Natural Synthesis of Productive Forms from Structured Descriptions of Sign Language

John Mcdonald, Michael Filhol

Natural Synthesis of Productive Forms from Structured Descriptions of Sign Language

John McDonald[1], Michael Filhol[2]
[1]DePaul University, Chicago, IL, USA, jmcdonald@cs.depaul.edu
[2]LISN, CNRS, Univ. Paris–Saclay, France, michael.filhol@cnrs.fr

2021

Abstract

Natural animation of Sign Language directly from linguistic descriptions continues to be a challenge especially in cases where the forms involved are more productive, such as geometric depictions. Prior work laid the foundation for natural Sign Language synthesis with the Paula animation system directly from AZee linguistic descriptions. This paper considers more elaborate discourse, composed of several clauses linked together by the overall meaning and involving largely productive signing. We make the case that one of the keys to natural animation of such discourse lies also in the segments between the typically annotated signs, in other words on the segments traditionally termed "transitions". By studying an example discourse video and the corresponding motion capture, we progressively build an efficient linguistic description of it and specify how to animate it naturally. Sign Language Avatar Transitions Discourse Motion Controls

1 Introduction

Signing avatars continue to be an active area of research due to their potential in a variety of applications for Deaf-hearing communication including:

- an output target for translation in short, scripted situations where interpreters are generally not available (Lancaster et al. 2003);
- a display to hide the identity of a signer in online communication (Kipp et al. 2011);
- a tool for Sign Language education (Jamrozik et al. 2010).

For an avatar to serve in these capacities, it must be able to produce the full range of motions and linguistic structures in Sign Language with a naturalness that does not distract the viewer from the message being communicated. Unfortunately, the naturalness and breadth of communicative ability of signing

avatars remains a challenge due to a variety of factors that cause the avatar to be judged as robotic or strange.

A primary cause of robotic synthesis is the lack of subtleties in the movements of the human body which contribute to the richness of the language in both linguistic and bio-mechanical capacities. These subtleties include

- the position and orientation of joints;
- motion details including acceleration profiles of joints;
- relative timing and coordination of the movements of body parts.

In fact, the pacing and dynamics of the body's motions can be influenced by both the emotions of the signer and by the grammatical structure of the discourse (Johnson and Liddell 2011; Wilbur 2017). Reproducing such dynamic subtleties is critical for synthesis to be judged natural. The avatar's linguistic input must represent, and the animation system must reproduce, such variations in timing and movement.

Recent efforts in both synthesis and representation have focused on including prosodic features in an effort to improve the naturalness of synthesized Sign (Adamo-Villani and Wilbur 2015; Filhol et al. 2017), but many avatars still do not incorporate such features. In addition, while many posture and motion details are captured by linguistic descriptions, others will necessarily be simplified in the process of encoding linguistic meaning. Yet the avatar must include such details in order to sign naturally.

2 The principle of *the coarser the better*

In an effort to increase the communicability of a signing avatar for both the range of Sign processes supported and the subtleties in pose, motion and timing of the avatar, this work builds on a principle, articulated a few years ago for Sign Language synthesis: animation will tend to be more natural when built from larger segments of discourse, e.g. a lexical sign or an entire non-manual process, rather than from a sequence of phonetic or articulatory constraints (Filhol et al. 2017). Examples of this can be seen in the two most common methods of synthesis, namely motion capture and keyframe animation.

Motion-capture synthesis, in which signing is recorded from human signers, can provide more natural animation when larger phrases are played back as recorded without alteration. But this is rarely possible when synthesizing novel utterances, and active research is focused on stitching together small segments of motion capture on the entire body, and on discrete parts of the body to layer non-manual processes (Gibet 2018). Such systems rarely build signing from phonetic descriptions, though generating motion capture based sign from has long been a goal as in (Gibet et al. 2011). Most often, however, this approach is forced to work with very large segments due to the nature and density of the data. Motion capture data does, however have another use in the pursuit of signing avatars. It provides a wealth of highly detailed measurements of human

motion that can be used to build procedural models of sign processes. In the present work, motion capture data is used exclusively in this capacity.

Keyframe animation has resulted in highly natural signing when the entire discourse has been directly animated by a human (Lombardo et al. 2011). However, keyframe animation computed directly from phonetic descriptions, while more flexible, is far more robotic with little coordination between body parts (Ebling and Glauert 2016). The advantage of keyframe data is that it is sparser than that of motion capture, and thus easier to edit and combine larger segments of animation, such as pre-animated fixed signs, into longer discourse (Wolfe et al. 2011). Coupled with a linguistic description that provides grammatical and prosodic context, the resulting synthesis can be more natural than when driven from phonetic descriptions (Filhol et al. 2017).

While working with larger animation segments is a useful goal, much of signing is highly productive and resists efforts to synthesize with large segments. Proform constructs such as classifier placements and size and shape specifiers have highly context sensitive and flexible movements that begin and end at any point in signing space, with the signer's hand and arm in arbitrary orientations (Schembri 2003; Woll 2007). When signers describe entire scenes that incorporate the size, shape, and movement of objects and actors in the scene, it is clearly impossible to either record or pre-animate all of the possibilities. An avatar must animate and combine the movements in such a way that they can be interpreted with all the right linguistic distinctions, for example, between communicating:

- placing or moving an object in sign space;
- placing collections of objects in space relative to each other.
- describing the size and shape of objects or scenery;

Prior work in this area has centered on *classifier predicates* and has focused mainly on synthesizing positional variability in such constructs (Huenerfauth et al. 2006; López-Colino and Colás 2011; Filhol and McDonald 2018). However, capturing the subtleties of context, shape and action in the situations above requires variation in posture, motion and timing that avatars have heretofore struggled to capture.

The present work will build on prior efforts to use the Paula avatar to synthesize signing directly from AZee linguistic descriptions, extending the capabilities of the synthesizer to encompass the cases listed above, by breaking these motions into smaller animatable units. These two systems are appropriate for this study as they allow the flexible specification and scheduling of signing processes in a multilinear fashion, i.e. multiple tracks with arbitrary unsyncrhonized timing, (McDonald et al. 2017; Filhol et al. 2017), however the lessons gleaned from the implementations presented are applicable to any system that strives to animate proform constructs.

Regardless of the size of the animatable units, whether they be single classifier placements, complete recordings of lexical items, or longer discourse, the

system must combine and coordinate them using grammatical and prosodic information from a linguistic description coupled with knowledge of human motion. To exemplify these issues, this paper presents a detailed case-study of an actual signed depiction of a scene, and develops a full linguistic description for it by observing the details of the motion that we must reproduce on the avatar.

3 Case study: describing a dining room table scene

Let us study the full discourse utterance given in the video titled "LSF-table-scene-signed.mp4" at `http://sltat.cs.depaul.edu/2019/mcdonald.mp4`, in which a table is presented on a rug, set with various items on and around it. The arrangement described is displayed in figure 1, the example excerpt being restricted the description of the rug and table.

Figure 1: Elicitation image with the example table scene (Benchiheub et al. 2016)

3.1 Overview of the example Sign production

Readers familiar with Sign Language will identify six consecutive segments comprising the whole discourse, each introducing one or more objects to the scene. We list and label these segments below, in order of appearance, together with the contents they introduce:

S1 a rug on the floor;

S2 a table on the rug;

S3 four chairs around the table;

S4 four plates on the table;

S5 four glasses on the table;

S6 four pairs of cutlery on the table.

Figure 2: S1 sequence (rug): fixed sign, shape, placement

For example, S1 begins with a fixed sign (RUG), then presents the shape of the rug by drawing its outline low on the ground and finally reaffirming its position in the scene (see figure 2).

In a similar fashion, the next segment S2 places a table on top of the rug, see figure 3. It is placed using a classifier to indicate a large flat rectangular object positioned and oriented parallel to the floor. To disambiguate it from any other flat rectangular object, such as a painting or a rug, a fixed sign TABLE precedes it, just like the fixed sign RUG at the start of S1.

Figure 3: S2 sequence (table): fixed sign, placement

All 6 segments follow the same construction pattern, made of two parts. The first gives the kind of object about to be placed, while the second explicitly places those objects in the scene, possibly with other information such as orientation, shape, etc.

The first parts in each of these examples consists of frozen signs, e.g. RUG in S1. The only exception is S6, which contains a combination of two of frozen signs (FORK + KNIFE) to mean cutlery more generally. The second halves are more variable in content but all involve one or more placements, often oriented, as follows:

- in S1, the rug is given a shape, drawn with the fingers, and a position low in the scene;
- in S2, the table is placed as a flat shape above the rug (around the point labelled P in the schematic representation of figure 4);
- in S3, two pairs of chairs facing each other are placed, one pair after the other;

- in S4, four plates are placed in a rapid sequence on the table (at $P_1, ..., P_4$);
- in S5, four glasses are placed near the same points, two-by-two in a way similar to the chairs;
- in S6, four pairs of cutlery (introduced as fork & knife) are placed near the same points again, both hands working simultaneously to place each pair.

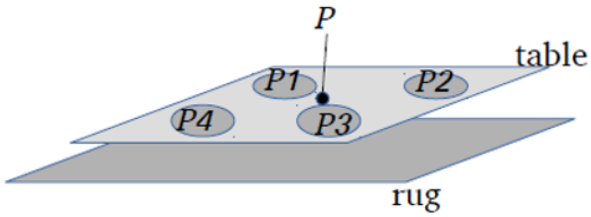

Figure 4: Diagram of the layout of the table scene

Every "placement" in the description above is performed with a classifier construction. Some placements are performed simultaneously with a one-hand proform on each hand (2 pairs of chairs, 2 pairs of glasses, 4 knife & fork pairs), while some use both hands for a single proform placement (rug, table, each of the 4 plates). Also, some of the placements are oriented in an iconic fashion (chairs facing the table and each other, all cutlery items pointing inwards on the table), whereas others just take a default orientation (rug, table, plates, glasses).

Looking at the production in finer detail, one might notice that eye gaze is directed to signing space in each of the second halves of the segments listed above, in contrast to the respective first part, where it is directed to the addressee. Differences in the dynamics are apparent as well, and will be a focus of later study in the present work. All of these consideration will serve one main goal: to replicate this whole scene through an avatar with an entirely automatic process.

3.2 Traditional annotation

Figure 5 shows how the utterance would traditionally be represented, labelling units ("glosses") and assuming "transitions" in between. In the figure, the classifier names have been abbreviated to keep the diagram as short as possible.

Simply animating these units sequentially will produce less than ideal results. Analogous to early speech synthesis systems that chained word recordings, or synthesised phonemes, one after the other and produced unnatural vocalizations, animation systems that have relied on such chaining of lexical productions also result in robotic output. Changes in pacing, timing, interpolation and in the magnitude of motion must occur in the avatar as it expresses these gestural

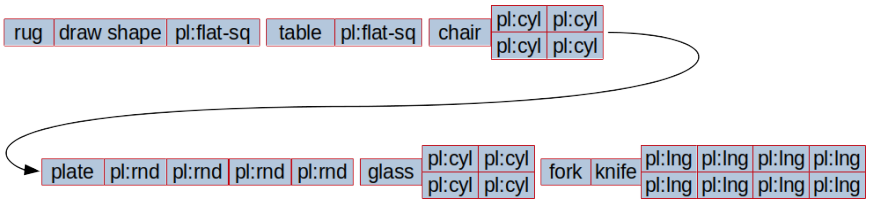

Figure 5: Traditional linear arrangement of labeled units

units, and will depend on their motivation in meaning and in what context they are signed. It is these aspects that make up the rhythm of signing that is the hallmark of natural motion as opposed to a uniform robotic production.

3.3 AZee

AZee is a formal approach to SL description built entirely on the linking of visible articulations (forms) and semantic functions (meaning). Signed forms in the native AZee representation are expressed using necessary and sufficient constraints on body articulations and on their synchronisation on the time line. The former include any relevant articulator and motion specification (no fixed set of, say, manual parameters); the latter includes duration of movements and transitions, but also between them if relevant and necessary.

In addition to signed forms, AZee allows to write functional values, i.e. functions that can be applied to arguments to produce a return (result) value. A function in AZee is called an *AZop* ("AZee operator"), and its arguments are named. Like any other value, an AZop can be named for later use.

The AZee approach to formally describing a Sign Language is to identify the meaning–form mappings in the language, and to write an AZop for each one. Such an AZop, parameterised by its meaningful variations, specifies the form part of the mapping as its return value, and is named after its meaning. This creates what is otherwise called a "production rule" for the target language. For example, the previously published production rule "`restaurant`" is defined as an AZop, with an optional argument *loc* (for location) since it is relocatable and the induced variation is meaningful. The return value (signed form) is the gestural form of the corresponding sign, possibly relocated according to *loc*.

If a set of enough production rules is made available to cover the target language, one can write expressions from it to represent signed utterances of any length, from a single sign to an entire discourse. Such AZee expressions all evaluate to *forms*, and therefore can be input to an animation system to render a video output. They are built with AZops carrying *meaning*, which is the point we wish to emphasise here.

Building expressions from the rule set indeed enforces that any sequence be generated by an AZop that was given meaning. In other words it becomes impossible to write a sequence for no semantically-grounded reason. Therefore we avoid the loophole described above where avatars seem to transition from one

piece to the next without conveying meaning overall. Any so-called transition generated from an AZee expression will be the result of a meaningful operation and its form will be controlled, timed and accompanied by auxiliary gestures as necessary. In short, and referring back to our example, the spaces between the labelled (blue) units in the diagram are no longer blank time fillers, but rather segments under the same degree of motion control.

The next sections explore the commonalities in the forms of our use case video, and in their meaning, introducing AZee production rules where appropriate to represent their association. AZop by AZop, we build an AZee expression that represents the entire use case video. As explained above, it will cover the forms of both the labelled units and the transitions in between. We will see that it is accomplished with a very small set of rules, a statement on the economy of the approach that we return to at the end of the paper.

4 Representing the discourse with AZee for more natural animation

In order to represent this discourse with AZee, we begin with an exploration of how each of the labelled units in figure 5 can be represented as AZops. We then look at the forms and meaning of the transitions to find the linguistic processes that knit those units together into a description of the whole table scene. Encoding these processes with AZops will give meaning and purpose to the transitions while simultaneously defining the necessary non-manual signals and timing adjustments within the labelled units, all of which are needed for the avatar to produce natural signing.

4.1 Labelled units

Some of the labelled units have a fixed and almost invariable form that can be encoded as such. We deal with those in a first section. In contrast, others units behave with much more internal composition of highly variable geometric elements, which we address afterwards.

4.1.1 Fixed units

Some of the labelled units are fixed in their articulated form, for example "RUG", "KNIFE", "FORK", etc. The form of each such entry is invariable and has a consistent meaning. This very fact defines an AZee linguistic "production rule", like the `restaurant` case mentioned above. In such cases the name is comparable to an ID-gloss (Johnston 2010), and the form specification is a conjunction of articulatory constraints (orientation, positioning), sometimes expressed around an optional location argument like `restaurant` or `table`. For instance, the first sign of the utterance can be represented by an AZop named `rug` and containing a set of articulatory constraints that produce the (manual) gesture in the left image of figure 2.

Then, the AZee expression below, which is a simple application of `rug` with no arguments[1], which will result in the correct articulations specified for the avatar to render:

```
rug()
```

To animate it however, a direct application of a set of articulatory constrains is precisely what what we are trying to avoid since it tends to produce robotic motion. But since the form of such an expression is fixed, the synthesis system is free to shortcut on them, bypassing the form specification to substitute a pre-existing animation of the sign, resulting in a more natural motion. This mechanism was proposed in earlier work, and called a *shortcut* (Filhol et al. 2017).

4.1.2 Geometrically productive units

In contrast to the fixed units, this section addresses those whose contained movement is variable in space, highly dependent on context like the classifier placements. The obvious problem in terms of animation will be that such units are too unlikely to be reused in other discourse exactly to be usefully treated with coarse pre-animated blocks.

The classifier placements are performed through a small downward movement we call "settle" (Filhol and McDonald 2018), and a chosen proform, for example `prf-flat-square-large` for large oblong surfaces. While the location changes in space, there is nevertheless a great deal about the form of each placement that is constant, including an eye gaze towards the object that precedes the settling movement. The meaning for each of these cases is consistent: placement of an entity in the scene, of the type indicated by the chosen proform, at the location where the movement settles.

Justified by the form–meaning association above, the AZop `place-proform`, with arguments *prf* (proform) and *loc* (location), was introduced in our work (ibid.). Its block diagram is given in figure 6, where "eg:loc" specifies the eye gaze directed towards *loc*.

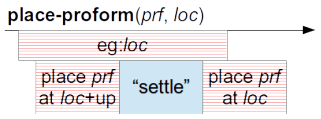

Figure 6: Block diagram for `place-proform(prf,loc)`

For example, the table placement at the end of S2 would be generated by the following expression, where `prf-flat-square-large` is the two-handed proform with two L-shaped hands in the same plane, and P the table center as illustrated in figure 4, around which the proform settles:

[1]The sign being relocatable, `rug` would accept a *loc* argument like `restaurant`. But in our video, it is applied without relocation as it is performed generically. The rug entity is nonetheless placed with what follows in the utterance.

`place-proform(`*prf*`=prf-flat-square-large,` *loc*`=`P`)`

When two placements are performed together like with the chairs (twice), the glasses (twice), or the cutlery (four times), the appropriate AZop to bring into play is `simultaneous`. Its form is simply to perform its arguments at the same time (in parallel, as shown figure 7), which is iconic of its meaning, namely that they are true or happen at the same time.

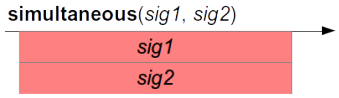

Figure 7: Block diagram for AZop `simultaneous(`*sig1*`,`*sig2*`)`

For example, the expression below places a knife at point P_k and a fork at P_f, simultaneously as two small elongated objects using proform `prf-elongated` for both:

`simultaneous(`*sig1*`=place-proform(`*prf*`=prf-elongated,`
loc`=`P_k`),`*sig2*`=place-proform(`*prf*`=prf-elongated,` *loc*`=`P_f`))`

These `place-proform` expressions are not fully fixed in form, and indeed are infinite in number because their *loc* argument takes its value from a continuous space. Thus to animate them, full shortcuts as for fixed signs are not possible. But the dynamics being constant, some fixed elements of motion can be triggered with the templating system introduced in the referenced work.

In this templating process, the AZee system is providing the classifier and the point in signing space at which it should be placed. The classifier itself most often indicates the shape that the hand will take[2]. But it is important to note that the specification of that handshape also influences overall posture including the height of the elbow, and the shoulder. The avatar synthesis system may draw on a pre-defined artist pose to set this additional posture information before retargeting the settle motion at the specified point.

The final gestural unit that we encounter in this description is the tracing of the outline of the rug. The signer performs this motion with a pointing handshape on each hand, and with the hands starting near the center line of the body. The hands then simultaneously trace, in a mirrored fashion, the four sides of the rug with an accompanying eye-gaze towards the object being depicted. Gestural units such as these are very similar to the shape deployments recently addressed, which make use of an AZop named `deploy-shape` (Filhol and McDonald 2020).

4.2 Transitions between labelled units

Now that we have covered the labelled blocks, explaining how they can be individually represented and animated, the next step is to define how these units

[2]Classifiers also can specify wider shapes on the body as, for example, when placing a tree. In that case the entire forearm and hand become the tree to be placed.

should be placed on the timeline to build the narrative. The annotated figure 8 reveals a total of 21 transitions, loosely grouped in categories enumerated below, where the indicated markings match those in the figure:

1. the transitions between the segments S1–S6 (5 solid vertical lines);
2. the transition between the two parts of every segment (6 dotted vertical lines in the figure), i.e. between the type/kind of item and the position/orientation of the placed items;
3. transitions between items that are grouped items into collections, e.g. of 4 plates, glasses, etc. (8 transitions marked with a diamond symbol);
4. the last two remaining (marked with a star), i.e. the transition between the size and shape specifier for the rug and its placement, and the transition in the juxtaposition of the fixed signs FORK and KNIFE.

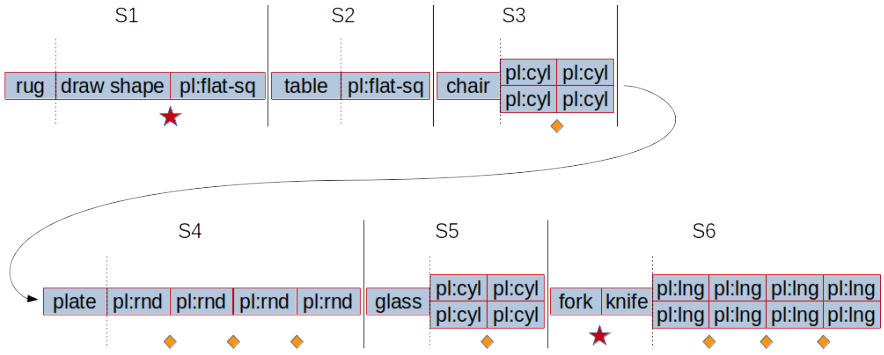

Figure 8: Annotated linear arrangement

As stated earlier, a sequence can only be generated from AZee production rules conveying meaning. We must therefore choose which AZops to combine. We do this by investigating the dynamics, as recorded by motion capture and video data, over the transitions.

4.2.1 Category and Side-Info

Considering the signer's motion during each of the six segments, a clear pattern emerges. The transition between the two parts on either side of the dotted vertical lines in figure 8 is fast (approximately .2 s on average), and the signer systematically raises and tilts her head with an accompanying raising of the eyebrows. This pattern has been consistently observed in other signers as well and is unique to this particular type of transition. If the avatar is to correctly communicate this contextual meaning, it will have to incorporate this head and eyebrow signal, regardless of whatever else is happening simultaneously in the discourse.

The consistency of this form, featuring two juxtaposed parts which we call *cat* and *elt*, and the consistent association with the meaning that "*elt* is to be understood as of category *cat*", is captured by the AZop `category`, with arguments *cat* and *elt*. A block diagram for `category` is given in figure 9. For example, S2 is represented by the following expression:

```
category(cat=table(),
elt=place-proform(prf=prf-flat-square-large, loc=P))
```

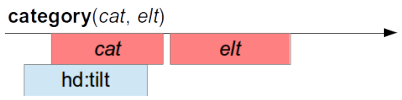

Figure 9: Block diagram for AZop `category(cat,elt)`

This single AZop is sufficient to handle all six of the dotted line transitions in figure 8. Either argument in a `category` expression may be a single glossable unit, such as a fixed sign like the *cat* argument in S2 above, or a compound statement, such as the collection of plates in S4 or the pair in the second part of S1, whose inner transition is starred in figure 8.

This transition between the shape depiction of the rug and its placement, is similar in form to that of `category`. It is a fast one, and involves a similar head/brow movement. However, the head motion is synced differently, as it immediately precedes the second of the two blocks rather than the first. To understand the difference and see what subtleties the avatar will need to capture, consider the actions during the production of the shape specifier and the classifier placement for the rug. We notice several things:

- the signer's gaze is down towards the shape being drawn during the deployment;
- the time between the two signals is about 5 frames (.2 s);
- the signer's head and gaze rise to the addressee for the second signal, and the eyebrows raise;
- the two signals are both signed at a normal pace.

This juxtaposition indicates that additional information is being appended. Note that tracing with the fingers is enough to define the size and shape of the object, which the `category` rule identified as a rug, but then to firmly anchor that rug in signing space, the signer provides additional information: it is placed "here".

This is exactly the form produced by the AZop `side-info`, whose arguments are *focus* and *info* and whose meaning is that "*focus* is given an additional, although linguistically non-focused, information *info*". The generic form it produces when applied is represented in the block diagram of figure 10. This interpretation is consistent with the signing in the video. The second half of S1

can therefore be captured with a straight-forward application of the production rule `side-info`:

```
side-info(focus=[draw rug], info=[place rug])
```

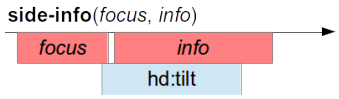

Figure 10: Block diagram for AZop `side-info(`*focus*`,`*info*`)`

4.2.2 Grouping items in collections

Other transitions group multiple items into a collection or list. Our discourse contains several such lists, identified in section 3.1:

L_1 two pairs of chairs facing each other across the table;

L_2 four plates on the table, one at each place setting;

L_3 two pairs of glasses;

L_4 four sets of cutlery.

Examining the motion capture data for the signer's wrist height in the first four placement lists L_1 through L_4, reveals that there are two very different kinds of motion present. The first placement, for the chairs on either side of the table, is deliberate and has a significant pause between the two pairs of classifier placements. The other placements are quicker with a "bouncing" motion between the placements. The motion curves (in height over time) for these actions in figure 11 clearly demonstrate the differences.

The first transition in the chair placement is very different from the others since between the placements is a clear hold or pause, and a hint of a blink, or at least a raising of the eyelids between the placements. In addition, the placement of the chairs is noticeably slower than the others. Measuring the total duration of the placements (the upward and downward motions in these graphs) yields average timings per placement of chairs (13 ms), plates (10 ms), glasses (11 ms) and cutlery (10 ms).

Thus, the placement of the chairs is on average about 25–30% slower in its placement, and feels more deliberate because of both the pacing and the pause between the placements. It is also worth noting that while the performance of the placements of the glasses is a little softer than that of the plates and cutlery, the glass placements are clearly not as slow as the chairs, and the motion curves do not show a clear plateau between the motions. Whether the slightly slower placement of the glasses is due to normal human variation in production or carries linguistic meaning will be investigated in a future study.

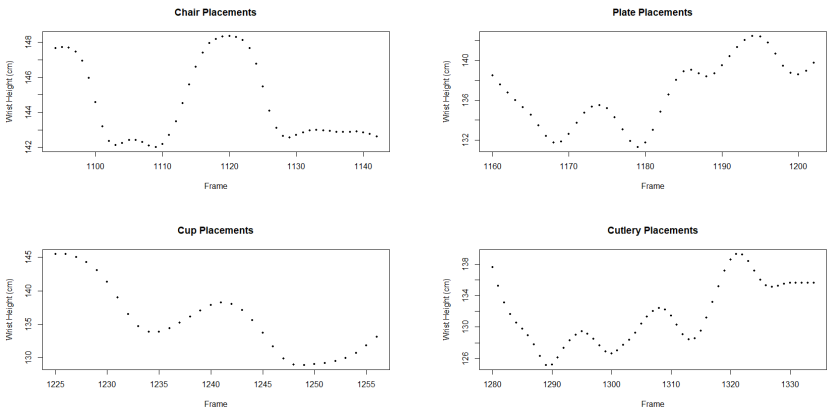

Figure 11: Vertical movement during classifier placements

In all, we see two kinds of collections presented here, one with holds and slower movements, which we will call form A, and the other with squeezed or faster item placement and short transition times with no hold, denoted as form B. Interestingly, the signer also holds and blinks at the end of S3 through S6. This is the same form A as observed for the list of chairs for example, only at a higher level of the discourse structure. Moreover, the meaning is compatible with the interpretation of a collection as well.

The meaning conveyed each time with form A is that of a closed enumeration of items, each given equal specific focus, without precedence or emphasis on any particular one. In AZee, this consistent form–meaning association is supported by an AZop called `each-of`, whose generic form A can be represented by the block diagram in figure 12.

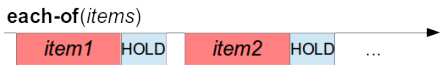

Figure 12: Block diagram for AZop `each-of`

The enumeration made of S3 through S6 is written as follows, where `list` is the native AZee operator for extensional list construction:

E_{tbl}: `each-of(`*items*`=list(S3, S4, S5, S6))`

In each exhibited occurrence of form B, the meaning is also to form an exhaustive collection like A. In contrast, to A however, none of the individual contents is emphasised, rather it acts to focus on the set. For example, four plates are placed around the table (and no more), and no emphasis is placed on any of them in particular.

The same form B also appears elsewhere, not involving placements, namely in the first half of S6, i.e. between fixed signs FORK and KNIFE on the second

starred transition of figure 8. The evidence that these form a list in the pattern of B is two-fold:

- the two signs are performed with no hold between them, as shown in figure 13;
- the duration of each sign is shorter than with isolated signs of the same form (FORK has a similar motion as GLASS but FORK lasts merely 9 frames as compared with 12 for GLASS; KNIFE is even truncated).

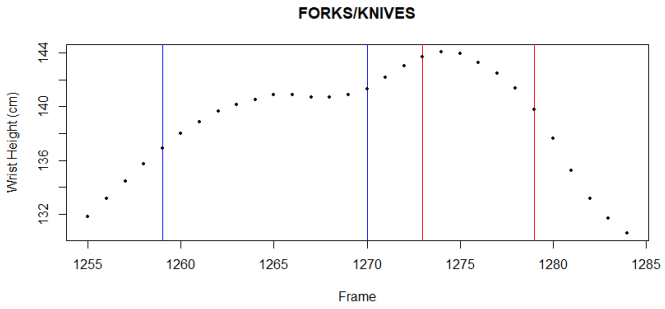

Figure 13: height of the wrist during the knives(blue)/forks(red) pair

In terms of meaning, we understand that there are exactly two items (closed enumeration), but the focus is on the set (pair) as a whole, not on the contained items. This is consistent with what follows as only pairs are placed afterwards without any detail on which is which. This consistent meaning–form association is supported by the AZop `all-of`, whose generic output form B is illustrated in figure 14. The four plates in the second half of S3 can be generated with

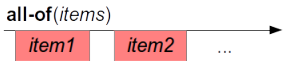

Figure 14: Block diagram for AZop `all-of`

the expression below, using the native `for` operator to generate a list of four similar placements by iterating on a list of points (see $P_1..4$ in fig. 4):

```
all-of(items=for p in list(P1,P2,P3,P4):
      place-proform(prf=prf-flat-round-large, loc=p))
```

The first half of S6 is encoded in the following AZee expression:

```
all-of(items=list(knife(), fork()))
```

Animating these types of collections requires that we investigate both the linguistically described forms and additional motion controls to transform the repetition of "settle" movements. There remain two transitions left unaccounted for, which we will return to afterwards.

5 Tuning Motion from Linguistic Descriptions

Our prior work (Filhol and McDonald 2018) explored the functionality necessary for the Paula avatar to use artist templates to allow natural placement and movement of classifier constructs from these kinds of AZee expressions. This section will explore the motion controls necessary to capture the dynamic differences between these simple placements, such as in the placements of the rug and table, and those in the each-of and all-of lists described in the last section. The motion controls that we will be using are similar to those animators have been manually using for a long time (Thomas et al. 1995), including:

- the shape of the motion path;
- the abruptness or ease in which a body part approaches or leaves a target;
- the synchronization in the timing of torso movement with the rest of the motion;
- other coordinated body motions that affect the perception of the movement.

However, the goal here is to trigger such features automatically from the linguistic descriptions in the previous section.

5.1 Isolated placements

In order to highlight the differences here, we recall, in detail, the motion generated from a simple "settle" placement. This is exemplified in the placement of the flat rectangular object to indicate the table's location on top of the rug as in figure 15a. In this situation, the avatar system can shortcut directly on the known "place-proform" process while filling in the necessary body postures and motions from an artist template, i.e. a pre-animated pose built by an artist as described in (Filhol and McDonald 2018). It then can use the timing and duration information directly from AZee to coordinate the manual and gaze processes.

As described in the previous work, the artist template provides several important cues for producing a natural posture and settling motion. A schematic plot of the motion is displayed in figure 16 where the horizontal axis is time and the vertical axis is the height of the signer's wrists. It is important to note that the hold at the end here is not intrinsically a part of this settle movement, but the ease, or softness with which the hands settle, is. The hold may come from an AZop that indicates the end of a phrase or clause.

5.2 Placements with each-of

The each-of list can be exemplified with a set of four plates on a table, say at points $p_1..p_4$, each anchored in its own positions using the AZee expression below. Note that this is a constructed AZee example, *not* faithful to the video content.

(a) Table (b) Plate

Figure 15: Single Frame from placements of objects

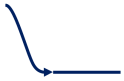

Figure 16: Diagram of a settle movement

```
each-of(items=for p in list(P1, P2, P3, P4):
        place-proform(prf=prf-flat-round-large, loc=p))
```

This expression takes a list of signed productions as an argument, and conveys the fact that each placement is applied in space, with no importance or precedence. The expression specifies the resulting forms to render, which consists of the expected sequence, with a specific holding time at the end of each item, allowing the interpretation of the above meaning. Figure 15b shows a single frame produced by this rule.

Since this process is a repeated application of the same movement at different spatial locations, the avatar system can simply apply the artist template as before several times with the additions of the holds specified by AZee. Figure 17 shows a schematic diagram of the resulting motion in this case.

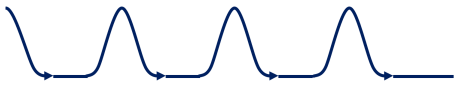

Figure 17: Diagram of movement for "each-of"

5.3 Placements with all-of

Things get a little more complicated when approaching the placement of the set of four plates. This time the signer uses quicker movements between subsequent items. In terms of meaning, the focus shifts from the individual items to the set formed by all of them together as expressed by the AZee expression already given in §4.2.2, and duplicated below. The resulting form specified by this new rule is a shorter duration or squeeze for each of the items, and does not specify hold blocks between them.

```
all-of(items=for p in list(P1, P2, P3, P4):
       place-proform(prf=prf-flat-round-large, loc=p))
```

In this case, however, analysis of corpus examples shows that the motion is altered in more ways than those provided by AZee. The downward placement actually ceases to "settle" and becomes a distinct bouncing between the placements. The top-down short-cutting system allows the avatar to distinguish the difference between the each-of and all-of. So, Paula is free to alter the motion within the bounds of the linguistic constraints to produce this bounce. This application of the "coarser the better" principle is in fact necessary here to provide the correct motion allowing the avatar to:

1. cause the arm's approach to the target point to be more abrupt instead of easing-in;
2. start the next cycle abruptly to complete the bouncing of the arm at the target point;
3. depending on the geometry of the classifier and the amount of arm motion involved, to shorten the stroke of the cycles to compensate for the squeezed timing;
4. alter the timing of the signer's eye and head movement in synchronization with the actions on the hands, with a more continuous progression.

The effects of all of these can be seen in figure 18 where the path bounces instead of coming in tangentially and the heights of the cycles are somewhat shorter than before. It is important to note here that this bouncing action really only makes

Figure 18: Diagram of movement for `all-of`

sense in the case of placements, the default case being only a squeeze. In fact, it is not at all clear that the all-of list connecting the FORK and KNIFE fixed signs exhibits this kind of discontinuity in velocity. This means that it is up to the animation system to decide when the bounce happens since it cannot be specified for us linguistically. The hierarchical description coupled with the templated shortcut system gives the synthesizer the needed freedom to do this. The synthesizer knows that it is building an `all-of` list and also that each of the items is a classifier placement and so can trigger the bouncing action.

6 Animation from the full hierarchical description

In the sections above, we have dealt with almost every transition exhibited in the example discourse. The only two left aside are the first two solid vertical

lines following S1 and S2 respectively. They do look much like the other inter-segment transitions, and in terms of meaning, they could be interpreted as part of the same list of objects, although not on/around the table but in the room. However, S1 and S2 both contain placements on which the following discourse sections depend:

- S1 provides the anchor point (rug on the floor) for the table that follows, which we interpret as placed on the rug;
- S2 provides the anchor point (table surface delimited around P) for the list of objects that follow, whose placements are all interpreted as relative to it (on/around the table).

Besides, while the difference is subtle in form, S1 and S2 each exhibit an ending hold duration that is slightly longer before the next segment begins.

A more appropriate AZee operator to represent the post-S1 and -S2 transitions is `context`, whose arguments are *ctxt* and *proc*. It produces the two in sequence while adding the (somewhat longer) hold at the end of *ctxt* (see block diagram in figure 19). Its interpretation is that the signed *proc* is in the context of *ctxt*. In this case, it is a signing space context. So, in our specific example,

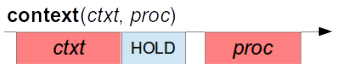

Figure 19: Block diagram for AZop `context(`*ctxt*`,`*proc*`)`

this AZop is applied twice at the very top level of the expression representing the entire utterance. Once S1 is the context (*ctxt*) for what follows, the top-level *proc* is itself divided in context S2 for the rest of the utterance:

`context(`*ctxt*`=S1, `*proc*`=context(`*ctxt*`=S2, `*proc*`=`E_{tbl}`))`

Piecing together the whole example discourse presented in §3.1, nesting the various expressions presented throughout this paper in one another leads us to a single expression for the whole utterance. Its recursive (hierarchical) structure can be represented graphically in the form of a tree, as is given in figure 20. For brevity, the figure only exhibits the rules generating transitions. Those generating the geometric units (proform placements) are abbreviated with the same labels as in figure 5.

The full AZee expression is available on request from the authors and the full animation produced by the Paula syntheszer is online at `http://sltat.cs.depaul.edu/2019/mcdonald.mp4`. A part of the animation score covering segments S1 and S2 is shown in figure 21. Recall that Paula's animation is scheduled by a system of tracks that can each control anything on the body (McDonald et al. 2017) and whose generated motions are seamlessly blended by the avatar:

- *Pre-Anim* and *Pre-Anim 2* that control both sides of the body based on pre-animated shortcuts and templates;

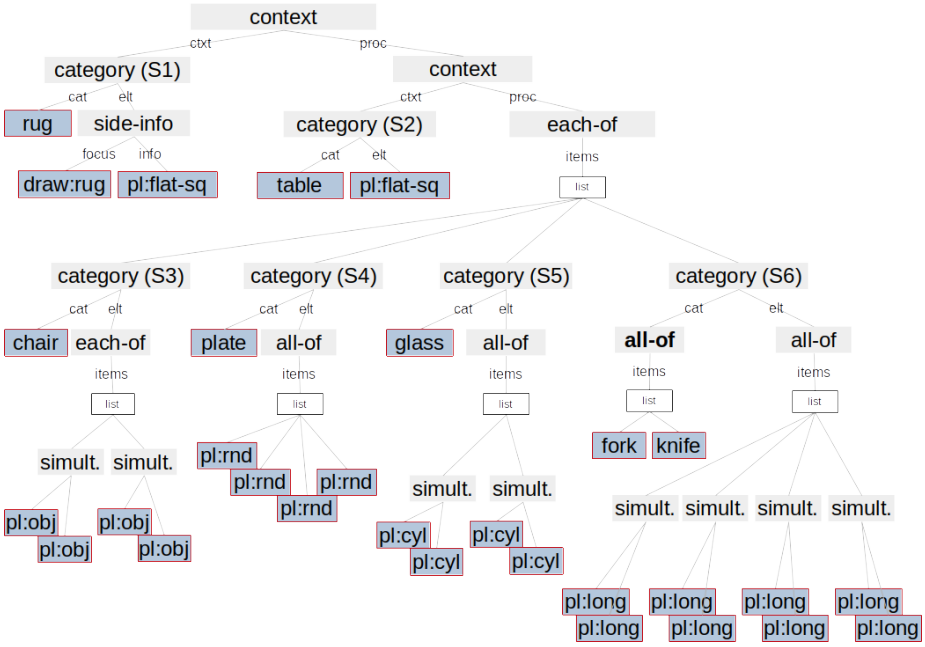

Figure 20: Table scene description in tree layout

- *Head Mvt* for head, torso and facial movement;
- *Blink* for scheduling blinks;
- *Gaze* for the avatar's eye-gaze.

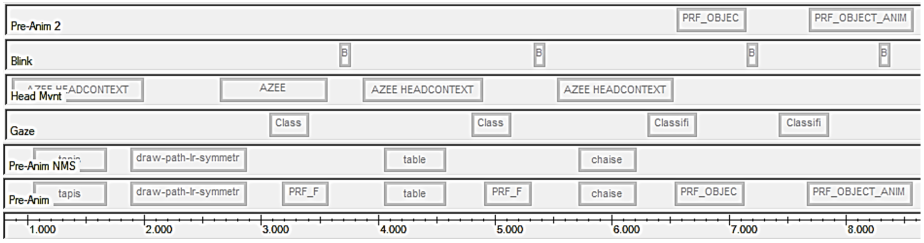

Figure 21: Paula avatar score built from Azee tree

It is worth mentioning again that this video has been produced directly from the AZee description with no intervention from an animator other than the construction of the shortcut animation dictionary and templates for the classifier placements. All of the relative timings between elements in the animation are scheduled directly from the AZee output.

To compare the results of the synthesis, figure 22 contains the relevant frames of the avatar's motion that compare to the signer's movement shown initially

in figures 2 and 3, in addition to the placement of later objects on the table. Notice that the non-manual signals in the fixed signs are consistent with the signer's as is the raising of the eyebrows for the rug placement after she draws the outline of the rug.

The signed depiction of the table setting is a very rich combination of elements that need to be each animated and then combined using the appropriate prosodic elements. From the description of the table scene in section 3.1, we can see that the overall depiction of the table is organized as several major groups of items, which are placed in relationship to each other. From the synthesis system's perspective, AZee is ideal as a description system for such discourse, because it organizes the description of the signing hierarchically so that the synthesis can shortcut at a variety of levels depending on the animation services at its disposal. In addition, it provides prosodic information including relative timing information for each process that will enable the animation system to coordinate all of the simultaneous signals in the discourse.

This entire scene description has been described linguistically with the following short list of AZops (identified and named linguistic production rules) that connect the labelled units:

- category(*cat*, *elt*), meaning *elt* as an instance of *cat* (used here at the top-level of every segment $S_{1..6}$);
- side-information(*focus*, *info*), meaning *focus*, about which additional information $info$ is provided;
- context(*ctxt*, *proc*), meaning *proc* occurring in the context established by *ctxt* (in our use case, anchoring a location relative to which items are positioned in *proc*);
- simultaneous(*sig1*, *sig2*), meaning *sig1* and *sig2* occur simultaneously;
- each-of(*items*), meaning the collection of *items* in the list, each with equal focus;
- all-of(*items*), meaning the set of *items* (most often of the same kind) as a single discourse entity.

This list largely overlaps the one already presented in (Filhol and McDonald 2020), which focused on complex shape deployments. The only additions here are *category*, *side-information* and *context*, which have been previously published in (Filhol and Hadjadj 2016). The parsimony of this AZop system is extremely encouraging, as it indicates that a relatively small set of AZee operators can generate extremely sophisticated signing, including the infinitely variable productive units, given the appropriate dictionary of pre-animated signs and proform templates.

7 Conclusion

The results in this paper expand on the ability of the AZee and Paula systems to represent and synthesize complex discourse through leveraging larger structures in sign language to link smaller units together in the discourse. The results are also not limited to these two systems but may be seen as a case study on how other linguistic and synthesis systems may be structured in order to achieve similar results.

This approach takes us from the traditional flat, linear paradigm of "whitespace" transitions between units that follow one another to a hierarchical, recursive approach able to represent connections between arbitrary chunks of signing. Transitions are now subject to motion control like the rest, and are no longer mere padding in the signing stream between relevant units of an assumed sequence.

The connecting AZee expressions do not just add meaning to transitions, but they also add forms, both of which are crucial to animation. Every rule that places one chunk after another is accompanied with a combination of gaze, head tilts, blinks, etc. All of these processes participate in the naturalness of the animation since they link various parts of the body during the discourse, so the avatar does not have a fixed stare at the camera or a rigid torso/shoulder line. Access to the meaning also gives the animation system necessary hints to add bio-mechanical forms when appropriate, for instance if there is a so-called head tilt, there might be spine involved as well, or in the case of the all-of list where motion controls are added to give the feel of a bounce when necessary.

In the future, we will be working to expand Paula's capabilities to leverage such hierarchical descriptions, and thus the types of discourse that can be synthesized with the combined systems. This will include expanding the rules that AZee offers and the types of templates and shortcuts that Paula can implement to animate them naturally.

In this study, our methodology was to synthesize signing that replicated that of a real signer, and it has succeeded in producing all of the linguistic elements present in the source Sign Language discourse. But it is important to note that the output of such a synthesis system must be tested for understanding, grammatically and naturalness with native signers and testing these AZee rules. Thus, testing the resulting animations with native signers will be critical to ongoing development and refinement of the system.

References

N. Adamo-Villani and R. B. Wilbur. Asl-pro: American sign language animation with prosodic elements. In *International Conference on Universal Access in Human-Computer Interaction*, pages 307–318. Springer, 2015.

M. Benchiheub, B. Berret, and A. Braffort. Collecting and analysing a motion-capture corpus of french sign language. In *Language Resources and Evaluation*

Conference (LREC), Representation and Processing of Sign Languages, Portorož, Slovenia, 2016.

S. Ebling and J. Glauert. Building a swiss german sign language avatar with jasigning and evaluating it among the deaf community. *Universal Access in the Information Society*, 15(4):577–587, 2016.

M. Filhol and M. N. Hadjadj. Juxtaposition as a form feature; syntax captured and explained rather than assumed and modelled. In *Language Resources and Evaluation Conference (LREC), Representation and Processing of Sign Languages, Portorož, Slovenia*, 2016.

M. Filhol and J. McDonald. Extending the azee-paula shortcuts to enable natural proform synthesis. In *Workshop on Representation and Processing of Sign Language, International Conference on Language Resources and Evaluation (LREC)*, pages 45–52, 2018.

M. Filhol and J. McDonald. The synthesis of complex shape deployments in sign language. In *Workshop on Representation and Processing of Sign Language, International Conference on Language Ressources and Evaluation (LREC)*, 2020.

M. Filhol, J. McDonald, and R. Wolfe. Synthesizing sign language by connecting linguistically structured descriptions to a multi-track animation system. In *International Conference on Universal Access in Human-Computer Interaction*, pages 27–40. Springer, 2017.

S. Gibet. Building french sign language motion capture corpora for signing avatars. In *Workshop on Representation and Processing of Sign Language, 8th Internationnal Conference on Language Resources and Evaluation (LREC 2018)*, pages 45–52, 2018.

S. Gibet, N. Courty, K. Duarte, and T. L. Naour. The signcom system for data-driven animation of interactive virtual signers: methodology and evaluation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1(1):6, 2011.

M. Huenerfauth, M. Marcus, and M. Palmer. *Generating American Sign Language classifier predicates for English-to-ASL machine translation.* PhD thesis, University of Pennsylvania, 2006.

D. G. Jamrozik, M. J. Davidson, J. C. McDonald, and R. Wolfe. Teaching students to decipher fingerspelling through context: a new pedagogical approach. In *Proceedings of the 17th National Convention Conference of Interpreter Trainers, San Antonio, TX*, pages 35–47, 2010.

R. E. Johnson and S. K. Liddell. A segmental framework for representing signs phonetically. *Sign Language Studies*, 11(3):408–463, 2011.

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

M. Kipp, Q. Nguyen, A. Heloir, and S. 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.

G. Lancaster, K. Alkoby, J. Campen, R. Carter, M. J. Davidson, D. Ethridge, J. Furst, D. Hinkle, B. Kroll, R. Layesa, et al. Voice activated display of american sign language for airport security. *Proceedings of the 18th Annual International Conference on Technology And Persons With Disabilities*, 2003.

V. Lombardo, C. Battaglino, R. Damiano, and F. Nunnari. An avatar-based interface for the italian sign language. In *Complex, Intelligent and Software Intensive Systems (CISIS), 2011 International Conference on*, pages 589–594. IEEE, 2011.

F. López-Colino and J. Colás. The synthesis of lse classifiers: From representation to evaluation. *Journal of Universal Computer Science*, 2011.

J. McDonald, R. Wolfe, S. Johnson, S. Baowidan, R. Moncrief, and N. Guo. An improved framework for layering linguistic processes in sign language generation: Why there should never be a "brows" tier. In *International Conference on Universal Access in Human-Computer Interaction*, pages 41–54. Springer, 2017.

A. Schembri. *Perspectives on Classifier Constructions in Sign Languages*, chapter Rethinking 'classifiers' in signed languages, pages 3–34. Psychology Press, 2003.

F. Thomas, O. Johnston, and F. Thomas. *The illusion of life: Disney animation*. Hyperion New York, 1995.

R. Wilbur. The linguistic description of american sign language. In *Recent perspectives on American sign language*, pages 7–31. Psychology Press, 2017.

R. Wolfe, J. McDonald, and J. Schnepp. Avatar to depict sign language: Building from reusable hand animation. *International Workshop on Sign Language Translation and Avatar Technology (SLTAT)*, 2011.

B. Woll. The linguistics of sign language classifiers: phonology, morpho-syntax, semantics and discourse. *Lingua: International Review of General Linguistics*, 117(7):1159–1353, July 2007.

S1 (rug):

S2 (table):

S3 (chairs):

S4 (plates):

S5 (glasses):

S6 (cutlery):

Figure 22: Still shots from the synthesis of the table scene expression