Retrospective of Kazakh-Russian Sign Language Corpus Formation

Alfarabi Imashev[1], **Aigerim Kydyrbekova**[1], **Medet Mukushev**[1], **Anara Sandygulova**[1], **Shynggys Islam**[2], **Khassan Israilov**[3], **Aibek Makazhanov**[2], **Zhandos Yessenbayev**[2]

[1]Department of Robotics Engineering, Nazarbayev University, Kazakhstan
[2]Computer Science Lab, National Laboratory Astana, Nazarbayev University, Kazakhstan
[3]Public Association "Kazakh Deaf Society", Astana branch, Kazakhstan
{alfarabi.imashev, aigerim.kydyrbekova, mmukushev, anara.sandygulova}@nu.edu.kz,
sislam@alumni.nu.edu.kz, {aibek.makazhanov, zhyessenbayev}@nu.edu.kz

Abstract

Sign language (SL) is a mode of communication that, in most cases, relies on visual perception exclusively and uses the visual-gestural modality. The advent of machine learning techniques has expanded the range of potential applications, not only in industry but also in addressing societal needs. Previous research has already demonstrated encouraging outcomes in developing sign language recognition systems that are both quite accurate and resilient. Nevertheless, the effectiveness and use of algorithms are impacted not only by their accessibility but also, at times to a greater extent, by the presence of substantial quantities of pertinent data. At the start of the local sign language corpus collection in 2015, there was a notable deficit of local Kazakh-Russian Sign Language data available for computer vision and machine-learning tasks. There were already corpora of another lexically close language, Russian Sign Language, but they were aimed at and tailored for lingustic research. We initiated the procedure by collecting data appropriate for machine-learning purposes. The subsets have been incorporated into the principal corpus and will be subject to future enhancements and refinements. This paper provides an overview of the collected components of the Kazakh-Russian Sign Language Corpus and the resulting outcomes derived from them.

Keywords: sign language, dataset collection, overview

1. Introduction

The emergence of machine learning approaches and techniques has broadened the scope of possible applications, not just in business or industry but also in meeting social demands. Previous research undertaken before 2015 has already shown promising results in the development of sign language recognition systems that are both highly accurate and durable. However, the efficiency and use of algorithms are influenced not only by their availability but also, often to a greater extent, by the existence of significant amounts of relevant data.

The government of Kazakhstan offers each deaf individual 60 hours per year of free sign language interpreting service support. These hours can be spent on medical, legal, or other communication requirements. The scarcity of interpreters per capita and the lack of remunerated interpreting services raise the necessity of supplementary alternative instruments for sign language recognition, translation and generation, which require datasets to train on. Regrettably, in 2015 there was not any dataset on local Kazakh-Russian Sign Language (K-RSL); there were corpora of similar Russian Sign Language (RSL) from Novosibirsk and Saint-Petersburg, but they were focused on linguistic research.

Thus, we decided to start collecting relevant data of local K-RSL suitable for machine learning applications. The sign language used by the deaf signers' community in Kazakhstan is not indigenous and is closely related to RSL as well as other sign language within the CIS. All of these sign languages have their roots in the Soviet Union's centralized language policy, which established the signing system. While no formal study comparing K-RSL with RSL was conducted, the expertise of interpreters, and our observations indicate a significant similarity in vocabulary and frequent mutual intelligibility.

Nevertheless, the deaf community in Kazakhstan has already assimilated distinctive and unique themes into the local sign language, such as native musical instruments, regional delicacies, famous sites, significant figures, traditional beliefs, and more. Note that although RSL and K-RSL share many lexical similarities, it is uncertain if this extends to other linguistic aspects of both languages.

This paper provides a concise overview of the collected components of the Kazakh-Russian Sign Language Corpus aiming at applying machine learning approaches, and the resulting outcomes derived from them within the last decade.

The following section provides brief overview on related datasets existed in 2015. Section 3 offers a summary of subsets present in the current corpus, focused on several linguistic properties often seen in most sign languages, such as phonological minimum pairings, sign variability, and sign polysemy. Section 4 explores potential alternative methods for acquiring new types of sign language datasets.

Proceedings of the 11th Workshop on the Representation and Processing of Sign Languages, pages 111–122
25 May 2024.

Table 1: Hand Image Datasets (VS: Vocabulary Size, NP: Number of Participants)

Dataset	Volume	VS	NP	Resolution
NUS-I (Kumar et al., 2010)	480	10	24	160x120
NUS-II (Pisharady et al., 2013)	2000+750	10	40	160x120, 320x240
Polish Sign Language - I (Kawulok et al., 2013)	899	25	12	174x131 to 640x480
Polish Sign Language - II	85	13	3	4672x3104
Polish Sign Language - III	574	32	18	3264x4928
ASL Finger Spelling Dataset (Pugeault and Bowden, 2011)	48,000	24	5+4	128×128
J. Triesch I (Triesch and Von Der Malsburg, 1996)	720	10	24	128×128 (gray, 8 bit)
J. Triesch II (Triesch and Von Der Malsburg, 2001)	1000	12	19	128x128 (color)
MU ASL dataset (Barczak et al., 2011)	2524	36	5	high-res

Table 2: Video Datasets (VS: Vocabulary Size, NP: Number of Participants)

Dataset	Volume	VS	NP	Resolution
ASLLVD (Neidle et al., 2012)	9,800 tokens	3,300	1-6	640x480, 60fps 1600x1200, 30fps
BosphorusSign22k (Özdemir et al., 2020)	22,542 (19h)	744	6	1920x1080, 30fps
CSL-1 (Huang et al., 2018)	25,000 (100h)	178	50	1920x1080, 25 fps
RWTH-PHOENIX-Weather (Forster et al., 2012)	21,822 (1,980 sent.)	911	7	210x260, 25 fps
Purdue RVL-SLLL (Martinez et al., 2002)	2,576	39	14	640x480
DEVISIGN (Chai et al., 2014)	24,000(21.87h)	2000	30	640x480
SIGNUM (Von Agris et al., 2008)	33,210 (55.3h)	450, 780	25	776x578, 30 fps
RWTH-BOSTON (Athitsos et al., 2008)	843	406	5	324x242, 30 fps
DGS - KORPUS (Nishio et al., 2010)	50h (public)	530	330	640x360, 50 fps

2. Related Work

The task of finding a database that is optimal for machine learning and creating a model is specific and individual, for each particular task posed by the researcher. At the beginning of the study, we encountered several dataset containing images of the hands. We mostly did not take into account datasets designed for Kinect or Key-glove like devices, as they do not fulfill the necessary criteria of our goal, which is the ability of the system to operate with K-RSL without the need for any extra costly technological equipment. After reviewing which ML algorithms to test, we decided to revise the following image (see Table 1) and video (see Table 2) datasets available to figure out the best practices of dataset collection taking place at that moment (before 2015 and in 2020).

3. Collected Datasets

This section provides a brief account of the progressive growth of the K-RSL corpus, encompassing all datasets gathered for it from 2015 until the present day.

At the outset of our research, none of the sign language datasets mentioned in the literature followed any strict established requirements for recognizing continuous sign language that is not dependent on a signer. In contrast to voice recognition, there was no pre-existing standard, baseline, or reference point. Therefore, we have tried to collect a dataset that is anticipated to assist research efforts for scholars who exhibit interest in the sign language recognition area. We believe that this dataset has the potential to become a benchmark for researchers who are studying advanced sign language recognition algorithms. It is signer-independent and suitable for continuous recognition. Furthermore, it includes cases of sign variability, polysemy (where the meaning of a sign is determined by mouthings), and phonological minimal pairs, which are very similar in performance. These factors make the task of automatic recognition more challenging and increase the complexity of the problem.

It is noteworthy that the deaf and hard-of-hearing community in Kazakhstan exhibits a high degree of insularity. Regrettably, according to Kazakhstan Deaf Society authorities and interpreters' experience, these issues arose due to instances of fraudulent activities perpetrated against individuals, including internet fraud, property crimes, violations of contracts, and lower wages, along with several instances of being involved in sects. All these negative experiences were deposited in memory and deeply ingrained in the local deaf culture, as was evident in how they viewed all outsiders. This led to the situation where interpreters and the state or non-profit deaf organizations became the primary conduits for establishing first communications and collaborations.

At the moment when our research began, there was a dominance of descriptors and feature extraction approaches in computer vision, and therefore, we also relied on the well-known ones and could cooperate with four sign language interpreters only for our first attempt.

One major limitation of the sign language recognition field, when we started our research, was that all trustworthy and reputable video data sources consisted of video data, which was entirely created in a controlled "laboratory" setting. In such settings, the camera remains stationary, the background is uniform and consistent, and the lighting conditions are usually predetermined and unchanging. This was the reason why we decided to collect 1/3 of our first dataset outside the lab (Figure 2).

Based on previous linguistic and applied research, as well as the increasing availability of technologies that can extract coordinates of the human body and facial features, such as MediaPipe[1] (see Figure 1) and OpenPose[2], we have identified several data types to collect for our dataset. These technologies, developed between 2017 and 2019, provide the opportunity to analyze and validate the unique characteristics of sign communication in different emotional states, as well as for questions or statements. It inspired us to specifically collect sentences with grammatical sentence type marking and marking of emotions to study the role of non-manual in recognition, collecting minimal pairs of signs as potentially challenging for recognition tasks. In the end, we collected quite a wide variety of data types, which are discussed in detail below.

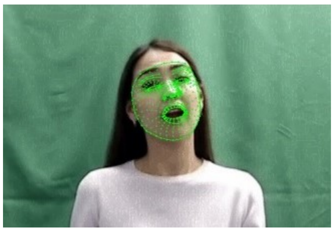

Figure 1: Face landmarks with MediaPipe.

3.1. Healthcare videos (2015-2017)

A survey conducted among representatives of the deaf community in Astana and practicing interpreters indicated that deaf signers primarily require accurate interpretations verified by experts for healthcare-related circumstances. Consequently, the initial demand from the community was to establish a comprehensive database for machine learning dedicated to the healthcare domain. All of this involved the development, formation, and collection of a sign language database that encompasses sentences comprising frequently employed medical phrases and terminology.

[1]https://developers.google.com/mediapipe

[2]https://github.com/CMU-Perceptual-Computing-Lab/openpose

Figure 2: Frames of healthcare dataset.

Interpreters who have accompanied deaf individuals in medical settings have collaborated to create a list of essential vocabulary terms. The reference interpreter and researchers then constructed sentences to ensure a balanced inclusion of signs in the dataset. Subsequently, we recorded the reference interpreter's performance of these sign sequences, ensuring that the hands, head, and face remained inside the camera's field of view and were well-lit. Afterward, we informed the other interpreters that we needed them to replicate his sign sequences since the output videos were for machine-learning algorithms. They agreed to reproduce the sign sequences in full, following the example of the reference interpreter. All 8846 videos were recorded using the website's tool,which stored them directly on the server. Once the entire dataset had been collected, interpreters were given the task of assessing each other and providing annotations for their colleagues (see Figure 3).

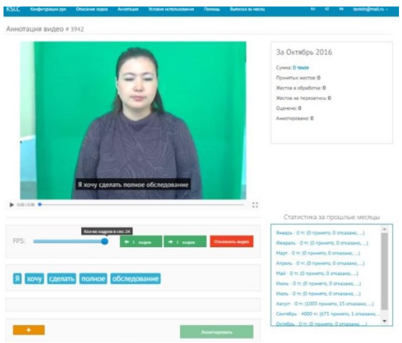

Figure 3: Annotation tool.

We ended up with approximately 148 unique sentences, choosing the top 5 repetitions based on performance quality. Unfortunately, basic CNNs and the Weka tool (Thornton et al., 2013) exhibited a relatively low recognition rate of approximately 53%. The involvement of only four interpreters, three recording modes, and storing videos on the website's server at 320x240 resolution undoubtedly impacted the output.

3.2. Healthcare images (2015-2017)

Revising outputs and drawbacks - we decided to extract images of the most frequent hand configurations to obtain a hand image dataset for training purposes. The idea was to extract cropped images of handshapes (as shown in Figure 4), which will be used for training purposes later.

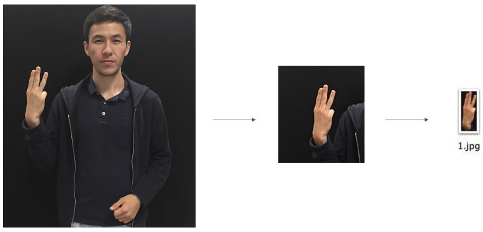

Figure 4: The frame, the ROI, and the element of the dataset.

At first, we decided to try it on a well-known dataset. We downloaded the NCSLGR handshapes videos dataset[3]. We took each 5th frame from videos, which let us obtain hand configurations of various angles. Using a simple hand detector, we extracted configurations by saving ROIs as images - we obtained the set of hand images. Then made the same for our videos.

Next, using HOG (Dalal and Triggs, 2005)+KMeans (MacQueen et al., 1967) clustering, we distributed the same configurations from different subsets to the separate folders for further training (see Figure 5).

Figure 5: Obtained hand configuration images dataset.

With this technique, we obtained 27 configurations (folders) of the highest inclusion numbers. We implemented a similar HOG+KMeans approach later in Mukushev et al. (2020a) too.

During that period, approaches associated with the generation of supplementary artificial data for training purposes seemed unrealistic. So we made

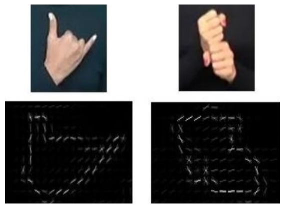

Figure 6: The HOG descriptor performance.

research and tests on various detectors and descriptors available at those time, such as local invariant descriptors: SIFT (Lowe, 1999), SURF (Bay et al., 2006), RootSIFT (Arandjelović and Zisserman, 2012); Binary descriptors: ORB (Rublee et al., 2011), BRISK (Leutenegger et al., 2011), and HOG descriptor. Considering all the advantages and disadvantages of the aforementioned descriptors, we have chosen to utilize the HOG descriptor (see Figure 6) in conjunction with the classification algorithm SVM (Boser et al., 1992) since SVM is reported to exhibit higher performance in cases where there is a lack of data.

Figure 7: Hand configurations from Polish, American and local SL dataset (merged dataset).

We also added images of the same configurations from the Polish SL dataset and got the merged dataset (see Figure 7). After that, we selected 10 configurations with 100 samples and implemented HOG+SVM, results and rates described in Imashev (2017).

3.3. Six emotions

The origins of theories regarding fundamental emotions can be traced back to ancient Greece and China as stated by Russell (2003). The fundamental idea of emotions has exerted significant influence for over fifty years. According to the current basic emotion theory, humans have a finite set of emotions that are considered biologically and psychologically "basic" (Wilson-Mendenhall et al., 2013). These emotions exhibit regular recurrence of consistent patterns (Russell, 2006). Researchers in Ekman et al. (2013) revealed evidence of prevalence for six specific emotions: anger, fear, sadness, happiness, surprise, and disgust combined with contempt.

We adhered to the conventional roster of six emotions, except one: five emotions (anger, fear, sadness, happiness, surprise) and "sorry".

[3] https://www.bu.edu/asllrp/cslgr/pages/ncslgr-handshapes.html

We compiled a list of sentences that are semantically compatible with each of the emotions, in collaboration with K-RSL interpreters. During the recording, the sentences were represented as sequences of glosses via a separate monitor in front of them. Each interpreter performed sentences in different order depending on the emotion. The list of sentences is in Appendix A.

Figure 8: The six emotions in our dataset.

3.4. Phonological minimal pairs

Analogous to the existence of phonological minimal pairs in spoken languages, a comparable phenomenon is observed in sign languages (Sandler, 2012; Thompson et al., 2013). In sign language, a minimal pair is a pair of signs with distinct meanings that are distinguished by only one of the major parameters, such as hand configuration, orientation, movement, or non-manual features. Minimal pairs can pose potential problems for recognition tasks, as they are formally similar but semantically different.

There are precedents in the literature for building datasets that specifically target minimal pairs for recognition purposes. As an example, the LIBRAS-UFOP (Cerna et al., 2021). This dataset contains 56 classes of minimal pairs of Brazilian Sign Language. The data was collected using a Microsoft Kinect V1 sensor, which provided comprehensive skeleton data. The dataset was evaluated for recognition using Convolutional Neural Networks (CNN) and long short-term memory (LSTM). The highest accuracy achieved was 74.25%.

The initial reference to phonological minimum pairs in Kazakh-Russian Sign Language was documented in Imashev et al. (2020).

Here are sentences and visual representations for phonological pairs such as RIGHT - MAY (see Table **??** and Figure 9 upper row), and BLUE - WEDNESDAY(v1) (see Table **??** and Figure 9 lower row). Figure 9 also shows two variants for the concept of WEDNESDAY. Note that WEDNESDAY(v1) and WEDNESDAY(v2) are examples of lexical variability, but only one of them forms a minimal pair with the sign BLUE. This serves as an illustrative example of a case where one sign can be part of a phonological minimal pair and a case of variability simultaneously.

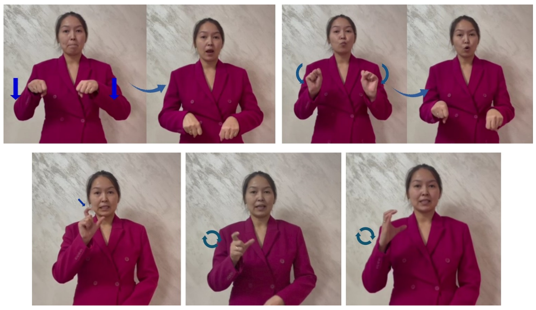

Figure 9: RIGHT(legal) - MAY (upper row), BLUE - WEDNESDAY(v1) - WEDNESDAY(v2) (lower row).

Overall, we collected sentences and videos of 8 pairs and 3 triplets.

3.5. Question vs. Statement

Question signs in K-RSL, like question words in spoken/written Kazakh and Russian languages, can be employed not only in interrogative sentences, but also in declarative sentences: “The place **where** sun never sets” and “**Where** are you going?”. Thus, any question sign can occur either with non-manual question marking (eyebrow rise, sideward or backward head tilt) or without it. Furthermore, question marks are accompanied by the mouthing articulation of the related word (see Figure 10).

Question signs are distinguished based on manual aspects, but additional information is obtained through mouthing, which aids recognition. Hence, the two categories of non-manual indicators, namely eyebrow and head position versus mouthing, have distinct functions in recognition. The former aids in distinguishing between statements and questions, while the latter assists in distinguishing between different question signs. To test and confirm, we selected ten question words and constructed twenty phrases: 10 questions and 10 sentences for each word for this dataset (see sentences for WHO in Table **??**).

Five interpreters were given them in written form on a screen in front of them one by one to perform (Imashev et al., 2020), the outputs of sign language recognition implementation with this dataset are described in Mukushev et al. (2020b).

Figure 10: A - WHEN, B - WHEN in question; C - HOWMUCH, D - HOWMUCH in question; E - WHERE(location), F - WHERE(location) in question; G - HOW, H - HOW in question; I - WHICH, J - WHICH in question; K - WHATFOR (reason), L - WHATFOR (reason) in question; M - WHICHONE , N - WHICHONE in question, O - WHERE(direction); P - WHERE(direction) in question; Q - WHO, R - WHO in question; S - WHAT/THAT, T - WHAT/THAT in question.

3.6. Statements, polar and content questions

For this task, we composed 10 sequences as statements, polar, and wh- questions (see Table **??**). We requested interpreters to perform all of them with emotions (in a neutral, surprised, and angry manner) to figure out how emotions and grammatical marking interact in the non-manual features. As mentioned before, deaf communities are quite gated, and this was the first contact and involvement of local native deaf signers in research: several of them (half of the individuals who appeared in this dataset) performed these sentences. Several other deaf signers requested to evaluate and try to recognize emotions (see Figure 11), the results described by Kimmelman et al. (2020). Besides, Kimmelman et al. (2020) is specifically about studying how eyebrow position is affected by sentence type marking and emotions.

3.7. K-RSL-173 (Nov. 2019-2020)

After completing a collection of several narrow-purposed subsets, we returned to the idea of collecting a dataset that contains a wide range of concepts used in everyday life. Taking into account the shortcomings of such datasets as PHOENIX (only 9 signers, and a narrow vocabulary about weather and regions of Germany) and DEVISIGN (the participants' performance looked a little unnaturally slow, and the gaze often looked like the

Figure 11: A statement, polar and wh- questions performed in three mood states.

performer did not know the meaning of the signs performed) provide us hints on how to collect our linguistically rich dataset with general, everyday life sentences performed mainly by native signers, fluent signers of different ages, and also filmed in different conditions. By gradually disseminating information about our research, working closely with interpreters for several years, and thereby increasing the level of trust in us from the deaf community, we were able to gather a sufficient number of deaf signers who agreed to participate in data collection and understand the importance for the community.

Initially, we composed 246 sentences, which were revised and narrowed down to 173 sentences with feedback from the reference interpreter, Khassan Israilov. For example, a sentence like 'A doctor told me I needed to remain in bed' (DOCTOR TOLD ME I NEED REMAIN BED REST REGIME), deaf signers will probably perform in a simplified manner as DOCTOR TOLD BED. We recorded these sentences produced by 50 signers (32 deaf, 6 hard of hearing, also 9 hearing CODA, and 3 hearing SODA, including 7 of them are also interpreters).

For sentence translation, we recorded translations of the most proficient (recognized by interpreters and the community) reference interpreter, who made his translations from written sentences, which were composed of spoken language in the manner closest to glosses to avoid any miscommunication. Initially, participants were asked to repeat sign after sign after him from videos. The first few people repeated this but said that they wanted to perform it differently. The next few people were given complete freedom; as a result, the translations of one sentence were completely different from each other (for example: MAY YOU PLEASE SAY TIME vs. just performing sign TIME with question face). This led to the fact that we could not collect the required number of sign inclusions for these participants. Therefore, we decided to allow the participants partial freedom with the opportunity to add any clarifications that they consider necessary or change the order of signs.

We detected sign variability at the start of the data collection process mode when participants had partial freedom. After reviewing videos from several initial participants, it was evident that there would be more variability occurrences in the dataset. It presented the opportunity to find specific examples of sign variability in the less explored K-RSL.

It also provided the basis for identifying the variability of signs — one of the reasons for dissatisfaction and arguments like "I do not want to perform signs the same"; there were also formulations like "I used to perform this sign differently". It helped us identify a certain number of cases of sign variability. See also Kimmelman et al. (2022) for a study on the lexical variability of isolated signs in RSL conducted in partnership with the Garage Museum of Contemporary Art.

Regarding sign variability, consider one of the concepts with several options that was detected in the current dataset. Three configurations used for LEISURE are in Figure 12 also may differ in motions (see Figure 13).

Figure 12: Three variants of **LEISURE** detected in the Dataset.

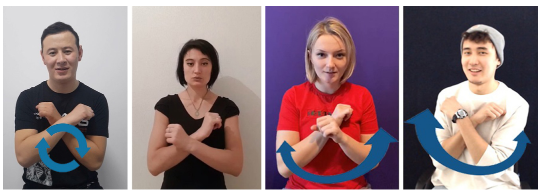

Figure 13: Different motions used for LEISURE.

It is noteworthy that all professional interpreters and several native deaf signers performed sign **LEISURE** in the same manner: the hands intersected in the wrist region. The dorsal sides of the clenched fists are in opposition to each other. This configuration rotates in a circular motion in front of the chest (see Figure 14). This observation may indicate the establishment of standardization, at least in the context of interpreting. Alternatively, it could reveal that these participants share a common geographical or educational background that sets them apart from other signers.

Figure 14: All interpreters performed in the same manner.

Another interesting phenomenon we have observed in the dataset is the presence of polysemic signs, more specifically, those that are distinguished by mouthing. Figure 15 displays different lexical variants of the sign SPOUSE, organized in columns and combined with the mouthing for WIFE or HUSBAND, arranged in rows.

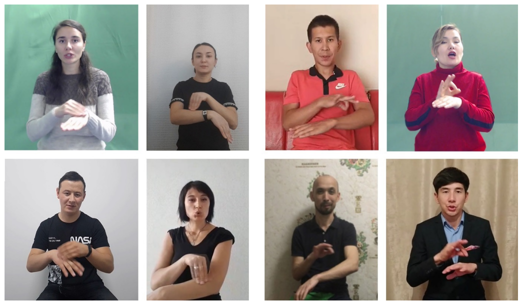

Figure 15: SPOUSE variants in handshapes and performance.

An example of a similar phenomenon case is described in Antonakos et al. (2015), German Sign Language Corpus The SIGNUM contains videos for concepts BRUDER and SCHWESTER which utilize the same sign but differ in mouthing (see Figure 16).

Figure 16: 'die Geschwister' sign used for both meanings 'Bruder' (brother) and 'Schwester' (sister) (Von Agris et al., 2008; Konrad et al., 2020).

We also discovered two neologisms in the dataset one resulting from the combination of two signs (see Figure 17 a) and the other arising from the combination of two concepts (see Figure 17 b).

In the end, we detected 43 cases of variability (2-6 variants each) and 2 cases of polysemy appearing in the dataset, all of the aforementioned

Figure 17: a) Instagram, b) Facebook.

nuances make it closer to natural sign language performance and more challenging for recognition tasks (Mukushev et al., 2022b).

4. Unpublished Datasets and Future Work

Since deaf individuals often communicate in public settings, the actions of others or external circumstances can disturb the background view. Algorithms that exhibit high accuracy rates under controlled laboratory conditions may perform worse when confronted with unpredictable real-world conditions. Given the difficulty of collecting a dataset in natural environments like parks or public places such as shopping malls, researchers should consider utilizing pre-existing video datasets with uniform backgrounds for keying purposes (see Figure 18). By training algorithms to achieve higher recognition rates in scenarios resembling crowded locations, this approach has the potential to improve sign recognition rates in real-world conditions.

Figure 18: Possible dataset keying.

Priorly acquired datasets can also be utilized as the foundation for generating datasets of 3D signing motion models. For instance, reusing our datasets to get 3D motion files from videos could be expanded to initiate a 3D Signing Dataset (see Figure 19).

Incidentally, amidst the circumstances posed by COVID-19 restrictions, A. Kydyrbekova diligently collected online school lessons aired on National TV, which broadcasted with sign language support

Figure 19: Data-driven signing agent (avatar).

(Mukushev et al., 2022a). Besides, a vocabulary dataset has been collected with 4 interpreters. This dataset contains topics like groceries, household items, also local notions and concepts such as musical instruments, dishes, etc. These two datasets will be available and provided at a later time.

5. Acknowledgements

We are deeply indebted and would like to express our deepest appreciation to the interpreters Gulmira Baizhanova, Khassan Israilov, Aidana Otegenova, Viktoria Antonishina, Samal Nurym, and Shyryn Kozhbanova for their knowledge and support, without which this research would not have been possible. We would also like to extend our deepest gratitude to Dr. Kimmelman for consistently staying connected and offering insights from a sign language perspective. Special thanks to Zhangeldy Bekbatyrov, Qyzylgul Sembina, Shynggys Islam, Azat Kassymgaliyev, Meir, Muslima Karabalayeva, Adai Shomanov, and Aigerim Kydyrbekova. Grateful acknowledgment to Aibek Makazhanov, Zhandos Yessenbayev, and Anara Sandygulova for supporting research on Kazakh-Russian Sign Language throughout the past decade. Research grant awards: 1) The targeted program O.0743 (0115PK02473) of the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan (2015-2017); 2) Nazarbayev University Faculty Development Competitive Research Grant Program 2019-2021 “Kazakh Sign Language Automatic Recognition System (KSLARS)”. Award number is 110119FD4545; 3) Nazarbayev University Faculty Development Competitive Research Grant Program 2022-2024, “Kazakh-Russian Sign Language Processing: Data, Tools, and Interaction”; the award number is 11022021FD2902.

6. Bibliographical References

Epameinondas Antonakos, Anastasios Roussos, and Stefanos Zafeiriou. 2015. A survey on mouth modeling and analysis for sign language recognition. In *2015 11th IEEE International Conference* *and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–7. IEEE.

Relja Arandjelović and Andrew Zisserman. 2012. Three things everyone should know to improve object retrieval. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2911–2918. IEEE.

Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. 2008. The american sign language lexicon video dataset. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8.

ALC Barczak, NH Reyes, M Abastillas, A Piccio, and Teo Susnjak. 2011. A new 2d static hand gesture colour image dataset for asl gestures.

Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, pages 404–417. Springer.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

Lourdes Ramirez Cerna, Edwin Escobedo Cardenas, Dayse Garcia Miranda, David Menotti, and Guillermo Camara-Chavez. 2021. A multimodal libras-ufop brazilian sign language dataset of minimal pairs using a microsoft kinect sensor. *Expert Systems with Applications*, 167:114179.

Xiujuan Chai, Hanjie Wang, and Xilin Chen. 2014. The devisign large vocabulary of chinese sign language database and baseline evaluations. In *Technical report VIPL-TR-14-SLR-001. Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS)*. Institute of Computing Technology.

Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE.

Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. 2013. *Emotion in the human face: Guidelines for research and an integration of findings*, volume 11. Elsevier.

Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus H Piater, and Hermann Ney. 2012. Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In *LREC*, volume 9, pages 3785–3789.

Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. 2018. Video-based sign language recognition without temporal segmentation. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 2257–2264. AAAI press.

Alfarabi Imashev. 2017. Sign language static gestures recognition tool prototype. In *2017 IEEE 11th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–4. IEEE.

Alfarabi Imashev, Medet Mukushev, Vadim Kimmelman, and Anara Sandygulova. 2020. A dataset for linguistic understanding, visual evaluation, and recognition of sign languages: The k-rsl. In *Conference on Computational Natural Language Learning*.

Michal Kawulok, Tomasz Grzejszczak, Jakub Nalepa, and Mateusz Knyc. 2013. Database for hand gesture recognition.

Vadim Kimmelman, Alfarabi Imashev, Medet Mukushev, and Anara Sandygulova. 2020. Eyebrow position in grammatical and emotional expressions in kazakh-russian sign language: A quantitative study. *PloS one*, 15(6):e0233731.

Vadim Kimmelman, Anna Komarova, Lyudmila Luchkova, Valeria Vinogradova, and Oksana Alekseeva. 2022. Exploring networks of lexical variation in russian sign language. *Frontiers in psychology*, 12:740734.

Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. 2020. Meine dgs – annotiert. öffentliches korpus der deutschen gebärdensprache, 3. release / my dgs – annotated. public corpus of german sign language, 3rd release.

P Pramod Kumar, Prahlad Vadakkepat, and Ai Poh Loh. 2010. Hand posture and face recognition using a fuzzy-rough approach. *International Journal of Humanoid Robotics*, 7(03):331–356.

Stefan Leutenegger, Margarita Chli, and Roland Y Siegwart. 2011. Brisk: Binary robust invariant scalable keypoints. In *2011 International conference on computer vision*, pages 2548–2555. IEEE.

David G Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. IEEE.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

A.M. Martinez, R.B. Wilbur, R. Shay, and A.C. Kak. 2002. Purdue rvl-slll asl database for automatic recognition of american sign language. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 167–172.

Medet Mukushev, Alfarabi Imashev, Vadim Kimmelman, and Anara Sandygulova. 2020a. Automatic classification of handshapes in russian sign language. In *Proceedings of the LREC 2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*. European Language Resources Association (ELRA).

Medet Mukushev, Aigerim Kydyrbekova, Vadim Kimmelman, and Anara Sandygulova. 2022a. Towards large vocabulary Kazakh-Russian Sign Language dataset: KRSL-OnlineSchool. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 154–158, Marseille, France. European Language Resources Association.

Medet Mukushev, Arman Sabyrov, Alfarabi Imashev, Kenessary Koishibay, Vadim Kimmelman, and Anara Sandygulova. 2020b. Evaluation of manual and non-manual components for sign language recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).

Medet Mukushev, Aidyn Ubingazhibov, Aigerim Kydyrbekova, Alfarabi Imashev, Vadim Kimmelman, and Anara Sandygulova. 2022b. Fluentsigners-50: A signer independent benchmark dataset for sign language processing. *PLoS ONE*, 17.

Carol Neidle, Ashwin Thangali, and Stan Sclaroff. 2012. Challenges in development of the american sign language lexicon video dataset (asllvd) corpus. In *5th workshop on the representation and processing of sign languages: interactions between corpus and Lexicon, LREC*.

Rie Nishio, Sung-Eun Hong, Susanne König, Reiner Konrad, Gabriele Langer, Thomas Hanke, and Christian Rathmann. 2010. Elicitation methods in the dgs (german sign language) corpus project. In *sign-lang@ LREC 2010*, pages 178–185. European Language Resources Association (ELRA).

Oğulcan Özdemir, Ahmet Alp Kındıroğlu, Necati Cihan Camgoz, and Lale Akarun. 2020. BosphorusSign22k Sign Language Recognition Dataset. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*.

Pramod Kumar Pisharady, Prahlad Vadakkepat, and Ai Poh Loh. 2013. Attention based detection and recognition of hand postures against complex backgrounds. *International Journal of Computer Vision*, 101:403–419.

Nicolas Pugeault and Richard Bowden. 2011. Spelling it out: Real-time asl fingerspelling recognition. In *2011 IEEE International conference on computer vision workshops (ICCV workshops)*, pages 1114–1119. IEEE.

Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. IEEE.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

James A Russell. 2006. Emotions are not modules. *Canadian Journal of Philosophy Supplementary Volume*, 32:53–71.

Wendy Sandler. 2012. The phonological organization of sign languages. *Language and linguistics compass*, 6(3):162–182.

Robin L Thompson, David P Vinson, Neil Fox, and Gabriella Vigliocco. 2013. Is lexical access driven by temporal order or perceptual salience? evidence from british sign language. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Chris Thornton, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. 2013. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 847–855.

Jochen Triesch and Christoph Von Der Malsburg. 1996. Robust classification of hand postures against complex backgrounds. In *Proceedings of the second international conference on automatic face and gesture recognition*, pages 170–175. IEEE.

Jochen Triesch and Christoph Von Der Malsburg. 2001. A system for person-independent hand posture recognition against complex backgrounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1449–1453.

Ulrich Von Agris, Moritz Knorr, and Karl-Friedrich Kraiss. 2008. The significance of facial features for automatic sign language recognition. In *2008 8th IEEE international conference on automatic face & gesture recognition*, pages 1–6. IEEE.

Christine D Wilson-Mendenhall, Lisa Feldman Barrett, and Lawrence W Barsalou. 2013. Neural evidence that human emotions share core affective properties. *Psychological science*, 24(6):947–956.

7. Appendix A. Sentences composed for six emotions dataset

Table 3: Sentences on 6 selected emotions

Anger	**Sadness**
People's anger	My memories of the past are sad
There is no need to rush - you will become angry	Sad face
Patience, you do not need to be angry	Sad eyes
Anger - is a strong feeling	They are sad
Anger prevents thinking rationally	I hear his voice is sad
Strong anger	There is no need to be sad
Anger helps to win	Sadness ends soon
When he is angry, everyone is scared	Happy and sad
Old people are angry	Looked away with a sad look
They are angry for no reason	Why are you sad
Fear	**Surprised**
Fear of the dark	Childhood is when everything is surprising
People struggle with their fears	Their knowledge is surprising
Fear is hard to hide	Are you surprised?
We are afraid of many things	Kazakhstan's nature is surprisingly beautiful
There is no need to be scared	The boy looked surprised
Fear has big eyes	Fairytales are surprising
Fear helps the enemy	The athletes' records are surprising
Very scary movie	Surprised faces
Grandmother fears the future	These discoveries are surprising for us
She was afraid of heights	They looked into the distance in surprise
Sorry	**Happy**
I'm sorry, and I'm suffering	Well-being is the source of happiness
You are always feel sorry	Serene happiness
Being able to be sorry is important for the future	True happiness
I feel sorry for him; that's why crying	I'm happy
Grandma always feels sorry for everyone	This is the reason for happiness
People must be kind and be able to feel sorry for each other-otherwise, the world has no future	Happy face
I'm sorry for the thrown-away books	A happy man
I'm really sorry	I found a job - I'm happy
I feel sorry for the animals	They are happy that they came
I'm sorry - I left	We are happy that we left