	1064	1M sentences	1,060
How2Sign (Duarte et al., 2021)	ASL	11	15,686	35,000 sentences	80
KRSL-OnlineSchool	**KRSL**	**7**	**20,000**	**1M sentences**	**890**

Table 1: Datasets used for Continuous Sign Language Recognition. This list excludes datasets of isolated signs. For KRSL-OnlineSchool vocabulary we counted unique words with at least 20 samples in transcripts

The remainder of this paper discusses related work, followed by descriptions of our methodology for the data collection. We then introduce the data itself and provide some statistics. The paper concludes with guidelines for future work utilizing collected dataset.

2. Related Work

Sign language datasets are critical for progressing the goals of Sign Language Recognition. RGB datasets captured with conventional cameras, for example, have practical use in real-world scenarios. These collections include videos of either isolated or continuous signing. Table 1 presents an overview of the most commonly used sign language datasets that are appropriate for the problem of Continous Sign Language Recognition with an inclusion of KRSL-OnlineSchool.

RWTH-Phoenix-Weather-2014T (Camgoz et al., 2018) is a German Sign Language (DGS) dataset used as a benchmark for most recent works in SLP. It features nine signers who performed sign language translations of the weather forecast on TV broadcasts. RWTH-Boston-400 (Dreuw et al., 2008) is one of the first CSLR benchmark datasets for American Sign Language (ASL). But it has only four signers present in the videos. In contrast, Video-Based CSL (Chinese Sign Language) (Huang et al., 2018) provides a large number of participants (n=50) involved in collecting the dataset. At the same time, they are all recorded in the same recording settings, and most participants seem to be unfamiliar with sign language as they sign in slow and artificial ways without involving any facial expressions. SIGNUM (**?**) is a signer-independent CSLR dataset of DGS with all participants being fluent in DGS and are either deaf or hard-of-hearing. However, all videos were shot with a single RGB camera in a supervised condition with the same lighting and uniform blue background.

These concerns of existing datasets limit the accuracy and robustness of the models developed for SLR and their contribution to the challenges of real-world signing. More recent datasets aim to address most challenges of the previous datasets: BSL-1K (Albanie et al., 2020) provides the largest number of annotated sign data, while How2Sign (Duarte et al., 2021) provides the largest vocabulary size. Similar to older datasets, they were either recorded in a controlled lab environment or extracted from the TV broadcast. From this perspective, KRSL-OnlineSchool is the sign language dataset that includes large vocabulary size and extensive gloss annotation needed for training recognition models.

3. Dataset Collection

The KRSL-OnlineSchool dataset consists of phrases and sentences in KRSL recorded as a synchronous interpretation of online lessons on various subjects for various grades (1-11 grades of primary, secondary and high school). KRSL is the sign language used in the Republic of Kazakhstan. KRSL is closely related to Russian Sign Language (RSL). While no official research comparing KRSL with RSL exists, they show a substantial lexical overlap and are entirely mutually intelligible (Kimmelman et al., 2020). Figure 1 shows the overview of our data collection methodology.

3.1. Video Collection Process

Every working day during the academic year, from September 2020 to May 2021, the El-arna national TV channel was broadcasting video lessons for grades from 1 to 11 with sign language translation. Lessons were live broadcast both on TV and channel's website from 9 AM till 5 PM, with an average duration of 10-12 minutes per lesson. We set up a computer with screen recording software and were recording online classes for 9 months. Table 2 shows a total number of collected lessons divided into subjects category.

The next step included a need to crop signers' region from the extracted videos and splitting videos into lessons by subjects and grades. We utilized OpenCV library for video processing and wrote custom scripts to perform this task. At the end we collected 890 hours of clean videos divided into 4,547 video lessons. Extracted videos have a resolution of 230x264 pixels. English lessons had to be discarded as they had no sign language translation.

3.2. Annotation Process

We have collected two types of annotations for our dataset, full text transcriptions of lessons and gloss annotation of short clips. Table 3 shows a total number of collected lessons presented by grade level and their number of transcripts.

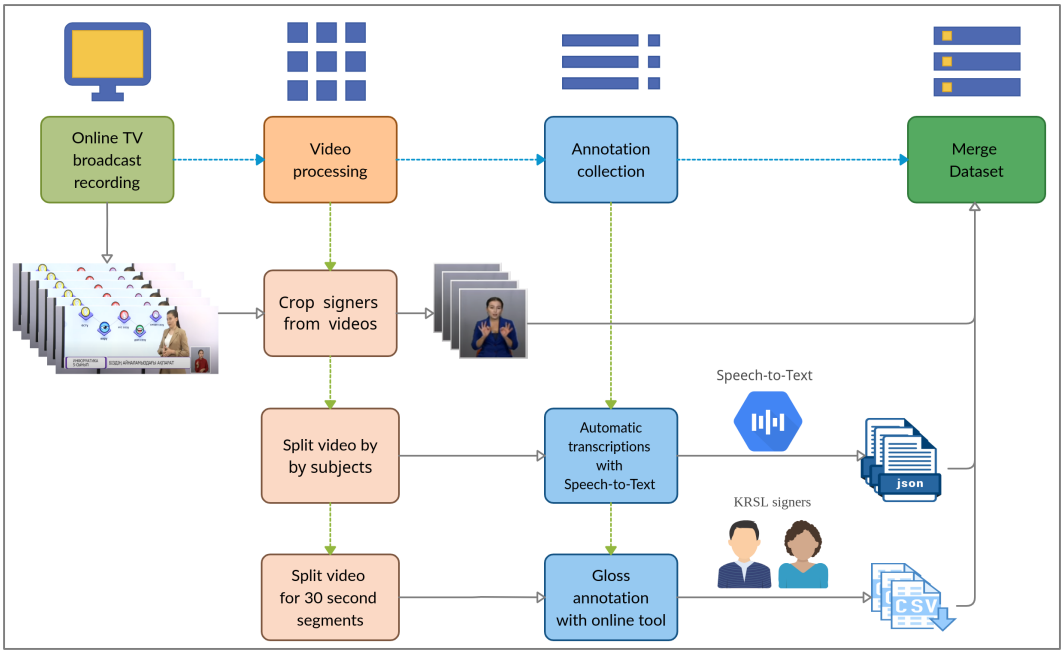

Figure 1: Dataset collection methodology

	Subject name	Videos
1	Literacy education	76
2	Math	602
3	Second language	794
4	Natural science	129
5	World science	91
6	Digital literacy	43
7	History	357
8	Kazakh language	538
9	World history	216
10	Algebra	298
11	Informatics	178
12	Geography	248
13	Chemistry	193
14	Literature	100
15	Geometry	185
16	Physics	263
17	Biology	236
Total		**4547**

Table 2: List of subjects in dataset

At first we utilized Kaldi ASR library (Povey et al., 2011) to collect full text transcriptions of lessons. However, this approach was not very convenient, as it required to extract audio streams from videos and splitting them into small segments. Later these segments were passed to automatic speech recognition algorithm, which then provided transcriptions for each segment. We then decided to utilize YouTube's captioning software which automatically recognized and

Grade	Videos	Transcripts
1	249	205
2	318	257
3	334	288
4	325	282
5	366	349
6	344	292
7	484	441
8	513	457
9	584	522
10	518	468
11	506	448
Total	**4547**	**4009**

Table 3: Number of lessons by grade

synced captions for each video. We wrote custom script using the YouTube API to download transcriptions for all lessons.

For gloss annotation, we divided videos into small 30-seconds clips in order to make the gloss annotation process simpler for deaf annotators. To date, a total of 325 hours of videos or 39,000 segments were uploaded and annotated using a custom web-based annotation tool.

We realized that it was necessary to send videos to annotators, to receive their annotations as well as keep track of their progress and time spent for monetary compensation (8 USD per hour). It was decided to implement a web-based annotation tool (https:surdobot.kz). Annotators were provided with login and password to enter the system. The tool has a simple user interface, which shows a random clip and a

text area to enter recognized glosses. Functionality of the tool also includes options to play, stop video clip, change playback speed, and submit annotation. Annotators also have options to view all clips they have processed and edit them if needed. Videos were divided and uploaded into online annotation tool as soon as they were processed. Thus, we have first annotated videos recorded in September, October and November of 2020. For annotation process we hired 8 annotators, 5 of whom are deaf and 3 are professional KRSL translators.

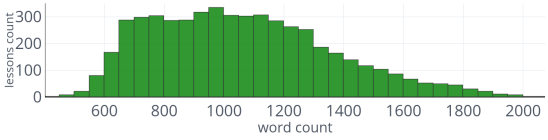

Figure 2: Full text transcripts length for each lesson

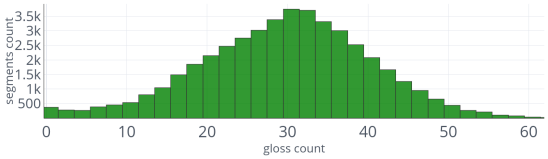

Figure 3: Gloss annotation length for each 30 second clip

4. Results

We collected 890 hours of video lessons divided into 4547 lessons. Around 325 hours of videos were annotated with glosses and 4,009 lessons out of 4,547 were transcribed with automatic speech-to-text software. For KRSL-OnlineSchool vocabulary we counted unique words with at least 20 samples in transcripts which give us a size of vocabulary of more than 20,000 words.

Figure 2 shows a word count in full text transcriptions of the lessons. An average word count of one lesson is around 1,000 words. Transcripts shorter than 800 words were mostly lessons for primary school classes, as they had shorter duration.

Figure 3 shows a gloss count in 30-seconds clips. An average gloss count of one clip is around 30 glosses. There are more than 1,000 unique glosses with at least 150 repetition for each. We are currently continuing the gloss annotation process with an aim to fully annotate all 890 hours of videos.

We have extracted 25 most frequently used words and glosses from both annotations. Figures 4 (words) and 5 (glosses) demonstrate these results. As we can see, there are some samples that appear in both charts. For example, most frequent token in both cases is “this”. Also, “minus”, “equal”, “today”, “correct”, “exercise”, “number”, “words”, “answer”, “need”, “watch” tokens are common for two charts. This shows us that both

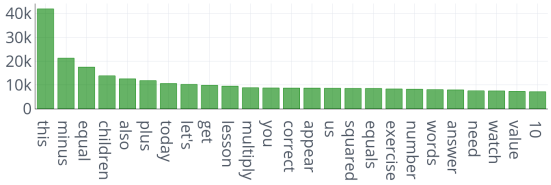

Figure 4: Top words in full text transcripts

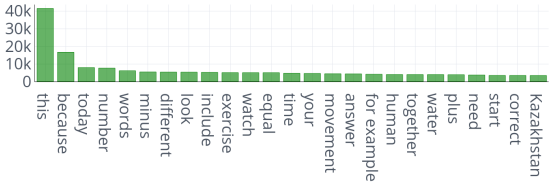

Figure 5: Top glosses in gloss annotations

automatic transcriptions and manual gloss labeling can be used for dataset annotation. We believe that number of correlating tokens will increase when the rest of the dataset is annotated with glosses.

Additionally, we have extracted 20 most frequently used 2-grams from both annotations. Figures 6 (words) and 7 (glosses) shows these results. There were fewer matching examples compared to top words-glosses charts. Some matching examples include “lesson today”, “correct answer”, “next assignment”, “equals minus” were common for both charts.

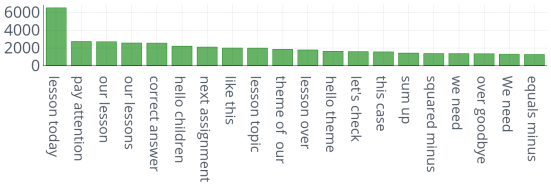

Figure 6: Top 2-grams for text transcriptions

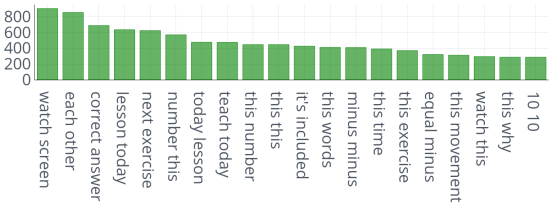

Figure 7: Top 2-grams for gloss annotations

5. Conclusion

We have presented a new dataset for Kazakh-Russian Sign Language created for the purposes of Sign Language Processing. It is a large-scale dataset that includes a large vocabulary size and extensive gloss annotation needed for training recognition models. It is one of the largest collected sign language dataset with more than 890 hours of videos, 325 of which are manually annotated with glosses and 1 million sentence tran-

scripts. This dataset can be utilized for experiments on weakly supervised Sign Language translation models by training a large teacher model with the help of gloss annotated data, which can later be evaluated on transcribed data.

6. Acknowledgements

This work was supported by the Nazarbayev University Faculty Development Competitive Research Grant Program 2019-2021 "Kazakh Sign Language Automatic Recognition System (K-SLARS)" under award number 110119FD4545.

7. Bibliographical References

Agris, U. v. and Kraiss, K.-F. (2010). SIGNUM Database: Video Corpus for Signer-Independent Continuous Sign Language Recognition. In Philippe Dreuw, et al., editors, *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 243–246, Valletta, Malta, May. European Language Resources Association.

Albanie, S., Varol, G., Momeni, L., Afouras, T., Chung, J. S., Fox, N., and Zisserman, A. (2020). BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *European Conference on Computer Vision*, pages 35–53. Springer.

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., et al. (2019). Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31. ACM.

Camgoz, C. N., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018). Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Dreuw, P., Neidle, C., Athitsos, V., Sclaroff, S., and Ney, H. (2008). Benchmark Databases for Video-Based Automatic Sign Language Recognition. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 1–6, Marrakech, Morocco, May. European Language Resources Association.

Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., and Giro-i Nieto, X. (2021). How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2735–2744.

Huang, J., Zhou, W., Zhang, Q., Li, H., and Li, W. (2018). Video-based sign language recognition without temporal segmentation. *arXiv preprint arXiv:1801.10111*.

Kimmelman, V., Imashev, A., Mukushev, M., and Sandygulova, A. (2020). Eyebrow position in grammatical and emotional expressions in Kazakh-Russian Sign Language: A quantitative study. *PLOS ONE*, 15(6):1–16, 06.

Koller, O. (2020). Quantitative Survey of the State of the Art in Sign Language Recognition. *arXiv preprint arXiv:2008.09918*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December.