easier

intElligent Automatic
SIgn languagE tRanslation

D2.3 REPORT ON AVATAR REPRESENTATION

Expanding existing standards for avatar representation

Revision: v.1.1

Work package	WP 2
Task	Task 2.3
Due date	30/11/2023
Submission date	30/11/2023
Deliverable lead	CNRS
Version	1.1
Authors	Michael Filhol (CNRS), Rosalee Wolfe (ATHENA), Fabrizio Nunnari (DFKI)
Reviewers	Thomas Hanke (UHH), Sarah Ebling (UZH)

Abstract	This report describes three new standards for avatar representation, supporting affect-augmented sign language representations. The standards propose formats for the output of automatic translation and can serve as the basis for post-editing systems. In addition to a novel method for specifying affect, the salient features of the standards are the representation of co-occurring processes in sign languages including their timing and coordination. The affect augmentations consist of emotive elements of communication. The report presents a set of edge/corner cases to test an avatar's ability to synthesize sign language. These test cases will be published as a Signing Avatar Challenge Dataset to promote future avatar development.
Keywords	Avatar representation, EASIER Notation, AZee

Grant Agreement No.: 101016982
Call: H2020-ICT-2020-2
Topic: ICT-57-2020
Type of action: RIA

WWW.PROJECT-EASIER.EU

Document Revision History

Ve rsion	Date	Description of change	List of contributor(s)
V1.0	10/11/2023	1st version of draft for comments	Michael Filhol (CNRS) Rosalee Wolfe (ATHENA)
v1.1	28/11/2023	After reviews	Michael Filhol (CNRS)

DISCLAIMER

The information, documentation and figures available in this deliverable are written by the "Intelligent Automatic Sign Language Translation" (EASIER) project's consortium under EC grant agreement 101016982 and do not necessarily reflect the views of the European Commission.

The European Commission is not liable for any use that may be made of the information contained herein.

COPYRIGHT NOTICE

Project co-funded by the European Commission in the H2020 Programme		
Nature of the deliverable:	**R**	
Dissemination Level		
PU	Public, fully open, e.g. web	✓
CL	Classified, information as referred to in Commission Decision 2001/844/EC	
CO	Confidential to EASIER project and Commission Services	

* R: Document, report (excluding the periodic and final reports)

DEM: Demonstrator, pilot, prototype, plan designs

DEC: Websites, patents filing, press & media actions, videos, etc.

OTHER: Software, technical diagram, etc.

EXECUTIVE SUMMARY

This report describes three new standards for avatar representation, supporting affect-augmented sign language representations. The standards propose formats for the output of automatic translation and can serve as the basis for post-editing systems. In addition to a novel method for specifying affect, the salient features of the standards are the representation of co-occurring processes in sign languages including their timing and coordination. The affect augmentations consist of emotive elements of communication. The report presents a set of edge/corner cases to test an avatar's ability to synthesise sign language. These test cases will be published as a Signing Avatar Challenge Dataset to promote future avatar development.

Funded by the Horizon 2020
Framework Programme of the European Union

TABLE OF CONTENTS

Funded by the Horizon 2020
Framework Programme of the European Union

ABBREVIATIONS

DGS	Deutsche Gebärdensprache (German Sign Language)
LREC	Language Resources and Evaluation Conference
LSF	Langue des signes française (French Sign Language)
PAD	Pleasure, Arousal, Dominance

Funded by the Horizon 2020 Framework Programme of the European Union

1 INTRODUCTION

Although there have been recent “shared tasks” created for Sign Language recognition [1], there has yet to be a challenge for sign language generation and display. At the 2023 edition of the Sign Language Translation and Avatar Technology workshop [2], a forum explored this issue.

In the lively and cordial discussion that ensued, several ideas regarding theme and format surfaced. One suggestion was to create challenges with rotating topics. Each topic could focus on a different aspect of sign language portrayal. Possibilities ranged from handshape realism, inclusion of sublinguistic physiology when portraying a linguistic parameter (effect of a moving articulator position on torso), prosodic issues and nonmanual signals. Ideas about format included a proposal that an announcement of a challenge should be accompanied by assets, and a period of time, with a “due date” perhaps coinciding with a SLTAT workshop. Each competing group participating in the challenge would specify the technology used and the amount of time expended, in addition to furnishing animations. Another suggestion would be to have live demonstrations to demonstrate generative capacity.

There were concerns about ensuring a level playing field for any competition. Because different research groups focus on different signed languages, choosing one sign language over another for the purposes of a challenge would give some teams an unfair advantage.

In terms of judging, both objective measures and subjective feedback would be useful to the competing teams. Because it is such a small community, recruiting judges for the competition may prove challenging, as it is highly likely that all those qualified for the role would also be on a competing team. Further, the appearances of current avatars are so distinctive that it would be impossible to conceal the identities of the competing teams, so a truly blind review would be difficult.

In conclusion, no consensus arose from the forum, except that it was too soon to create a standard for avatar representation. However, one suggestion that seemed to have a modicum of support was a call for a sample set of challenge data that would inspire ideas on how to create such future challenges. To meet this need, this report suggests three types of avatar challenge data as a starting place for future signing avatar challenges. These are generating affect via PAD (Please, Arousal, Dominance) notation, generating sentences via the EASIER notation, a gloss-oriented authoring system, and generating discourse via the AZee framework.

2 PAD REPRESENTATION

Affect carries information that can often be lost when a spoken utterance is transcribed in writing. It is the same with signed languages. In fact, Sign Language interpreters take care to convey the emotional state of an utterance when interpreting from spoken-to-signed languages and vice versa. In contrast to discrete representations of affect, which distinguish a set of basic labeled emotions combined to represent more complex emotions as proposed by Ekman [3], componential representations such as PAD (Please, Arousal, Dominance) represent emotions in three continuous dimensions [4].

Pleasure, alternately called Valence, is a mood state that ranges from extreme sorrow (value 0) to extreme happiness (value 1). Arousal describes the level of excitement or engagement, also with a range of zero to 1. Dominance describes the perception of control over the surrounding environment, which can range from sense of utter helplessness to complete

Funded by the Horizon 2020 Framework Programme of the European Union

mastery of the situation. The PAD model provides a more consistent method to identify emotion via automatic recognition [5] or to produce emotion via automatic generation, such as in video games [6].

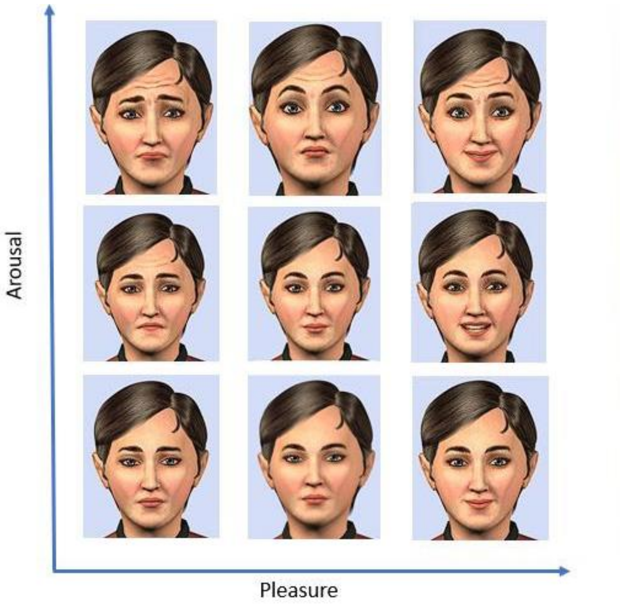
The EASIER project utilises PAD representation for generating facial portrayal of affect in their signing avatar Paula. This is the first time that such a representation was applied to a signing avatar and the results are encouraging.

The challenge data for the PAD representation consists of settings for nine discrete regions within the Pleasure and Arousal dimensions (Table 1), and a matrix of synthesized facial expressions corresponding to the settings (Table 2). Because the effects of the Dominance dimension primarily occur on the spinal column and arms in signed language, the facial portrayal incorporates Pleasure (Valence) and Affect exclusively.

TABLE 1: DISCRETISATION FOR THE PA SPACE INTO NINE REGIONS.

P,A = (0.0, 1.0)	P,A = (0.5, 1.0)	P,A = (1.0, 1.0)
P,A = (0.0,.0.5)	P,A = (0.5, 0.5)	P,A = (1.0, 0.5)
P,A = (0.0, 0.0)	P,A = (0.5, 0.0)	P,A = (1.0, 0.0)

TABLE 2: REFERENCE FACIAL EXPRESSIONS FOR EACH OF THE NINE PA DISCRETE REGIONS.



3 EASIER NOTATION

The EASIER Notation is a specification of a gloss-based scripting language that facilitates the authoring of signed-language content to be portrayed by an avatar. Because there is no widely accepted writing system for sign language, many people use glosses, although using glosses alone will not capture all processes present. For example, many automatic spoken-to-signed translation systems produce glosses to represent sign language [7]. Further, many human translators use glosses in their written notes in preparing to record a translation. Thus, a system that could accept output from machine translation as well as providing for human post-processing would be a useful tool.

Funded by the Horizon 2020
Framework Programme of the European Union

The EASIER Notation is text-based, and is human-readable, but provides a way to represent critical features of sign language that are not possible to specify through glosses alone. These features include:

- grammatical markers for yes-no questions, wh-questions and negation;
- prosodic markers for phrases and sentences;
- affect, expressed as PAD triples.

For a complete list of linguistic processes represented by the EASIER Notation, please see [8]. The EASIER Notation can express co-occurring processes, such as the spreading of WH-question marker over a partial or complete statement. The only limitation is that a specification for the beginning or ending of a co-occurring process is always synchronised to the onset or the conclusion of a lexical item.

A set of challenge data in the EASIER notation consists of a set of utterances in the notation system, and a corresponding set of synthesised sentences. See Table 3 for a set of sentences in DGS (German Sign Language), specified in EASIER Notation. Note the incorporation of PAD representation for affect information. These sentences have been generated automatically via the Paula avatar and a zip file of the animations utterances is available at:

https://drive.google.com/drive/u/0/folders/1T7won0DeM3Fyi12hbBLik6_S9v6q-Xt6

Table 3: A set of Sentences in DGS, specified in Easier Notation.

1. <affect=PAD(0,1,0)> ICH BALD ZIEHEN.
2. HALLO JETZT BEREIT ANFANG.
3. BITTE DU NOCHMAL <YN-q> GEBARDEN
4. <affect=PAD(-1,1,0)> ENTSCHULDIGUNG ICH FALLEN.
5. <affect=PAD(0,1,0)> BITTE WARTEN, BALD ANTWORT.
6. <affect=PAD(1,1,0)> DANKE, DU APP BENUTZEN. <affect=PAD(1,1,0)> HAPPY_TIPPEN.
7. <yn-q>BRINGEN DU PIZZA?
8. <affect=PAD(1,1,0)> #JOHN #MCDONALD BEKANNT.
9. ICH MAG MEIN KATZE
10. WARTEN BITTE, <affect=PAD(1,1,0)> PERSON KANN JETZT EINTRETEN.
11. <affect=PAD(-1,-1,0)> GESCHICHTE BEKANNT. <YN-q> DU VERSTEHEN.
12. <neg> WARTEN BITTE </neg>. PERSON KANN EINTRETEN.

4 AZEE NOTATION

AZee is a formal representation approach to SLs [9]. It is based on the essential notion of production rule, from which discourse expressions are built. Such a rule is a mapping between a meaning, e.g. “cold, winter”, and an articulated (signed) form, e.g. that given in Figure 1.

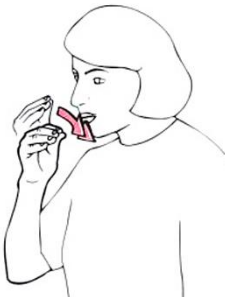


FIGURE 1: SIGNED FORM FOR "COLD, WINTER" IN LSF (FRENCH SIGN LANGUAGE).

A defined production rule can then later be *applied* like a regular programming language function to generate its output form. This is written using a colon prefix before the name, e.g., “`:cold`” with our example. The output is usually a *score*, i.e., a timeline specifying the required articulations, movements, postures together with the necessary timing constraints to synchronise them. Such AZee-generated scores can then be used by an avatar to render the production. This has been done for example with Paula, the avatar used in EASIER.

Production rules can also be parameterised. For example, the rule “`almost-reaching`” carries the meaning “almost *sig* but just not quite” and produces the form *sig* with a controlled overlap of the facial expression shown in Figure 2. Argument “*sig*” can be any signed score with its own meaning, itself recursively built with nested rules. For instance in our corpus [9], about a second championship title almost won but lost by a thin margin, a signer produces the following expression, in which “`second (in order)`” carries the eponymous meaning (“after first”):

```
:almost-reaching
    'sig
    :second (in order)
```

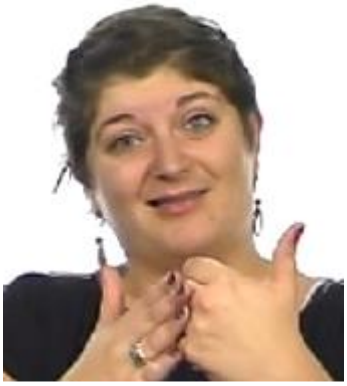


FIGURE 2: FACIAL EXPRESSION PRODUCED BY "`almost-reaching`".

Funded by the Horizon 2020 Framework Programme of the European Union

Notice the AZee indented syntax, where arguments of an applied rule are named with a quote mark prefix on the preceding lines. It is also possible and equivalent to use the pipe character to pack expressions on one line:

```
:almost-reaching| 'sig| :second (in order)
```

In the AZee paradigm, sign sequence is explained by the recursive application of rules that are parameterised with more than one argument and whose produced form is the sequence of their arguments. For example, rule "`info-about`" has two named arguments "*topic*" and "*info*", and generates the timeline of Figure 3. Its meaning is: "*info* given about *topic*". The phrase "[the/this] house is pretty" could therefore be rendered in LSF using the input expression (assuming the eponymous rule "`pretty`"):

```
:info-about
   'topic
   :house
   'info
   :pretty
```

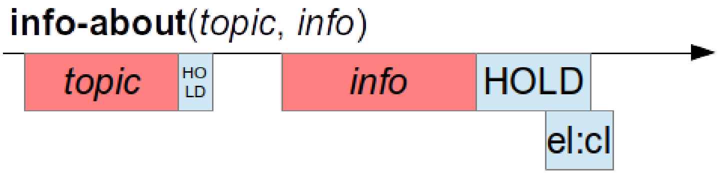


FIGURE 3: TIMELINE PRODUCED BY "`info-about`".

For full-size examples of AZee expressions representing actual signed discourse, we invite the reader to visit the AZee directory of the 40 brèves corpus. It contains 120 entries of 30 seconds' worth of signing on average, totalling 1 hour of LSF covered with AZee. This is what we suggest can be used as challenge data.

https://www.ortolang.fr/market/corpora/40-breves

Notice that non-manual features accompany the sequence (in this instance an eye blink), as well as timed holding blocks. We observe that rules never consist in plain sequences only. Incidentally, sync rules controlling the articulations over time do not have to rely on sign (gloss) boundaries, which contrasts with any other system at this point.

The advantage is that avatars become livelier, animations more natural as they part from the typical robotic rhythm induced by plain concatenations of units with similar transitions between them. Meaning is always inherently given to (and the origin of!) an observed sequence in AZee, and resulting productions always filled with head tilts, shoulder line rotations, eye blinks, eye gaze orientations, and manual holds. Thus, AZee supports co-occurring processes that are not necessarily tied to the onset or termination of a manual production (lexical item).

5 CONCLUSIONS

These three formats offer alternatives for specifying aspects of signed languages that attempt to expand the descriptive coverage of signed languages in a way that facilitates its portrayal via signing avatar. Perhaps it is too soon to designate any of these formats as a standard, but we plan to submit these as a paper to the 2024 LREC Sign Language Workshop with the hope

Funded by the Horizon 2020 Framework Programme of the European Union

that they will spark future discussion and development of automatic sign language generation through avatar technology.

REFERENCES

[1] De Sisto, M., Vandeghinste, V., Egea Gómez, S., De Coster, M., Shterionov, D., & Saggion, H. (2022). Challenges with sign language datasets for sign language recognition and translation. In Calzolari N, et al., editors. LREC 2022, 13th International Conference on Language Resources and Evaluation; 2022 June 20-25; Marseille, France. Paris: European Language Resources; 2022. 10 p.. European Language Resources Association.

[2] Wolfe, R., Braffort, A., Efthimiou, E., Fotinea, E., Hanke, T., & Shterionov, D. (2023). Special issue on sign language translation and avatar technology. Universal Access in the Information Society, 1-3.

[3] Ekman, P., & Friesen, W. V. (1978). Facial action coding system. Environmental Psychology & Nonverbal Behavior.

[4] Mehrabian A. (1996), Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament, Current Psychology 14:261–292.

[5] Rios, A., Reichel, U., Bhuvaneshwara, C., Filntisis, P., Maragos, P., Burkhardt, F., Eyben, F., Schuller, B., Nunnari, F., & Ebling, S. (2023). Multimodal Recognition of Valence, Arousal and Dominance via Late-Fusion of Text, Audio and Facial Expressions.

[6] Huang, M., Ali, R., & Liao, J. (2017). The effect of user experience in online games on word of mouth: A pleasure-arousal-dominance (PAD) model perspective. Computers in Human Behavior, 75, 329-338.

[7] Dreuw, P., Stein, D., Deselaers, T., Rybach, D., Zahedi, M., Bungeroth, J., & Ney, H. (2008). Spoken language processing techniques for sign language recognition and translation. Technology and Disability, 20(2), 121-133.

[8] Hanke, T., König, L., Konrad, R., Kopf, M., Schulder, M., & Wolfe, R. (2023, June). EASIER Notation–a proposal for a gloss-based scripting language for sign language generation based on lexical data. In 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW) (pp. 1-5). IEEE.

[9] Challant, C., & Filhol, M. (2022, June). A first corpus of AZee discourse expressions. In Language Resources and Evaluation Conference.

Funded by the Horizon 2020
Framework Programme of the European Union