Towards Automated Sign Language Production: A Pipeline for Creating Inclusive Virtual Humans

Lucas Bernhard
lucas.bernhard@uni-a.de
University of Augsburg
Augsburg, Germany

Fabrizio Nunnari
fabrizio.nunnari@dfki.de
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany

Amelie Unger
amelie.unger@ergosign.de
Ergosign GmbH
Hamburg, Germany

Judith Bauerdiek
Christian Dold
Marcel Hauck
Alexander Stricker
bauerdiek@charamel.com
dold@charamel.com
hauck@charamel.com
stricker@charamel.com
Charamel GmbH
Cologne, Germany

Tobias Baur
Alexander Heimerl
Elisabeth André
Melissa Reinecker
tobias.baur@uni-a.de
alexander.heimerl@uni-a.de
elisabeth.andre@uni-a.de
melissa.reinecker@uni-a.de
University of Augsburg
Augsburg, Germany

Cristina España-Bonet
Yasser Hamidullah
Stephan Busemann
Patrick Gebhard
cristinae@dfki.de
Yasser.Hamidullah@dfki.de
stephan.busemann@dfki.de
patrick.gebhard@dfki.de
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany

Corinna Jäger
Sonja Wecker
Yvonne Kossel
Henrik Müller
corinna.jaeger@yomma.de
sonja.wecker@yomma.de
yvonne.kossel@yomma.de
henrik.mueller@yomma.de
yomma GmbH
Cologne, Germany

Kristoffer Waldow
Arnulph Fuhrmann
Martin Misiak
kristoffer.waldow@th-koeln.de
arnulph.fuhrmann@th-koeln.de
martin.misiak@th-koeln.de
TH Köln
Cologne, Germany

Dieter Wallach
dieter.wallach@ergosign.de
Ergosign GmbH
Hamburg, Germany

ABSTRACT

In everyday life, Deaf People face barriers because information is often only available in spoken or written language. Producing sign language videos showing a human interpreter is often not feasible due to the amount of data required or because the information changes frequently. The ongoing AVASAG project addresses this issue by developing a 3D sign language avatar for the automatic translation of texts into sign language for public services. The avatar is trained using recordings of human interpreters translating text into sign language. For this purpose, we create a corpus with video and motion capture data and an annotation scheme that allows for real-time translation and subsequent correction without requiring to correct the animation frames manually. This paper presents the general translation pipeline focusing on innovative points, such as adjusting an existing annotation system to the specific requirements of sign language and making it usable to annotators from the Deaf communities.

CCS CONCEPTS

• **Human-centered computing** → *Human computer interaction (HCI)*; **Accessibility systems and tools**; *Accessibility technologies*; • **Computing methodologies** → *Machine translation.*

KEYWORDS

sign language production, annotation, corpus, motion capture, automatic translation.

ACM Reference Format:
Lucas Bernhard, Fabrizio Nunnari, Amelie Unger, Judith Bauerdiek, Christian Dold, Marcel Hauck, Alexander Stricker, Tobias Baur, Alexander Heimerl, Elisabeth André, Melissa Reinecker, Cristina España-Bonet, Yasser Hamidullah, Stephan Busemann, Patrick Gebhard, Corinna Jäger, Sonja Wecker,

PETRA '22, June 29-July 1, 2022, Corfu, Greece
 ACM
ISBN 978-1-4503-9631-8/
https://doi.org/10.1145/3529190.3529202

Yvonne Kossel, Henrik Müller, Kristoffer Waldow, Arnulph Fuhrmann, Martin Misiak, and Dieter Wallach. 2022. Towards Automated Sign Language Production: A Pipeline for Creating Inclusive Virtual Humans. In *The15th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '22), June 29-July 1, 2022, Corfu, Greece.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3529190.3529202

1 INTRODUCTION

Sign languages are complex in a sense that they include various visual elements such as hand shapes and trajectories, gestures, facial expressions, and body posture. There are approximately 70 million Deaf or Hard of Hearing (DHH) People worldwide, who communicate in sign language [35, 36]. Across the globe, each country has its own sign language, which can differ significantly from others. But even within the same country, there may be several dialects, in which different signs represent a distinct meaning (e.g. [22] for American Sign Language).

There is a misconception that providing a written form of spoken language to DHH individuals is sufficient for barrier-free communication or information. However, sign languages have their own grammar, syntax and linguistic complexity and are therefore languages in their own right, separated from any spoken language which is seen by many DHH individuals as a foreign language who, instead, consider their respective local sign language as their native language [16]. Thus, written text poses a potential communication problem for these people and requires appropriate means of translation into sign languages in order to facilitate communication between hearing and Deaf People.

A common approach to meet this requirement are videos in which an interpreter translates the respective content from text or speech into sign language (e.g., various television broadcasts). However, this approach is not always practical, seeing that the videos are expensive to produce, static, and usually contain only a portion of the original information. If parts of the content change, the provider is forced to have entirely new videos created by sign language interpreters. In addition, there are only few sign language interpreters available which hinders the widespread implementation of digital barrier-free content. Due to this conundrum, research approaches have examined the automatic generation of sign language translations for some time, usually using a 3D avatar for visualization [12, 17, 21].

However, automatic translation between sign language and spoken language is a highly complex task. This is partly due to the complexity of sign languages as they are not simply a concatenation of basic signs, as described in [4]: There are a multitude of ways to "inflect" a sign by applying modifications to the manual and non-manual components on order to change the meaning of a sentence. Non-manual components include, among others, mouth, eyes, nose, eyebrows and movement of the torso. For example, facial expressions are used to express emotions and they can take the role of adjectives and adverbs. Non-manual components can also have grammatical meaning, e.g., in German Sign language a yes/no-question is expressed by tilting the head, lifting the eyebrows and a wide opening of the eyes. As for the manual components, a sign can be "relocated", meaning that the position or trajectory of the hands are altered during a sign. For example, the sign for "walking" can specify the direction by adjusting the start and end position of the sign. On top of that, the task of automatic sign language production (SLP) is made even more difficult due to data on sign language being comparatively scarce. As a result, SLP remains an open research issue to this day.

The collaborative project AVASAG [2] is developing a real-time controlled 3D sign language avatar for the automatic translation of German text into German Sign Language with the project's scope focusing primarily on the field of travel information and services, transport, and tourism. Within this application domain, meaningful and barrier-free communication is of essential importance, seeing that inaccessible information when traveling can lead to severe consequences in regard to traveling plans, as well as the emotional state of the affected.

Thus, our project aims to allow real-time translation of text input into sign language, presented by the 3D avatar. In addition, a user interface is being implemented which enables human operators to make corrections to the automatically produced translation of the avatar. This mode of manual correction by users can therefore be used, for example, to exchange wrongfully placed signs or fine-tune specific movements.

To successfully achieve these objectives, our project team involves partners from various domains within computer science, including avatar animation, machine learning, motion capturing, and user experience (UX) design. Additionally, the team also includes DHH individuals as sign language experts, thereby ensuring high quality of the displayed translation into sign language.

Besides pursuing a human-centred approach with results that match the expectations of the target group, we provide the following technical contributions:

- A translation pipeline that supports sufficient naturalness of animation while still allowing for low effort correction if the automatic translation fails.
- The creation of a sign language corpus that includes video and motion capture (MoCap) data as well as a detailed annotation including information on subtle movements and grammar, besides the signs that are performed.
- An innovative simplified "boolean-based" manual annotation process where annotators only need to mark "meaningful differences" between recordings and the exact values are computed automatically.

This paper presents the architecture of the AVASAG project. In the next section we discuss related work in SLP. We describe the conduct and results of focus groups with sign language speakers which were held at the beginning of the project in Section 3. In Section 4, we give an overview of the project and address its individual components in detail, with a focus on the annotation.

2 RELATED WORK

Approaches to sign language production (SLP) typically use 3D avatars. Many of them, e.g., [12, 17], are based on a transcription of sign language, such as HamNoSys [14]. Such transcriptions usually describe individual signs as a composition of discrete-valued elementary components (e.g., hand shape, facial movements). In terms of transcription of sign language the work [21] is perhaps the closest work to ours. The authors encode signs using a default animation which is then altered by a set of modifiers that must be

manually adjusted based on the context. The modifiers here can control facial expressions, gaze, execution speed of a sign and how the location or trajectory of a sign is changed. A full sentence then is created by concatenating the resulting animations for the individual signs. The advantage of transcription-based solutions is that it is comparatively easy to create or modify animations without having to manually create the frames for the avatar skeleton. However, the results of such methods are often described as robotic [20], which leads to negative perceptions by Deaf users.

Apart from avatar-based approaches, some studies investigate SLP from a different perspective: In their recent work Stoll *et al.* [30] use a Generative Adversarial Network [13] to directly generate the pixels of the sign language video. This allows for photorealistic results, which are difficult to achieve with avatars. However, since the neural network generates the pixel data, subsequent editing of the translation is limited using this approach.

In [28] Saunders *et al.* present the first approach of an end-to-end translation, which they improved in [29]. Here, a written sentence of spoken language is translated into a sequence of 3D skeletal poses. The resulting 3D skeleton can subsequently be used to control an avatar or directly produce a video, similar to the above mentioned approach in [30]. This strategy has the potential to produce very fluid and realistic motion. However, it would also make subsequent correction difficult as one would have to manually adjust the animation itself.

In our project, we take an approach that enables post-translation correction while at the same time allowing for sufficiently fluent movements: As individual signs are recorded with MoCap, the animation of individual signs itself is fluent. Moreover, in the annotation, additional modifications of signs are given as continuous values rather than discrete ones. If incorrect signs are displayed in the translation, only the instruction as to which sign should be shown needs to be replaced. To correct inflections, "3D gizmos" will be developed for the user interface.

3 REQUIREMENT ANALYSIS AND INVOLVEMENT OF DHH USERS

In order to ensure that the quality of the avatar-design as well as that of the sign language performed by the avatar meets the requirements and expectations of Deaf community members, a close integration of the latter is of central importance for the success of the project. Such an integration helps not only to establish more appropriate approaches and results within the project, but also aids hearing project members to better understand the situation, pain points and actual needs of Deaf or Hard of Hearing (DHH) individuals, thus building a bridge between DHH and hearing communities. We therefore aim primarily at a direct exchange with DHH individuals throughout all phases of the project, establishing, in addition, research-based appropriate representations of Deaf People's abilities and needs, while at the same time securing a continuous supervision of the project by sign language experts. To do so, we rely on methods from within the domains of user experience (UX) and service design. Thus, we focus on analysing and understanding our target group with their abilities and needs beforehand while also evaluating our results with members of the target group throughout and after the project's duration.

3.1 Focus groups

As the project focuses primarily on the context of travel information and services, our first objective in regard to analysing the target group was to understand the type, origin and effects of the barriers that Deaf People face here as well as the emotional state these people are in when doing so. Thus, we conducted focus group sessions with 10 members of Deaf communities, moderated by a Deaf member of the research team and evaluated by hearing UX design experts [32]. The sessions were conducted in a remote setting as video conferences (due to COVID-19 restrictions), each lasting between 90 and 120 minutes including a short break of 15 minutes. A brief discussion guide with questions, goals, and time frames for each part of the session was prepared beforehand. Seeing that previous research has discovered strong tendencies of rejection from DHH communities towards hearing researchers [1, 23] we were careful to always address participants in their native language, i.e., German Sign Language (GSL). Thus, all further documents made available to the participants (i.e., email messages, consent forms, questionnaires, and video conference instructions) were prepared beforehand both in written German and as recorded GSL-videos, while the actual sessions themselves were carried out entirely in GSL. Finally, each session was 1) recorded, 2) translated, and 3) transcribed as written text for the content to be easily accessible for non-GSL speakers and available for documentation, publications, or other forms of reference. Close attention was paid to having all manuscripts double-checked by at least one hearing and one deaf interpreter. With this sign-language-centered approach we wanted to allow participants to express themselves in the language they feel most comfortable with, thereby receiving a more truthful impression of their perspectives.

Translating our entire approach into GSL did, however, cause challenges that were unforeseen to us beforehand, with many of them being grounded in the fundamental difficulty of translating content from sign language into written forms of spoken language. This concerns both, the communication towards participants and the communication from participants towards researchers, raising concerns regarding e.g., budgets and duration when translating material or the matter of the assessing, documenting, and storing user-content produced in GSL:

> *"With regard to further utilization such as quoting participants within articles or assessing qualitative or quantitative measures, text works well, because it can easily be stored, retrieved, and quickly scanned. However, in the case of content produced in SL, textual representations are difficult to establish, seeing that any SL is based on visual images which can—similar to a picture or photograph—be described through text but hardly captured completely without the loss or distortion of the information contained. If, for instance, we want to quote a participant from our focus groups within a journal article, we are forced to rely not on the original statement, but on its interpretation in a verbal language—with the interpretation being highly subjective as it is commonly the case when describing images."* [32, S. 4]

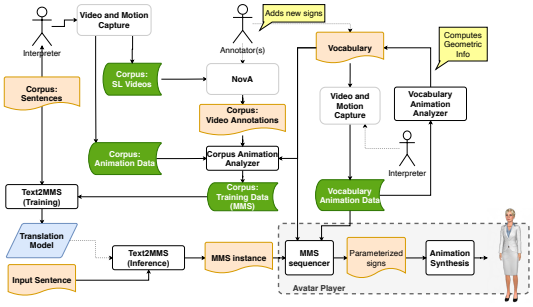

Figure 1: Overview of the off-line training (top) and real-time translation (bottom) pipelines.

Conducting these focus groups taught us not only valuable lessons in regard to participatory and collaborative research between Deaf and hearing people [31], but also highlighted moments of stress and exclusion experienced by Deaf People when traveling. Examples of these include the lack of non-auditive access to real-time information, uninformed and untrained staff in regards to the needs of and communication with DHH travelers, or the text-heavy design of booking and information platforms. Thus, our results range from such reports on a broad spectrum of experienced barriers within different types of public transportation to technological needs and requirements, both inside and outside the traveling context—results documented and analysed, amongst other, in the form of personas [7, 8, 19, 27] and user journeys [15] within the AVASAG project. This exchange with the target group allowed us to understand their needs and requirements in regard to supportive technologies such as a sign language avatar which are directly integrated and considered within the project's implementation. In addition, our collected insights form the basis not only for optimising further exchange with the target group in general, but especially for finding accessible and appropriate procedures when evaluating the project's progress and results with Deaf community members. Further information on our approach, results and main lessons learned from a hearing researchers point of view can be found in [32].

3.2 Future involvement of the DHH community

While the above described focus groups allowed us an analytic insight into the user population's perspectives, further integration of and cooperation with DHH communities will focus on evaluating the project's activities and results. Based on the experiences collected so far, we need to find ways of conducting sign-language accessible ways of formative as well as summative evaluation procedures. These feedback assessments are to be integrated both, in regard to the annotation mechanisms as well as the design, movement and signing of the final avatar. The following section on the system overview provides a more detailed discussion on these and all other components of the AVASAG concept.

4 SYSTEM OVERVIEW

4.1 General architecture

The central concept for the production of sign language in the project is to encode sentences in sign language into a sequence of signs and modifications, or "inflections" (as explained in Section 1), for these signs. Figure 1 shows a diagram of the offline training and real-time translation phases.

The real-time translation process can be described as follows: Written sentences are translated by a neural network into an intermediate representation, which we call MMS (multi modal sign stream). MMS consists on the one hand of the glosses[1] for the signs that have to be executed and on the other hand of a multitude of information about inflections of signs that are important for comprehensibility and naturalness of sign language. This intermediate representation is then used to animate the avatar. To do this, animation data from previously recorded signs in the "vocabulary" is played back in sequence (connected by transitions) and modified by the inflections.

With this goal in mind, the offline training process is designed as follows: A group of domain experts list a set of sentences for the specific application domain (e.g., traveling information). Then, expert interpreters translate those sentences into sign language while being captured via both video recording and MoCap sensors. Furthermore, each time a new sign is encountered, it is recorded again separately, without the context of the sentence ("non-inflected"), and added to the vocabulary. The video recordings of the sentences are then annotated with our annotation tool NOVA. Subsequently, the MoCap recordings of the sentences as well as the signs in the vocabulary are used by the "Corpus Animation Analyzer" to compute the inflection parameters that convert the non-inflected form of the respective sign in the vocabulary to the way it appears in the sentences. The glosses together with the inflection parameters form the MMS. This is finally used to train a neural network that translates sentences into MMS.

In the following sections, we will discuss each of the components in more detail.

4.2 Motion Capture Dataset

The corpus of the AVASAG project, which is currently under development, contains frontal Full-HD videos of signed sentences as well as synchronously recorded motion capture (MoCap) data for full body, hands and face. Each video contains one sentence or a few related sentences in German Sign Language on the topic of travel information. The recordings so far are between approximately 9 and 40 seconds long.

The MoCap setup is realized with an optical *Optitrack* [26] capturing system with twelve *Flex 13* optical cameras. Thereby, a spherical shape is used to increase tracking accuracy in the important area in front of the person (Figure 2). The signer wears a special suit with reflective markers attached to it that are recorded in a frequency of 120 Hz. Finally, each frame contains the 3D information for every captured point that is used to reconstruct a human skeleton including hand and body movement. In order to reduce errors and inaccuracies we focused the motion capturing on the upper body

[1] A gloss is a term for a sign in written form.

Figure 2: The motion capture setup for precise tracking information at close distances. An *Optitrack* system with 12 cameras in a spherical shape is used for capturing the movement.

and used a upper-body baseline marker set with 25 markers. For the facial expression recording, existing technologies with rigs are not applicable due to the performed gestures near the face. We evaluated depth cameras especially for the face, which did not work optimally in our needed range. For this reason, we decided to use a video-based machine learning approach.

The videos are annotated with glosses, including start and end time, as well as with original sentences of the spoken language in German and translations in English. Additional information about the presented sign language such as movements and grammatical information is also provided. Furthermore, the corpus is coupled with a vocabulary in which all signs occurring in the sentences are entered. Each entry includes video and MoCap data (recorded again separately) as well as syntactic information about the sign (e.g., body contacts). Details about the annotation and the vocabulary will be discussed in the next section.

4.3 Annotation

In this section we will discuss the annotation scheme and the annotation process. The first segment covers the vocabulary, which contains entries for all signs. After that, we will look at the annotation scheme of the sentences. Since this is relatively complex, we look at the annotation of two signs taken from a sentence as examples. We then take a look at our annotation tool NOVA and how we have adapted it to speed up the annotation of sign language. In the last segment, we explain how the manual annotation is refined by analyzing the animation data (which finally results in the MMS which we use to animate the avatar).

4.3.1 Vocabulary. The vocabulary is a collection of entries for all the signs that appear in the signed sentences. For each entry, the sign is recorded individually (i.e., in addition to the recording of full sentences) via video and MoCap to have it available in its non-inflected form. These MoCap recordings are used to drive the avatar and they are needed for the annotation of the sentences, as explained within the following sections.

For each sign, a gloss, or gloss-ID, is defined in the vocabulary to uniquely identify it, following the principle of the AUSLAN annotation scheme [18]. The gloss-ID consists of one or more words of spoken language trying to describe a common meaning of the sign to make it easily readable, for example “ARRIVE”, “TRAIN”, “YOU-ALL”. Additionally, gloss-IDs can have trailing numbers to distinguish between signs that have similar or the same meaning but different executions, for example “PASS-THROUGH”, “PASS-THROUGH2”. Finally, for certain types of signs, the glosses are also given a particular prefix to clarify their meaning: E.g., numbers are given the prefix “num:” and gestures that are also common in spoken language and are not exclusive to sign language are given the prefix “gest:”, for example “num:11”, “gest:OH-WELL”.

We note here that signs do not have an official name or written identification. This is because sign languages do not have written forms that are used for regular communication [18] and so defining the gloss-IDs is a non-trivial task. In our project the gloss-IDs are frequently revised by our Deaf members. Sometimes gloss-IDs are changed to give a more accurate description in spoken language. Other times, gloss-IDs need to be removed if there is another one that describes the same sign. This is rather time consuming as one has to compare the corresponding videos and it becomes more and more difficult as the number of different signs grows in the corpus.

In addition to the gloss-ID, for each entry syntactic information is given:

- The number of used hands. (Certain signs are performed using one hand only [4].)
- If it is relocatable in space.
- If there is contact with other body parts or between hands.
- If mouthing or mouth gestures are present. (In sign language, mouthing refers to forming a word with the mouth silently [4].)
- References to all appropriate WordNet synsets [24]. (In short, in WordNet groups words according to their possible meanings.)

These WordNet will be used for data augmentation (see Section 4.4) and the other information is needed to ensure that the animation plays correctly after translation. E.g., contacts must be preserved even if a sign is changed by inflections or retargeted to avatars of different body proportions. And if a sign contains mouthings or mouth gestures, they have precedence over facial expression that convey a specific mood (e.g., smiling).

4.3.2 Annotation scheme for sentences. The main tier of the annotation scheme is the **gloss** tier, which consist of the time segmentation to identify the beginning and end of a sign. Within each time slot, the annotator inserts the gloss-ID of the vocabulary. The closely related tiers **dominant hand** and **non-dominant hand** must be annotated only if the hand configurations and/or trajectories are performing a movement that differs from the non-inflected animation. Independently from the other, one hand can be annotated with another *gloss-ID* (e.g., when performing two signs in parallel), or as *hold* if it keeps the last position of the previous gloss, or with a specific *hand-shape-ID* when performing a classifier.

The largest group of tiers describe differences in the execution of a sign during a sentence in comparison with the sign in the vocabulary. Here, the labels in these tiers have the same time segmentation as the tiers mentioned above. The tiers must be checked with a boolean *true* only if *meaningful differences* with the vocabulary are noticed. When there is no meaningful difference is apparent, no label is set instead. For the manual elements the tiers are **dominant hand reloacted** and **non-dominant hand reloacted**. For non-manuals the tiers are **torso**, **shoulders**, **head**, **mouth/mouthings**,

cheecks, **eyes**, **eyebrows**, **facial expression**. We define "meaningful differences" as *differences* that can be perceived at the motor level in the execution of the sign, and which the annotator can recognize as *intentional*, with the goal of conveying extra or distinct meaning compared to the sign in the vocabulary. The categories for differences to be specified are similar to the ones in [21]. However, the presented scheme here is much simpler as we do not need to quantify the differences using discrete values. Instead flagging a difference will trigger the automatic computation of continuous *inflection parameters* (see Section 4.3.4).

Finally, explicit *grammar roles* are also annotated (i.e., **why-question**, **yes/no-question**, and **negation**). They can span over multiple glosses in the sentence and are typically related to well-encoded motions. Notice that the time segmentation of the grammar roles is independent from the other tiers. If a negation is present in the sentence, but no head-shake has been annotated in the *head* tier, then a warning of inconsistency will be raised.

To make this concept clearer, let us look at concrete examples. After recognizing the signs, the annotator needs to enter the associated glosses, which are defined in the vocabulary. Then one has to compare how the sign in the sentence differs from the recording in the vocabulary. For this purpose, Figure 3 shows the recordings of two signs, each in non-inflected form in the top row and in the possibly inflected form as performed during a sentence in the bottom row.

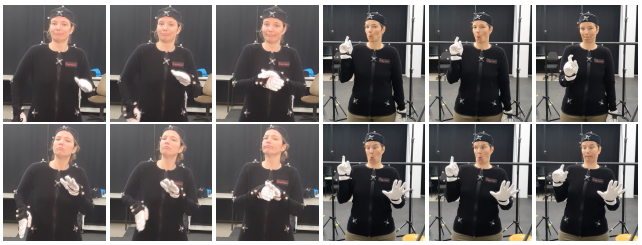

Figure 3: Examples of recorded signs. The left side shows the sign with gloss "AUFBRUCH" (engl. "DEPARTURE") and the right side shows the sign with the gloss "KOMMEN(-von-hinten)" (engl. "COME(-from-behind)"). The upper row is the non-inflected form from the vocabulary and the lower row shows an inflected form during a sentence.

For the sign "DEPARTURE" on the left side of the figure, the annotation of the inflections is as follows: First, in the *eyebrows* tier, a label of *true* was entered to mark a meaningful difference in the position. There is also an additional subtle head movement in the video of the sentence which was labeled in the *head* tier. Finally, apart from these differences between the recordings, a final *true* label was placed in the *why-question* tier because this sign is part of a rhetorical where-question that is asked in the sentence. All other tiers are left blank.

For the sign "COME(-from-behind)" on the right side of the figure, meaningful differences in the *eyebrows*, the *eyes*, and the *torso* were marked. Furthermore, in this sign the non-dominant (left) hand is in a different position in the recording of the sentence than it is in the vocabulary recording. This is because the signer left this hand in the same position as it was at the end of the previous sign and so a *hold* label was entered in the tier for the non-dominant hand.

This "boolean-based" annotation simplifies the manual process, as otherwise, in the example above, the annotators would have to indicate the direction in which the torso is tilted and how much it is tilted (e.g., on a fixed scale, from slightly tilted to strongly tilted). Instead, the exact inflection values will be calculated automatically using the MoCap data (see Section 4.3.4).

4.3.3 Annotation tool. Our annotation tool NOVA [3] supports a collaborative annotation process by maintaining a database backend, which allows users to load and save annotations from and to a MongoDB [25] database running on a central server. This gives annotators the possibility to immediately commit changes and follow the annotation progress of others. The user interface has been designed with a special focus on the annotation of continuous recordings involving multiple modalities.

Different annotation schemes are supported such as *discrete* and *free* schemes. Discrete annotations consist of a list of labelled time segments. Each segment has a start and and end time and label name. An annotator has to choose one name from a predefined list of label names for each label. Free annotations are like discrete annotations, but here annotators are free to choose the label names. This is useful if an annotation task can not easily be reduced to a few labels, for example in case of spoken speech transcriptions.

As we need annotators who know sign language, the amount of possible contributors is limited and accommodating those who are willing to participate is essential. In order to provide Deaf annotators with easy access to our annotation software we developed a browser-based version of our annotation tool. A screenshot of a loaded recording session is shown in Figure 4.

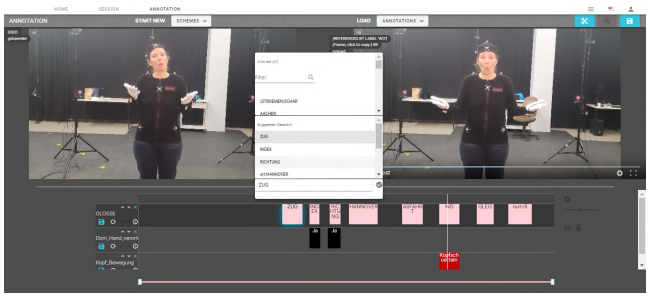

Figure 4: At the top full-body videos of recording sessions are displayed. The left one is the recording of a sentence, the right one is the recording of a single sign. The bottom half shows different annotation tiers, the top one being the gloss tier. When editing a gloss label, a menu is displayed, which can be seen here in the middle. The menu is divided into three sections. In the top section is a list of all glosses that have appeared in our recordings so far, in the middle is a list of glosses for the loaded session, and at the bottom is a text input field for manual input.

For the project, most annotation tiers can be realized using a discrete scheme as we have a predetermined list of values. For example, some tiers only need a single value which is used to

express a difference in execution, as mentioned before. For the gloss tier we cannot define a list of possible values beforehand as we can not wait for all recording sessions to be finished before starting to annotate. So a free scheme seems suitable at first. However, since there is no "correct" name for a given sign (see Section 4.3.1) free scheme as described above is not very practical, too. To ensure that we always get the correct label name for a sign, the annotators would have to search through the vocabulary and copy the right gloss each time. In the long run, this is a very tedious and time-consuming workflow.

Because of this, we decided to extend the user interface for the free scheme, see Figure 4: First, we added the possibility to load a list of label names that are independent from the video currently annotated from a column of a table and display them to the annotator. The annotator can simply click on one of these names to apply it to a label. Generally, this can be used to define a list of allowed or suggested labels which can updated quickly and accessed easily by the annotators. In our project, we use this to show a list of all glosses that were added to the vocabulary (which is currently stored in a table). This can be seen at the top of the menu shown in Figure 4. The second feature again allows to load label names from a table, but here they associated with a specified recording name inside the table. Meaning, in the table two columns are needed, one containing the names of the recordings and the other containing lists of associated label names. This is perhaps less generally applicable, but more useful for our case: The translations of the spoken sentences into glosses are prepared before the video and MoCap recordings are made to ensure a consistent language and save time during recording sessions. This means that we already have a list of glosses that are specific to the loaded video before the annotation begins. These glosses are displayed to the annotators as well which can be seen at the middle of the menu shown in Figure 4. This means that the annotators usually only have to pick a label form a small list of suggestions. In case that a label appears to missing or wrong, they can search through the displayed list of all glosses in the vocabulary instead. Also, the usual text input field for free labels is still present in case they spot a typing error. The annotators are advised to not use these two options in other cases. But they are available to not slow on the annotation process in case of an error.

Another feature we added is the option to load videos via a URL based on the name of a label. For this, in the table a column of (certain) label names and one ore more columns with corresponding links of videos to be shown to the annotator are needed. This feature could be used to show a tutorial or explanation for certain labels. In our case, we use it to load non-inflected signs. As explained earlier, the annotators often have to compare the inflected form of a sign during a sentence with the non-inflected form. With this feature, they simply select a label and press a hotkey to load the non-inflected form. Further, we give annotators the option to play the regular loaded video and the referenced video in sync for the duration of the label. This means that the annotators can watch the sign performed during the sentence and the non-inflected sign in a synchronized way. The synchronisation is achieved by adjusting the playback speed of the referenced video depending on the length of the label and the length of the reference video itself.

Google Sheets documents are currently supported as tables to load label values or reference videos from. These features (and various smaller improvements to the user interface) were designed in consultation with the Deaf members of our team.

4.3.4 Animation data analysis. Because of our "boolean-based" approach the manual annotation is not sufficient to directly animate an avatar. And so, after a video was manually annotated, inflection parameters that transform signs from their non-inflected form into the way they appear in the sentences must be calculated.

The inflection parameters are computed by comparing and measuring the differences between the performance of the sign in the sentence vs. its execution in the vocabulary. Such differences are computed on four levels: hand trajectories, torso shift and rotation, head rotation, and facial expression. For the first two categories, the difference is expressed in terms of non-rigid 3D spatial transformations (4x4 matrices) that transform the lines traced in space by the hand palms or by the torso center with translation, rotation, scaling, and shearing. For the head, the difference will be a pure rotational transformation (3x3 matrix) measured at the neck joint. For the facial expression, the difference will be a vector with the difference of the weights of all the blend-shapes realizing the facial skin motion (approximately, each blend-shape corresponds to the activation of one, or more, facial muscles). The implementation of the measurement of such differences will be based on trajectories transformation (e.g., [9]) and mesh registration (see [33] for a survey).

The gloss labels from the manual annotation together with the computed inflection parameters form the MMS. Using the MMS, an Avatar can be animated: The gloss labels are used to play back non-inflected signs from the vocabulary in sequence. Then, the animation data is modified using the computed inflection parameters. Finally, information from the vocabulary is used to apply certain restrictions, e.g., preserving contacts between hands (as described in Section 4.3.1).

4.4 Automatic translation

The Text2MMS is a machine learning module in charge of the conversion between written text and the MMS abstraction. For the task, we design a neural architecture that takes sequences of words as input and outputs the most probable sequence of glosses from the vocabulary. Inflection parameters are predicted during generation as continuous real numbers. The key components of this architecture (Fig. 5) are: Two encoders, a mapping module and a decoder. To deal with multimodal inputs, two different encoders are used, one for texts and another one for MoCap. As MMS contains glosses which convey the meaning of the sentence, we add a gloss supervision block in the text encoder to facilitate the training. The mapping block constitutes a latent space for the different types of encoded features. Notice that MoCap data can only be used during training since it will not be available at inference time, but it is important for the system to be able to relate text and movement.

Considering the fact that even for simpler machine learning tasks large amounts of data are needed, we will use pre-trained language models that will be fine-tuned to perform our task [10]. For data augmentation we will generate synthetic data using the

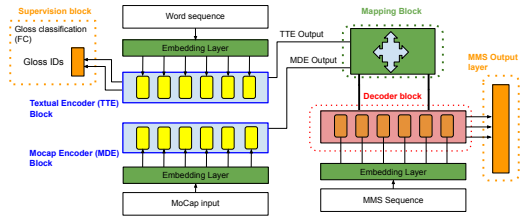

Figure 5: Overview of the envisaged architecture to convert from plain text to MMS. The training will benefit from both MoCap and textual data in a multimodal setting.

Figure 6: Our current avatar performing a sign.

relations in WordNet [24], word classes, and our vocabulary joined with unsupervised methods when possible [11, 34].

4.5 Avatar

We are currently using one of our existing "human-like" avatars after adjusting specific aspects to fit our needs (see Figure 6): The mesh was modified to match the body measurements of the MoCap actor to avoid errors while retargeting between the MoCap data and the avatar's skeleton. For visibility and perceptibility, it is important that the model has a high contrast between skin, clothes and background color and that careful lighting with shadows for a 3D effect is chosen [20]. Thus, we chose dark clothes and grey background for our avatar after discussing it with the Deaf members of our team.

For the animation synthesis, we use the cloud-based Charamel software VuppetMaster [6], which supports a 3D real-time rendering engine based on WebGL standard, thus making it possible to run the avatar on all known devices (including browsers) without any additional installation. Inverse kinematic chains were implemented to allow relocating signs in the signing space and to prevent the mesh from intersecting with itself as the animation are partly procedural.

By the end of the project, we want to offer a photorealistic avatar in addition to the human-like one. For the photorealistic avatar, we made a first test by scanning a human subject. The highly detailed mesh and associated texture of the photorealistic avatar were generated via hybrid photogrammetry approach using eight cameras, different filter systems and stochastic pattern projections. To animate the face, we created fifty-one facial action units on the avatar's mesh. After the photorealistic avatar is finished, a human-like will be modeled using the photorealistic as the basis to compare their acceptance by users.

To allow users to make changes to the generated animation defined by the MMS via a user interface, we are adapting the authoring tool VM Storybuilder [5].

5 CONCLUSIONS

We presented a system for automatic translation of written text into sign language, which aims to overcome typical problems of avatar-based approaches, such as lack of naturalness. We use a neural network approach to translate German text into an intermediate representation, which in turn is used to control the avatar. The representation is very detailed, with information about which sign should be executed, as well as about subtle movements that are important for the comprehensibility of sign language. Furthermore, as a point of innovation, this representation allows easy manual post-correction of the automatic translation, which is not or only partially possible with other methods in automatic sign language production. To train the neural network, we created a corpus for automated sign language synthesis – an area where machine learning approaches are limited by the scarcity of data. The corpus includes synchronous video and MoCap data of signed sentences as well as individual signs, and detailed annotation that goes well beyond specifying glosses. To facilitate manual annotation work, we rely on a novel annotation scheme in which annotators only need to mark meaningful subtle movements and exact values are then computed automatically from the MoCap data. In order to be as responsive as possible to the needs of the Deaf, we conducted focus groups with members of the Deaf communities. In addition, the sign language experts of the research team were involved in the design and development process from the very beginning and provided valuable feedback on all components of the system.

So far, we have several hundred recordings of sentences and different individual signs. We are currently working on the algorithms for computing the inflection parameters and a robust way of recognizing facial action units in our recordings. After that, we will perform a systematic evaluation of the signing avatar with members of the Deaf communities.

ACKNOWLEDGMENTS

This work has received funding from the German Federal Ministry of Education and Research (BMBF), Grant Numbers 16SV8488, 16SV8489, 16SV8490, 16SV8492, 16SV8493.

REFERENCES

[1] Melissa L Anderson, Timothy Riker, Stephanie Hakulin, Jonah Meehan, Kurt Gagne, Todd Higgins, Elizabeth Stout, Emma Pici-D'Ottavio, Kelsey Cappetta, and Kelly S Wolf Craig. 2020. Deaf ACCESS: Adapting Consent Through Community Engagement and State-of-the-Art Simulation. *The Journal of Deaf Studies and Deaf Education* 25, 1 (2020), 115–125.

[2] AVASAG consortium. 2021. *AVASAG project web page.* Charamel GmbH. Retrieved August 20, 2021 from https://avasag.de

[3] Tobias Baur, Alexander Heimerl, Florian Lingenfelser, Johannes Wagner, Michel F. Valstar, Björn Schuller, and Elisabeth André. 2020. eXplainable Cooperative Machine Learning with NOVA. *KI - Künstliche Intelligenz* 34, 2 (2020), 143–164.

[4] Penny Braem. 1995. *Einführung in die Gebärdensprache und ihre Erforschung.* Signum-Verlag, Hamburg, Germany.

[5] Charamel GmbH. 2021. *Structure of StoryBuilder.* Charamel GmbH. Retrieved December 23, 2021 from https://vuppetmaster.de/documentation/docs/storybuilder/2_structure/

[6] Charamel GmbH. 2021. *VuppetMaster web page*. Charamel GmbH. Retrieved August 20, 2021 from https://vuppetmaster.de

[7] Torkil Clemmensen. 2004. Four approaches to user modelling—a qualitative research interview study of HCI professionals' practice. *Interacting with Computers* 16, 4 (2004), 799–829.

[8] Alan Cooper. 2004. *The inmates are running the asylum: Why high-tech products drive us crazy and how to restore the sanity* (2 ed.). Sams Publishing, Indianapolis.

[9] Arie Croitoru, Peggy Agouris, and Anthony Stefanidis. 2005. 3D trajectory matching by pose normalization. In *Proceedings of the 2005 international workshop on Geographic information systems - GIS '05*. ACM Press, Bremen, Germany, 153.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[11] Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data Augmentation with a Generation Approach for Low-resource Tagging Tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6045–6057. https://doi.org/10.18653/v1/2020.emnlp-main.488

[12] Sarah Ebling and John Glauert. 2016. Building a Swiss German Sign Language avatar with JASigning and evaluating it among the Deaf community. *Universal Access in the Information Society* 15, 4 (2016), 577–587.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014), 2672–2680.

[14] Thomas Hanke. 2004. HamNoSys—Representing sign language data in language resources and language processing contexts. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Vol. 4. ELRA, Lisbon, Portugal, 1–6.

[15] Tharon Howard. 2014. Journey mapping: A brief overview. *Communication Design Quarterly Review* 2, 3 (2014), 10–13.

[16] Matt Huenerfauth and Vicki Hanson. 2009. Sign language in the interface: access for deaf signers. *Universal Access Handbook. NJ: Erlbaum* 38 (2009), 14.

[17] Vince Jennings, Ralph Elliott, Richard Kennaway, and John Glauert. 2010. Requirements for a signing avatar. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*. ELRA, Valletta, Malta, 133–136.

[18] Trevor Johnston. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International journal of corpus linguistics* 15, 1 (2010), 106–131.

[19] Plinio Thomaz Aquino Junior and Lucia Vilela Leite Filgueiras. 2005. User modeling with personas. In *CLIHC '05: Proceedings of the 2005 Latin American conference on Human-computer interaction*. ACM Press, Cuernavaca, 277–282.

[20] Michael Kipp, Quan Nguyen, Alexis Heloir, and Silke Matthes. 2011. Assessing the deaf user perspective on sign language avatars. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*. ACM, ACM Press, New York, USA, 107–114.

[21] Vincenzo Lombardo, Cristina Battaglino, Rossana Damiano, and Fabrizio Nunnari. 2011. An Avatar-based Interface for the Italian Sign Language. In *2011 International Conference on Complex, Intelligent, and Software Intensive Systems*. IEEE Computer Society, Washington, DC, USA, 589–594. https://doi.org/10.1109/CISIS.2011.97

[22] Ceil Lucas, Clayton Valli, and Robert Bayley. 2002. Sociolinguistic Variation in American Sign Language. *Bibliovault OAI Repository, the University of Chicago Press* 24, 4 (2002).

[23] Michael McKee, Deirdre Schlehofer, and Denise Thew. 2013. Ethical Issues in Conducting Research With Deaf Populations. *American Journal of Public Health* 103, 12 (2013), 2174–2178.

[24] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[25] MongoDB, Inc. 2021. *MongoDB*. MongoDB, Inc. Retrieved December 23, 2021 from https://www.mongodb.com/

[26] NaturalPoint. 2021. *OptiTrack*. NaturalPoint. Retrieved December 23, 2021 from https://optitrack.com/

[27] John Pruitt and Tamara Adlin. 2006. *The Persona Lifecycle*. Elsevier, San Francisco.

[28] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Progressive Transformers for End-to-End Sign Language Production. In *European Conference on Computer Vision*. Springer, Berlin/Heidelberg, Germany, 687–705.

[29] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Networks. *International Journal of Computer Vision* 129, 7 (2021), 2113–2135.

[30] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2Sign: Towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision* 128, 4 (2020), 891–908.

[31] Amelie Unger, Dieter Wallach, and Nicole Jochems. 2021. Lost in Translation: Challenges and Barriers to Sign Language-Accessible User Research. (2021). To be presented at The 23rd International ACM SIGACCESS Conference on Computers and Accessibility.

[32] Amelie Unger, Dieter P. Wallach, and Nicole Jochems. 2021. Lost in Translation: Challenges and Barriers to Sign Language-Accessible User Research. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3441852.3476473

[33] Oliver van Kaick, Hao Zhang, Ghassan Hamarneh, and Daniel Cohen-Or. 2011. A Survey on Shape Correspondence. *Computer Graphics Forum* 30, 6 (2011), 1681–1707.

[34] Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 6382–6388.

[35] World Federation of the Deaf. 2021. *Our Work*. World Federation of the Deaf. Retrieved September 27, 2021 from https://wfdeaf.org/our-work/

[36] World Health Organization. 2021. *Deafness and hearing loss*. World Health Organization. Retrieved September 27, 2021 from https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss