Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology (SLTAT 7), pages 59–66
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022

KoSign Sign Language Translation Project: Introducing The NIASL2021 Dataset

Mathew Huerta-Enochian[1] Du Hui Lee[1] Hye Jin Myung[2] Kang Suk Byun[2] Jun Woo Lee[2]
[1]EQ4ALL
11, Nonhyeon-ro 76-gil, Gangnam-gu, Seoul. Republic of Korea.
{mathew, scottlee}@eq4all.co.kr
[2]Kangnam University
40, Gangnam-ro, Giheung-gu, Yongin-si, Gyeonggi-do. Republic of Korea.
{acq57rep, byunkang-suk, knudeaf}@kangnam.ac.kr

Abstract

We introduce a new sign language production (SLP) and sign language translation (SLT) dataset, NIASL2021, consisting of 201,026 Korean-KSL data pairs. KSL translations of Korean source texts are represented in three formats: video recordings, keypoint position data, and time-aligned gloss annotations for each hand (using a 7,989 sign vocabulary) and for eight different non-manual signals (NMS). We evaluated our sign language elicitation methodology and found that text-based prompting had a negative effect on translation quality in terms of naturalness and comprehension. We recommend distilling text into a visual medium before translating into sign language or adding a prompt-blind review step to text-based translation methodologies.

1. Introduction

In this paper, we introduce a new Korean and Korean Sign Language (KSL) translation dataset, NIASL2021, containing 201,026 paired Korean-KSL samples from the emergency alert message and weather broadcast domains. NIASL2021 was created to support KoSign, a sign language translation (SLT) and sign language production (SLP) development project, and can thus be used for SLT and SLP and serve as a reference for avatar development. We also present a critical evaluation of the translation methodology used in NIASL2021 to inform future collection methodologies.
Our contributions are:

- Introduction of the complete NIASL2021 dataset
- Quantitative evaluation of the translation methodology used in NIASL2021, which revealed that text-free prompting produced better translations than text-based prompting.

In section 2 we briefly review relevant research and development projects before introducing the new translation dataset in section 3. We then present a quantitative evaluation of our translation methodology in section 4 and present our conclusions in section 5.

2. Background

The primary language for many Deaf and hard of hearing (DHH) individuals is their region's sign language. While hearing people can easily access a wide variety of news sources, DHH signers are usually limited to a handful of deaf news services or must consume media through text. Though using an interpreting service is reasonable for large events and critical news broadcasts, it is usually impractical to do so for daily news, weather reports or non-critical alert messages. We suggest that an automatic sign language translation engine targeting this domain would be highly impactful to DHH signers as a supplement to existing interpreting services, underscoring the need for new emergency-situation translation datasets.

2.1. Translation Data Collection

Translation datasets are multilingual datasets with a semantic alignment between each language. A common trend in collection methodologies for monolingual datasets is to prompt for expressions in the informants native language or in a neutral medium (like images) to reduce the influence of a foreign language as is mentioned in (Filhol and Hadjadj, 2018), (Nishio et al., 2010), and (Hong et al., 2009). However, for translation datasets, a non-native language prompt is usually used to create translations. Even when employing professional translators, an increase in so-called translationese is unavoidable. See (Koppel and Ordan, 2011) for a discussion. If the training data is intended to be non-directional, a common method to reduce translationese imbalance is to collect an equal proportion of source data from each language as in (Bojar et al., 2018), where 50% of language A is translated into language B and 50% of language B is translated into language A for every language pair A and B in the dataset. Source language texts are usually collected from existing material.
Since sign languages are extremely low-resource, existing sign language source material for a given translation topic will be insufficient. Therefore, the above 50-50 solution must be abandoned or data must be manually generated from structured, semi-structured, or unstructured interviews for sign language datasets. Unstructured interviews will yield inconsistent content

while structured interviews that allow fine control over content will be subject to unwanted language influence and translationese. We are not aware of any accepted solution to this problem, and most projects assume that using professional interpreters will minimize the severity of translationese.

The two most common benchmark translation datasets for sign languages are RWTH-PHOENIX-Weather 2014T from (Camgoz et al., 2018) and How2Sign (Duarte et al., 2020). RWTH-PHOENIX-Weather 2014T contains German and German Sign Language (DGS) translation pairs from weather broadcasts while How2Sign contains English and American Sign Language (ASL) translation pairs from a variety of domains. Both feature text, sign video translations, and single-channel gloss annotations. Recently, (Camgöz et al., 2021) introduced several news and weather broadcast sign language datasets with an order of magnitude more data than in RWTH-PHOENIX-Weather 2014T.

Sign language datasets use the terms sign, type, and gloss to encode and explain a signed passage. We refer to (Johnston and Schembri, 1999)'s definition of a sign: signs are "a relatively stable, identifiable visual-gestural act with an associated meaning which is reproduced with consistency by native signers and for which, consequently, particular agreed values can be given for hand shape, orientation, location, and movement." Types are a fixed naming system for signs, and each type is distinct in appearance or in meaning. We refer to (Konrad et al., 2020) for further discussion of types. Finally, glosses are the text representations or annotations of a sign.

2.2. Sign Language Production

Though there is some overlap in the usage of "sign language translation" (SLT) and "sign language production" (SLP), literature is becoming clearer in using SLT to refer to translating sign into text or speech (a natural extension of sign language recognition) and SLP to refer to translating text or speech into sign language. However, SLP also covers topics of avatar generation and how to digitally express signing.

2.3. The KoSign Project

KoSign is an ongoing SLT and SLP engine development project that started in 2021 and is funded by the Korean Ministry of Trade, Industry, and Energy.[1] The project is a collaboration between five domestic member organizations: EQ4ALL, KETI, KAIST, TestWorks, and the Korean Association of the Deaf. To support continued development, we secured additional funding for a large-scale Korean-KSL translation data collection project (see section 3) and are continuing to acquire funding for other projects in support of KoSign. The scope of this project is two-fold:

- Research machine-learning-based SLT and SLP (including relevant avatar technologies)
- Develop a usable, bi-directional Korean-KSL translation engine

We are leading the project and conducting SLP research and development. We utilize transformer-like models to predict type tokens and sign timing data, decoding into a multi-channel signing space. We are conducting human evaluations for our models and will release our results in the future.

A brief overview of our avatar player was provided in (Kim et al., 2022). We divide our avatar into five channels: left hand, right hand, body, lower face, and upper face. We then use inverse kinematics (IK) and animation composition to model each channel and combine them into one animation. This method is a simple way to expand a limited number of animations into a large set of complex animations.

2.4. Other Sign Language Production Projects

There are a number of ongoing projects of similar size and scope to KoSign. (EASIER, Accessed 2022 04 04) and (SignON, Accessed 2022 04 04) are two projects funded by the EU's Horizon 2020 research program. Both projects aim to create models for automatic translation between sign languages and spoken/written languages. Both projects target multiple European sign and spoken languages. (AVASAG, Accessed 2022 04 04) (Avatar-basierter Sprachassistent zur automatisierten Gebärdenübersetzung) on the other hand is a research project focusing on developing a real-time controlled avatar for translating German texts into sign language.

3. NIASL2021

We[2] introduce NIASL2021,[3] a new Korean-KSL translation dataset, collected over the domains of Korean government emergency alert messages and weather broadcasts. Collection was a multi-organization effort and native signers were intimately involved in the process.

NIASL2021 contains 201,026 unique data samples (segmented at the Korean sentence and multi-sentence level) and can be used to train both SLT and SLP (gloss-, pose-, or video-generating) models. KSL translations use 7,989 unique types, and all samples feature a single signer only. Data samples are organized into one of forty-three categories: weather and forty-two emergency alert categories. There are many similar categories, and since multiple disaster events often co-occur, there is significant overlap between categories.

[1] 산업통상자원부 in Korean.

[2] In this section, we use "we" to refer to our work and the passive voice for work conducted by other parties.

[3] The project was funded by the Korean National Information Society Agency (NIA). The dataset will be released in late 2022, accessible through https://aihub.or.kr/; we will host an in-depth user guide at https://eq4all-data.github.io from the fourth quarter of 2022.

For example, the landslide and flooding categories have overlap with heavy rain and typhoon categories.

Each sample in the dataset has five components: metadata about the sample, Korean text, a KSL video translation of the text, gloss annotations, and automatically-extracted keypoint estimations. For simplicity, we bundle the metadata, Korean text, gloss annotations, and keypoint data together in a JSON file so that each sample can be expressed with only a video file and a human-readable data file.

Since there is an abundance of emergency alert and weather broadcasts available in Korean and none originally in KSL, KSL videos in every sample are translated from the associated Korean text. As discussed in 2.1, this may introduce undesired translationese in the KSL samples, but we took as many steps as possible to reduce this risk.

Note that a subset of NIASL2021 was used in (Kim et al., 2022).

3.1. Korean Text Data

Korean text was initially scraped from government alert and news websites to create a raw Korean text dataset. This dataset had extreme class imbalance. Categories related to recent issues like Covid-19 had many samples, but other categories like terrorism had few or none. Additional samples were manually created based on government text outlines for categories with too few samples.

This raw text dataset was split into two subsets, one subset set aside for the final dataset and one subset used to train a series of GPT2-like natural language generation models for offline-augmentation as in (Kumar et al., 2020). Using these models, each category was oversampled (except for weather broadcasts and the infectious diseases alert categories, which already had a sufficient number of samples). Generated sequences were then reviewed based on grammar and similarity with training samples to ensure that synthetic data was in distribution. Synthetic samples were then combined with the unused text subset to create the final set of Korean text. Note that sample metadata indicates if it is a synthetic or original sample.

3.2. KSL Video and Annotation Data

Based on feedback from KSL experts, we allowed multiple translations to be made for each Korean source text. For each source, KSL experts determined how many sign language translations should be prepared, ranging from one to three translations. Researchers should be aware of this detail when using the dataset as over one-fourth of the data is made up of one-to-many translations. If needed, researchers can reduce the dataset to a 148,984 sample subset of one-to-one translations.

3.2.1. Translation and Video Capture

After translation duplicity was determined for a source text, we would assign the text to a translator and an evaluator three days before a scheduled translation filming date. We instructed translators and evaluators to research each sample and prepare for translation and evaluation, respectivley, during this three-day period. On the day of filming, evaluators would review the prepared translations. Translations that required little or no correction could be filmed, and translations judged as insufficient were corrected right away or returned to the translator for improvement. Based on initial discussion with KSL experts, we felt that this method should be effective for producing high-quality translations.

Translations were filmed either in a studio with two or five cameras or were crowd-sourced and filmed with phone cameras or web cams. Of the 201,026 samples in the dataset, 127,624 (63.49%) samples were created in-studio and 73,402 (36.51%) were crowd-sourced. The multi-camera setups captured one frontal view of the signer and one or four 45° angled view(s) of the signer (45° views were offset from above, down, left, and right for the five-camera setup and left for the two-camera setup).

All translators and evaluators were native signers and had previous experience translating Korean into KSL. Official translation videos may feature the translator or may be filmed with a different signer who re-signed the prepared translation exactly. Translator, evaluator, and signer IDs were all collected in sample metadata.

Though all signers and evaluators were native signers, we received feedback from participants that the crowd-sourced videos may be of a lower quality than in-studio translations. This is to be expected from crowd sourcing but also indicates the need for more strict review of crowd-sourced translations in the future.

3.2.2. Annotation

Filmed translations were annotated by hand with 90-95% of samples annotated by deaf participants and the remaining 5-10% by hearing signers. Additionally, our type system was created and managed by deaf participants.

A single-channel gloss list would not sufficiently preserve the meaning of the KSL translations in this domain. For example, one common translation pattern was a disaster event like a fire that would be expressed with one hand while the other hand explained what to do about the event (take a detour, go the opposite way, etc.). After consulting with KSL experts, we decided to annotate the dominant hand[4] and non-dominant hand separately, as well as eight types of non-manual signals (NMS): puffed cheeks (denoted Ci), head shake (Hs), eye brow furrow (EBf), head nod (Hno), mouthings (Mmo), rounded lips (Mo1), tongue out (Tbt), and smile (Mctr). We refer to these ten different annotation types as tiers. All annotations are time aligned to the corresponding translation video.

Following the convention from (Kita et al., 1997), hand signs can be segmented into four movements: prepara-

[4] All recorded signers self reported as right-handed.

tion, stroke, hold, and retraction. The movement most associated with a sign is the stroke. Preparation and retraction are more akin to inter-sign movements and hold is an optional movement where the articulator is held in the sign or gesture's final position. We instructed annotators to align annotations with the start of the stroke and the end of the hold.

Each annotation in the sign tiers was from one of four categories: type, dynamic number (signs combining number hand shapes with gestures to express certain quantities, such as dates, times, durations, and ages), fingerspelling (FS), and number. We annotated FS and numbers separately since a series of digits and a multi-digit number need to be expressed differently (for example, 555 can be either "five five five" or "five hundred and fifty five"), and annotating groups of FS numbers together significantly eased the annotation burden given the frequent phone numbers, addresses, and quantity expressions in the dataset.

Though existing annotation tools like ELAN (Wittenburg et al., 2006) are well-developed, we designed our own webtool to have more control over the annotation interface and for better integration into our online data pipeline. This allowed us to create a separate annotation insertion menu for each of the annotation categories, streamlining the user interface.

In addition to the manual gloss annotations, pose data was automatically extracted from each KSL video using OpenPose. For videos filmed from more than one angle (the in-studio five-camera and two-camera videos), OpenPose-generated 2D keypoints from two separate camera angles were used to calculate 3D keypoints for each frame using MATLAB. Since crowdsourced videos only have a single view, they contain 2D keypoint data.

3.3. Challenges

3.3.1. Signing Dates

In KSL, the day of the month cannot be signed without also signing the month. For example, "the 11^{th}" cannot be signed by itself in KSL, but "the 11^{th} of January" can be signed. However, it is common to express only the day of the month in Korean, especially in emergency alert messages and weather broadcasts since these sources are not intended to be relevant outside of a small temporal window. To create realistic training data, we included samples with this pattern and instructed translators to denote the month using the zero value hand shape when translating. We also added a flag in sample metadata so researchers can choose to remove these data points or find some other work around.

3.3.2. Translating Unclear Context

One of the biggest hurdles was translating low-context and unclear phrases into KSL. There were two root causes for this ambiguity: differing context requirements between Korean and KSL and poor Korean source text segmentation.

The first problem refers to when something in Korean can be expressed with ambiguity, but any translation to KSL (as with most sign languages) is highly context-dependent.

Since recording long sequences increases the need for multiple takes and increases signer fatigue, source text was intentionally segmented into short sequences. Additionally, most of the synthetic text data (see section 3.1) was generated at the sentence level. This led to the second problem mentioned above. Many such cases were removed, but we allowed some to be translated since it was not always clear what samples reflected real-world data (because of the first problem above) and what samples were vague due to processing error.

For future projects, we recommend segmenting at a higher level or assigning consecutive samples to the same translator.

3.3.3. Annotating Productive Signs

Following (Johnston and Schembri, 1999), we differentiate between two classes of signs in NIASL2021: established and productive signs. Established signs are simply signs collectively known to users of a sign language. Productive signs are created through a novel combination of sign building-blocks (known as phonomorphemes) or the selective modification of one or more established signs or phonomorphemes. These are new or modified signs spontaneously expressed based on the signing context.

We annotated productive signs by labeling them with the most similar type (referred to as its "parent type") and adding up to three special symbols and an optional string identifier. We added a "#" character to the end of every productive sign annotation, and optionally added a short explanatory string after the "#" character. If the sign terminated prematurely, we added a "@" character after the "#" and optional string. Finally, when the hand shape varied from the hand shape of the parent type, we added a "&" character to the beginning of the annotation.

For example, if the signer indicates that a car turns left using a productive sign derived from the parent type "car1", then we might annotate the type as "car1#turnleft". If the hand is shaped a little tighter to indicate that the car is small, it will be annotated as "&car1#turnleft". Finally, if the signer indicates that the car starts to turn left but stops the sign abruptly (perhaps to indicate that left turns are not allowed), the annotation would be "car1#turnleft@". Note that actual types are in Korean.

4. Translation Methodology Evaluation

Anecdotally, we noticed that some of the KSL translations were unclear without checking the Korean source. Based on qualitative review, we tentatively identified two reasons for low quality signing: unclear Korean source passages (see section 3.3) and spoken language influence on translations (see section 2.1). We can mitigate source ambiguity by aligning longer segments, but

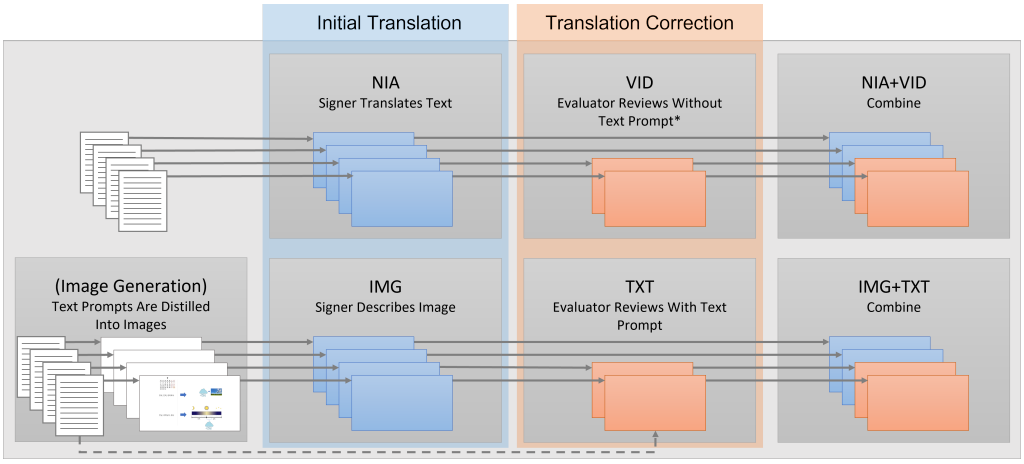

Figure 1: Overview of evaluation video generation. Best viewed in color.
*Source text is made available after initial review.

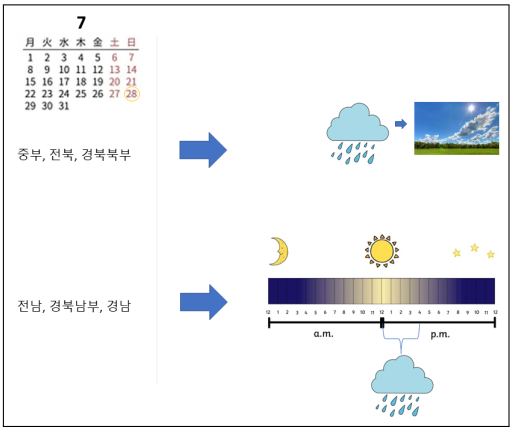

Figure 2: Example of an image prompt created from part of a weather report. Only location names and morning/evening abbreviations are expressed as text.

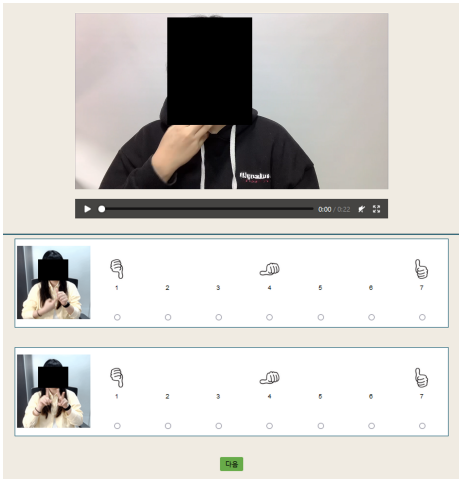

Figure 3: An example from our evaluation web tool.

avoiding spoken language influence will require a new translation methodology.

To evaluate translation quality and to explore the influence of spoken language in prompted sign language translation, we designed two new translation methodologies: NIA+VID and IMG+TXT. Both are two step methodologies with an initial translation (what we call NIA and IMG video translations, respectively) and a translation correction (VID and TXT corrected translations, respectively). Thus, NIA+VID and IMG+TXT videos refer to corrected videos and any initial translations that are not corrected.

NIA+VID uses the NIASL2021 translation methodology as the initial translation (for convenience, we use translations from the dataset) and an initially prompt-blind evaluation step. For IMG+TXT, prompts are first converted into image representations. Signers then describe the image as the initial translation. The signer is then shown the original prompt and given the option to update their initial translations. See figure 1 for a visual overview of the two methodologies.

We further define signing quality as the aggregate of signing naturalness and comprehensibility, evaluated on a likert scale, and make the following hypotheses:
H_1: TXT $<$ IMG Text-aware correction decreases the signing quality.
H_2: NIA $<$ VID Text-unaware correction increases the signing quality.
H_3: NIA $<$ IMG Image-prompted translations are of a higher quality than text-prompted translations.
H_4: NIA+VID $<$ IMG+TXT Image-prompted translations are of a higher quality than text-prompted translations, even with corrections.

Finally, it is important that we validate the adequacy of all sign videos with respect to the source texts as there are likely trade offs in adequacy, naturalness, and comprehensibility.

4.1. Methodology

We sampled fifty source sentences from NIASL2021 and worked with four native signers to generate video translations for each sample following the two procedures outlined above. The four signers do professional work related to sign language.

To measure the effects of prompting, it was important that no signer translated the same source text for both NIA+VID and IMG+TXT, so we used a round-robin assignment method.

In total, signers created 148 videos: 50 NIA videos, 9 VID videos, 50 IMG videos, and 29 TXT videos. We then had two native signers review the videos to find cases where video quality or lack of signer preparation may interfere with evaluations. These videos were re-signed exactly (including hand signs and NMS) according to the original video but with a more stable camera and with the signer having practiced before filming.

We then arranged for nine native signers to evaluate the videos. Three of the evaluators work professionally in sign language translation and annotation with us, one is involved in sign language research, and five work in fields unrelated to sign language. Similar with the translation procedure assignment above, it was crucial that evaluators not review multiple videos corresponding to the same source sentence since this could affect comprehensibility. We used the latin-square method to balance evaluator assignments and guarantee that each video was reviewed at least two times.

We required evaluators to watch an introductory video of a native signer explaining the goal of the research, the importance of honest feedback, and how to interpret the likert items. We also worked with our sign language team to design an online evaluation tool for deaf users. To encourage evaluations without influence from written or spoken language, we removed as much text from the evaluation interface as possible. We replaced the standard likert text prompts with video prompts that play when activated by the mouse cursor. Using text was reported as too confusing and hard to look at, and using continuous video prompts was reported as being too distracting. The likert scale was also based on significant user feedback. Rather than text labels, we used three symbols to augment number labels: a thumbs down over 1, a horizontal thumb over 4, and a thumbs up over 7. The naturalness and comprehensibility prompts translate as "the signing in this video is natural" and "the signing in this video is understandable", respectively. The scale values range from 1 for strongly disagree to 7 for strongly agree. The evaluation interface can be seen in figure 3.

After videos were evaluated, we became aware of a possible quality difference between crowd-sourced translations and in-house translations (see section 3.2.1). To avoid introducing bias into our analysis, we removed samples that used crowd-sourced translations from NIA and VID. This removed a total of nine videos and twenty-seven evaluations from our analysis.

We also arranged for two professional interpreters to evaluate all 148 videos in terms of adequacy with the source texts (i.e., source-based direct assessment). This evaluation used two two-point likert items and one four-point likert item for each video. The first prompt translates to English as "Compared to the Korean, the KSL translation has added content" with a true/false response. The second prompt translates similarly as "Compared to the Korean, the KSL translation has missing content" with identical response values. Finally, the third prompt translates as "The main points of the Korean and the KSL translation are…" with a response of 1 for the same, 2 for slightly different but acceptable, 3 for different and unacceptable, and 4 for very different and unacceptable.

4.2. Results

We collected a total of 304 likert scale evaluations for naturalness and comprehensibility. Raw likert results are summarized in table 1.

We calculated Cronbach's alpha for the two likert items to be 0.889. According to (Nunnally, 1994)'s interpretation for applied research, this is a sufficient level of reliability between the two indicators, and we combined the scores into one aggregate quality score. For hypothesis testing, we applied ordinal logistic modeling with mixed effects to measure the effect of video type on signing quality. For tests between IMG and TXT and between NIA and VID, we limit IMG and NIA to videos matching TXT and VID, respectively. We also present quality z-scores normalized over evaluators in table 2 to build intuition.

Treating video type as a fixed effect and evaluator and source sentence as random effects produced the best fitting model for all four tests. We used Holm-Bonferroni correction for multiple hypothesis testing to recalculate p value thresholds. Models were implemented using the "ordinal" R package, and we used likelihood ratio tests to calculate p values as per (Christensen, 2019).

For H_1, we restricted analysis to IMG (encoded as 0) and TXT (encoded as 1) videos. For H_2, we restricted analysis to NIA (encoded as 0) and VID (encoded as 1) videos. For H_3, we restricted analysis to NIA (encoded as 0) and IMG (encoded as 1) videos. For H_4, we used the combined video sets NIA+VID (encoded as 0) and IMG+TXT (encoded as 1). See table 3 for results.

Regarding adequacy scores, IMG+TXT videos scored higher than NIA+VID on average, but no statistically significant differences could be found, and the estimated effect size (based on Cliff's Delta) is below the minimal small threshold according to both (Vargha and Delaney, 2000) and (Romano et al., 2006).

4.3. Discussion

The mode of scores for all translation videos is six or seven for both likert items. By subdividing our scale into disagreement (responses 1, 2, or 3), neutral (response 4), and agreement (responses 5, 6, and 7), we found that, for naturalness, NIA videos had a 66.33%

Video	Total	1	2	3	4	5	6	7	5+6+7
NIA	101	5.94%	2.97%	6.93%	17.82%	18.81%	**26.73%**	20.79%	66.33%
VID	12	8.33%	0.00%	0.00%	8.33%	25.00%	**33.33%**	25.00%	83.33%
IMG	127	0.00%	3.15%	6.30%	12.60%	19.69%	25.98%	**32.28%**	77.95%
TXT	64	4.69%	1.56%	9.38%	15.63%	18.75%	**29.69%**	20.31%	68.75%
NIA+VID	101	6.93%	1.98%	2.97%	17.82%	20.79%	**26.73%**	22.77%	73.29%
IMG+TXT	133	2.26%	3.01%	9.02%	10.53%	20.30%	**27.82%**	27.07%	75.19%
NIA	101	1.98%	6.93%	5.94%	13.86%	22.77%	23.76%	**24.75%**	71.28
VID	12	0.00%	0.00%	0.00%	16.66%	25.00%	**33.33%**	25.00%	83.33
IMG	127	0.79%	0.00%	8.66%	11.81%	15.75%	23.62%	**39.37%**	78.74
TXT	64	1.56%	6.25%	10.94%	7.81%	12.50%	**32.81%**	28.13%	73.44
NIA+VID	101	1.98%	5.94%	4.95%	12.87%	23.76%	23.76%	**26.73%**	74.25
IMG+TXT	133	1.50%	3.01%	9.02%	10.53%	14.29%	27.82%	**33.08%**	75.19

Table 1: *Top*: Naturalness likert results. *Bottom*: Comprehension likert results. VID and TXT are included for reference, but NIA+VID and IMG+TXT are more informative for comparison. Mode response values are in bold.

Type	Total	mean	std
NIA	101	-0.3051	1.1148
IMG	127	**0.2432**	0.8292
NIA (matched)	12	-0.5439	1.5002
VID (matched)	12	**0.2450**	0.8091
IMG (matched)	64	**0.1923**	0.7584
TXT (matched)	64	-0.0471	1.0399
NIA+VID	101	-0.2114	1.0430
IMG+TXT	133	**0.1257**	0.9746

Table 2: Signing quality z scores (calculated over evaluator). For comparison, scores are grouped by translation step, and high scores are presented in **bold**.

rate of agreement while VID and IMG (both created from text-free prompts) had an agreement rate of over 75%. Furthermore, NIA agreement for naturalness increased to over 73% after text-free correction was introduced (NIA+VID). On the other hand, IMG agreement dropped slightly to 75.19% when the text-aware correction was introduced (IMG+TXT). While agreement for comprehensibility scores follows the same trend, it did not vary as drastically.

Based on the above and on user-normalized z-scores for the aggregate signing quality score, all of our hypotheses seem to be supported. However, statistical tests revealed that we can reject the null hypotheses only for H_3 and H_4 and not for H_1 or H_2.

Given that there was no loss in adequacy, it is clear that text-free prompting produced better translations then text-based prompting (H_3: NIA $<$ IMG), and the IMG+TXT procedure produced better translations than those from the NIA+VID procedure (H_4: NIA+VID $<$ IMG+TXT). Both produced better translations on average than NIA translations.

5. Conclusion

We introduced NIASL2021, providing an overview of the dataset, the collection methodology, and challenges. We then provided an evaluation of the translation methodology used for NIASL2021. We found that text-free prompting produced better translations than text-based prompting. We recommend the following for future data collection projects:

1. Prompting from visual media. Text-to-image distillation can be used for small projects or when a standardized rubric can be developed.

2. (If text-based prompts are used) introducing an evaluation step where the evaluator does not have access to the source text.

6. Acknowledgements

This work was supported by the Bio Industry Core Technology Development Project funded by the Korean Ministry of Trade, Industry, and Energy (MOTIE, Korea) [Grant Number: 20014406], supported by the Data Construction Business for AI funded by the National Information Society Agency (NIA, Korea) [Grant Number: 69], and supported by the Citizen-Customized Life Safety Technology Development Program funded by the Ministry of the Interior and Safety (MOIS, Korea) [Grant Number: 2021-MOIS61-003].

7. Bibliographical References

AVASAG. (Accessed: 2022-04-04). Avatar-basierter sprachassistent zur automatisierten gebärdenübersetzung. *https://www.avasag.de/* by AVASAG 2022.

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Koehn, P., and Monz, C. (2018). Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages

H_i	Coeff	p-value	Threshold	Cliff's Delta*	Reject NH†
H_1	-0.400	$= 0.259$	0.05	0.1474 (small)	No
H_2	1.528	$= 0.0854$	0.025	0.4000 (medium)	No
H_3	0.986	$= 0.0001$	0.0125	0.1885 (small)	Yes
H_4	0.6445	$= 0.0105$	0.0167	0.0830 ($<$ small)	Yes

Table 3: Regression results.
*Interpretation based on (Romano et al., 2006). †If the null hypothesis is rejected, we conclude that H_i is correct.

272–303, Belgium, Brussels, October. Association for Computational Linguistics.

Camgoz, N. C., 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 (CVPR)*, June.

Camgöz, N. C., Saunders, B., Rochette, G., Giovanelli, M., Inches, G., Nachtrab-Ribback, R., and Bowden, R. (2021). Content4all open research sign language translation datasets. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5. IEEE.

Christensen, R. H. B. (2019). A tutorial on fitting cumulative link mixed models with clmm2 from the ordinal package. *Tutorial for the R Package ordinal https://cran.r-project.org/web/packages/ordinal*, 1.

Duarte, A. C., Palaskar, S., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., and Giró-i-Nieto, X. (2020). How2sign: A large-scale multimodal dataset for continuous american sign language. *CoRR*, abs/2008.08143.

EASIER. (Accessed: 2022-04-04). Intelligent automatic sign language translation. *https://www.project-easier.eu/* by EASIER PROJECT 2021-2023.

Filhol, M. and Hadjadj, M. N. (2018). Elicitation protocol and material for a corpus of long prepared monologues in sign language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Hong, S.-E., Hanke, T., König, S., Konrad, R., Langer, G., and Rathmann, C. (2009). Elicitation materials and their use in sign language linguistics. In *Poster presented at the Workshop "Sign Language Corpora: Linguistic Issues" in London*.

Johnston, T. and Schembri, A. C. (1999). On defining lexeme in a signed language. *Sign language & linguistics*, 2(2):115–185.

Kim, J.-H., Hwang, E. J., Cho, S., Lee, D. H., and Park, J. C. (2022). Sign language production with avatar layering: A critical use case over rare words. In *Language Resources and Evaluation Conference*.

Kita, S., Gijn, I. v., and Hulst, H. v. d. (1997). Movement phases in signs and co-speech gestures, and their transcription by human coders. In *International Gesture Workshop*, pages 23–35. Springer.

Konrad, R., Hanke, T., Langer, G., König, S., König, L., Nishio, R., and Regen, A. (2020). Öffentliches DGS-Korpus: Annotationskonventionen / Public DGS Corpus: Annotation conventions. Project Note AP03-2018-01, DGS-Korpus project, IDGS, Universität Hamburg, Hamburg, Germany.

Koppel, M. and Ordan, N. (2011). Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, page 1318–1326, USA. Association for Computational Linguistics.

Kumar, V., Choudhary, A., and Cho, E. (2020). Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

Nishio, R., Hong, S.-E., König, S., Konrad, R., Langer, G., Hanke, T., and Rathmann, C. (2010). Elicitation methods in the dgs (german sign language) corpus project. In *sign-lang@ LREC 2010*, pages 178–185. European Language Resources Association (ELRA).

Nunnally, J. C. (1994). *Psychometric theory 3E*. Tata McGraw-hill education.

Romano, J., Kromrey, J. D., Coraggio, J., Skowronek, J., and Devine, L. (2006). Exploring methods for evaluating group differences on the nsse and other surveys: Are the t-test and cohen'sd indices the most appropriate choices. In *annual meeting of the Southern Association for Institutional Research*, pages 1–51. Citeseer.

SignON. (Accessed: 2022-04-04). Sign language translation mobile application and open communications framework. *https://signon-project.eu/* by SignON PROJECT 2021-2023.

Vargha, A. and Delaney, H. D. (2000). A critique and improvement of the cl common language effect size statistics of mcgraw and wong. *Journal of Educational and Behavioral Statistics*, 25(2):101–132.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In Nicoletta Calzolari, et al., editors, *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559, Genoa, Italy, May. European Language Resources Association (ELRA).