AVASAG: A German Sign Language Translation System for Public Services

Fabrizio Nunnari[3] **Judith Bauerdiek**[1] **Lucas Bernhard**[4] **Cristina España-Bonet**[3]
Corinna Jäger[6] **Amelie Unger**[2] **Kristoffer Waldow**[5] **Sonja Wecker**[6]
Elisabeth André[4] **Stephan Busemann**[3] **Christian Dold**[1] **Arnulph Fuhrmann**[5]
Patrick Gebhard[3] **Yasser Hamidullah**[3] **Marcel Hauck**[1] **Yvonne Kossel**[6]
Martin Misiak[5] **Dieter Wallach**[2] **Alexander Stricker**[1]

[1]Charamel GmbH, Cologne, Germany (Project-Coordinator)
[2]Ergosign GmbH, Hamburg, Germany
[3]German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus D3.2, Saarbrücken, Germany
[4]Human-Centered Artificial Intelligence, University of Augsburg, Germany
[5]TH, Köln, Germany
[6]yomma GmbH, Hamburg, Germany

Abstract

This paper presents an overview of AVASAG; an ongoing applied-research project developing a text-to-sign-language translation system for public services. We describe the scientific innovation points (geometry-based SL-description, 3D animation and video corpus, simplified annotation scheme, motion capture strategy) and the overall translation pipeline.

1 Introduction

The development of software solutions able to translate (bi-directionally) from spoken language to sign-language (SL) has received a lot of attention during the last years. In Europe, the involvement of the public institutions in such line of research culminated with the funding, under the H2020 program, of two 3-year long research projects, namely EASIER [12] and SignON [13, 15].

In this paper, we present the architecture of project AVASAG [11] (Avatar-basierter Sprachassistent zur automatisierten Gebärdenübersetzung = Avatar-based speaking assistant for the automated translation of sign language), which is a project funded by the German ministry for education and research (BMBF) aiming at deploying a commercial system able to automatically translate text to sign language in various domains of public services (e.g., announcements for railway stations, airports, harbors, and hygiene warnings). The implementation choices are driven by the following requirements and constraints:

1. The system is devoted to off-line translation services. Hence, translation does not need to be necessarily in real-time, but rather offer the possibility to human operators (likely trained interpreters) to finalize the animation through manual editing, and approve it before delivery;
2. The avatar animation will be tuned to maximize comprehensibility, while at the same time maintaining a sufficient level of acceptance in terms of naturalness of the animation. This

can be seen as the compromise set to initial synthetic voices used for public services;

3. In order to approach the market within a reasonable time frame, the project focuses on realizing at first well recognized forms of inflection of lexical signs (e.g., sign relocation, interrogative forms, role shifts, classifiers), but still open to the realization of more creative iconic gestures in future extensions;
4. The system is engineered to scale with time as new signs are added to its vocabulary to support more application domains.

From a scientific point of view, the project aims at the following innovation points.

First, the project is developing a translation system that goes beyond classic symbolic representation of SL. Existing *SL-description*s range from mere un-contextualized GLOSSES, to more sophisticated formats specifying hand (shape, orientation, location, trajectory) and facial movements (e.g., Stokoe [16], HamNoSys [5]). This gives the opportunity to human operators for the corrections of sentences when the text-to-SL-description fails. However, compared to data-driven approaches, animations driven by SL symbolic descriptors are generally judged as generating non-believable unnatural animations (see [10] for an overview). On the other hand, recent end-to-end data-driven approaches are moving towards the generation of 3D sign pose sequences [14] that could be used to animate an avatar from a kinematic level, but hinders the possibility of a manual correction of the translation result.

In this project, we try to merge the advantages of data-driven animation, which leads to more natural looking results, while leaving the capability of post-translation manual correction. Given a vocabulary of motion captured signs, the inflection of signs within the context of a sentence will be realized by transforming a sign data using: i) non-rigid 3D transformation (translation, rotation, scaling, shearing) of the hand trajectories and torso movement, ii) corrective blend-shapes on the facial animation, and iii) time-warping functions controlling the dynamic of the execution. All of those inflection transformations are driven by numerical parameters that can be manipulated by a human through the use of “3D gizmos” in a dedicated editing GUI.

As an advantage with respect to performing end-to-end translation (directly from text to avatar animation data), predicting only inflection parameters significantly reduces the size of the target output, thus likely allowing for the preparation of accurate models with much less training material.

Second, the project will deliver to the scientific community, with public unrestricted access, a **corpus of sentences** with parallel data whose entries are composed of:

- The ***text of the sentence*** in natural language (both in the host language and in its English translation);
- a ***GLOSS transcription*** of the sentences using the philosophy of the gloss-ID [6];
- the ***3D motion capture (MoCap) data*** of the corresponding SL translation, captured in a high-class motion capture studio, for full body, hands, and facial animation;
- ***Full-HD*** video of the interpreter during the same motion capture session, hence synchronized at frame level with the 3D MoCap data;
- the ***annotation*** of the sentence videos on different tiers (see the third innovation point for details) performed with a cross-check procedure by native deaf and SL interpreters;
- the ***inflection parameters***, i.e., the values for the (3D) transformation inflecting the signs.

In addition, the corpus will be paired with a **vocabulary of signs** (aka signary), indexed by gloss-IDs, where each sign will contain: the MoCap data and video of non-inflected signs executed in the same settings as for the sentences, syntactic information of the sign, such as

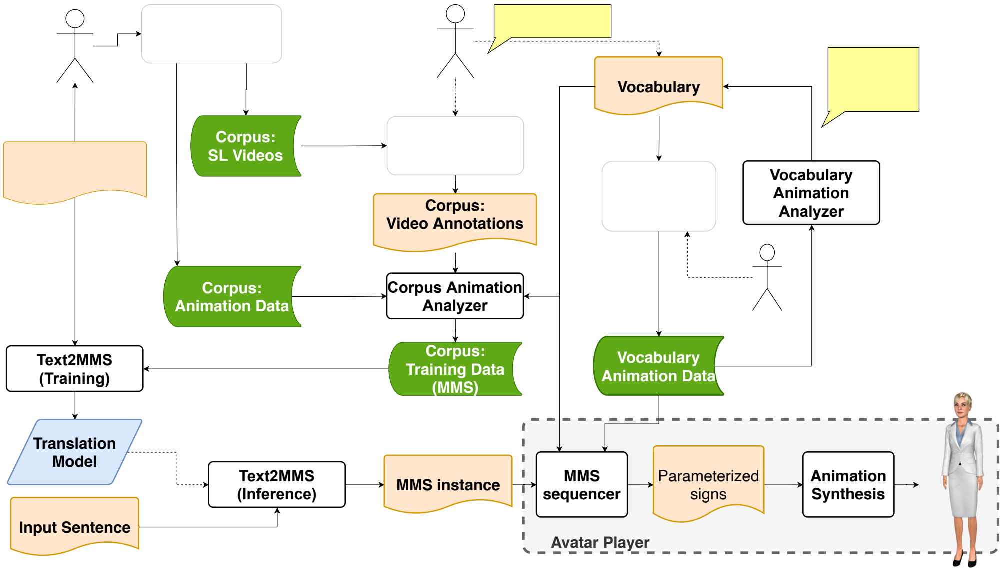

Figure 1: Overview of the off-line training (top) and real-time translation (bottom) pipelines.

symmetry, number of hands, use of mouthing, body contacts, and finally the references to all the possible semantic meanings in WordNet [9].

Third, the annotation process will follow an innovative “boolean-based” simplified annotation scheme, where annotators on each tier must check a flag *only if* the execution of a sign in the sentence shows *meaningful differences* with respect to its lexicalized form in the vocabulary. Here, by “meaningful”, we mean deliberate inflections (such as sign relocation, eyebrows movement, body shifts, head movement, facial expressions, and the like) of the sign in order to convey additional meaning. The extraction of the exact values of those differences, i.e., the magnitude of the inflection parameters, is delegated to a procedural analysis of the 3D animation data. Such a simplified annotation strategy is supposedly faster than existing schemes, where annotators must select values from closed lists or insert free text.

Fourth, we are employing a MoCap system that combines different data streams together, like real-time streamed point clouds and multiple (depth) cameras. These different data sources are then combined and processed inclusively to create a matching animation.

2 System Overview

Figure 1 shows a diagram of the offline and realtime phases of the proposed architecture.

Motion Capture The corpus creation starts from a set of written sentences that are translated into sign language and recorded with both a video camera and a full-body motion capture system (fingers, hands, arms, torso, head, and face). The recording is performed simultaneously, so to have a perfect match between the video and the animation data recording.

Annotation The video material will be annotated using the annotation tool NOVA [1]. Here, the data is stored in a collaborative annotation database, so that the annotation work can be divided among several users. In addition, machine learning methods can be integrated into

NOVA with the goal of generating annotations in part automatically. We will train different neural networks for this purpose in an attempt to reduce the workload of the annotators.

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. Two additional tiers annotate if the dominant or non-dominant hand are performing another sign, holding the previous one, or placing a classifier in the signing space. Each of the remaining tiers must be checked with a boolean *true* only if *meaningful differences* with the vocabulary are noticed for the movement of the **manual elements** (configuration, location, orientation), **non-manuals** (torso, shoulders, head, mouth/mouthings, cheecks, eyes, eyebrows, facial expression); flagging a difference will trigger the automatic computation of *inflection parameters*. Finally, explicit **grammar roles** (wh-question, yes/no question, negation) are also annotated.

Vocabulary creation The vocabulary is created interactively during the annotation of the segmentation tiers. Each time a new sign is encountered, the annotator will insert a new entry in the vocabulary. Each new sign will be then motion-captured in its non-inflected form and both the video and the MoCap data associated to the vocabulary entry. Each entry is completed with annotations about the number of used hands, if it is relocatable in space, if there is contact with other body parts or between hands, mouthing or mouth gestures, and references to all appropriate WordNet synsets [9].

Animation data analysis The Corpus Animation Analyzer computes the inflection parameters that transform signs from their non-inflected form into the way they appear in the sentences. The implementation is based on trajectories transformation (e.g., [2]) and mesh registration (see [17] for a survey). As a result, inflection parameters will take the form of 4D matrices for non-rigid 3D transformations or vectors for corrective blendshape weighting. The output of the analysis–the MMS (multi-modal signstream)–consists of the annotated sentences augmented with the sign inflection parameters.

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 will train a neural network that takes sequences of words as input and outputs the most probable class for each element in the vocabulary. Inflection parameters will be predicted as continuous real numbers. Given that machine learning heavily depends on the amount of data used for training, and the corpus might not achieve consistent sizes in the short term, we will adopt both transfer learning and data augmentation techniques. In the first case, we will use pre-trained language models that will be fine-tuned to perform our task [3]. In the second case, we will generate synthetic data using the relations in WordNet, word classes, and our vocabulary joined with unsupervised methods when possible [19, 4].

Avatar creation For the character creation, a state-of-the-art 3D computer graphic program (e.g., Autodesk 3ds Max) will be used. For the development of the photorealistic avatar a 3D photo-scan system for generating high level realistic face textures will be build up. To avoid errors potentially introduced while retargeting between the MoCap data and the avatar's skeleton, the avatar is tuned according to body measurements on the actor. As suggested in previous research [8], we will apply high contrast between skin, clothes and background color, and will provide careful lighting with shadows for a 3D effect.

Avatar animation The avatar animation consists of parsing MMS sequences, and play back the resulting animation data. For the animation synthesis, we use the cloud-based Charamel software VuppetMaster [18], 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).

The animation of hands and torso will be driven by inverse kinematic chains. Fifty-one facial action units can be used for creating expressive facial animations, which is fundamental for the comprehension of sign language [7].

User Evaluation With pursuing a human-centered approach within the project, ensuring a focus on the needs of those who are supposed to use and understand the avatar, is essential. Evaluations by and exchange with the target group is thus integrated into the entire process—starting with elaborate measures of user research in early phases (requirements analysis such as personas, scenarios, and user stories) that form the foundation for formative and summative evaluations conducted within usability labs later on. In this way, we want to achieve not only a high acceptance and quality of the avatar, but also strengthen the acceptance and support within the SL-community towards our project's approach.

Ensuring the sign language quality An essential part of project is the continuous checking of the sign language quality of the avatar to be developed. This is achieved through the collaboration with a team of sign language experts and professional interpreters who supervise the annotation process and ensure a high quality standard of the avatar with regard to the representation of sign language. In this way, representatives of the future user group work actively within the realization of the avatar and influence the development according to their requirements.

3 Current Status and Future Work

At the moment of writing, the project completed its initial investigation stage and it is at the beginning of its development stage. The corpus structure has been finalized. The MoCap environment has been tested, finalized, and was used to capture the first sentences of the corpus. The annotation tool has been configured and sign language experts are using it. Scripts to automatize the processing of the corpus (e.g., extraction of facial animation data from videos, consistency check) are under development. The avatar animation engine can playback motion captured sentences and non-inflected signs with body and hands. As soon as facial animation is supported, the avatar will undergo the first user evaluation. Tools for the analysis of the animation data (such as trajectory transformation and mesh registration) are under investigation.

As a first goal, the system shall enable translations for public services, which is characterized by a formal communication register. In the future, the system will be extended to be applied in different contexts, where more complex sign language features, such as iconicity, pose higher challenges for the whole translation pipeline.

References

[1] Tobias Baur, Alexander Heimerl, Florian Lingenfelser, Johannes Wagner, Michel F. Valstar, Björn Schuller, and Elisabeth André. eXplainable cooperative machine learning with NOVA. *KI - Künstliche Intelligenz*, January 2020.

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

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[4] Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq R. Joty, Luo Si, and Chunyan Miao. Daga: Data augmentation with a generation approach for low-resource tagging tasks. *CoRR*, abs/2011.01549, 2020.

[5] Thomas Hanke. HamNoSys-representing sign language data in language resources and language processing contexts. In *LREC*, volume 4, pages 1–6, 2004.

[6] Trevor Johnston. From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15(1):106–131, April 2010.

[7] Michael Kipp, Alexis Heloir, and Quan Nguyen. Sign language avatars: Animation and comprehensibility. In *International Workshop on Intelligent Virtual Agents*, pages 113–126. Springer, 2011.

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

[9] George A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, November 1995.

[10] Lucie Naert, Caroline Larboulette, and Sylvie Gibet. A survey on the animation of signing avatars: From sign representation to utterance synthesis. *Computers & Graphics*, 92:76–98, November 2020.

[11] AVASAG project web page. https://avasag.de, 2021.

[12] EASIER project web page. https://www.project-easier.eu, 2021.

[13] SignON project web page. https://signon-project.eu, 2021.

[14] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks. *Int. J. Comput. Vis.*, 129(7):2113–2135, 2021.

[15] Dimitar Shterionov, Vincent Vandeghinste, Horacio Saggion, Josep Blat, Mathieu De Coster, Joni Dambre, Henk Van den Heuvel, Irene Murtagh, Lorraine Leeson, and Ineke Schuurman. The signon project: a sign language translation framework. In *Proceedings of the 31st Meeting of Computational Linguistics in The Netherlands (CLIN 31)*, July 2021.

[16] William Stokoe. *Sign language structure: An outline of the visual communication systems of the American deaf.* Univ. of Buffalo, Buffalo, NY, 1960.

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

[18] VuppetMaster web page. https://vuppetmaster.de, 2021.

[19] Jason Wei and Kai Zou. 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)*, pages 6382–6388, Hong Kong, China, November 2019. Association for Computational Linguistics.