Automatic Annotation Elaboration as Feedback to Sign Language Learners

Alessia Battisti and **Sarah Ebling**
Department of Computational Linguistics
University of Zurich, Switzerland
{battis,ebling}@cl.uzh.ch

Abstract

Beyond enabling linguistic analyses, linguistic annotations may serve as training material for developing automatic language assessment models as well as for providing textual feedback to language learners. Yet these linguistic annotations in their original form are often not easily comprehensible for learners. In this paper, we explore the utilization of GPT-4, as an example of a large language model (LLM), to process linguistic annotations into clear and understandable feedback on their productions for language learners, specifically sign language learners.

1 Introduction

Annotating linguistic data is a complex task, presenting ongoing challenges such as interpreting ambiguities and accounting for annotators' perceptions (Basile et al., 2021). In the context of sign languages, this complexity is increased by the absence of common writing systems and codified grammars (Baker et al., 2016), along with the challenge posed by the simultaneous production of manual and non-manual components[1] in expressing information.

Annotating sign language data is still a human-based and extremely time-consuming process. As evidenced by the ongoing German Sign Language (*Deutsche Gebärdensprache*, DGS) Corpus project, the annotation task proves to be highly labor-intensive, demanding approximately up to 600 minutes to transcribe and annotate a single minute of signing (Hanke, 2017).

These challenges slow down advancements in (semi-)automatic annotation of sign language data. No computational tools at a production stage are currently capable of supporting the process of generating (high-quality) annotations as part of a semi-automatic setting.

Recent years have seen the rise of Large Language Models (LLMs), enabling the annotation of large textual datasets. LLMs have proven effective in reliably annotating data by supporting human annotators (Gilardi et al., 2023). This concept of LLM-based annotation has extended to language teaching and assessment (Kasneci et al., 2023; Mahlow, 2023). In this area, efforts have been made to provide language learners with formative feedback by processing data annotations, although only tested with written learner data (Caines et al., 2023).

In this paper, we align with this latter research area and explain our idea of using linguistic annotations for providing feedback to sign language learners with the assistance of a large language model. First, we introduce in Section 2 the process of annotating sign language (learner) data as well as the first approaches to using LLMs for providing feedback to learners based on previous written annotation. In Section 3, we outline the annotation process and the annotation scheme. We provide a brief explanation of how our annotation process works, to make it accessible to a non-expert audience. Our goal is to illustrate the steps necessary to generate formative feedback.

Then, in Section 4, we present an experiment by employing GPT-4 (OpenAI, 2023) to process the linguistic annotations from data of our corpus of continuous sign language learner productions. Leveraging GPT-4 as an instance of LLMs, our goal is to transform the linguistic annotations into clear and understandable feedback to sign language learners. An evaluation of the feedback with sign language learners is also presented. The findings suggest that our initial approach holds promise in aiding sign language learning.

[1] Manual and non-manual components represent the phonological linguistic units that differentiate signs. The manual components comprise four main parameters: hand shape, orientation, position, and movement. Non-manual components consist of the movement of eyebrows, mouth, nose, gaze, and position and movement of the head and upper body.

Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII), pages 46–60

2 Related Work

2.1 Annotating Sign Language (Learner) Data

Sign language annotation is a human-based process that consists of two steps: transcription, where a written version of the signed production is created, and annotation, which enriches video data with additional information, such as linguistic features (Konrad, 2011). For this task, in the context of sign languages, expert annotators employ annotation software such as ELAN (Crasborn and Sloetjes, 2008), iLex (Hanke and Storz, 2008), or SignStream (Neidle et al., 2001). These tools do not offer automatic annotation, that is, they do not automatically segment and label the video stream.

Glosses are commonly used as semantic labels of signs, written in capital letters and corresponding to the base form of a word in the surrounding spoken language[2] (Johnston, 2010). They find extensive use in (automatic) Sign Language Processing (SLP), particularly in the domain of Sign Language Translation (SLT) (Müller et al., 2023). The Hamburg Notation System (HamNoSys; Prillwitz, 1989) is a transcription system designed for representing the form of signs, employing approximately 200 symbols to depict the phonetic parameters of signs.

While recent years have witnessed various efforts in SLP introducing methods for automatic data annotation, these approaches are often language-dependent or target only one specific aspect of annotation (e.g., De Sisto et al., 2021; Mukushev et al., 2022, for sign segmentation; Bull et al., 2020 for sentence segmentation; Varol et al., 2021, for sentence alignment; Östling et al., 2015, for part-of-speech tagging; Chaaban et al., 2021 for non-manual segmentation and sign segmentation). Nevertheless, these tools currently either remain unavailable or are not well-suited for automated data annotation without extensive human post-editing.[3]

None of these studies specifically address the processing of sign language data originating from language learners. In recent years, research into second language acquisition of sign languages (SSLA) has increased, proposing various datasets from non-native signers (L2) (Schönström, 2021). The L2 data undergoes an initial annotation comparable to the process applied to native signer (L1) data, including the addition of glosses and HamNoSys information (see above), for example. Following this, error annotation is applied to highlight deviations from canonical forms or disfluencies, a common practice also used in the study of spoken language learning (Gilquin and De Cock, 2011).

On the one hand, deviations are annotated and analyzed at a single sign level, focusing on individual glosses and manual errors (Rosen, 2004; Ortega and Morgan, 2015; Ebling et al., 2021). On the other, deviations are labeled and analyzed at a sentence level, highlighting the need for annotating non-manual components (Mesch and Schönström, 2020; Gulamani et al., 2020).

2.2 LLMs Applied to L2 Data

LLMs have demonstrated their effectiveness in reliably supporting human annotators (Gilardi et al., 2023). This capability of LLMs for annotation has been extended to the domain of language teaching and assessment, demonstrating the potential to enhance the teaching and learning experience across various education levels (Kasneci et al., 2023; Mahlow, 2023).

For example, LLMs can automatically annotate and evaluate learners' written work by assigning scores. As highlighted in previous studies, they play a crucial role in providing immediate feedback and explanations of errors (Nagata et al., 2021; Caines et al., 2023). This immediacy has been proven to be more effective for student learning (Steiss et al., 2023). However, it is important to note that the focus of all these studies is limited to writings and automated essay scoring.

Focusing on the use of corpora in language classes, teachers commonly extract insights into both correct and incorrect usage of terms and linguistic constructs from annotated corpora to provide formative feedback to learners. Nagata et al. (2020) present an approach which leverages annotation in existing feedback comments to automatically generate new feedback comments.

The idea of offering sign language learners automatic and immediate feedback during sign language learning assignments is a recent topic that has not been explored to a great extent (Huenerfauth et al., 2015; Hassan et al., 2022).

Research suggests that sign language learners prefer feedback that not only identifies the precise

[2] In this study, the term "spoken language" refers to any language that is not signed, whether expressed in written or oral form.

[3] For a general introduction to sign language processing, refer to Bragg et al. (2019). To explore existing sign language corpora, consult Kopf et al. (2021) (where, albeit, sign language learner corpora are not included). Additionally, for an overview of the various annotation formats employed in signed corpora, refer to Kopf et al. (2022).

moment of an error in a video but also includes detailed written feedback alongside visual cues (Huenerfauth et al., 2017; Hassan et al., 2022).

Existing systems, limited in number, utilize annotations to analyze learner inputs but typically offer binary correct/incorrect feedback or, at most, assign a numerical score (Tarigopula et al., 2022).

In contrast to the research presented above, our work explores the use of GPT-4 for elaborating existing annotations created for sign language linguistics research, reframing them as feedback for sign language learners. The motivation behind this is that individuals not used to linguistic annotations may find them challenging to understand without additional explanation.

3 Annotating the Corpus Data

In this section, we offer an overview of our annotation process, outlining the annotation scheme. This scheme encompasses the features extracted from the annotated data, which are then used to input into the GPT-4 model for our experiments (Section 4).

For our experiment, we utilized data from a Swiss German Sign Language (*Deutschschweizerische Gebärdensprache*, DSGS) corpus including a longitudinal sign language learner sub-corpus and a corresponding sub-corpus of native signers. The L2 data was gathered from the same learners at four collection points separated by six-month intervals between March 2022 and November 2023. Both L1 and L2 videos underwent post-processing and were imported into iLex (Section 2.1).

3.1 Annotators

In the case of low-resourced languages such as sign languages, human expert annotators are not widely available (Mehta and Srikumar, 2023). Our team of annotators comprises two deaf expert annotators, bringing years of experience in teaching and researching sign language to the task, along with two annotators in training. All of them, including the annotators in training, are project team members.

To ensure data integrity, we performed continuous validation and cross-checking of annotated data. Annotations adhere to a four-eyes principle, subject to cross-checks by the two expert annotators. Difficult and divergent cases are discussed in the presence of a sign language linguist. Annotations by annotators in training undergo a double-check and corrections if necessary.

Since the annotations by annotators in training are still under review, we have decided to consider only the sentences annotated by the experts for the experiment introduced in Section 4.

3.2 Transcription and Annotation Scheme

Table 1 presents the features included in the full transcription and annotation scheme underlying the creation of our DSGS corpus. Each feature corresponds to a tier within an iLex transcript. Tiers, in this context, are distinct layers used to encode the simultaneous usage of various information channels or features, such as hands, eyebrows, and mouthing. Figure 3 in Appendix C shows an example of a sentence produced by a DSGS learner, annotated in iLex.

The scheme was designed so as to include annotation at both a segmental and suprasegmental[4] level, including the non-manual components. The scheme aims at capturing the complexity and nature of the co-occurrence of features, including information at higher levels, as indicated in the "Additional information" column (cf. Table 1).

Item refers to an exercise used as elicitation task for the creation of our corpus, such as picture or video retelling, along with exercise boundaries and name. Information in this tier was automatically annotated based on the starting and ending times registered by the video recording software.

Each *Item* contains one to n sentences marked in the *Sentence* tier. Specifically, this tier contains a segmentation of the video into sentence-like units.

Each sentence is then segmented into manual and non-manual components. Annotation of the manual components involves inserting *glosses* and describing the sign form with the four parameters of hand shape, orientation, position, and movement using *HamNoSys* (Ebling et al., 2018).

For non-manual components, annotation consists of labeling linguistic facial or upper body form and movement. The labels assigned to each non-manual feature were based on schemes employed in previous sign language studies (Gabarró-López and Meurant, 2014; Lackner, 2019). These labels were adapted to suit DSGS and the objectives of the corpus. In sum, we defined 81 labels for mouth gestures, comprising 57 labels for lip form and movement, twelve for identifying cheek movement and shape, and 22 for tongue movement (e.g., *upper lip left raised*). Concerning the nose, we defined

[4]The term "suprasegmental" is employed as these components constitute a layer atop the segmental layer (Pfau, 2017).

Levels	Manual components	Non-manual components (nmc)	Additional information	Error annotations
Item	Gloss: right hand (rh)	Mouthing	Topic/focus	Parameter(s) different (rh) + acceptability
Sentence	Gloss: left hand (lh)	Mouth gesture	Function	Parameter(s) different (lh) + acceptability
	Gloss: both hands (bh)	Nose	Prosody	Parameter(s) different (bh) + acceptability
	HamNoSys: rh	Shoulders + Upper body	Role	Parameter(s) different (nmc) + acceptability
	HamNoSys: lh	Head	*Comment*	Sentence Problem + acceptability sentence
	HamNoSys: bh	Gaze		*Comment*
	HamNoSys: variance rh	Eye lids		
	HamNoSys: variance lh	Eyebrows		

Table 1: Tiers in the transcription and annotation scheme of our DSGS corpus. Each column corresponds to a main annotation block. Each block contains a list of tiers or features.

seven labels (e.g., *wrinkled nose*). We identified 19 labels for upper body and shoulder movement, and 20 labels for head movements (e.g., *nodding head*). Additionally, we established 30 labels for eye gaze, ten for eyelids, and eight for the eyebrow movements (e.g., *raised eyebrows*).[5]

For L2 data, the annotation scheme was expanded to incorporate the tiers outlined in the *Error annotations* column in Table 1. These tiers provide information about deviations from the canonical form for manual and non-manual components, as well as any deviation occurring at the sentence level. For each deviation, annotators assigned a degree of (non-)acceptability (*not acceptable, acceptable, fully acceptable*). This value indicating whether the deviating feature remained comprehensible and to what extent it affected the overall comprehensibility of the sentence. Appendix A presents two versions of a sample signed sentence, one judged as acceptable and the other deemed not acceptable.

4 Using GPT-4 for Sign Language Annotation Elaboration

The primary aim of this experiment was to leverage the ability of GPT-4[6] to produce coherent textual feedback from keywords representing linguistic annotations, intended to be presented to sign language learners. Our final scenario is to incorporate the feedback generated through this process into a prototype sign language assessment system, providing additional feedback alongside visual hints. Figure 1 shows a predecessor prototype system giving feedback on lexical items, i.e., on individual signs. Note that our aim is to provide feedback on the production of continuous sign sentences, thereby naturally also touching on the correctness of production of non-manual components of signing.

[5]For further details on the annotation scheme and labels, please contact the first author.

[6]https://openai.com/gpt-4

4.1 Data

We retrieved the error annotations from iLex, randomly selecting 100 annotated sentences, and structured the input prompt in a JSON format suitable for the GPT-4 model. Each sentence contains between 0 and 12 errors (μ 4.7, σ 2.1).

We chose a direct prompting approach, that is, we provided the model with a single prompt that included comprehensive task details. This contained the task definition, purpose, and the intended target user, simulating how a person would instruct another person to explain a list of errors to a learner. Since the annotations are in German and the generated feedback needs to be in German, we formulated the prompt in the same language. The prompt specifically instructed the model to elaborate on annotations, categorizing them into three types: annotations regarding manual components, non-manual components, and annotations at the sentence level.

We requested the exclusion of any information about the language or input sentence. This was motivated by our observation that GPT-4 had previously consistently provided explanations about sign languages in general. We established precise terminology for the term “sign” to address semantic ambiguity. This ensured that the model used the German word “Gebärde” instead of “Zeichen”, as both are homonyms for “sign” in English. Likewise, we specified the use of the term “Komponenten” instead of “Signale” to refer to non-manual components. The final prompt template is provided in Table 2.

To restrict inappropriate inferences for our purpose, we set the temperature parameter to 0.2. This value allows for prompt rephrasing without becoming repetitive or excessively creative. We kept the other parameters unchanged and set the seed to 42 to guarantee consistent generations.

Tables in Appendix D report examples of prompts and generated outputs.

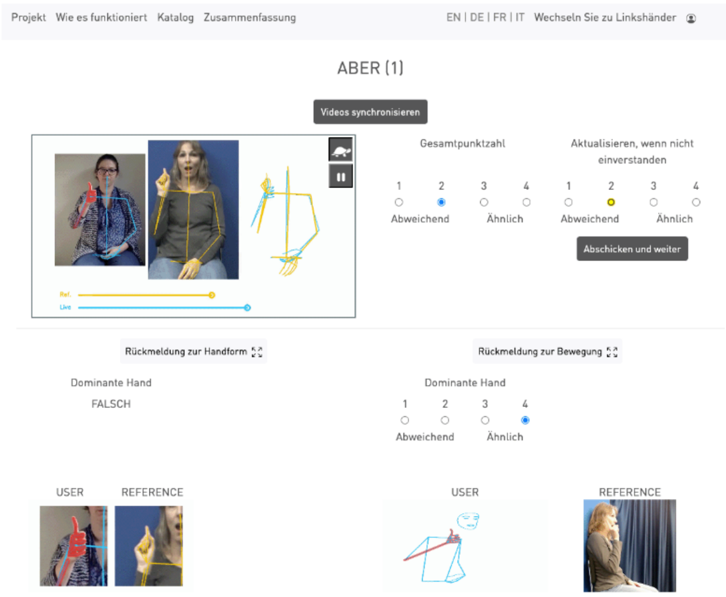

Figure 1: Screenshot displaying the earlier prototype system offering feedback on hand shape (*Handform*) and movement (*Bewegung*) at sign level.

DE: *Analysiere Fehler in einem in Deutschschweizerischer Gebärdensprache (DSGS) produzierten Satz ohne Sprach- oder Satzdetails. Verfasse eine klare, ganz kurze Erklärung wie für einen DSGS-Lernenden. Nutze ‚Gebärde' statt ‚Zeichen', ‚Komponente' statt ‚Signale'. Bitte duzen, auf Förmlichkeiten verzichten. Verzichte auf zusätzliche Informationen. Hier sind die Annotationen: %. Manuelle Fehler: %. Nicht-manuelle Fehler: %. Problem auf Satzebene: %. Akzeptabilität ganzer Satz: %.*
EN: 'Analyze errors in a sentence produced in Swiss German Sign Language (DSGS) without language or sentence details. Write a clear, very short explanation as for a DSGS learner. Use 'sign' instead of 'sign', 'component' instead of 'signals'. Please use first names, avoid formalities. Do without additional information. Here are the annotations: %. Manual errors: %. Non-manual errors: %. Problem at sentence level: %. Whole sentence acceptability: %.'

Table 2: Basic prompt template, where % is a placeholder for annotations. German version is on top, English translation done using DeepL on bottom. Note the wrongly translated sentence 'Use sign instead of sign' due to the semantic ambiguity explained in the section.

4.2 Accuracy Evaluation

To assess the performance, we conducted two distinct evaluations. The first analysis aimed at identifying the primary challenges for the model. In this evaluation, we manually examined the generated text for each prompt, checking the included error information to count the number of *true positives*, i.e., instances in which the model correctly included the error information from the prompt in its output; *false positives*, i.e., cases in which the model incorrectly included information in the output that was not present in the original prompt; and *false negatives*, where the model failed to include the error information from the prompt in the formulated output.

We counted individual error instances, not the overall generated text, since we wanted to observe how the model treated and explained each single error. We then computed precision, recall, and F1-score.

4.3 Learner Evaluation

The second evaluation aimed at assessing the quality of the generated feedback texts from the perspective of the target users, namely sign language learners. Specifically, the goal was to determine the comprehensibility of the generated texts for sign language learners and evaluate their subjective usefulness, with the expectation that our approach will be integrated into the sign language assessment prototype. To achieve this, we invited sign language learners to read and judge the generated texts. We enrolled five volunteers, each with beginner to intermediate level of DSGS.

For every generated text, evaluators were tasked with comparing the output to the errors included in the prompts. They were asked to respond to three questions (translated into English in the following): *Q1*. Is the generated output correct and does it include the expected information? *Q2*. Is the generated output readable? *Q3*. Is the generated output understandable?

Each question was designed to assess a specific aspect of the generated text. The first question focused on investigating the accuracy and completeness of the text. Especially, it considered an output as accurate only when it encompassed all the anticipated information. The second question targeted its fluency, and the third evaluated its clarity in the error presentation. The evaluators had the possibility to add a comment about the text if they deemed it necessary.

All evaluators were instructed to select a value on a 5-point Likert scale, where 1 represented a poor rating (*strongly disagree*) and 5 indicated a very good rating (*strongly agree*). The decision to use an odd scale was taken to allow respondents to choose the midpoint in cases of neutral understanding, without being compelled to extreme evaluations. The anonymized input information and generated outputs, including the number of errors, as well as the evaluations of the outputs are published on Zenodo.[7]

Annotator agreement was measured on the 100 items annotated by all evaluators separately. We calculated the pairwise raw percentage as well as Gwet *AC2*[8] for ordinal data. We decided to use Gwet *AC2* as measure of inter-rater reliability for ordinal and interval measurement because it addresses the limitation of label's distribution of the Krippendorff's α (Feng, 2014; Gwet, 2014). In our data, for each question, evaluators agreed to the extent of 72% to 90% on a single label, that is the label of value 5 (Figure 2). This imbalanced distribution can result in meaningless α coefficients; hence our choice of Gwet *AC2*.

4.4 Results

Accuracy Results By manually analyzing the generated output text, we noticed that in general the texts were correct, fluent and readable, but the model sometimes failed to generate informative elaborations suitable to the given task.

As reported in the Table 3, out of all of the error information, 99% was correctly inserted in the generated output, showing a high level of accuracy.

The model achieved a recall of 93%. While this is a high number, it still indicates that the model missed to insert or wrongly inserted some information on errors. We investigated this value and identified that the system incorrectly provided information about non-manual errors for 32 instances: In general, it stated that the non-manual components were produced correctly instead of acknowledging an error (Example 57 in Table 9). The errors made by GPT-4 in generating the explanations and the non-manual error annotation show a positive correlation that is statistically significant (ρ = 0.16, p-value < 0.001). This implies the existence of a certain ambiguity either in the prompt or in the error annotation, which proves challenging for the LLM to manage.

In cases where the prompt did not contain any error, the model generated irrelevant texts by explaining the range of possible errors in sign language learning (Example 35 in Table 10). This might be due to the limited context of the prompt. In two cases, for example, the model included a description of a manual error that was not originally present in the prompt (Example 44 in Table 8).

Overall, GPT-4 achieved an F1-score of 0.96, suggesting a good trade-off between precision and recall in elaborating on the error annotations to produce a textual feedback for learners.

Precision	0.99
Recall	0.93
F1	0.96

Table 3: Evaluation scores of the GPT-4-generated outputs calculated on the single error instances.

Learner Evaluation Results To better understand the perceived fluency and the clarity of the texts on the part of the sign language learners, we calculated the percentage of the Likert values assigned by the learner evaluators to each question.

As shown in Figure 2, the three aspects obtained high percentages for the highest values for all three aspects of correctness, fluency, and clarity. In particular, 90.3% of the answers to the second aspect obtained a value of 5. Consistent with these percentages, the Pearson correlations between the investigated aspects were highly significant. This suggests that the texts with the correct representations of the content were also evaluated as being more readable and comprehensible by the learners (*Q1-Q2*: ρ = 0.32, p-value < 0.001; *Q1-Q3*: ρ = 0.60, p-value < 0.001; *Q2-Q3*: ρ = 0.34, p-value < 0.001).

Overall reliability was 0.88 (Gwet's *AC2*), calculated as the mean of the values for the individual

[7] https://zenodo.org/doi/10.5281/zenodo.10567377

[8] https://irrcac.readthedocs.io/en/latest/index.html

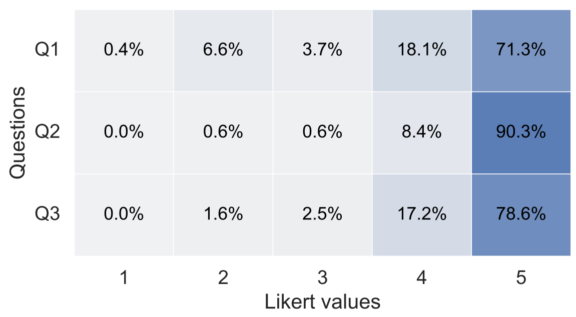

Figure 2: Percentage values of the answers given by the evaluators. Q1: accuracy and completeness of the text; Q2: fluency of the text; Q3: clarity in error presentation.

aspects: 0.83, 0.95, and 0.87 for *Q1*, *Q2*, and *Q3*, respectively. The average pairwise raw percentage estimated on the evaluated items was 58.6% for *Q1*, 79.7% for *Q2*, and 64% for *Q3*. The lowest label, i.e., 1, was never assigned to *Q2* and *Q3*. This label was assigned only by one evaluator to *Q1*. Most discrepancies in the evaluations were observed between labels 4 and 5, indicating that the nuances between a value of 4 and 5 were poorly defined. Specific scores between each pair of evaluators can be found in Tables 4, 5, and 6 in Appendix B.

A qualitative analysis of the comments revealed common patterns in the positive reception of supportive statements such as *Keep up the good work!* present in the output generated by the model (Example 4 in Table 9). Even though we did not specifically ask for such comments in the initial prompt, the model produced them. This observation led us to reflect on the importance of incorporating a more personal touch in the feedback generated for learners.

Less appreciated by the evaluators were phrases that asked learners to pay attention to the correct repetition of a hand movement or shape of the sign as shown in an (invented) example video. Similarly, they criticized the lack of specificity in the description of errors concerning non-manual components. Besides, the evaluators negatively commented on the length of the outputs (length in sentences: μ 7.9, σ 3.2; length in tokens: μ 105.4, σ 43.8). During the creation of the prompt and the experimentation to find suitable parameters for our experiment, we had attempted to limit the number of output tokens; however, the model tended to cut off the feedback, especially when there were numerous errors to explain within the prompt.

5 Conclusion and Outlook

In this paper, we have presented the multifaceted challenges associated with annotating sign language data. We have covered the steps included in the annotation process and summarized the main features in a sign language annotation scheme.

We have also reported on an experiment using GPT-4 to elaborate on linguistic error annotations from the scheme to provide more comprehensible feedback to sign language learners (cf. Example 2 in Table 7). The rationale behind this experiment was that the annotation reprocessing offers an opportunity to offer immediate feedback to sign language learners, enhancing their learning experience.

The evaluation results demonstrated the successful application of GPT-4 in this task. They also obviated a need for more concision, evidence, and specificity in the generated error elaborations. The level of details of our annotation scheme may have constrained the quality of the generated feedback, leading to inconsistencies in GPT-4's treatment of non-manual errors, as shown in Examples 4 and 57 in Table 9. A preliminary cleaning step and paraphrasing of the linguistic annotations could potentially address these issues.

In a practical scenario, GPT-4 can be guided to provide additional explanations or customized support by using clear prompts that incorporated human-written explications. Within the linguistic annotations of the corpus, we offer insights into errors tied to specific grammatical constructs. A future enhancement could involve instructing GPT-4 to generate feedback comments explaining linguistic concepts and grammatical structure, moving beyond solely error-focused elaborations. This shift could contribute to a more comprehensive and educational feedback system, aiding sign language learners in grasping the underlying linguistic principles. This aspect could be evaluated by performing a study comparing feedback generated by sign language experts with that produced by GPT-4.

In future research, we also aim to explore few-shot methods to improve the model's capability of elaboration. We tried different promptings and adjusted settings. Even though we aimed for a clear prompt, the choice of the 0.2 temperature parameter was somewhat arbitrary. Future work could explore different parameter settings.

Lastly, in this paper, the numbers related to annotators, the amount of annotated data and evaluators may seem small compared to the data commonly used in Natural Language Processing (NLP) tasks employing deep learning methods. Yet this paper aims at enhancing knowledge and awareness of annotation efforts in sign languages, presently recognized as under-resourced languages (Joshi et al., 2020). It sheds light on how the community addresses the demands for large, high-quality, and annotated datasets required for technological progress.

Ethics Statement

Ensuring fair and respectful collaboration is a fundamental aspect of our ethical commitment to this research initiative. Within this project, expert annotators are members of the team and receive fair compensation for their valuable contributions. The evaluators who assessed the GPT-4 outputs participated voluntarily, retaining the option to withdraw at any point and the freedom to reconsider their involvement in the evaluation work.

Acknowledgments

This work was funded through the Swiss National Science Foundation (SNSF) Sinergia project "Scalable Multimodal Sign Language Technology for Sign Language Learning and Assessment II" (SMILE II) (grant agreement no. CRSII5_193686) and the Swiss Innovation Agency (Innosuisse) flagship IICT (PFFS-21-47). The authors would like to thank the learners for their assistance in evaluating the GPT-4-generated texts and the reviewers for providing valuable feedback.

References

A. Baker, B. van den Bogaerde, and R. Pfau. 2016. *The Linguistics of Sign Languages: An Introduction*. John Benjamins Publishing Company.

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a perspectivist turn in ground truthing for predictive computing. *CoRR*, abs/2109.04270.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 16–31, New York, NY, USA. Association for Computing Machinery.

Hannah Bull, Michèle Gouiffès, and Annelies Braffort. 2020. Automatic Segmentation of Sign Language into Subtitle-Units. In *Computer Vision – ECCV 2020 Workshops*, Lecture Notes in Computer Science, pages 186–198, Cham. Springer International Publishing.

Andrew Caines, Luca Benedetto, Shiva Taslimipoor, Christopher Davis, Yuan Gao, Oeistein Andersen, Zheng Yuan, Mark Elliott, Russell Moore, Christopher Bryant, Marek Rei, Helen Yannakoudakis, Andrew Mullooly, Diane Nicholls, and Paula Buttery. 2023. On the application of large language models for language teaching and assessment technology.

Hussein Chaaban, Michèle Gouiffès, and Annelies Braffort. 2021. Automatic annotation and segmentation of sign language videos: Base-level features and lexical signs classification. In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2021, Volume 5: VISAPP, Online Streaming, February 8-10, 2021*, pages 484–491. SCITEPRESS.

Onno Crasborn and Han Sloetjes. 2008. Enhanced ELAN functionality for sign language corpora. In *Proceedings of LREC 2008, Sixth International Conference on Language Resources and Evaluation*.

Mirella De Sisto, Dimitar Shterionov, Irene Murtagh, Myriam Vermeerbergen, and Lorraine Leeson. 2021. Defining meaningful units. Challenges in sign segmentation and segment-meaning mapping (short paper). In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 98–103, Virtual. Association for Machine Translation in the Americas.

Sarah Ebling, Necati Cihan Camgoz, Penny Boyes Braem, Katja Tissi, Sandra Sidler-Miserez, Stephanie Stoll, Simon Hadfield, Tobias Haug, Richard Bowden, Sandrine Tornay, Marzieh Razavi, and Mathew Magimai-Doss. 2018. SMILE Swiss German sign language dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4221–4229, Miyazaki, Japan. European Language Resources Association (ELRA).

Sarah Ebling, Katja Tissi, Sandra Sidler-Miserez, Cheryl Schlumpf, and Penny Boyes Braem. 2021. Single-parameter and parameter combination errors in L2 productions of Swiss German Sign Language. *Sign Language & Linguistics*, 24(2):143–181.

Charles Feng. 2014. Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 1:1–10.

Sílvia Gabarró-López and Laurence Meurant. 2014. When nonmanuals meet semantics and syntax: a practical guide for the segmentation of sign language discourse. In *6th Workshop on the Representation and*

Processing of Sign Languages: Beyond the Manual Channel, Reykjavik, Iceland.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Gaëtanelle Gilquin and Sylvie De Cock. 2011. Errors and disfluencies in spoken corpora: Setting the scene. *International Journal of Corpus Linguistics*, 16(2):141–172. Publisher: John Benjamins Type: Journal Article.

Sannah Gulamani, Chloë Marshall, and Gary Morgan. 2020. The challenges of viewpoint-taking when learning a sign language: Data from the 'frog story' in British Sign Language. *Second Language Research*, 38(1):55–87.

Kilem Li Gwet. 2014. *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters*, fourth edition edition. Advances Analytics, LLC, Gaithersburg, Md.

Thomas Hanke. 2017. Wörterbuch ohne Wörter? Zum Entstehen eines Wörterbuches der Deutschen Gebärdensprache. In *Jahrbuch der Heidelberger Akademie der Wissenschaften für 2016*, pages 84–88. Universitätsverlag Winter, Heidelberg, Germany.

Thomas Hanke and Jakob Storz. 2008. iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC)*, pages 64–67, Marrakesh, Marocco. ELRA.

Saad Hassan, Sooyeon Lee, Dimitris Metaxas, Carol Neidle, and Matt Huenerfauth. 2022. Understanding ASL Learners' Preferences for a Sign Language Recording and Automatic Feedback System to Support Self-Study. In *The 24th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–5, Athens Greece. ACM.

Matt Huenerfauth, Elaine Gale, Brian Penly, Sree Pillutla, Mackenzie Willard, and Dhananjai Hariharan. 2017. Evaluation of Language Feedback Methods for Student Videos of American Sign Language. *ACM Transactions on Accessible Computing*, 10(1):2:1–2:30.

Matt Huenerfauth, Elaine Gale, Brian Penly, Mackenzie Willard, and Dhananjai Hariharan. 2015. Comparing methods of displaying language feedback for student videos of american sign language. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, ASSETS '15, page 139–146, New York, NY, USA. Association for Computing Machinery.

Trevor Johnston. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora:. *International Journal of Corpus Linguistics*, 15(1):106–131. Publisher: John Benjamins Publishing Company.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Reiner Konrad. 2011. *Die lexikalische Struktur der Deutschen Gebärdensprache im Spiegel empirischer Fachgebärdenlexikographie*. Ph.D. thesis.

Maria Kopf, Marc Schulder, and Thomas Hanke. 2021. Overview of Datasets for the Sign Languages of Europe.

Maria Kopf, Marc Schulder, Thomas Hanke, and Sam Bigeard. 2022. Specification for the Harmonization of Sign Language Annotations.

Andrea Lackner. 2019. Describing Nonmanuals in Sign Language. In Andrea Lackner, editor, *Grazer Linguistische Studien*, volume 91, pages 45–103. University of Graz, Graz, Austria.

Cerstin Mahlow. 2023. Large Language Models and Artificial Intelligence as Tools for Teaching and Learning Writing. *Osnabrücker Beiträge zur Sprachtheorie*, 101:175–196.

Maitrey Mehta and Vivek Srikumar. 2023. Verifying annotation agreement without multiple experts: A case study with Gujarati SNACS. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10941–10958, Toronto, Canada. Association for Computational Linguistics.

Johanna Mesch and Krister Schönström. 2020. Use and acquisition of mouth actions in L2 sign language learners: A corpus-based approach. *Sign Language & Linguistics*, 24(1):36–62. Publisher: John Benjamins Publishing Company.

Medet Mukushev, Aigerim Kydyrbekova, Vadim Kimmelman, and Anara Sandygulova. 2022. 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.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.

Ryo Nagata, Masato Hagiwara, Kazuaki Hanawa, Masato Mita, Artem Chernodub, and Olena Nahorna. 2021. Shared task on feedback comment generation for language learners. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 320–324, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Ryo Nagata, Kentaro Inui, and Shin'ichiro Ishikawa. 2020. Creating corpora for research in feedback comment generation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 340–345, Marseille, France. European Language Resources Association.

Carol Neidle, Stan Sclaroff, and Vassilis Athitsos. 2001. SignStream: A tool for linguistic and computer vision research on visual-gestural language data. *Behavior Research Methods, Instruments, & Computers*, 33(3):311–320.

OpenAI. 2023. Gpt-4 technical report.

Gerardo Ortega and Gary Morgan. 2015. Phonological Development in Hearing Learners of a Sign Language: The Influence of Phonological Parameters, Sign Complexity, and Iconicity: Phonological Development in Sign L2 Learners. *Language Learning*, 65(3):660–688.

Robert Östling, Carl Börstell, and Lars Wallin. 2015. Enriching the Swedish Sign Language corpus with part of speech tags using joint Bayesian word alignment and annotation transfer. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 263–268, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

Roland Pfau. 2017. Non-manuals and tones : a comparative perspective on suprasegmentals and spreading. *Linguística : Revista de Estudos Linguísticos da Universidade do Porto*, 11:19–58.

S. Prillwitz. 1989. *HamNoSys Version 2.0. Hamburg Notation System for Sign Languages: An Introductory Guide*. Intern. Arb. z. Gebärdensprache u. Kommunik. Signum Press.

Russel S. Rosen. 2004. Beginning L2 production errors in ASL lexical phonology: A cognitive phonology model. *Sign Language & Linguistics*, 7(1):31–61.

Krister Schönström. 2021. Sign languages and second language acquisition research: An introduction. *Journal of the European Second Language Association*.

Jacob Steiss, Tamara P Tate, Steve Graham, Jazmin Cruz, Michael Hebert, Jiali Wang, Youngsun Moon, Waverly Tseng, and mark w uci. 2023. Comparing the quality of human and chatgpt feedback on students' writing.

Neha Tarigopula, Sandrine Tornay, Skanda Muralidhar, and Mathew Magimai-Doss. 2022. Towards accessible sign language assessment and learning. In *International Conference on Multimodal Interaction, ICMI 2022, Bengaluru, India, November 7-11, 2022*, pages 626–631. ACM.

Gul Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2021. Read and Attend: Temporal Localisation in Sign Language Videos. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16852–16861, Nashville, TN, USA. IEEE.

A Annotation Example

Below, we provide an sample sentence, "I am not hard-of-hearing", produced by two DSGS learners. In the first version, the sentence was deemed acceptable because both head and eyebrow movements were executed correctly. In the second version, non-manual components are either missing or used incorrectly (e.g., the head is moved in correspondence with the adjective, not with manual negation), rendering the sentence incorrect.

```
Version 1:
      shaking head    shaking head
                    (furrowed eyebrows) ||
IX-1  NICHT ++  |   SCHWERHOERIG    ||
I      not          hard-of-hearing

Version 2:
                       shaking head
                                    ||
IX-1  NICHT ++  |   SCHWERHOERIG    ||
I      not          hard-of-hearing

DE: Ich bin nicht schwerhörig
EN: I   am  not   hard-of-hearing
```

B Percentage Agreement

	A_0	**A_1**	**A_2**	**A_3**	**A_4**
A_0	-				
A_1	**68.7**	-			
A_2	60.7	73.9	-		
A_3	54.0	62.6	59.6	-	
A_4	51.5	49.0	44.3	46.5	-

Table 4: Raw percentage agreement across evaluators on *Q1*, the question about accuracy and completeness aspect.

	A_0	**A_1**	**A_2**	**A_3**	**A_4**
A_0	-				
A_1	97.0	-			
A_2	84.3	85.2	-		
A_3	96.0	**99.0**	86.5	-	
A_4	66.7	69.4	67.0	70.7	-

Table 5: Raw percentage agreement across evaluators on *Q2*, the fluency aspect.

	A_0	**A_1**	**A_2**	**A_3**	**A_4**
A_0	-				
A_1	**81.8**	-			
A_2	71.9	76.1	-		
A_3	54.0	52.5	53.9	-	
A_4	74.7	78.6	67.0	53.5	-

Table 6: Raw percentage agreement across evaluators on *Q3*, the question about the clarity aspect.

C iLex Example

In the Figure, we show an example of a DSGS sentence produced by a learner, annotated in iLex. The columns represents the tiers, namely the features of Table 1.

D Output Examples

In the following tables, we will display some prompts and their corresponding generated outputs. The German versions are presented at the top. The English translations of the prompts and the generated outputs were automatically done using DeepL[9] (at the bottom).

[9] https://www.deepl.com/translator

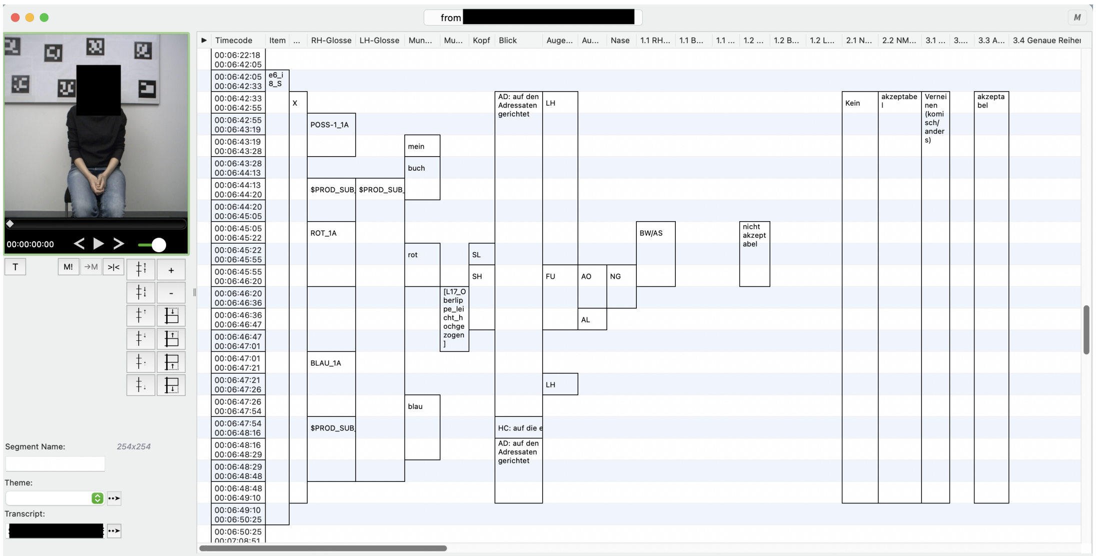

Figure 3: Example of a sentence annotation in iLex, with the transcript name and participant's face anonymized.

	Prompt	**Output**
2	*Analysiere Fehler in einem in Deutschschweizerischer Gebärdensprache (DSGS) produzierten Satz ohne Sprach- oder Satzdetails. Verfasse eine klare, ganz kurze Erklärung wie für einen DSGS-Lernenden. Nutze ‚Gebärde' statt ‚Zeichen', ‚Komponente' statt ‚Signale'. Bitte duzen, Förmlichkeiten verzichten. Verzichte auf zusätzliche Informationen. Hier sind die Annotationen: Manuelle Fehler: 1. Glosse SALAT_1A ist akzeptabel, Fehler in der Handbewegung. Nicht-manuelle Fehler: Fehler in Augenbrauenbewegung, Mundbild, Mundform, Kopfbewegung ist nicht akzeptabel. Problem auf Satzebene: Kein. Akzeptabilität ganzer Satz: akzeptabel.*	*Beim Gebärden des Satzes mit der Gebärde SALAT_1A ist die Handbewegung nicht ganz richtig. Achte auch auf deine Augenbrauen, Mundbild, Mundform und Kopfbewegung – diese sind nicht korrekt und sollten verbessert werden. Der gesamte Satz ist trotzdem verständlich.*
	'Analyze errors in a sentence produced in Swiss German Sign Language (DSGS) without language or sentence details. Write a clear, very short explanation as for a DSGS learner. Use 'sign' instead of 'sign', 'component' instead of 'signals'. Please use first names, avoid formalities. Do without additional information. Here are the annotations: Manual errors: 1st gloss SALAT_1A is acceptable, error in hand gesture. Non-manual errors: Errors in eyebrow movement, mouth image, mouth shape, head movement is not acceptable. Problem at sentence level: None. Acceptability of whole sentence: acceptable.'	'When signing the sentence with the SALAT_1A gesture, the hand movement is not quite correct. Also pay attention to your eyebrows, mouth image, mouth shape and head movement - these are not correct and should be improved. The whole sentence is still understandable.'

Table 7: Example of generated elaborations judged as good explanation by the learner annotators.

	Prompt	Output
44	*Analysiere Fehler in einem in Deutschschweizerischer Gebärdensprache (DSGS) produzierten Satz ohne Sprach- oder Satzdetails. Verfasse eine klare, ganz kurze Erklärung wie für einen DSGS-Lernenden. Nutze ‚Gebärde' statt ‚Zeichen', ‚Komponente' statt ‚Signale'. Bitte duzen, Förmlichkeiten verzichten. Verzichte auf zusätzliche Informationen. Hier sind die Annotationen:* ***Manuelle Fehler: Kein****. Nicht-manuelle Fehler: Fehler in Augenbrauenbewegung, Kopfbewegung ist nicht akzeptabel. Problem auf Satzebene: Geschwindigkeit oder Fluss des produzierten Satzes. Akzeptabilität ganzer Satz: akzeptabel.*	*In deinem Satz in DSGS gibt es ein paar Fehler:* ***1. Manuelle Fehler: Achte darauf, dass jede Gebärde korrekt geformt und ausgeführt wird.*** *2. Nicht-manuelle Fehler: Deine Augenbrauenbewegung und Kopfbewegung passen nicht zum Inhalt des Satzes. Diese Komponenten sind wichtig, um den Satz richtig zu betonen und zu strukturieren. 3. Problem auf Satzebene: Die Geschwindigkeit oder der Fluss deiner Gebärden ist nicht gleichmässig. Versuche, einen natürlichen Rhythmus zu finden, damit der Satz flüssig wirkt. Obwohl es diese Fehler gibt, ist der Satz insgesamt akzeptabel. Übe weiter, um die Genauigkeit zu verbessern!*
	'Analyze errors in a sentence produced in Swiss German Sign Language (DSGS) without language or sentence details. Write a clear, very short explanation as for a DSGS learner. Use 'sign' instead of 'sign', 'component' instead of 'signals'. Please use first names, avoid formalities. Do without additional information. Here are the annotations: **Manual errors: None**, non-manual errors: error in 1st eyebrow movement, head movement is not acceptable. Sentence level problem: speed or flow of sentence produced. Acceptability of whole sentence: acceptable.'	'There are a few errors in your sentence in DSGS: **1. Manual errors: Make sure that each gesture is formed and executed correctly.** 2. non-manual errors: Your eyebrow movement and head movement do not match the content of the sentence. These components are important to emphasize and structure the sentence correctly. 3. problem at sentence level: the speed or flow of your gestures is not even. Try to find a natural rhythm to make the sentence flow. Although there are these mistakes, the sentence is acceptable overall. Keep practicing to improve accuracy!'

Table 8: Example a of generated hallucination. It includes a manual error (in bold), even though it was not in the original prompt.

	Prompt	Output
4	*Analysiere Fehler in einem in Deutschschweizerischer Gebärdensprache (DSGS) produzierten Satz ohne Sprach- oder Satzdetails. Verfasse eine klare, ganz kurze Erklärung wie für einen DSGS-Lernenden. Nutze ‚Gebärde‘ statt ‚Zeichen‘, ‚Komponente‘ statt ‚Signale‘. Bitte duzen, Förmlichkeiten verzichten. Verzichte auf zusätzliche Informationen. Hier sind die Annotationen: Manuelle Fehler:* ***Nicht-manuelle Fehler: Fehler in Augenbrauenbewegung, ist akzeptabel.*** *Problem auf Satzebene: Kein. Akzeptabilität ganzer Satz: voll akzeptabel.*	*Da keine spezifischen Fehler in den manuellen oder nicht-manuellen Komponenten angegeben sind und der Satz als "voll akzeptabel" bewertet wurde, gibt es nichts zu korrigieren. Deine Gebärden sind korrekt, und die Augenbrauenbewegung ist auch akzeptabel. Mach weiter so!*
	'Analyze errors in a sentence produced in Swiss German Sign Language (DSGS) without language or sentence details. Write a clear, very short explanation as for a DSGS learner. Use 'sign' instead of 'sign', 'component' instead of 'signals'. Please use the first name, avoid formalities. Do without additional information. Here are the annotations: Manual errors: **non-manual errors: error in eyebrow movement, is acceptable.** Sentence-level problem: None. Whole sentence acceptability: fully acceptable.'	'Since there are no specific errors indicated in the manual or non-manual components and the sentence was rated as "fully acceptable", there is nothing to correct. Your gestures are correct, and the eyebrow movement is also acceptable. Keep up the good work!'
57	*Analysiere Fehler in einem in Deutschschweizerischer Gebärdensprache (DSGS) produzierten Satz ohne Sprach- oder Satzdetails. Verfasse eine klare, ganz kurze Erklärung wie für einen DSGS-Lernenden. Nutze ‚Gebärde statt ‚Zeichen‘, ‚Komponente‘ statt ‚Signale‘. Bitte duzen, Förmlichkeiten verzichten. Verzichte auf zusätzliche Informationen. Hier sind die Annotationen: Manuelle Fehler: 1. Glosse GESTERN_1A ist akzeptabel, Fehler in der Aufführungsstelle.2. Glosse ESSEN_1A ist nicht akzeptabel, Fehler in der Handbewegung.3. Glosse SCHLECHT_1A ist nicht akzeptabel, Fehler in der Handbewegung und Aufführungsstelle.4. Glosse IX_1A ist akzeptabel, Fehler in der Handbewegung.* ***Nicht-manuelle Fehler: Fehler in Augenbrauenbewegung, Mundform, ist akzeptabel.*** *Problem auf Satzebene: Kombination der Probleme. Akzeptabilität ganzer Satz: nicht akzeptabel.*	*In deinem Satz gibt es einige Fehler: 1. Bei der Gebärde für "GESTERN" ist die Position, wo du die Gebärde machst, nicht richtig. 2. Die Gebärde für "ESSEN" ist nicht korrekt, weil du die Handbewegung falsch machst. 3. Bei "SCHLECHT" machst du sowohl die Handbewegung als auch die Position falsch. 4. Die Gebärde, die auf etwas zeigt (IX), ist okay, aber die Bewegung deiner Hand ist nicht richtig.* ***Ausserdem hast du einen Fehler bei der Bewegung deiner Augenbrauen, aber deine Mundform ist in Ordnung.*** *Weil diese Fehler zusammenkommen, ist der ganze Satz leider nicht akzeptabel. Achte darauf, die einzelnen Komponenten der Gebärden und ihre Kombination im Satz zu verbessern.*
	'Analyze errors in a sentence produced in Swiss German Sign Language (DSGS) without language or sentence details. Write a clear, very short explanation as for a DSGS learner. Use 'sign' instead of 'sign', 'component' instead of 'signals'. Please use first names, avoid formalities. Do without additional information. Here are the annotations: Manual errors: 1. gloss GESTERN_1A is acceptable, error in performance part.2. gloss ESSEN_1A is not acceptable, error in hand movement.3. gloss SCHLECHT_1A is not acceptable, error in hand movement and performance part.4. gloss IX_1A is acceptable, error in hand movement. **Non-manual errors: error in 1st eyebrow movement, mouth shape, is acceptable.** Problem at sentence level: combination of problems. Acceptability of whole sentence: not acceptable.'	'There are a few mistakes in your sentence: 1. in the gesture for "GESTERN", the position where you make the gesture is not correct. 2. the gesture for "ESSEN" is not correct because you are making the hand movement incorrectly. 3. for "SCHLECHT" you are making both the hand movement and the position incorrectly. 4. the gesture pointing to something (IX) is okay, but the movement of your hand is not correct. You also have a mistake in the movement of your eyebrows, but your mouth shape is fine. Unfortunately, because these errors come together, the whole sentence is not acceptable. Make sure to improve the individual components of the gestures and their combination in the sentence.'

Table 9: Examples of generated elaborations wrongly formulating the errors in the non-manual components.

	Prompt	Output
35	*Analysiere Fehler in einem in Deutschschweizerischer Gebärdensprache (DSGS) produzierten Satz ohne Sprach- oder Satzdetails. Verfasse eine klare, ganz kurze Erklärung wie für einen DSGS-Lernenden. Nutze ‚Gebärde‘ statt ‚Zeichen‘, ‚Komponente‘ statt ‚Signale‘. Bitte duzen, Förmlichkeiten verzichten. Verzichte auf zusätzliche Informationen. Hier sind die Annotationen: Manuelle Fehler: Kein. Nicht-manuelle Fehler: Kein. Problem auf Satzebene: Kein. Akzeptabilität ganzer Satz: Kein.*	*Ohne konkrete Details zum Satz oder den spezifischen Fehlern kann ich keine spezifische Analyse durchführen. Allgemein gesagt, wenn in einem Satz in der Deutschschweizerischen Gebärdensprache (DSGS) Fehler auftreten, können diese in verschiedenen Bereichen liegen: 1. Manuelle Fehler: Diese betreffen die Ausführung der Gebärden mit den Händen. Zum Beispiel könnte die Handform, die Orientierung, die Bewegung oder die Position im Raum falsch sein. 2. Nicht-manuelle Fehler: Diese beziehen sich auf Ausdrücke, die nicht mit den Händen gemacht werden, wie Mimik, Kopfbewegungen oder Körperhaltung, die für die Bedeutung eines Satzes wichtig sein können. 3. Problem auf Satzebene: Hier geht es um die Struktur des Satzes, also wie die Gebärden zusammengesetzt sind. Es könnte ein Problem mit der Reihenfolge der Gebärden geben oder mit der Grammatik der DSGS. 4. Akzeptabilität ganzer Satz: Dies bezieht sich darauf, wie gut der gesamte Satz verstanden wird. Selbst wenn einzelne Gebärden korrekt sind, kann der Satz als Ganzes unklar oder missverständlich sein. Um Fehler zu vermeiden, achte darauf, dass jede Komponente der Gebärde korrekt ist, die nicht-manuellen Signale passend sind, die Satzstruktur der DSGS entspricht und der Gesamtsatz klar und verständlich ist.*
	‘Analyze errors in a sentence produced in Swiss German Sign Language (DSGS) without language or sentence details. Write a clear, very short explanation as for a DSGS learner. Use ’sign’ instead of ’sign’, ’component’ instead of ’signals’. Please use the first name, avoid formalities. Do without additional information. Here are the annotations: Manual errors: None. Non-manual errors: None. Sentence-level problem: None. Whole sentence acceptability: None.’	‘Without concrete details about the sentence or the specific errors, I cannot carry out a specific analysis. Generally speaking, when errors occur in a sentence in Swiss German Sign Language (DSGS), these can be in different areas: 1. Manual errors: These concern the execution of the gestures with the hands. For example, the hand shape, orientation, movement or position in space may be incorrect. 2. non-manual errors: These relate to expressions that are not made with the hands, such as facial expressions, head movements or body posture, which can be important for the meaning of a sentence. 3rd problem at sentence level: This concerns the structure of the sentence, i.e. how the gestures are put together. There could be a problem with the order of the signs or with the grammar of the DSGS. 4. acceptability of the whole sentence: This refers to how well the whole sentence is understood. Even if individual signs are correct, the sentence as a whole may be unclear or misunderstood. To avoid errors, make sure that each component of the gesture is correct, the non-manual signals are appropriate, the sentence structure conforms to DSGS, and the overall sentence is clear and understandable.’

Table 10: Example of a generated hallucination. This is a completely invented text due to the absence of errors in the prompt.