Advancing Annotation for Continuous Data in Swiss German Sign Language

Alessia Battisti[*], **Katja Tissi**[†], **Sandra Sidler-Miserez**[†], **Sarah Ebling**[*]
[*]University of Zurich
Andreasstrasse 15, 8050 Zurich
{battis, ebling}@cl.uzh.ch

[†]University of Teacher Education in Special Needs
Schaffhauserstrasse 239, 8050 Zurich
katja.tissi@hfh.ch | sandysidler@gmail.com

Abstract

This paper presents a transcription and annotation scheme introduced specifically for L1 and L2 continuous data of Swiss German Sign Language, with potential applicability to other sign languages. The scheme includes a novel way of annotating linguistic errors in L2 data, thereby contributing to a deeper understanding of sign language learning. An initial validation approach is outlined, revealing challenges and underscoring the necessity for a more comprehensive method for validating sign language (learner) data. The paper emphasizes the overarching goal of achieving interoperability among sign language corpora and research groups, particularly in advancing sign language data validation techniques.

Keywords: Sign language data, learner corpus, annotation scheme, inter-annotator agreement

1. Introduction

Transcribing and annotating sign language data represents a significant bottleneck in the development of sign language corpora, especially when aiming for substantially sized, well-annotated datasets for automated Sign Language Processing (SLP) tasks. Many challenges in SLP arise not only due to a scarcity of consistent and detailed annotations but also due to the variation in annotation standards and granularity across projects.

In sign language corpus creation, it is crucial for annotation schemes and guidelines to adopt a broader perspective, characterized as “holistic and forward-thinking” by Hodge and Crasborn (2022). In a “holistic” approach, both basic and detailed annotations are combined from the beginning of the annotation process. The former, comparable to transcription (Konrad, 2011), includes segmentation and tokenization, which involves identifying manual actions, usually at the level of lexical units. The latter enriches the transcription with a more detailed level of annotation, such as non-manual actions and potentially grammatical functions. A more comprehensive approach such as this promotes best practices and represents a step towards standardization of signed language corpora.

This paper presents the development of an annotation scheme integrating basic and detailed annotations, designed for multidisciplinary use in sign language linguistics, automatic sign language assessment, and SLP. The development of this scheme was an integral part in constructing a longitudinal corpus of Swiss German Sign Language (*Deutschschweizerische Gebärdensprache*, DSGS) second language (L2) learners, alongside a corpus of native/early learners (L1) of DSGS.

We summarize the process of annotating sign language (learner) data and present the annotation scheme. Given that the data is continuous signing that exceeds the level of individual signs, our scheme primarily focuses on the annotation of non-manual components that sometimes stretch across multiple manual signs. Furthermore, we address the annotation of L2 errors and suggest the potential of our scheme for future annotation of sign language (learner) data to enhance interoperability of datasets and thus facilitate cross-linguistic studies. Finally, we introduce an initial validation approach and preliminary results, highlighting the challenges encountered and the need for a comprehensive validation method for sign language (learner) data.

Section 2 introduces previous work in the area of sign language annotation, with a focus on inter-annotator agreement in sign language data. Section 3 summarizes our annotator process, while Section 4 describes the annotation scheme in details. In Section 5, we outline an initial validation approach on our annotated data.

2. Related Work

2.1. Annotations of Sign Language (Learner) Data

Several attempts have been made to define standards and best practices in sign language data

Proceedings of the 11th Workshop on the Representation and Processing of Sign Languages, pages 1–12
25 May 2024.

annotation (Nonhebel et al., 2004; Johnston, 2010; Schembri and Crasborn, 2010; Cormier et al., 2016). The selection of annotation scheme and the specificity of its labels are frequently influenced by the linguistic theories embraced by the researchers and by their research questions (Hodge and Crasborn, 2022). For instance, lexical frequency and morphosyntactic analysis guide the annotation scheme for the Auslan Corpus (*Australian Sign Language*) (Johnston, 2008), while phonetics and phonology shape the scheme for the NGT Corpus (*Nederlandse Gebarentaal*, Sign Language of the Netherlands; Crasborn et al., 2006-2017).

Kopf et al. (2022) delineates commonalities and differences between annotation conventions as applied to several publicly accessible sign language corpora. In the section dedicated to non-manual components, the authors point out that there are few studies describing the annotation of non-manual activities. Among the most recent works, Johnston (2019) provides detailed insights into the considerations made to annotate the form and the function of these components in Auslan, while Wallin and Mesch (2018) describe how they treated and annotated these activities in the corpus of Swedish Sign Language (*Svenskt teckenspråk*, STS).

Given the importance of these components at the sentence and discourse levels, Gabarró-López and Meurant (2014) explain how to use certain non-manual components, including head nod or movement, eye blink, and gaze, as criteria to facilitate sign language discourse segmentation in French Belgian Sign Language (*Langue des signes de Belgique francophone*, LSFB). Similarly, to describe the components' function at the sentence and discourse levels, Lackner (2019) illustrates their annotation and their potential configurations in Austrian Sign Language (*Österreichische Gebärdensprache*, ÖGS).

However, none of the aforementioned studies specifically address the annotation of manual and non-manual components in sign language learner data. Despite the increased interest in research focusing on sign second language acquisition (SSLA) and the creation of datasets from non-native signers (L2 signers) (Schönström, 2021), management and annotation of L2 data remains an understudied area (Mesch and Schönström, 2018). This is characterized by a lack of guidelines for annotating errors or L2 linguistic structures. In addition to basic or detailed annotations similar to those applied to L1 data, L2 data is typically enriched with annotations that highlight deviations from canonical forms or disfluencies, a common practice also employed in the studies of spoken language learning (Gilquin and De Cock, 2011).

For analyzing the Corpus in Swedish Sign Language as a Second Language (SSLC-L2), Mesch and Schönström (2018) proposed a method to annotate typical L2 structures, which includes conventions for annotating phenomena specific to L2 languages. The authors build upon their previous studies on annotations of non-manual components and errors (Schönström and Mesch, 2014; Mesch et al., 2016).

Until recently, research on SSLA has primarily focused on analyzing individual glosses and manual errors (Rosen, 2004; Ortega and Morgan, 2015; Ebling et al., 2021; Kurz et al., 2023). However, there has been a growing interest in investigating higher-level linguistic constructions, such as sentences or discourse, highlighting the need for annotating non-manual components also for L2. For example, Mesch and Schönström (2020) explored the use of mouth actions in SSLC-L2, while Gulamani et al. (2020) examined the adoption of different viewpoints in British Sign Language (BSL) learners.

2.2. Inter-annotator Agreement in Sign Language Data

None of the above-mentioned studies present an approach for the validation of annotated data. Studies on sign languages either do not report on reliability or provide only superficial ratings of inter-rater agreement (Schembri and Crasborn, 2010). For example, Hodge (2014) conducted a thorough examination of the annotation procedure, where additional annotators reviewed annotations of clause-like expressions by way of re-analysis.

Calculating agreement on sign language data annotations is a complex process that must consider multiple variables, such as the diversity of time spans and labels used.

In the context of annotations on behavioral studies, Andersson and Sandgren (2016) proposed a method called *temporally weighted overlap ratio*, to use with the ELAN annotation software (Wittenburg et al., 2006), to calculate agreement between two annotated events. Considering a certain time span, the authors search for an event in two different annotation transcripts. If an event is found and has the same label for Annotator A and Annotator B, an agreement is calculated based on the time overlap between the two events weighted by the maximum length of the event. This approach can also be applied to measure agreement between two events in a given time span in sign language data.

3. Annotation Process

As mentioned in Section 1, we devised the annotation process and scheme as part of constructing a longitudinal corpus of continuous DSGS L2 production, in parallel with an L1 control corpus. In total, 35 participants were recorded, resulting in approximately 70 hours of recorded data.

The L1 control corpus comprises recordings of ten deaf signers performing the same tasks as the DSGS learners. Examples of tasks include picture or video retelling. We enlisted deaf signers who use DSGS as their primary language and acquired the language at different ages (M=3.8, SD=6.1). Among the 25 L2 participants, 14 were students of a DSGS interpreter training program. We followed these students throughout their language learning journey by recording their language production four times over an 18-month period.

Annotation is carried out by a team comprising two L1 deaf expert annotators with extensive experience in teaching and researching sign language, alongside two L1 deaf annotators-in-training, all of whom are project members. The data is annotated using the iLex software (Hanke and Storz, 2008), allowing for the linkage of all sign tokens in the corpus to their corresponding sign types in the lexicon and propagating any changes to sign types across all transcripts.

Figure 1 illustrates the data processing steps, starting from raw data in the recording phase to the subsequent data annotation rounds. Initially, we pre-process the data and generate transcripts that include selected tiers for both manual and non-manual components, with task boundaries automatically annotated based on recording software timestamps.

The data then undergoes two main rounds of processing. The first round involves segmenting tasks into sentences and sign units, identifying manual and non-manual components for both L1 and L2 data, and labeling the time span for each identified feature. In the second round, deviations from the canonical form are identified and labeled in the L2 data. Additional tiers are added to the L2 transcripts to facilitate marking deviations for both manual and non-manual components. A third round involves cross-checking and validating annotated data applying the four-eyes principle, where 20% of annotated data are re-annotated by the two expert annotators to calculate agreement. Annotations by annotators-in-training undergo double-checking, with corrections made as needed. Disagreements between annotators are discussed with an expert sign language linguist to understand the disagreement factors and resolve differences.

Due to the comprehensive nature of the annotation task and the corpus's extensive volume, only selected tasks of the first two data collection points have been annotated thus far. On average, for both L1 and L2 data, annotators require 30 minutes to annotate a sentence containing six glosses.

Figure 2 displays a sample transcript in iLex for an L2 learner production, showing annotations from the first and second rounds.

4. Annotation Scheme

In developing the annotation scheme, we were faced with the challenge of determining the granularity of the annotation, which is dependent upon the intended application of the corpus.

In our scheme, we aimed to strike a balance between basic and detailed annotation to accommodate an array of future analyses. We have defined various labels for each feature or component and organized these labels into macro categories to establish a coarser annotation level. This coarser level is expected to facilitate SLP tasks and statistical linguistic analyses.

Table 1 presents the main blocks of features covered by our annotation scheme, with each block corresponding to a set of tiers within an iLex transcript. In the following sections, we provide detailed explanations of the tiers included in each main block.

Video
Item / Task
Sentence
Manual components
Non-manual components
Errors
Additional information
Comments

Table 1: Main blocks of tiers in the transcription and annotation scheme.

4.1. Task Level

The initial segmentation of the video stream involves automatically annotating the task starting and ending times, along with the task code, in the **Item** tier.

Following this, each task time span is segmented into sentence-like units, which are labeled within the **Sentence** tier. These units may encompass anywhere from one to *n* sentences.

The segmentation process is subsequently extended to manual and non-manual components within each sentence.

4.2. Manual Components

In general, the most basic level of corpus annotation is tokenization. Tokens pertaining to manual components are identified and segmented within the sentence adhering to a wider segmenting system (Hanke et al., 2012).

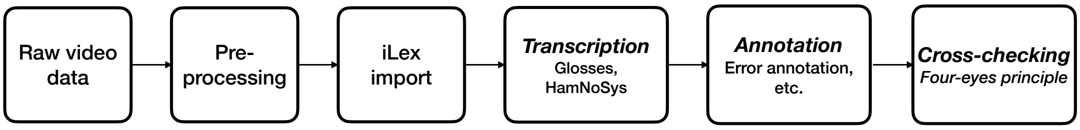
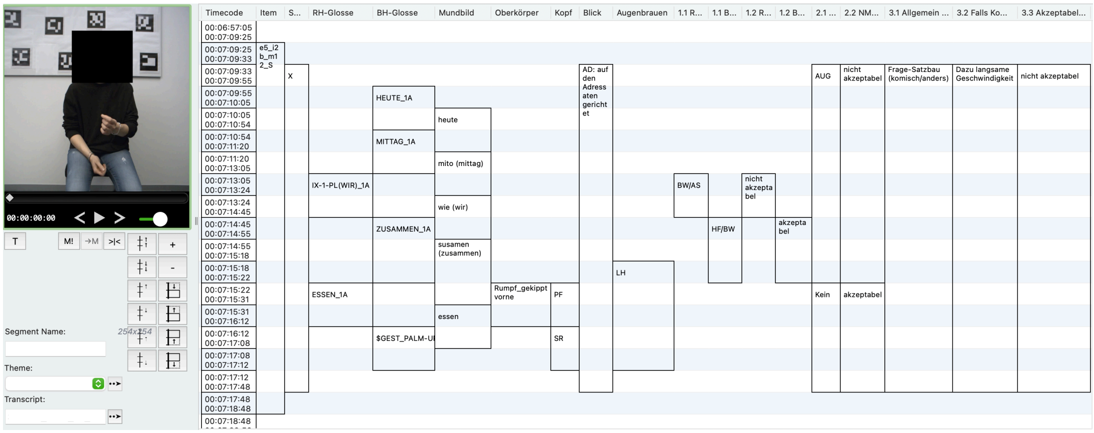


Figure 1: Visualization of the data process from raw data to data annotation.



Timecode	Item	S...	RH-Glosse	BH-Glosse	Mundbild	Oberkörper	Kopf	Blick	Augenbrauen	1.1 R...	1.1 B...	1.2 R...	1.2 B...	2.1 ...	2.2 NM...	3.1 Allgemein ...	3.2 Falls Ko...	3.3 Akzeptabel...
00:06:57:05 00:07:09:25																		
00:07:09:25 00:07:09:33	e5_i2b_m12_S																	
00:07:09:33 00:07:09:55		X						AD: auf den Adressaten gerichtet						AUG	nicht akzeptabel	Frage-Satzbau (komisch/anders)	Dazu langsame Geschwindigkeit	nicht akzeptabel
00:07:09:55 00:07:10:05				HEUTE_1A														
00:07:10:05 00:07:10:54					heute													
00:07:10:54 00:07:11:20				MITTAG_1A														
00:07:11:20 00:07:13:05					mito (mittag)													
00:07:13:05 00:07:13:24			IX-1-PL(WIR)_1A							BW/AS		nicht akzeptabel						
00:07:13:24 00:07:14:45					wie (wir)													
00:07:14:45 00:07:14:55				ZUSAMMEN_1A							HF/BW		akzeptabel					
00:07:14:55 00:07:15:18					susamen (zusammen)													
00:07:15:18 00:07:15:22									LH									
00:07:15:22 00:07:15:31			ESSEN_1A			Rumpf_gekippt vorne	PF							Kein	akzeptabel			
00:07:15:31 00:07:16:12					essen													
00:07:16:12 00:07:17:08				$GEST_PALM-UI			SR											
00:07:17:08 00:07:17:12																		
00:07:17:12 00:07:17:48																		
00:07:17:48 00:07:18:48																		
00:07:18:48																		

Figure 2: Sample transcript in iLex with manual, non-manual, and error annotation tier.

Table 2 outlines the tiers for the manual components included in our scheme. Following identification, manual components are annotated by inserting **identificative glosses** (ID glosses) as semantic notations, and described in their form using the **Hamburg Notation System for Sign Languages** (HamNoSys; Prillwitz, 1989). In using the iLex corpus lexicon system, we are assured of having consistent use of glosses by different annotators. The selection of glosses was motivated by their widespread usage as common semantic labels of signs. In addition, glosses are extensively employed in SLP, particularly in the domain of Sign Language Translation (SLT) (Müller et al., 2023).

In this phase, we distinguish between signs produced with the left or right hand as well as between one-handed and two-handed signs. The tier **Gloss Right Hand (RH)** is annotated for one-handed signs articulated on the right hand, while **Gloss Left Hand (LH)** is annotated for one-handed signs articulated on the left hand. Two-handed signs are annotated in **Gloss Both Hands (BH)**. The hand dominance of the signer is stored in the signer's metadata.

Non-conventionalized signs, like gestures, are annotated similarly to glosses and allocated to the tiers of the hand used for articulation, identified by the affix *GEST_*. Fingerspelling follows the same approach as single signs, annotated with the affix *FA_* to the gloss.

Qualifiers are combined with glosses to indicate variant forms, involving slight differences in the phonological parameters (Konrad et al., 2012). The form variance is reported in the corresponding **HamNoSys variance** tier. For glossing and qualifier addition, we adhere to the glossing conventions[1] of our iLex DSGS instance and those described in Konrad et al. (2012) and Ribeaud and Cicala (2019).

Manual Components
Gloss RH
HamNoSys RH
HamNoSys variance RH
Gloss LH
HamNoSys LH
HamNoSys variance LH
Gloss BH
HamNoSys BH
HamNoSys variance BH

Table 2: Tiers of the manual components. RH: right hand. LH: left hand. BH: both hands.

4.3. Non-manual Components

Non-manual activities undergo detailed annotation in our scheme. Labels for each feature were based on the scheme for non-manual components in Hanke (2001), then determined on the most frequently annotated forms in previous DSGS studies and compared with those in studies outlined in Section 2. Each label specifies the form, movement, or both of a specific facial or body part compared to a neutral position. All labels were assigned an identifying code and accompanied by an image or

[1] https://dsgs-handbuch.ch/information/

illustration available in iLex to facilitate the annotation process. At this stage, assignments of these labels to grammatical functions were not made. The complete annotation scheme for non-manual components is available in both German and English on Zenodo.[2]

Table 3 displays the tiers included in the non-manual components block of the annotation scheme. The **Mouthing** tier captures lip movements like those of spoken German words. As mouthings are often not exact pronunciations of words, the annotator inserts the letters representing what they observe during the lip movement of the signer displayed in the video. For example, in Figure 2, we can see how the mouthing "mito" was written for the word *Mittag* ('noon') because the final voiced velar consonant *g* does not involve any lip movement.

For **Mouth gestures**, annotators have the option to select from 81 labels. This is the most detailed part in our scheme, reflecting various nuances in the form and movement of mouth components such as lips, cheeks, teeth, tongue, and their combinations. These labels are grouped into nine macro categories based on the form rather than function of the labels, as was done for the Auslan corpus (Johnston, 2019).

Regarding the **Nose**, seven labels are defined and categorized as static or dynamic based on nose movement characteristics, such as *static wrinkled nose*.

In the **Upper body** tier, thirteen labels describe main movements, such as leaning or moving the torso in a specific way and subtly turning or rotating the torso so that it faces a particular direction. The direction is annotated from the signer's point of view. **Shoulders** can be annotated separately from the upper body when their movements seem crucial to be considered in isolation, featuring six labels grouped under the macro categories of the upper body.

Fundamental in defining the sentence function, **Head** movements are segmented into twenty labels, subdivided based on movement type or location. Table 9 in Appendix B provides the list of head component labels.

Eye-related movements, namely **Eye gaze**, **Eyebrow** movements, and **Eyelid** motion, are segmented and labeled separately. In most of the tasks, the participant gaze is straight on the camera (cf. tier "Blick" ('gaze') in Figure 2). The annotation of gaze direction is crucial for marking the position or differences in object location. Eight labels denote various eyebrow positions, mostly upwards or downwards, while ten eyelid labels distinguish eye aperture and motion.

Non-manual components
Mouthing
Mouth gesture
Nose
Upper body
Shoulders
Head
Eye gaze
Eyelids
Eyebrows

Table 3: Tiers of the non-manual components.

4.4. Error Annotation

The error annotation tiers aim to capture productions by DSGS learners that deviate from the canonical form (Table 4). They are divided into three main categories: manual components, non-manual components, and sentence level.

For **manual components**, we adopted error definitions and categories from Ebling et al. (2018). These tiers, connected to gloss tiers, annotate deviations related to phonological parameters and their combinations.

For **non-manual components**, deviations regarding eyebrow and head movements, mouthing, mouth gestures, and their combinations are annotated. These features play a crucial role in sentence function definition.

The third category addresses **sentence-level error definition**. Drawing from prior studies and our main annotators' long teaching experience, we defined a restricted list of error categories to start from: sentence construction, question construction, negation, affirmation, statement connection, indexing, verbs, signing space, tempo and fluency, combined issues, and others. Where one of the latter two categories is chosen, the annotators describe the corresponding errors in a free-text field of a separate tier.

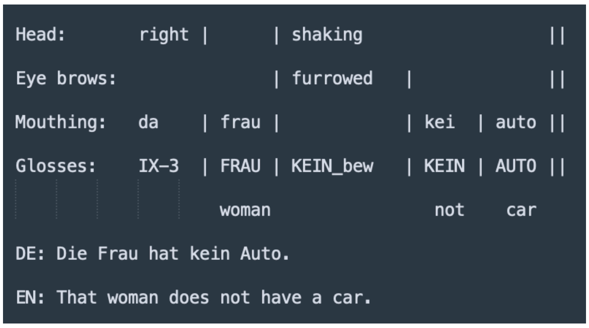
Each deviation receives a degree of **(non)-acceptability** (*not acceptable, acceptable, fully acceptable*), indicating severity of the deviating feature and impact on sentence comprehension. Additionally, the entire sentence receives an acceptability value, regardless of the number of annotated deviations. Figure 3 illustrates a simplified annotation example of a sentence deemed as "not acceptable" due to incorrect sentence construction, such as the use of the mouthing "da" ('there') and the improper use of eyebrows in the sentence.

The *acceptability of the sentence* tier is also annotated for L1 data. The rationale behind this decision is explained in the next section.

[2] https://doi.org/10.5281/zenodo.10669639

Error annotation
Deviations Gloss RH
Acceptability
Deviations Gloss LH
Acceptability
Deviations Gloss BH
Acceptability
Deviations NMC
Acceptability
Sentence problem
Sentence acceptability

Table 4: Tiers of error annotation. NMC: non-manual components.



```
Head:       right |      | shaking                  ||

Eye brows:               | furrowed   |             ||

Mouthing:   da    | frau |            | kei  | auto ||

Glosses:    IX-3  | FRAU | KEIN_bew   | KEIN | AUTO ||

                    woman                 not    car

DE: Die Frau hat kein Auto.

EN: That woman does not have a car.
```

Figure 3: Example of the annotation of a “not acceptable” sentence.

4.4.1. Why Annotate Acceptability?

Assuming a single “ground truth” in spoken and sign languages poses inherent challenges in achieving high agreement on language interpretation and understanding (Plank, 2022). Variations in annotation may arise from linguistic complexities, subjectivity, or instances where multiple interpretations are plausible (Plank et al., 2014; Manning, 2011; Rottger et al., 2022; Basile et al., 2021; Pavlick and Kwiatkowski, 2019; Nie et al., 2020). Sign languages are known to exhibit considerable structural variability (Bayley et al., 2015).

In the absence of a definitive ground truth, specifying acceptability values becomes more meaningful than assigning binary correct/incorrect values (Mehta and Srikumar, 2023). In the context of sign languages, the concept of acceptability of intuitive judgments was explored by Arendsen (2009) for the manual/phonological components of single signs in relation with iconicity. We thus designate sentences within an acceptable range from L1 data as correct, establishing them as the ground truth. Therefore, annotations of components in these acceptable sentences serve as a form of gold standard.

Having said this, we recognize that the annotation of acceptability values, like in error annotation, inherently entails a certain degree of subjectivity.

4.5. Additional Information

The additional tiers listed in Table 5 have not yet been systematically annotated at the current stage. This block of tiers is reserved for future rounds of annotations following preliminary linguistic analysis. In the interim, annotators may include comments in the *Comments* tier or annotate straightforward features. The **Translation** tier involves inserting a literal translation in German of individual signs and sentences. The **Functions**, **Topic/Focus**, **Prosody**, and **Role** tiers are designed to label various functions of annotated components, not only at the sentence level but also at the discourse level.

Additional information
Translation
Comments
Functions
Topic/Focus
Prosody
Role

Table 5: Additional tiers.

5. Validating the Annotation

As discussed in Section 3, our data undergoes a cross-checking step in which part of it is double-annotated. This step allows for the calculation of inter-annotator agreement (IAA) between the two expert annotators (Section 3), to assess the consistency of the (error) annotation labels, and to provide a quantitative evaluation of the complexity of the annotation task.

It is essential to recognize that agreement between annotators should not be mistaken with accuracy, as annotators may share possible biases present in the guidelines or cultural preconceptions (Basile et al., 2021; Plank, 2022).

5.1. Method

Incorporating different agreement metrics enabled us a thorough evaluation, considering various facets of annotation agreement. Applying Gwet’s *AC1* was motivated by specific limitations of Cohen’s κ (Cohen, 1960), particularly its tendency to underestimate coefficients for high-chance agreements and its lack of robustness against imbalanced categories (Feinstein and Cicchetti, 1990; Gwet, 2014).

In L1 data, we randomly extracted and duplicated 20% of the dataset, amounting to two transcripts. Each expert annotator annotated the transcript assigned to them and the counterpart annotated by the other expert. We then extracted the annotations from iLex and computed agreement using the following methods. First, in each transcript sentence, we examined annotated time spans sharing the same feature annotation, computed the overlap proportion of each feature and then calculated the temporally weighted overlap ratio, as described in Andersson and Sandgren (2016). We reported the formula for calculating the ratio along with the explanation and an example in Appendix C. As illustrated in Figure 4 in Appendix C, we treated all labels within the same feature as identical.

Second, we calculated Cohen's κ, Krippendorff's α (Krippendorff, 2019), and Gwet *AC1* score for nominal data across all labels for each transcript. This analysis utilized macro categories for each annotated component, disregarding the time variable.

For L2 data, we randomly selected 20% of the annotated L2 sentences for the first two data collection points, amounting to a set of 38 sentences. Within these selected sentences, we introduced new tiers for error annotation while deactivating the original error annotation tiers. The second annotator reviewed the annotation of manual and non-manual components performed by the first annotator in the first and second rounds, and then carried out a new error annotation using only their initialized tiers. We then extracted the annotations from iLex and assessed reliability using Cohen's κ, Krippendorff's α, and Gwet *AC1* for nominal data. Agreement concerning acceptability values was evaluated using Cohen's κ, Krippendorff's α, and Gwet *AC2* score for ordinal data.

For error annotation of non-manual components, adjustments to the time span were made depending on the alleged occurrence of a non-manual component. Thus, we computed the overlap ratio and temporally weighted overlap ratio for this category, as outlined in Appendix C. For glosses and sentence-level annotation, we focused solely on the annotation label without considering timing. This choice stemmed from the consistent timing across annotators, established through prior segmentation and linkage of tiers in iLex.

5.2. Results

We acknowledge that direct comparison of the results from these methods is not feasible due to their differences in computation. Nevertheless, this initial exploration represents our first step toward a comprehensive evaluation of our annotated data.

Below, we present our preliminary findings regarding the validation of the data.

5.2.1. L1 Data

On average, the annotation of manual and non-manual components in the L1 data achieved an overlap ratio of 0.18, encompassing cases for which the overlap duration is equal to 0. In instances of zero overlap, distinguishing missed events from misalignments was challenging. By excluding these events, the average overlap ratio increased to 0.62. Specifically, manual components attained an average of 0.64 (median: 0.88), while non-manual components averaged 0.45, ranging from 0.01 to 0.97. We calculate the temporally weighted overlap ratio for the events in each sentence. The average is 0.52, ranging from 0.29 to 0.96.

The agreement on labels is detailed in Table 6. Overall, the agreement between the two expert annotators did not reach high values. Considering both manual and non-manual components and excluding rows with zero overlap in time, the agreement yielded a κ score of 0.49 and a Gwet score of 0.52. Krippendorff's values closely align with the κ scores.

	κ	α	$Gwet$
manual	0.57	0.57	0.61
nmc	0.39	0.38	0.47
manual+nmc	0.49	0.44	0.52

Table 6: Reliability as measured by inter-annotator agreement using κ, α, Gwet *AC1*.

5.2.2. L2 Data

On average, the error annotation in the non-manual components of the L2 data achieved an overlap ratio of 0.35, ranging from 0.0 to 1 (median: 0.19). After excluding cases with zero overlap, the ratio increased to 0.55, ranging from 0.03 to 1 (median: 0.50). We calculated the temporally weighted overlap ratio for the events in each sentence obtaining an averaged score of 0.66.

Regarding the assigned labels, as presented in Table 7, agreement between the two expert annotators is modest. κ scores range from 0.16 for the error annotation of non-manual components to 0.52 for the error annotation of manual components, indicating a considerable degree of subjectivity in both annotation tasks. Krippendorff's values closely mirror the κ scores.

Interestingly, the acceptability values for the error annotation of non-manual components achieved a Gwet score of 0.60, suggesting moderate to high agreement between the two expert annotators in assessing the severity of deviation for non-manual features.

5.3. Discussion

The level of agreement depends on the task, complexity of the annotation scheme, and the number of annotators along with their degree of expertise.

	κ	α	$Gwet$
manual	0.52	0.53	0.56
accept_manual	0.32	0.33	0.34
nmc	0.16	0.15	0.25
accept_nmc	0.25	0.24	0.60

Table 7: Reliability as measured by inter-annotator agreement using κ, α, Gwet *AC1* (for components) or *AC2* (for acceptability). *Manual*: error annotation of the manual components; *nmc*: error annotation of the non-manual components; *accept*: agreement on the acceptability judgments.

Examining our results, the scores derived from our preliminary agreement calculations lead us to reflect on the primary factors contributing to disagreements.

Firstly, our findings underscore the inherent difficulty in achieving high agreement in tasks involving video stream segmentation. The accurate segmentation of signs presents challenges even for trained annotators, resulting in slight time variations in sign segmentation. However, these variations can cause discrepancies in calculations. In addition, the detailed nature of our annotation scheme, as described in Section 4, inherently amplifies disagreement among annotators. In general, studies analyzing sign language datasets refrain from reporting agreement scores, complicating efforts to benchmark our results within the broader landscape of sign language reliability assessments. The discrepancy between manual and non-manual component values (cf. Table 6 and Table 7) underscores the heightened challenge associated with annotating non-manual activities, possibly deriving from ambiguous guidelines or unclear instances of non-manual activity in videos.

Secondly, the complexity of the annotation task is reflected in the complexity of calculating agreement between annotators. Following the method outlined by Andersson and Sandgren (2016), which involves calculating the temporally weighted overlap ratio only between events with the same label, we do not assess whether there might be other annotated events occurring simultaneously but labeled differently. For instance, in cases where Annotator A annotated a time span with a label from the list of the “Eyelid” feature while Annotator B annotated the same time span with a label from the “Eyebrow” list, this could mean missing an event by one or both annotators. Considering the simultaneity of components in sign language, it is plausible that the time span involves both “Eyelid” and “Eyebrow” movements simultaneously. A next step would be to examine these “alternative classifications” with an aim to agree on one way of annotating and analyzing them.

As suggested by Schembri and Crasborn (2010), further exploration into agreement calculations for sign language data is needed. Establishing annotation standards would facilitate comparison of agreement values across different corpora, allowing for the development of a systematic method for calculating agreement in sign language data.

Despite the relatively modest agreement values, it is imperative not to perceive them as a limitation for dataset validation and subsequent use of these annotations. Widely debated in the context of spoken languages, human label variation (in other words, disagreement) offers valuable data insights to consider in the development of technologies, particularly those aimed at enhancing “technology which is by and for humans; inclusive and reliable” (Plank, 2022).

6. Conclusion and Outlook

We have presented the annotation process and scheme for L1 and L2 DSGS continuous data, focusing on the labeling of non-manual components. We have introduced a method for annotating and categorizing linguistic errors in L2 data, and proposed our idea of creating a ground truth encompassing variability. Viewing sentence acceptability as a facet of ground truth expands traditional notions, accommodating the inherent variability in sign language data analysis.

Our annotation scheme remains a work in progress, open to modification and adaptation. Statistical analyses are warranted to evaluate the scheme’s efficacy and the utility of macro categories. Refinement on higher levels of annotation, such as on the levels of sentence function and semantic roles (some tiers are described in Section 4.5), remains an area for future development.

While the scheme was created for DSGS, it can be adapted to other sign languages by adjusting the labels of each feature. To maintain the application of cross-linguistic comparisons, the adjustments would not change the content of the components but only the names that are assigned to these components.

We have described our first approach to data validation, illustrating difficulties given by the different variables to consider in the calculation. Agreement calculation methods, particularly considering time spans and labels, demand further exploration to systematically analyze annotated events and spot missed or erroneously annotated instances.

As we move forward, collaborative efforts and continued refinement of annotation practices will facilitate the advancement of sign language research.

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).

The authors would like to thank Regula Perrollaz, and Annemarie Büchli-Meier for their annotation work, and Penny Boyes Braem for her insightful assistance, comments, and reviews.

7. Bibliographical References

Richard Andersson and Olof Sandgren. 2016. ELAN Analysis Companion (EAC): A Software Tool for Time-course Analysis of ELAN-annotated Data. *Journal of Eye Movement Research*, 9(3).

Jeroen Arendsen. 2009. *Seeing signs: on the appearance of manual movements in gestures*. publisher not identified, Place of publication not identified. OCLC: 823212702.

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

Robert Bayley, Adam C. Schembri, and Ceil Lucas. 2015. *Variation and change in sign languages*, page 61–94. Cambridge University Press.

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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Kearsy Cormier, Onno Crasborn, and Richard Bank. 2016. Digging into Signs: Emerging Annotation Standards for Sign Language Corpora. In *7th Workshop on Representation and Processing of Sign Languages: Corpus Mining*, pages 35–40, Portorož, Slovenia. ELRA.

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.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

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.

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. 2001. Visicast deliverable d5–1: Interface definitions. technical report. visicast project. Technical report.

Thomas Hanke, Silke Matthes, Anja Regen, and Satu Worseck. 2012. Where Does a Sign Start and End? Segmentation of Continuous Signing. In *5th Workshop of the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (LREC 2012).*, Istanbul, Turkey. ELRA.

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.

Gabrielle Hodge. 2014. *Patterns from a signed language corpus: Clause-like units in Auslan (Australian sign language)*. Ph.D. thesis, Macquarie University.

Gabrielle Hodge and Onno Crasborn. 2022. Good practices in annotation. In Trevor Johnston, Julie A. Hochgesang, and Jordan Fenlon, editors, *Signed Language Corpora*, pages 46–89. Gallaudet University Press, United States.

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.

Trevor Johnston. 2019. Auslan Corpus Annotation Guidelines.

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

Reiner Konrad, Thomas Hanke, Susanne König, Gabriele Langer, Silke Matthes, Rie Nishio, and Anja Regen. 2012. From form to function. a database approach to handle lexicon building and spotting token forms in sign languages. In *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 87–94, Istanbul, Turkey. European Language Resources Association (ELRA).

Reiner Konrad, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, and Anja Regen. 2022. Public DGS Corpus: Annotation Conventions / Öffentliches DGS-Korpus: Annotationskonventionen.

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

Klaus Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology*, fourth edition edition. Thousand Oaks, California.

Kim B. Kurz, Geo Kartheiser, and Peter C. Hauser. 2023. Second language learning of depiction in a different modality: The case of sign language acquisition. *Frontiers in Communication*, 7.

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.

Christopher D. Manning. 2011. Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6608, pages 171–189. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

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, Krister Schönström, Nikolaus Riemer Kankkonen, and Lars Wallin. 2016. The interaction between mouth actions and signs in swedish sign language as an l2. In *Presented at the The 12th International Conference on Theoretical Issues in Sign Language Research (TISLR)*.

Johanna Mesch and Krister Schönström. 2018. From Design and Collection to Annotation of a Learner Corpus of Sign Language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

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.

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.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Annika Nonhebel, Onno A. Crasborn, and Els van der Kooij. 2004. Sign language transcription conventions for the echo project (version 9). 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.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Barbara Plank. 2022. The “problem” of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

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

Marina Ribeaud and Isabelle Cicala. 2019. *Handbuch Glossierung der Deutschweizerischen Gebärdensprache (DSGS)*, second edition. fingershop.ch.

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

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Adam Schembri and Onno Crasborn. 2010. Issues in creating annotation standards for sign language description. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, Valletta, Malta.

Marc Schulder, Sam Bigeard, Thomas Hanke, and Maria Kopf. 2023. The sign language interchange format: Harmonising sign language datasets for computational processing. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5.

Krister Schönström. 2021. Sign languages and second language acquisition research: An introduction. *Journal of the European Second Language Association*, 5(1):30–43.

Krister Schönström and Johanna Mesch. 2014. Use of nonmanuals in adult L2 signers in Swedish Sign Language – Annotating the nonmanuals. In *6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*, Reykjavik, Iceland. ELRA.

Katja Tissi. 2021. DSGS-Handbuch. Interkantonale Hochschule für Heilpädagogik Zürich HfH.

Lars Wallin and Johanna Mesch. 2018. Annoteringskonventioner för teckenspråkstexter : Version 7 (januari 2018). Technical report, Stockholm University, Sign Language.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*, Genoa, Italy. European Language Resources Association (ELRA).

8. Language Resource References

Crasborn, Onno and Zwitserlood, Inge and Ros, Johan and van Kampen, Annemieke. 2006-2017. *Collection Corpus NGT*. The Language Archive. PID https://hdl.handle.net/1839/8e5a77a3-8d1a-492a-bc86-9a3398b0809c.

Trevor Johnston. 2008. *Auslan Corpus*. Endangered Languages Archive. PID http://hdl.handle.net/2196/00-0000-0000-0000-D7CF-8.

A. Non-manual Components: Macro categories

B. Head Labels

The two-letter codes, an extension of HamNoSys to non-manuals, were initially defined in the ViSiCAST project (Hanke, 2001) and have been adapted by our annotators.

Feature	Labels	Macro categories
Mouthing	-	-
Mouth gesture	81	9
Nose	7	2
Upper body	13	2
Shoulder	6	1
Head	20	6
Eye gaze	30	6
Eyelids	10	3
Eyebrows	8	2

Table 8: Number of labels and number of macro categories in our scheme.

Head	Macro categories
NO: Head nod (up and down)	cat. 1 Nodding
NU: Simple head nod up [dynamic]	
ND: Simple downward head nod [dynamic]	
RL: Tilted to left or right nodding head	
SH: Head shaking (left and right)	cat. 2 Shaking
SS: Tilted to left or right shaking head	
NF: Tilted forward [static]	cat. 3 Front
PF: Shifted forward	
OG: Head tilted forward (nodding)	
NB: Tilted backwards	cat. 4 Back
PB: Shifted backward	
LN: Head nod (up and down) left (up and down)	
RN: Head nod (up and down) right (up and down)	
SL: Turned to the left	cat. 5 Lateral
SR: Turned to the right	
TL: Tilted to the left (static)	
TR: Tilted to the right (static)	
KD: Head rotation	cat. 6 Strongly dynamic
KK: Head tilt (dynamic)	
LI: Head movement coupled to gaze [dynamic]	

Table 9: Labels defined for the Head feature.

C. Temporally weighted overlap ratio

Equation 1 illustrates an example of the agreement calculation with two events in the L1 data, as illustrated in Figure 4. Column A represents two events for the feature "Blick" ('gaze') annotated by Annotator A in one sentence, while Column B represents the two events in the same sentence annotated by Annotator B. We have:

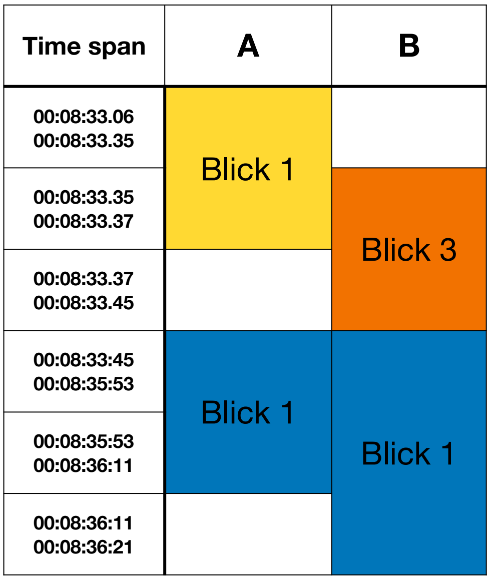
$$\begin{aligned} E &= \{\epsilon_1, \epsilon_2\} \\ T &= \{t_1 = 0.39, t_2 = 2.76\} \\ O &= \{o_1 = 0.05, o_2 = 0.95\} \end{aligned} \tag{1}$$

where E is the set of $n = 2$ events, each labeled by Annotator A and Annotator B; T is the set of maximum duration for each events in E, and O represents the set of overlap proportions for the events in E. The overlap proportion is calculated by dividing the duration of the overlap by the maximum temporal extent of the event.

The temporally weighted overlap ratio is then calculated as follow:

$$\frac{\sum_i^n O_i T_i}{\sum_i^n T_i} = \frac{(0.05 * 0.39) + (0.95 * 2.76)}{(0.39 + 2.76)} = 0.84 \tag{2}$$

If we were to consider only the overlap proportion without accounting for temporal duration, the calculation for the overlap ratio would be as follows: $0.05 + 0.95/2 = 0.5$, even though the length of the annotated overlap varies.



Time span	A	B
00:08:33.06 00:08:33.35	Blick 1	
00:08:33.35 00:08:33.37		Blick 3
00:08:33.37 00:08:33.45		
00:08:33:45 00:08:35:53	Blick 1	Blick 1
00:08:35:53 00:08:36:11		
00:08:36:11 00:08:36:21		

Figure 4: Simplified representation of two events in a same sentence, annotated by two annotators, Annotator A and Annotator B.

Please note that even if the annotators assigned two different labels for event ϵ_1, they both annotated the feature "Blick" ('gaze') in this timespan.