Univ Access Inf Soc (2008) 6:419–434
DOI 10.1007/s10209-007-0095-7

LONG PAPER

Generating American Sign Language animation: overcoming misconceptions and technical challenges

Matt Huenerfauth

Published online: 12 October 2007

Abstract Misconceptions about the English literacy rates of deaf Americans, the linguistic structure of American Sign Language (ASL), and the suitability of traditional machine translation (MT) technology to ASL have slowed the development of English-to-ASL MT systems for use in accessibility applications. This article traces the progress of a new English-to-ASL MT project targeted to translating texts important for literacy and user-interface applications. These texts include ASL phenomena called "classifier predicates." Challenges in producing classifier predicates, novel solutions to these challenges, and applications of this technology to the design of user-interfaces accessible to deaf users will be discussed.

Keywords American Sign Language · Deafness · Assistive Technology · Natural Language Processing · Machine Translation

Abbreviations

ASL	American Sign Language
NLP	Natural Language Processing
MT	Machine Translation
GUI	Graphical User Interface

M. Huenerfauth (✉)
Department of Computer Science, Queens College,
The City University of New York, 65-30 Kissena Boulevard,
Flushing, NY 11367, USA
e-mail: matt@cs.qc.cuny.edu

1 Introduction

Without aural exposure to spoken English during the critical language-acquisition years of childhood, many deaf adults have below-average levels of written English literacy. In fact, studies have shown that the majority of deaf high school graduates in the US have only a fourth grade English reading level [4]. This means that many deaf students age 18 and older have a reading level more typical of a 10-year-old student. Many of these people with low levels of English literacy are actually fluent in American Sign Language (ASL). ASL is the primary means of communication for approximately one half million deaf people in the United States, and it is a full natural language with a linguistic structure distinct from English [13, 14, 16, 18]. Thus, it is possible to have fluency in ASL without literacy in written English. These low levels of literacy have become an even more significant issue in recent decades as new information and communications technologies have arisen that place an even greater premium on English literacy skills.

Unfortunately, most deaf accessibility aids, like television closed-captioning or teletype telephones, require their user to have strong English literacy skills. Many computer software designers also incorrectly assume that written English text on a user-interface is accessible to deaf users since it is presented visually. An automated English-to-ASL machine translation (MT) system could make information and services accessible when English text captioning is too complex, an English-based user-interface is too difficult to navigate, or when live interpreting services are unavailable. This type of MT software could also be used to build new educational software for deaf children to help them improve their English literacy skills. The goal of this research project is to develop such an English-to-

ASL MT system, specifically one that can produce animations that include some important ASL phenomena called "classifier predicates."

This paper will explore how accessibility technology has been slow to address this literacy issue because of several misconceptions: the rate of English literacy among the deaf, the linguistic status of ASL, the importance of certain ASL phenomena called "classifier predicates," and the suitability of traditional natural language processing software to the special linguistic properties of ASL. As a subsequent step, the paper will describe the development of a machine translation (MT) system to translate English text into ASL animations—with a particular focus on those 3D spatial aspects of the language that have received little attention from previous researchers. In particular, this project has proposed several novel MT technologies to address the special linguistic challenges of ASL, and these technologies have had some exciting advantages for the development of tools for deaf users.

2 Common misconceptions

2.1 Literacy rate of deaf users

Many accessibility "solutions" for the deaf simply ignore part of the problem—often designers make the assumption that the deaf users of their tools have strong English reading skills. For example, television "closed captioning" converts an audio English signal into visually presented English text on the screen. However, the reading level of this text may be too high for many deaf viewers. While some content may be accessible with this approach, deaf users may be cut off from important information contained in news broadcasts, educational programming, political debates, and other broadcasts with a more sophisticated level of English language. Communication technologies like teletype telephones (sometimes referred to as telecommunications devices for the deaf or TDDs) similarly assume the user has English literacy. The user is expected to both read and write English text in order to have a conversation.

An issue that has become more significant in recent years is the accessibility of computer software and websites to people with disabilities. Unfortunately, few software companies have addressed the connection between deafness and literacy, and so few computer user-interfaces make sufficient accommodation for the deaf. Many software designers believe that if audio information is also presented as written English text, then the accessibility needs of a deaf user have been met.

A machine translation system from English text into ASL animations could increase the accessibility of all of these technologies. Instead of presenting written text on a television screen, telephone display, or computer monitor, each could instead display a small animated virtual human character performing ASL output. Researchers in computer graphics have built several animated models of the human body that are sufficiently articulate such that they can perform ASL [20]. Most animation systems use a basic instruction set to control the character's movements. Therefore, a translation system would need to analyze an English text input and produce a "script" in this instruction set specifying how the character should perform the ASL translation output. Systems have also been developed that use a sign-language-specific script to control an animated character [3].

2.2 ASL versus Signed English

Even when designers understand that presenting English text is not a complete solution for deaf users, confusion regarding the language status of ASL has delayed the creation of MT technology. Many researchers have assumed that the reason why many deaf people have difficulty reading English is that it is presented in the form of words written in Roman characters. Under this assumption, if every word of an English sentence were replaced with a corresponding ASL sign (the assumption is also made that such a correspondence always exists), then deaf users would be able to understand the text.

There is a common misconception that English and ASL have the same linguistic structure—that one language is merely a direct encoding of the other. In fact, the word order, linguistic structure, and vocabulary differences between English and ASL are comparable to those between many pairs of written languages. And while there are some signing communication systems that use English structure (such as Signed English, Signed Exact English, etc.), these are typically used in educational contexts and are not natural languages. In most cases, presentation of ASL signs in English word order (and without the accompanying ASL linguistic information contained in facial expressions, eye gaze, etc.) will not be understandable to a deaf user.

This confusion over the linguistic status of ASL has led some researchers to produce MT systems that produce Signed English and not ASL. There have been several projects that have simply transliterated English sentences word-for-sign using an English-to-ASL dictionary of video clips or animations. These systems produce output with identical structure and word order to the original English sentence [5]. The problem with such projects is that they rarely produce output, which is useful to deaf users who had difficulty understanding the structure and meaning of the original English text. Unfortunately, many of these

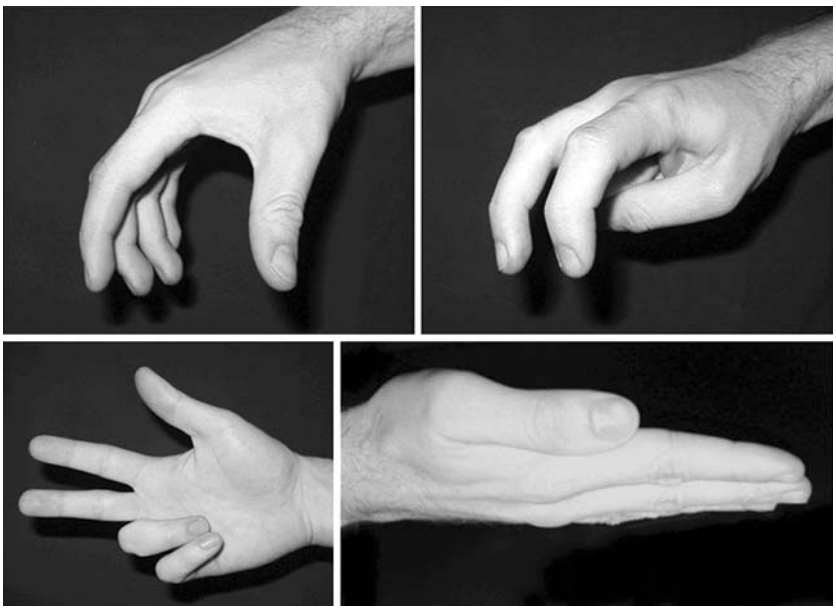

Fig. 1 Some handshapes used during classifier predicates: "Downward C", "Bent V", "Sideways 3," and "Flat B"

systems advertise themselves as "translation" systems and claim to produce ASL—thus misleading and disappointing users, as well as other researchers.

2.3 Ease of applying NLP Technology to ASL

The previous section has suggested that there are differences in the linguistic structure of English and ASL. In fact, the structure of ASL is quite different than most written/spoken languages, and its visual modality allows it to use phenomena not present in these languages [13, 14, 18]. In addition to using hands, facial expression, eye gaze, head tilt, and body posture to convey meaning, an ASL signer can use the surrounding space for communicative purposes. For example, signers can assign objects or people under discussion to locations in space, and later refer to them by pointing to these locations. The locations are not meaningful topologically, i.e., positioning an entity to the left of another in space does not mean it is to the left of the other in the real world.

Other ASL phenomena do make use of the space around the signer in a topologically meaningful way; these constructions are called "classifier predicates." During the performance of a classifier predicate, the signer's hands represent an entity in space in front of them, and they position, move, trace, or re-orient this imaginary object to indicate the location, movement, shape, or other properties of some corresponding real world entity under discussion. A classifier predicate consists of two simultaneous components: (1) the hand in a semantically meaningful shape and (2) a 3D path that the hand travels through space in front of the signer.

For example, to express "the car parked between the cat and the house," the signer could use three classifier predicates: (1) the non-dominant hand in a "Downward C" handshape would indicate a location in space where a miniature invisible house could be envisioned, (2) the dominant hand in a "Bent V" handshape would indicate a location in space where a miniature invisible cat could be envisioned, and (3) the dominant hand in a "Sideways 3" handshape would trace a path in space corresponding to a car driving and stopping between the "house" and "cat" locations in space. The "car" will park on top of a flat platform created by the non-dominant hand using the "Flat B" handshape (as shown in Fig. 1.)

Before each of these classifier predicates, the signer would also need to perform an ASL sign to indicate what object is being described with the classifier predicate. In this case, the signer would produce the ASL sign "HOUSE," "CAT," or "CAR" respectively before each of the three classifier predicates listed above. During the performance of each of these noun signs, the signer will look at the audience and raise his/her eyebrows. During the classifier predicates for the "house" and "cat," the signer will aim their eye-gaze at the location assigned to the "house" and the "cat." During the classifier predicate showing the motion path of the "car," the signer will follow the path of the "car" with his/her eyes. See Fig. 2 for a timeline representing the entire performance. In this figure, the top row represents the activity of the signer's eye gaze, the middle row represents the activity of the signer's right hand (a.k.a. "dominant hand"), and the bottom row represents the activity of the signer's left hand.

As part of research work on the machine translation of English text into ASL animations, a prototype system has

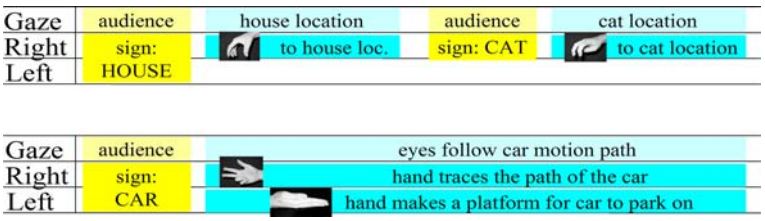

Fig. 2 A timeline of the performance of the ASL translation of the English sentence "the car parked between the cat and the house." This ASL performance contains three ASL nouns, each of which precedes an ASL classifier predicate

been implemented that can generate animations of ASL classifier predicates. Given a symbolic representation of the meaning of an English sentence, the system can produce an animation of a character with articulated head tilt, eye gaze, eyebrow raise, and arm/hand movement.

Figure 3 contains nine still images taken from an animation produced by the system—the individual images are labeled "(a)" to "(i)" in the figure. During the video, an animated human character performs a series of three ASL classifier predicates—the same classifier predicates that are described in Fig. 2. In image 3(a), the woman makes the ASL sign for "HOUSE" while looking at the audience with her eyebrows raised. In image 3(b), the woman blinks while she lowers her hands during the final part of the ASL sign "HOUSE." In 3(c), she moves her right hand into a position in space on her left side where the "HOUSE" object is located (using the "Downward C" handshape that is used for bulky objects). She aims her eye gaze at the location associated with the "HOUSE." In image 3(d), the woman makes the ASL sign for "CAT" while looking at the audience with her eyebrows raised. In 3(e), she moves her right hand into a position in space on her right sign where the "CAT" object is positioned (using the "Bent V" handshape for stationary animals). She aims her eye gaze at this location. In images 3(f) and 3(g), the woman makes the ASL sign for "CAR" while looking at the audience and raising her eyebrows. In 3(h), she uses her right hand (in the "Sideways 3" handshape that is used for motorized vehicles) to show the motion path of the car which moves between the location of the house and cat. Finally, in image 3(i), the right hand comes to rest on top of the open palm of the left hand—showing the car parking. During car's movement, the signer's eye gaze followed the motion path of the car.

Not every ASL sentence contains a classifier predicate, and apart from the non-topological "pointing" pronouns mentioned at the beginning of this section, many ASL sentences have a structure that looks similar to English or other written languages. The problem is that the subset of the language without classifier predicates has received a disproportionately large amount of attention from linguistic and MT researchers (because it is easier to analyze and generate since it is closer in structure to known written languages). Even when MT researchers appreciate the distinct language status of ASL and try to build translation systems, they have chosen to focus on these non-classifier-predicate parts of the language [5, 10].

This simplification has allowed several researchers to reuse translation technologies originally designed for written languages. Some have had success at producing ASL animations on this limited (non-spatial) portion of the language [21], and others have begun to address some spatial issues [19]. Unfortunately, these systems employ traditional computational linguistic approaches that do not model the spatial arrangement of objects in a 3D scene. Therefore, these systems are not able to produce classifier predicates from an English text [6]. No previous ASL MT system has proposed how to generate classifier predicates, and this aspect of the language has been ignored. The next section of this paper will discuss why this is not an acceptable simplification.

A further complication of ASL that has made it a difficult subject of machine translation research is that the language has no written form. There is no orthography commonly used by ASL signers, and therefore a first step in any MT project is to select some form of notation or symbolic representation to facilitate processing the language. This lack of a writing system has also made it difficult and expensive to collect large corpora of ASL with sufficient detail for computational research purposes. This has prevented many of the most popular statistical MT approaches from being applied to ASL since most would require large amounts of parallel English-ASL language data to train a machine learning algorithm.

2.4 Underestimating ASL classifier predicates

Omitting classifier predicates from the output of an English-to-ASL MT system is not an appropriate or desirable simplification for several reasons. The first is that classifier predicates are actually quite common in native ASL signing. Studies of sign frequencies show that classifier predicates occur once per minute and up to seventeen times per minute in some genres [17]. Further, classifier predicates are the only way to convey some concepts

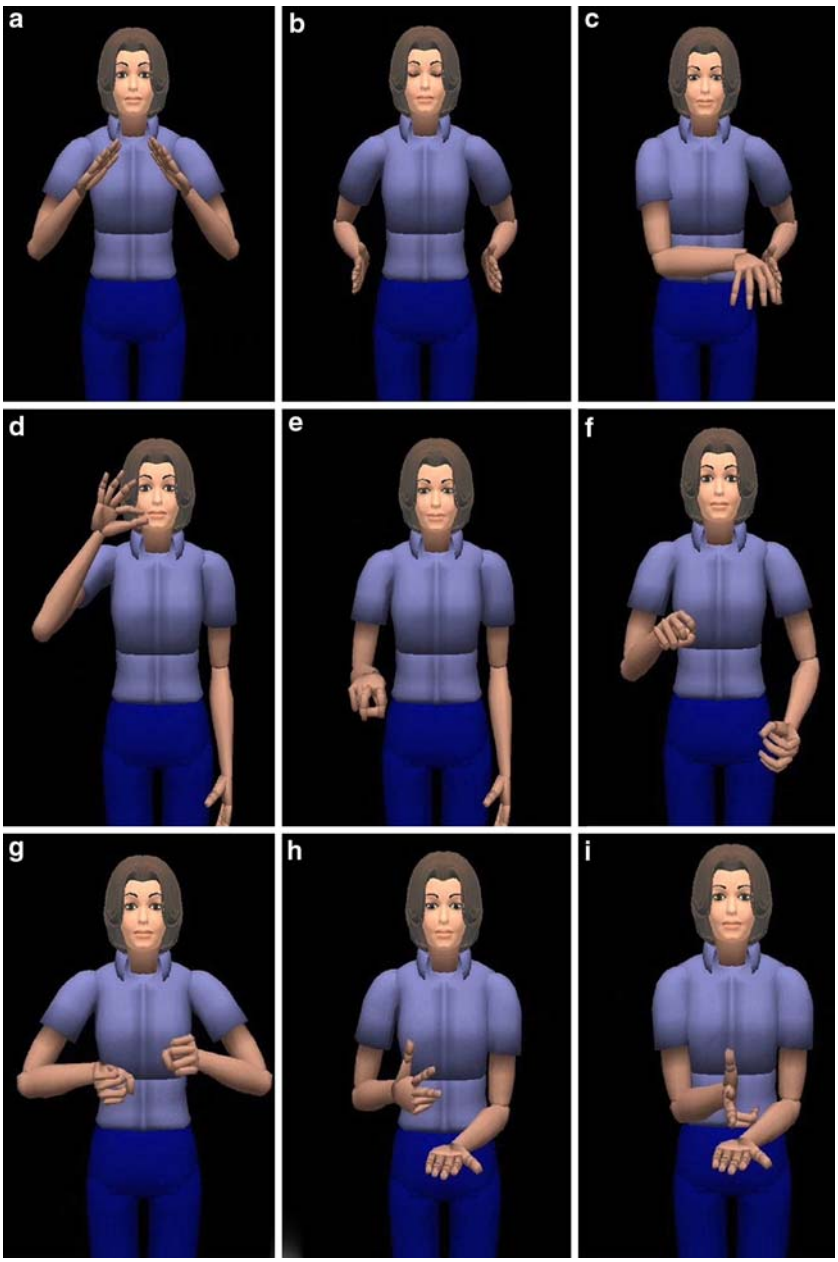

Fig. 3 Nine images taken from an animation produced by the ASL generation system created for this project

contained in English sentences. For example, to express information about spatial layout, arrangement, shapes, outlines, alignment, or movement in ASL, a signer will use classifier predicates. Finally, when the ASL equivalent of an English sentence uses a classifier predicate, then the structure of the two sentences is quite divergent—a lengthy English sentence may be expressed using a small number of meaningful spatial hand movements. This structural difference can make these English sentences difficult to understand for a deaf user and thus important for an MT system to translate. The translation of these sentences is therefore especially relevant from an accessibility perspective.

Classifier predicates are particularly important when producing an accessible user interface. Since ASL lacks a written form, any English on an interface would need to be translated into ASL and presented as a small animated character performing ASL on the screen. Clearly, a computer application that involved spatial concepts would require classifier predicates in the ASL output, but more generally, these predicates are important in an interface because they enable the animation to refer to other elements on the screen. Since the ASL cannot be statically "written" on elements of the interface, the dynamic animation performance will frequently need to refer to and describe elements of the surrounding screen. When

discussing a computer screen, a human ASL signer will typically draw an invisible version of the screen in the air with their hand and use classifier predicates to describe the layout of its components and explain how to interact with them. After the signer has "drawn" the screen in this fashion, individual elements are referred to by pointing to their corresponding location in the signing space. Making reference to the onscreen interface is especially important when a computer application must communicate step-by-step instructions or help. English-illiterate users of a computer application would likely also have limited computer experience; so, conveying this type of content may be especially important for them.

3 Translation problems and novel solutions

Some of the linguistic discussion above has suggested that ASL is a difficult language to produce using machine translation software. Beyond the misconceptions above, it has been this ASL-specific MT difficulty that has slowed the development of English-to-ASL software. This section will explore this issue in further detail. Specifically, several interesting challenges encountered during the design and development of the English-to-ASL MT system will be described. In each case, the adopted solution to the problem will be explained in order to illustrate how ASL has motivated several new MT technologies.

3.1 Extending current ASL Technology to CPs

A non-linguistic representation of an animation of a 3D character performing a classifier predicate would need to record a large number of parameters over time: all of the joint angles for the face, eyes, neck, shoulders, elbows, wrists, fingers, etc. If an MT system had to generate classifier predicates while considering all of these values, the task would be quite difficult. The goal of a good linguistic "phonological model" is to abstract away from some of the details of the language output and help make the generation process easier to describe. A good model will reduce the number of independent parameters needed to be specified by the generation process while still allowing it to produce a complete output. Previous ASL phonological models record how the handshape, hand location, hand orientation, movement, and non-manual elements of a signing performance change over time [2]. However, these models are ill-suited to the representation of classifier predicates. Not only do they record too little information about the orientation of the hand, but they record too much information about the handshape (only a limited number of shapes appear in classifier predicates). Finally, these models make it very difficult to specify the complex motion paths required for some classifier predicates (e.g., the various 3D motion paths that the "car" might have traveled in the earlier "parking car" example in Figs. 2, 3).

As a first step in producing ASL classifier predicates, a symbolic representation was selected for these phenomena that would serve as the output of the MT process. While a good representation should help to simplify and parameterize the signing animation movements, it should be sufficiently detailed that 3D animation software can still use it as input to produce a final output animation of the ASL performance.

Specifically, eye-gaze and head-tilt are represented as a pair of 3D points in space at which they are aimed [8]. This simplification is made because what is semantically meaningful in a classifier predicate about eye-gaze and head-tilt is the point at which they are aimed, not the exact details of neck or eyeball angles. Fortunately, the animation software to be used by this system can calculate head/eye positions for a virtual character given a 3D point in space. Therefore, this model is sufficient for producing an animation. The next section discusses how special invisible placeholder objects are arranged in the space in front of the signer. These placeholders serve as targets for the 3D coordinates of the eye-gaze and head-tilt, and so the model has a method of calculating their values.

In a classifier predicate, it is the position of the hand (not the whole arm or elbow) that is semantically meaningful, thus making possible another simplification. The locations in space of the dominant and non-dominant hands are recorded as another pair of 3D coordinates. The shape of each hand and the 3D orientation of the palm are also recorded. Given hand location and orientation values, there are algorithms for calculating realistic elbow/shoulder angles for a 3D virtual human character [15, 22]; so, the model is again sufficient for generating animation [8].

The specification of an ASL performance is therefore a stream of location, orientation, and handshape values for the animated character over time. Special data structures have been developed that represent the coordination timing relationships between the movements of various parts of the signer's body during the performance [11].

3.2 Calculating 3D motion paths

The model of ASL classifier predicate output described above needed to select 3D coordinates for parts of the body over time. In earlier work, several possible methods for calculating such 3D motion trails were compared [7]. The most simplistic approach considered was to pre-store a list of all possible pairs of English motion verbs and ASL classifier-predicate motion paths. However, due to the

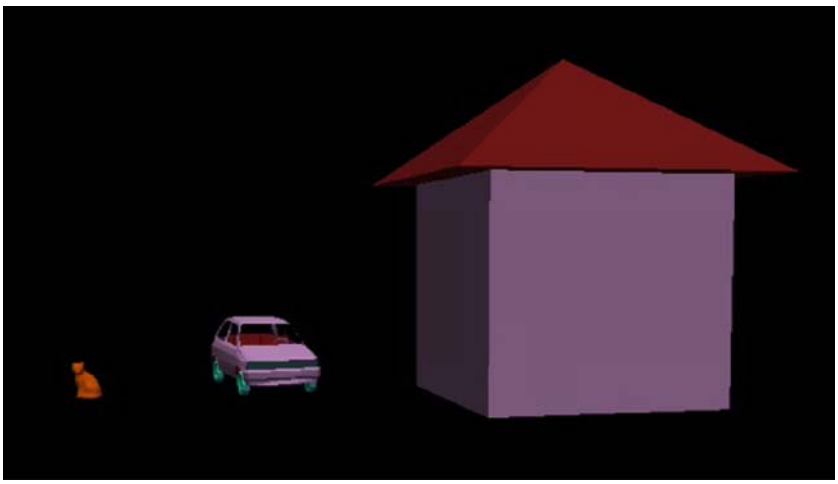

Fig. 4 Still image from a "scene visualization" animation of the English sentence: "The car parked between the cat and the house"

many possible arrangements of 3D scenes that would each require slightly different forms of classifier-predicate motion paths, this approach is combinatorially impractical (for example, consider all the different shapes and inclines of roads along which a car could travel). Other heuristic rule-based approaches to calculating motion paths were also discounted based on linguistic considerations [7]. What this comparison made clear was that in order to produce a classifier predicate, some method was needed for modeling the 3D layout of the objects in the scene being described by an English text.

The developed system can use existing "scene-visualization" software to analyze an English text describing the motion of real-world objects and build a 3D graphical model of how the objects mentioned in text are arranged and move [1]. This model is "overlaid" onto the volume in front of the ASL signer (Fig. 2). For each object in the model, a corresponding invisible placeholder is positioned in front of the signer. The layout of placeholders mimics the layout of objects in the 3D model. In the "car parked between the cat and the house" example, two miniature invisible objects representing a "house" and a "cat" are positioned in front of the signer's torso, and another object, with a motion path terminating between the "cat" and the "house," is added to represent the "car." The locations and orientations of the placeholders are later used to select the locations and orientations for the signer's hands while performing classifier predicates about them. So, the motion path calculated for the car will be the basis for the 3D motion path of the signer's hand during the performance of the classifier predicate describing the car's motion.

Figure 4 contains a still image from an animation of a 3D scene of a cat, car, and house that the developed system is able to describe in ASL. This is a 3D representation of the scene described by the English sentence: "The car parked between the cat and the house." There is a small "cat" on the left of the image, a "car" in the middle, and a "house" on the right side. This animation is typical of the output of the "scene visualization" software.

Fig. 5 The 3D scene from Fig. 4 is overlaid onto the space in front of the signing character

In Fig. 5, it is possible to see how the 3D scene from Fig. 4 is overlaid onto the volume of space in front of the animated signing character in the ASL system. Mapping the scene onto the "signing space" allows deciding how to position the hands of the signer during classifier predicates that will describe the layout of this scene. It is important to note that these miniature objects will not be visible during the final animation output of the system (seen in Fig. 3).

3.3 Visual details

One of the most difficult aspects of the generation of a 3D graphical scene from an English sentence is correctly producing all of the sizes, shapes, colors, and other visual details of the objects being represented. Details of the surrounding setting and presence of background objects/characters not directly mentioned in the English text are also particularly challenging for the scene-visualization software. For an ASL system, many of these visual details are not important for the production of classifier predicates. Rarely are extraneous details and background objects described using the hands during the performance of a classifier predicate, unless they are important to the action being discussed. Spending processing and development time on these parts of the 3D model is unnecessary.

Unlike some applications of the scene-visualization software—where entities described by the English text would need to be rendered on the screen—in this situation, the 3D objects would be transparent. Therefore, the MT system does not care about the exact appearance of the objects being modeled. Only the location, orientation, and motion paths of these objects in some generic 3D space are important, since this information will be used to produce classifier predicates. Details of size and shape are largely irrelevant; so, the system can use some form of general placeholder object instead of animating visually accurate 3D "cars," "cats," or "houses" for example.

Figure 6 contains an image of the placeholder objects floating in space in front of the signing character; the layout of these placeholders corresponds to the layout of the objects in Fig. 5. Many of the 3D animation details of the objects shown in Fig. 5 are not needed during the creation of the ASL animation output. Only the location (center of mass) and orientation of each object is important to record.

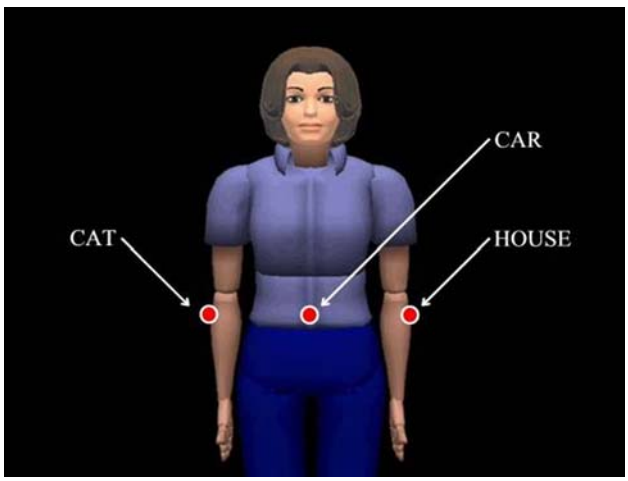

Fig. 6 Placeholder objects in front of the signer representing the objects shown in Fig. 5

These two pieces of information will affect how the signer's hands are positioned and oriented during the ASL performance. While these placeholders appear as "dots" in the figure, they are actually invisible during the sign language animation that is produced.

While there is less visual detail, there are some additional pieces of linguistic information that should be recorded in the 3D scene. Specifically, objects are described using different handshapes based upon the semantic class of the real-world entity being discussed. For example, the motion of motorized vehicles is shown using a "Sideways 3" handshape, the placement of stationary animals or seated humans is shown using a "Bent V" handshape, and the placement of bulky objects is shown using a "Downward C" handshape (as shown in Fig. 1). To facilitate the selection of the proper handshape in the animation output, the invisible placeholder objects will need to record which semantic categories they belong to.

Within this framework, the purpose of an ASL signer producing a classifier predicate can be regarded as an attempt to convey information to the audience about what the invisible placeholders are doing in space. The 3D model of these placeholders over time can thus serve as a loose "semantic" (underlying meaning) representation of a set of classifier predicates. In this light, the two classifier predicates in the "parking a car" example can be thought of as conveying that a bulky object occupied a point in space and a vehicle object moved toward it and stopped.

3.4 Encoding ASL grammar for classifier predicates

There are many rules of ASL grammar that govern how signers construct a linguistically correct classifier predicate performance. For instance, in the "parking a car" example, the signer had to show the location of the surrounding objects before showing the motion path of the main object, the "car." Before each of the three classifier predicates was performed, the object being described had to be the topic of conversation. Therefore, a noun phrase ("HOUSE," "CAT," or "CAR") was performed before each of the classifier predicates to identify which object was going to be positioned in space. Some element of the ASL animation system must record these rules for how to construct a classifier predicate performance and select the best combination of classifier predicates to convey the information in the 3D scene.

After calculating the 3D layout of the entities discussed in an English text, an approach is needed to generate animations of classifier predicates describing this scene. In [7] it is argued that a recent linguistic model of classifier predicate generation proposed by Scott Liddell [13, 14] can serve as a starting point for developing such an approach.

Fig. 7 Pseudocode for a classifier predicate planning template used in the ASL system

```
PARKING-A-VECHILE-REL-OBJECTS
Parameters:       de0 (type: Discourse Entity), dlist0 (type: List of Discourse Entities)
Resources:        dominant hand, non-dominant hand, eye-gaze
Restrictions:     de0 is a vehicle
Preconditions:    (de0 is topic) or (de0 has been positioned in space already)
                  for x in dlist0: (x has been positioned in space already)
Effects:          de0 is topicalized
                  de0 is positioned in space
                  express(verb:park-23, agent:de0, nearby-objects:dlist0)
Subplans:         MOVING-VEHICLE-TO-LOC( de0, resources: dominant-hand)
                  FINAL-PLATFORM( de0, resources: non-dominant-hand)
                  EYETRACK( de0, resources: eye-gaze)
Subplan Time:     Concurrent
```

In this model, signers have a mental image of a scene to be discussed (much like a 3D graphics specification) which they map onto the space around their body, and they use 3D information from this scene to select and fill templates for a classifier predicate from a template lexicon. For example, this lexicon may store a template for "parking a vehicle," but the exact 3D locations of the car are left as parameters. When the signer needs to produce an actual "parking a vehicle" predicate, the 3D locations of the "car" can be taken from the scene, the template will be instantiated, and a classifier predicate motion path is calculated. In this way, a single "parking a vehicle" template is used to produce all of the possible "parking" classifier predicates with different possible motion paths and locations. Unfortunately, the linguistic model does not provide much detail about the internal structure of these templates nor their selection/filling process [13] [14].

The MT project reported in this paper has developed new computational models for classifier predicate generation [8] within an English-to-ASL MT system that formalize and implement this linguistic account with some modifications. Figure 7 contains pseudocode for a classifier predicate planning template. This template specifies how to perform a classifier predicate that shows a vehicle parking between other surrounding objects. The Parameter "de0" is a variable representing a Discourse Entity being discussed, in the specific case, a "car." The Resources represent the parts of the body that this template can control. The Restrictions specify that the de0 must be a vehicle. Preconditions are like restrictions, except that the system can try to use other templates in its library to satisfy the preconditions of this template. In this case, the Preconditions require the "car" to already be the topic of conversation or to have already been positioned in space around the signer. Further, the surrounding objects must already have been positioned in space. The Effects specify the changes that result from performing this classifier predicate. Effects from one template could be used to satisfy Preconditions of a later template in an animation that is being created. The Effects field also specifies the English verb semantics that this template expresses. The Subplans specify other templates in the system's library that should be triggered as part of the performance of the current template (the templates listed in this subplan are shown in Fig. 8). The Subplan Time can specify whether these subplans should be performed concurrently or sequentially. If they are performed concurrently, then the resources passed to each of the subplans should not overlap. No two subplans should try to control the same part of the signer's body at the same time.

Figure 8 contains several templates listed as subplans in Fig. 7. MOVING-A-VEHICLE-TO-LOC produces a hand movement showing a car driving. The Actions field specifies how to control the location, orientation, and handshape of the dominant hand (the right hand of the signer). The location and orientation of the hand is set using the location and orientation values of the invisible placeholder for the object de0. The handshape is set using the value of the constant "Sideways 3," which specifies the hand configuration that produces the "Sideways 3" handshape in Fig. 1. FINAL-PLATFORM produces a horizontal platform with the non-dominant hand at the final location value of the object passed as parameter de0 (the term "non-dominant-hand" is abbreviated as "ndh" in the Actions field of this template). EYETRACK causes the signer's eye-gaze to track the location of the object passed as parameter de0.

These templates actually serve as planning operators for an artificial intelligence planning process—this means that each of these templates specifies a set of preconditions that must be satisfied prior to their classifier predicate being performed. Each planning operator also has a set of effects that it can accomplish in the planning process, and so the system can trigger additional classifier predicates to satisfy the preconditions of classifier predicates to be performed later in the animation. For instance, if there is a requirement that the background objects are positioned in space prior to the performance of the motion path of the main object being discussed, then the system can trigger additional classifier predicates to place these surrounding objects in space beforehand.

MOVING-VEHICLE-TO-LOC
Parameters: de0 (type: Discourse Entity)
Resources: dominant-hand
Restrictions: de0 is a vehicle
Preconditions: (de0 is topic) or (de0 has been positioned in space already)
Effects: de0 is topicalized
de0 is positioned in space
express(verb:drive-22, agent:de0)
express(verb:move-8, agent:de0)
Actions: dominant-hand.location = track_location (de0.location)
dominant-hand.orientation = track_orientation(de0.orientation)
dominant-hand.handshape = track_handshape("Sideways 3")

FINAL-PLATFORM
Parameters: de0 (type: Discourse Entity)
Resources: non-dominant-hand
Actions: ndh.location = get_to_location(de0.location)
ndh.orientation = get_to_orientation("Palm Up Knuckles Fwd")
ndh.handshape = get_to_handshape("Flat B")

EYETRACK
Parameters: de0 (type: Discourse Entity)
Resources: eye-gaze
Actions: eye-gaze.location = track_location(de0.location)

Fig. 8 Pseudocode for additional templates used as subplans in Fig. 7: MOVING-A-VEHICLE-TO-LOC, FINAL-PLATFORM, and EYETRACK

Another attractive design feature of these templates is that they enable the linguistic and the animation portion of the system's development to be partitioned. Inside of the templates, the motion paths of the hands are specified using symbolic functions that are parameterized on the locations and orientations of the invisible placeholder objects floating around the signer. The developers who design the templates do not need to know the actual 3D details of these motion paths; instead, they can focus on the linguistics of ASL while creating the template.

3.5 Some movement paths are linguistic

While some ASL classifier predicate motion paths can be directly taken from the motion paths of invisible placeholder objects, other classifier predicates display movements, which are less visually representative and more linguistically determined. Sometimes the motion path of the hands is not an exact representation of the 3D motion path of the placeholder objects in the scene. The hand moves in a manner, which is different than the motion path of the object it is conveying.

An example of a classifier predicate in which the path of hand motion does not match the path of motion of the placeholder object is the classifier predicate for "leisurely walking upright figure" [14]. To perform a classifier predicate showing how a person walks in a leisurely manner, an ASL signer would put his or her hand in a "Number 1" handshape (index finger pointing up, all other fingers closed). Then the signer would next bounce his or her hand up and down as it moves along the 3D path walked by the human being described. While the hand bounces, the meaning being conveyed is not that a human is bouncing, but that the person is walking leisurely. This bouncing quality of the movement is linguistically (not visually) determined.

The example above and other linguistic considerations [7] indicate that not all of the information needed to select the 3D motion path of the hand during a classifier predicate comes from the invisible placeholder objects. Some linguistic information must also be taken from the original English sentence to convey special forms of meaning, as in the case of the concept of "leisurely" in the example above.

The use of a template-based approach can solve this "linguistic movement" problem. Some of the information about the 3D path of the signer's hand is "hard-coded" inside of the template, and other information about the 3D motion path is taken from the invisible placeholder objects in front of the signer's torso. In this way, the system does not rely on the 3D scene for every 3D motion detail. Some portions can be specified ahead of time in the template, and complex motion paths can be calculated based on the location of the invisible placeholder (that might not be identical to the locations or movement paths of those placeholders).

In the "leisurely walking" example, the general path of the human's motion is taken from the invisible placeholder object, but the up-and-down bouncing is hard-coded inside of the template for "leisurely walking" [7]. When designing a template for this classifier predicate, the system would need to use a special 3D-motion-path transformation function that would change the motion path of the

"human" placeholder into the bouncing motion path of the hand for the classifier predicate.

3.6 High processing and development overhead

One problem with the planning-template-based translation approach illustrated in Fig. 7 is that it requires a template to be written for each English motion verb that will need to produce an ASL classifier predicate. This could potentially imply a lot of programming effort to produce a machine translation system that successfully processes a wide variety of input sentences. Another problem is that the calculation of the 3D graphical model coordinates and layout could require a lot of processing time, thus preventing real-time English-to-ASL translation.

However, the use of 3D animation software is not necessary to translate those English sentences that do not produce ASL classifier predicates. For these input sentences, the translation approach described above would be overly powerful, and it would be overly cumbersome to implement and process. For ASL sentences that do not produce classifier predicates, some of the traditional MT technologies originally developed for written languages and used by some of the previous systems mentioned at the start of this paper would be able to produce a successful translation.

Figure 9 illustrates the three different forms of output that the English-to-ASL machine translation system could produce. If the input is a spatially descriptive English sentence, then the classifier predicate (CP) software described above would be used to produce an ASL sentence containing a classifier predicate. The pathway for English inputs producing classifier predicates includes the scene-visualization software, but the pathway for other inputs does not. Other English input sentences could be processed by English-to-ASL machine translation software that is based on more traditional translation technology originally designed for written languages. This would produce ASL sentences that do not contain classifier predicates or other uses of 3D space around the signer's body to convey linguistic meaning.

Finally, if the English input sentence cannot be successfully processed by either of the other two pathways (i.e., it contains a word or grammar construct they cannot process), then a word-to-sign dictionary lookup process would be used to produce a Signed English sentence. This is a sentence in which signs are arranged in exact English word order without the use of any other grammatical features of ASL. Since most deaf signers have some familiarity with non-ASL English-like forms of signing, this word-for-sign transliteration may be partially understandable to the users. This English-like output would only

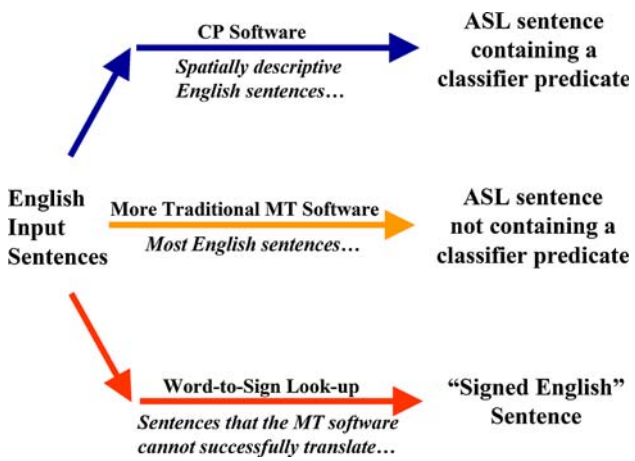

Fig. 9 Three different forms of output from the English-to-ASL machine translation system

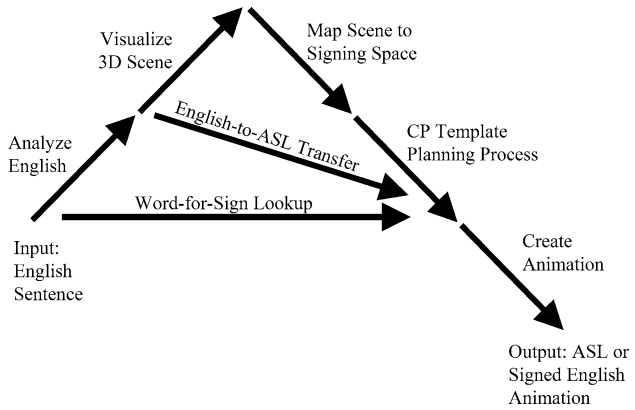

Fig. 10 The multi-pathway processing flow of the English-to-ASL machine translation system

be produced if the system would have otherwise been unable to produce any results.

The system will process an input sentence using the most sophisticated pathway for which sufficient linguistic resources exist and will "fall back" on simpler pathways as needed. This architecture is therefore able to blend deep 3D-processing and broad input-coverage in a single system [6]. Figure 10 contains a diagram of the architecture of a complete English-to-ASL translation system that uses this three-pathway design. The top pathway (up and down the pyramid shape) is the classifier-predicate generation pathway. The middle pathway (that follows the "English-to-ASL Transfer" arrow) produces ASL sentences not containing a classifier predicate. The bottom pathway (that follows the "Word for Sign Lookup" arrow) produces a series of signs in exact English word order.

The initial implementation focus of this project has been the translation of English text into ASL classifier predicates. Current work focuses on implementing the portion of

the system along the top pathway of Fig. 10. To handle a variety of input sentences, this classifier-predicate generation technology will be embedded within a complete English-to-ASL MT system that contains multiple processing pathways [6].

3.7 Mapping from English to ASL sentences

The "parking" example at the start of this paper illustrated how a single English sentence ("the car parked between the cat and the house") can produce multiple classifier predicates (one for the house, one for the cat, and one for the car). In fact, it is common for several classifier predicates to be needed to convey the semantics of one English sentence and vice versa. Even when the mapping is one-to-one, the classifier predicates may need to be rearranged during translation to reflect the scene organization or ASL-specific conventions on how these predicates are sequenced or combined. For instance, when describing the arrangement of furniture in a room, signers generally sequence their description starting with items to one side of the doorway and then circling across the room back to the doorway again. An English description of a room may be significantly less spatially systematic in its ordering.

Multiple classifier predicates used to describe a single scene may also interact with and constrain one another. The selection of scale, perspective, and orientation of a scene chosen for the first classifier predicate will affect those that follow it. Other times, the semantics of multiple classifier predicates may interact to produce emergent meaning. For example, one way to convey that an object is between two others in a scene is to use three classifier predicates: two to locate the elements on each side and then one for the entity in the middle. In isolation, these classifier predicates do not convey a spatial relationship, but in coordinated combination, this semantic effect is achieved. These linguistic considerations demonstrate that whatever approach is taken to generating ASL classifier predicates, it should be easy to link English verbs and ASL classifier predicates in one-to-one, many-to-one, one-to-many, and many-to-many manners. The generation approach should also make it easy to make decisions about multiple classifier predicates at the same time, and it should allow the effects of one classifier predicate to satisfy preconditions of later ones.

To address all of the above concerns, the developed system uses the same template-based formalism to represent the structure in-between and within classifier predicates. This approach simplifies the system, in that it allows a single formalism (and processing software) to be implemented to handle both inter-classifier predicate and intra-classifier predicate generation decisions. It also facilitates the non-one-to-one mappings of English verbs and ASL classifier predicates described above. It also gives the translation systems more flexibility, as there is no need to pre-commit to a fixed number of classifier predicates at the start of the generation process, and more can be added to the output as necessary to satisfy ASL-specific linguistic rules.

For example, the template in Fig. 7 uses its Subplans field to specify the internal structure of the classifier predicate performance. In order to satisfy the Preconditions of this template, other templates will need to be triggered from the system's library and scheduled earlier in the animation. For instance, to assign positions in space to the surrounding objects (listed in the parameter "dlist0" of Fig. 7), other classifier predicates that position each object in space will be triggered. Thus, the planning-based template formalism can specify both the arrangement of individual classifier predicates with respect to one another and how the individual elements of an animation are coordinated to perform a single classifier predicate.

3.8 Evaluation of ASL generation systems

The lack of a standard writing system for ASL, and the parallel nature of an ASL performance, during which multiple parts of the body move in a coordinated manner, can make it difficult to evaluate systems that produce ASL animation output using traditional automatic machine translation evaluation metrics [9]. Many machine translation metrics for written languages compare a string produced by a system to some human-produced "correct-translation" string. While an artificial ASL writing system could be invented for the system to produce as output for the purpose of evaluation, it is unclear whether human ASL signers could accurately or consistently produce written forms of ASL sentences to serve as "correct-translations" for such an evaluation. And of course, real users of an English-to-ASL machine translation system would see an animation of a human character—not an artificial string-based encoding of ASL—as the output. Thus, basing the evaluation methodology on a string-encoding of the ASL output is not ideal; it would be better for actual ASL animations to be shown as part of an evaluation process.

A more meaningful measure of an ASL system would be a user-based study in which ASL signers view animations produced by the system and evaluate them. The evaluation of the prototype classifier predicate generation system has followed this approach. Members of the deaf community who are native ASL signers viewed animations of classifier predicates produced by the system. As an upper baseline, they were also shown animations of classifier predicates produced using 3D motion capture technology to digitally

record the performance of other native ASL signers. Their evaluation of animations from both sources was compared to measure the system's performance. In fact, there was also a lower baseline used in the study—animations of Signed English transliterations of English sentences. This reflects the current state of the art in English-to-Sign translation technology. Some preliminary results of the evaluation study are discussed in the next section.

4 Prototype implementation and evaluation

This project has produced a detailed specification of the classifier predicate translation models [8], the generation approach [7], the multi-pathway machine translation architecture in which it will be situated [6], and a multi-channel timing representation for ASL performances [11]. A prototype implementation of the classifier-predicate-generation pathway of the system has been developed. This prototype consists of the linguistic data structures and processing architecture, and it has sufficient lexical and grammatical resources to support a limited linguistic repertoire of ASL classifier predicates.

A pilot evaluation study was conducted in which native ASL signers evaluated the output of the system and compared it to two types of animation baselines, namely animations of ASL produced by motion-capture of human signers and animations of Signed English sentences manually scripted by animators [9, 12]. This pilot study indicated that the overall software design of the system was capable of generating animations of ASL sentences that participants felt were grammatical, understandable, and natural-appearing. Specifically, participants rated the ASL animations produced by the system significantly higher in all three categories (grammaticality, understandability, naturalness) than several Signed English animations that they were shown as a lower-baseline [12]. As this was only an initial study of a prototype implementation, these results are preliminary; additional evaluation of the system will be conducted in future work.

5 User-interface applications

While the description of the ASL animation system above explains how scene-visualization software can be used to calculate the arrangement of the placeholders, it is possible that their layout could be calculated in other ways. If the English sentence to be translated is discussing objects whose spatial locations are known to the computer, then "scene-visualization" software is not needed to arrange the placeholders. For example, if an animated ASL signer were embedded in a computer user-interface in order to present the elements of the surrounding Graphical User Interface (GUI), then the system will need invisible placeholders in front of the signer representing the layout of the windows, buttons, and icons. These placeholders will be used to produce classifier predicates that describe or refer to these GUI elements.

In this scenario, the scene visualization software is not needed: the screen coordinates of the GUI elements can be used to directly determine how their corresponding invisible placeholders should be arranged in front of the signer's torso. When the ASL character is embedded in a user-interface, the current screen coordinates of the surrounding GUI elements can be used to instantiate a corresponding set of invisible placeholders in front of the signing character. The scene-visualization software is not necessary—the layout of the placeholders is a simple mapping process from the screen coordinates to the volume of space in front of the signer. If GUI elements change location, then the location of their corresponding placeholders can be updated automatically, and these changes can be reflected in the ASL animation produced.

Figure 11 contains a diagram representing the operation of the classifier predicate generation system when the scene-visualization software is used. This figure illustrates several of the stages in the translation of an English sentence into an ASL classifier predicate. Specifically, it shows how scene-visualization software is used to produce a 3D model of the arrangement of the objects mentioned in the text. Then this 3D model is overlaid onto the volume of space in front of the signing character. Finally, these objects are converted into invisible placeholders that record the location and orientation of the objects in the scene.

By comparison Fig. 12 shows how the system would work when the spatial layout of the objects is known to the computer, as is the case of GUI screen elements. This figure illustrates how the system can produce ASL classifier predicates that describe the layout of the Graphical User Interface (GUI). Unlike Fig. 11, a scene-visualization step is not needed in this process. The computer system can directly access the two-dimensional GUI screen coordinates of the windows, buttons, and icons on the screen, and it can use this information to set up invisible placeholders in front of the signing character. These placeholders will be used during the creation of classifier predicates that describe the layout of the computer screen.

For computer software developers who wish to make their programs accessible to ASL users, using an automatic ASL translation system to produce animations describing the user-interface is more practical than videotaping a human ASL signer. First of all, not every software company may devote the resources into making or updating such videos, and there is another challenge: variations in screen size, operating system, or user-configured options

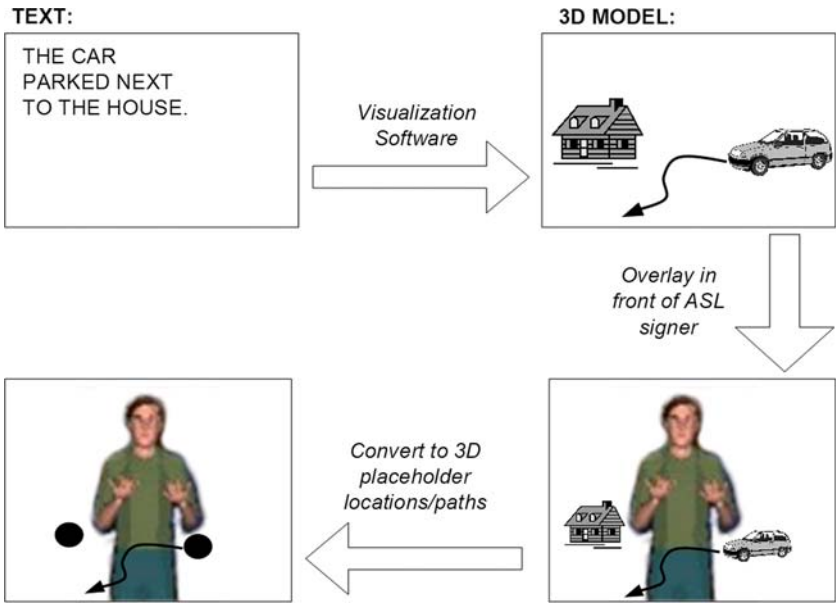

Fig. 11 Stages in the translation of an English sentence into an ASL classifier predicate

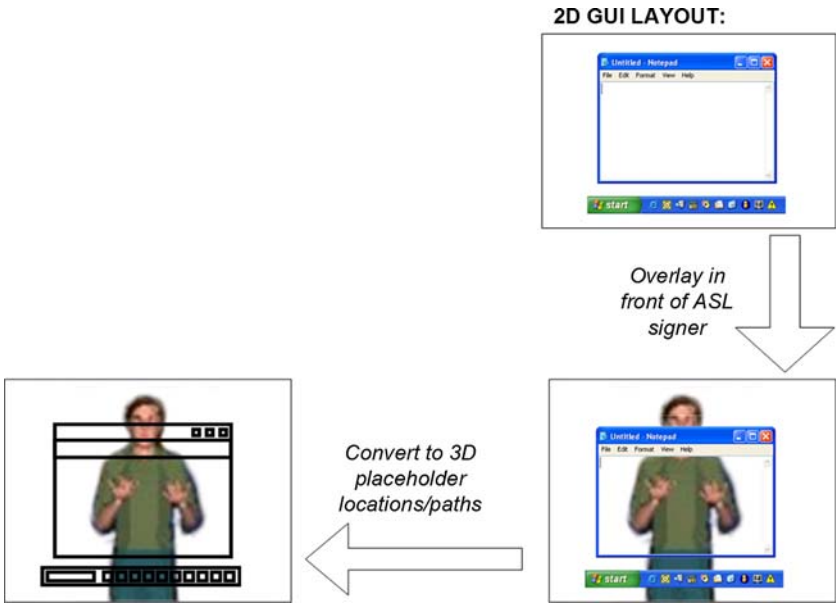

Fig. 12 American Sign Language classifier predicates used to describe the layout of a Graphical User Interface (GUI)

may cause the icons, frames, buttons, and menus on an interface to be arranged differently. A different layout of classifier predicates would be needed to describe each of these different screen configurations; producing a video of a human signer for each would be impractical. By using a translation system, the 3D placeholders could be updated dynamically to match the screen, and they can be used during generation of ASL classifier predicate animations to describe any GUI configuration.

The simplified style of writing often found in help-file text can make it easier to translate with an automatic English-to-ASL machine translation system. For instance, the consistent manner in which English help-file or instructional text refer to user interface elements can be exploited to simplify the translation process. In natural language text, there are often many different ways to refer to an object under discussion. For instance, all of the following could be used in a conversation to refer to the same object: "the blue car across the street," "the blue car," "the car," "the hatchback," etc. Pronouns may also be used (e.g., "it" or "that"). In order to successfully translate English text into ASL, a machine translation system would need to successfully determine that all of these phrases actually refer to the same object and which phrases in the text do not refer to this object. This can be a difficult task. Fortunately, the technical writers who create the English text in software help-files typically use a controlled vocabulary and consistent terminology when referring to

elements of the onscreen user interface. This consistent use of terminology can significantly simplify the task of *reference resolution* described above.

6 Conclusion and future work

This paper has illustrated how several misconceptions about the deaf experience, the linguistics of ASL, and the suitability of traditional MT technology to the language have delayed the creation of English-to-ASL machine translation software. Several of the important challenges in developing MT methods for ASL have also been described to show how studying ASL can push the boundaries of current MT methodologies. Both the special difficulty in translating classifier predicates and the familiarity some ASL signers have with Signed English have motivated this system's exploration of a multi-pathway architecture for MT. The spatial nature of classifier predicates motivated the integration of scene-visualization software to produce a 3D model of objects under discussion. The capabilities of the scene-visualization software motivated new representations of classifier predicate placeholder objects and output phonological models.

While this article has focused on English and ASL, many of the issues discussed, including many of the developed linguistic technologies for generating classifier predicate animations, are applicable to other international sign languages. These languages have their own lexical signs and grammatical structures distinct from ASL, and they use their own system of classifier predicates. Each of these other sign languages uses slightly different handshapes or motion path patterns. A 3D model serving as an intermediary between a written and a signed language could be used to translate Japanese to Japanese Sign Language, French to French Sign Language, Dutch to Sign Language of the Netherlands, etc.

In future work, additional evaluation trials will be conducted to determine optimal values for visual parameters of the animation: the lighting conditions of the animation, the camera angle (within the virtual animation), and several color settings. In addition, various performance parameters will also be evaluated: the speed of the signer's movement during different portions of the performance, the height of the signer's eye-brow raise, the amount of space to use in front of the signer to set-up the 3D scene, etc. The results of these trials will be used to guide the future development of the ASL animation system.

Another important aspect of future work is the expansion of the linguistic coverage of the system, in terms of both the variety of linguistic structures and vocabulary items that can be successfully translated from English and the variety of ASL phenomena and signs that can be generated in the ASL animation output. In parallel to the development of the core ASL technology, the design of new assistive technology applications that can benefit deaf users will also be explored. These new applications will motivate improvements in the ASL animation software, and the particular requirements of each application will drive and prioritize what linguistic components of the software should be the focus of development.

Acknowledgments This work was supported by a grant from the US National Science Foundation (Award #0520798 "SGER: Generating Animations of ASL Classifier Predicates," Universal Access Program, 2005). Software used in this project has been donated by Siemens UGS Tecnomatix and Autodesk. I would like to thank my collaborators at the Center for Human Modeling and Simulation at the University of Pennsylvania: Liming Zhao, Erdan Gu, and Jan Allbeck. I would also like to thank Mitch Marcus, Martha Palmer, and Norman Badler for their guidance and support during this work.

References

1. Bindiganavale, R., Schuler, W., Allbeck, J., Badler, N., Joshi, A., Palmer, M.: Dynamically altering agent behaviors using natural language instructions. In: Proceedings of the 4th International Conference on Autonomous Agents, AGENTS 2000, 3–7 June 2000, Barcelona, Catalonia, Spain (2000)
2. Coulter G (ed) Phonetics and phonology: current issues in American Sign Language Phonology. Academic, New York (1993)
3. Elliott, R., Glauert, J., Jennings, V., Kennaway, J.: An overview of the SiGML Notation and SiGML Signing Software System. In: Streiter, O., Vettori, C. (eds.), Proceedings of the Workshop on the Representation and Processing of Signed Languages, 4th International Conference on Language Resources and Evaluation: LREC 2004. 30 May 2004, Lisbon, Portugal, pp. 98–104 (2004)
4. Holt, J.: Demographic stanford achievement test—8th edition for deaf and hard of hearing students: reading comprehension subgroup results (1991)
5. Huenerfauth, M.: A survey and critique of American Sign Language natural language generation and machine translation systems. Technical Report MS-CIS-03–32, Computer and Information Science, University of Pennsylvania (2003)
6. Huenerfauth, M.: A multi-path architecture for machine translation of English text into American Sign Language animation. In: Proceedings of the Student Workshop of the Human Language Technologies conference/North American chapter of the Association for Computational Linguistics annual meeting: HLT/NAACL 2004. Boston, MA, USA (2004)
7. Huenerfauth, M.: Spatial representation of classifier predicates for machine translation into American Sign Language. In: Proceedings of the Workshop on the Representation and Processing of Signed Languages, 4th International Conference on Language Resources and Evaluation: LREC 2004. Lisbon, Portugal (2004)
8. Huenerfauth, M.: Spatial and planning models of ASL classifier predicates for machine translation. In: Proceedings of the 10th international conference on theoretical and methodological issues in machine translation: TMI 2004, Baltimore, MD, USA (2004)
9. Huenerfauth, M.: American Sign Language generation: multimodal NLG with multiple linguistic channels. In: Proceedings of the Association for Computational Linguistics, 43rd Annual Meeting, Student Research Workshop, Ann Arbor, MI, USA (2005)

10. Huenerfauth, M.: American Sign Language spatial representations for an accessible user-interface. In: Proceedings of the 3rd international conference on universal access in human-computer interaction, Las Vegas, NV, USA (2005)
11. Huenerfauth, M.: Representing coordination and non-coordination in an American Sign Language Animation. In: Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2005), Baltimore, MD, USA (2005)
12. Huenerfauth, M: Generating American Sign Language classifier predicates for English-to-ASL Machine Translation, Ph.D. Dissertation, Computer and Information Science, University of Pennsylvania (2006)
13. Liddell, S.: Grammar gesture and meaning in American Sign Language. Cambridge University Press, Cambridge (2003)
14. Liddell, S.: Sources of meaning in ASL classifier predicates. In: Emmorey, K. (eds.) Perspectives on classifier constructions in sign languages. Workshop on Classifier Constructions, La Jolla (2003)
15. Liu, Y.: Interactive reach planning for animated characters using hardware acceleration. Doctoral Dissertation, Computer and Information Science, University of Pennsylvania (2003)
16. Mitchell, R. How many deaf people are there in the United States. Retrieved June 28, 2004 from Gallaudet Research Institute, Graduate School and Professional Programs, Gallaudet University Web site: http://www.gri.gallaudet.edu/Demographics/deaf-US.php (2004)
17. Morford, J., MacFarlane, J.: Frequency characteristics of American Sign Language. Sign Lang. Stud. **3**(2), 213–225 (2003)
18. Neidle, C., Kegl, J., MacLaughlin, D., Bahan, B., Lee, R.: The syntax of American Sign Language: functional categories and hierarchical structure. MIT, Cambridge (2000)
19. Sáfár, É., Marshall, I.: The architecture of an English-Text-to-Sign-Languages translation system. In: Angelova, G. (ed.) Recent advances in natural language processing (RANLP). Tzigov Chark, Bulgaria, pp. 223–228 (2001)
20. Wideman, C., Sims, M.: Signing avatars. In: Proceedings of the Technology and Persons with Disabilities Conference, March 15–20, 1999, Los Angeles, CA, USA (1998)
21. Zhao, L., Kipper, K., Schuler, W., Vogler, C., Badler, N., Palmer, M.: A machine translation system from English to American Sign Language. In: Proceedings of the 4th conference of the association for machine translation in the americas on envisioning machine translation in the information future, lecture notes in computer science, 1934, Springer, Heidelberg, London, pp. 54–67 (2000)
22. Zhao, L., Liu, Y., Badler, N.I.: Applying empirical data on upper torso movement to real-time collision-free reach tasks. In: Proceedings of the SAE Digital Human Modeling Conference, Iowa City (2005)