American Sign Language Spatial Representations for an Accessible User-Interface

Matt Huenerfauth

University of Pennsylvania
Philadelphia, PA 19104
matt@huenerfauth.com

Abstract

Several common 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 paper will discuss these issues and will trace the progress of a new English-to-ASL MT project that has made translating texts important for literacy and user-interface applications a priority. These texts include ASL phenomena called "classifier predicates." Some challenges in producing classifier predicates – and this project's novel solutions to these challenges – will be described. Finally, the way in which this English-to-ASL machine translation system can be used to develop user-interface more accessible to deaf users will be discussed.

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 U.S. have only a fourth grade English reading level (Holt 1991). (Students age 18 and older have a reading level more typical of a 10-year-old student.) The primary means of communication for approximately one half million deaf people in the United States is American Sign Language (ASL), a full natural language with a linguistic structure distinct from English (Neidle et al., 2000, Liddell, 2003; Mitchell, 2004). Thus, it is possible to have fluency in ASL without literacy in written English. This literacy issue has become more significant 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 software designers incorrectly assume that written English text in a user-interface is always accessible to deaf users. 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 project is to develop such an English-to-ASL MT system, specifically one that can produce ASL animations that include important phenomena called "classifier predicates."

2 Several Misconceptions

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 computational linguistic software to the special linguistic properties of ASL. This paper will then describe the work of this project: to develop a machine translation (MT) system to convert from 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.

2.1 Misconception: All deaf users are English-literate

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 programming 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. Communications 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. A recent issue is that 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 needs of the user are met.

A machine translation system from English text into American Sign Language 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 articulate enough to perform ASL (Wideman and Sims, 1998). Most animation systems use a basic instruction set to control the character's movements; so, 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 (Elliot et al., 2004).

2.2 Misconception: ASL is manually performed English

Even when designers understand that presenting English text is not a complete solution for deaf users, confusion within the Natural Language Processing research community over 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 we were to replace every word of an English sentence 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, these are often limited to use in English classrooms for the deaf. 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 computational linguistic 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 (Huenerfauth 2003). 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 systems advertise themselves as "translation" systems and claim to produce ASL – thus misleading and disappointing any accessibility technology researcher who may be interested in making use of true ASL translation software.

2.3 Misconception: MT technology designed for written languages will work well for ASL

The previous section has already 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 seen in these languages (Neidle et al. 2000; Liddell 2003). 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 doesn't 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" (or CPs). During a CP, 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 CP 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 next to the house," the signer could use two 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 and (2) the dominant hand in a 'sideways 3' handshape would trace a path in space corresponding to a car driving and stopping next to the 'house' location in space. (See Figure 1.) There are also important elements of facial expression, eye gaze, and head tilt which convey meaning during a CP, but they are omitted from the current discussion.

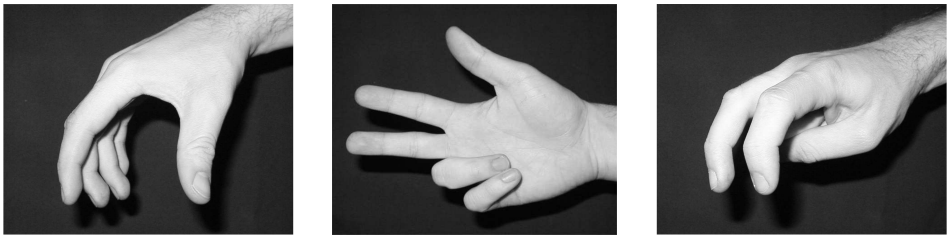

Figure 1: CP Handshapes: 'Downward C', 'Sideways 3', and 'Bent V' (discussed later).

Not every ASL sentence contains a classifier predicate, and if we ignore the non-topological "pointing" pronouns, then many ASL sentences have a structure that looks similar to English or other written languages. The problem is that this non-CP subset of the language 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 exclusively on these non-CP parts of the language (Huenerfauth 2003).

This simplification has allowed them to reuse translation technologies originally designed for written languages. Some MT researchers have had success at producing ASL animations on this limited (non-spatial) portion of the language (Zhao et al. 2000; Sáfár and Marshall 2001). Unfortunately, since all of these systems employ only traditional linguistic approaches and none of them attempt to model the spatial arrangement of objects in a 3D scene being discussed, none of them are able to produce classifier predicates from an English text (Huenerfauth 2004a). No previous ASL MT system has proposed how to generate classifier predicates; this aspect of the language has simply 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 linguistic research. This has prevented many of the most popular stochastic 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 Misconception: It's OK to ignore CPs

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) (Morford & MacFarlane 2003). Further, CPs are the only way to convey some concepts 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 therefore important for an MT system to translate. (These are the sentences that we would especially want an MT system to translate 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, he or she can refer to individual elements 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-file text. English-illiterate users of a computer application would likely also have limited computer experience; so, conveying this type of content is 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 in this MT system's development will be described. In each case, our solution to the problem will be explained in order to illustrate how ASL has motivated several new MT approaches.

3.1 Problem: ASL phonological models not suited to CPs

A non-linguistic representation of an animation of a 3D character performing a CP 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 a 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 us 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 (Coulter 1993); however, these models are ill-suited to the representation of CPs. 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 CPs). Finally, these models make it very difficult to specify the complex motion paths required for some CPs (consider the various 3D motion paths that the "car" may travel in our earlier "parking" example).

3.2 Solution: A new phonological model of ASL CPs

As a first step in producing ASL CPs, this project had to select a symbolic representation for these phenomena that would serve as the output of the MT process. While a good representation should help to simplify and parameterize the output, it should be sufficiently detailed that 3D animation software can still used it as input to produce a final output animation of an ASL performance.

Specifically, eye-gaze and head-tilt are represented as a pair of 3D points in space at which they are aimed (Huenerfauth 2004c). This simplification is made because what is semantically meaningful in a CP 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; so, this model is sufficient for producing an animation. (We'll later discuss how special invisible placeholder objects will be arranged in the space in front of the signer. These placeholders will 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 CP, the position of the hand (not the whole arm or elbow) is semantically meaningful; so, the model can make another simplification. The locations in space of the dominant and non-dominant hands are recorded as another pair of 3D coordinates. (We also record the shape of each hand and the 3D orientation of the palm.) Given hand location and orientation values, there are algorithms for calculating realistic elbow/shoulder angles for a 3D virtual human character (Liu 2003); so, the model is again sufficient for generating animation (Huenerfauth 2004c).

3.3 Problem: Calculating 3D motion paths is difficult

The model of ASL classifier predicate output described above needed to select 3D coordinates for parts of the body over time. In a previous paper, several possible methods for calculating such 3D motion trails were compared (Huenerfauth 2004b). The most simplistic approach considered was to pre-store a list of all possible pairs of English motion verbs and ASL CP motion paths. However due to the many possible arrangements of 3D scenes that would each require slightly different forms of CP 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 (Huenerfauth 2004b). 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.

3.4 Solution: Use of Scene-Visualization Software

This system uses 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 (Badler et al., 2000). This model is "overlaid" onto the volume in front of the ASL signer (Figure 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 next to the house" example, a miniature invisible object representing a 'house' is positioned in front of the signer's torso, and another object (with a motion path terminating next to 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 CPs 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 classifier predicate describing the car's motion.

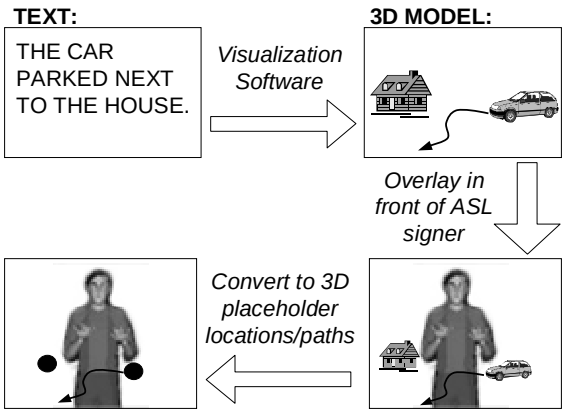

Figure 2: Converting English text into placeholders.

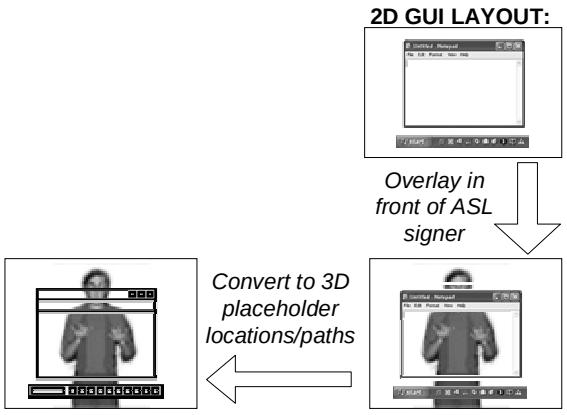

Figure 3: Arranging placeholders for GUI elements.

3.5 Problem: Visual Details are Difficult to Get Correct

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 eventual action being discussed. Spending processing and development time on these parts of the 3D model is unnecessary.

3.6 Solution: Placeholder Objects are Invisible & Abstract

Unlike some applications of the scene-visualization software – where entities described by the English text would need to be rendered to 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" or "houses" for example.

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. (See Figure 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, we can regard the purpose of an ASL signer producing a classifier predicate 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 CPs. In this light, the two classifier predicates in the "parking" example can be thought of as conveying that a bulky object occupied a point in space and a vehicle object moved toward it and stopping.

While the description 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 and must discuss the elements of the surrounding 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 (Figure 3).

3.7 Problem: 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. An example of a classifier predicate in which the path of hand motion does not match the path of motion of the placeholder object being described would be the predicate for "leisurely walking upright figure" (Liddell 2003b). To show the path that person walks in a leisurely manner, the signer can put the hand in a "number 1" handshape (index finger pointing up, all other fingers closed). Then the signer bounces the 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 he or she is walking leisurely. This bouncing quality of the movement is linguistically (not visually) determined.

This example and other linguistic considerations (Huenerfauth 2004b) have led this project reject the idea that 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 (like the concept of "leisurely" in the example above).

3.8 Solution: Templated Approach to CP Generation

After calculating the 3D layout of the entities discussed in an English text, an approach is needed to generate animations of CPs describing this scene. We have argued (Huenerfauth 2004b) that a recent linguistic model of classifier predicate generation (Liddell 2003) can serve as a starting point for developing such an approach. In Liddell's model, signers have a mental image of a scene to be discussed (much like a 3D graphics model) 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).

Unfortunately, Liddell (2003) does not provide detail about the internal structure of these templates nor their selection/filling process. This MT project has developed new computational models for classifier predicate generation (Huenerfauth 2004c) within an English-to-ASL MT system that formalize and implement this linguistic account (with some modifications). The use of a templated approach solves the "linguistic movement" problem described above. 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 taken from the invisible placeholder objects in front of the signer's torso. In this way, we don't rely on the 3D scene for every 3D motion detail. We can specify some portions ahead of time in the template, and we can calculate complex motion paths 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" (Huenerfauth 2004b).

3.9 Problem: High Processing and Development Overhead

One problem with the translation approach described above 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 be 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 3D graphical model coordinates and layout could require a lot of processing time (thus preventing real-time English-to-ASL translation).

3.10 Solution: Use of a Multi-Path MT system

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 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. For this reason, the MT approach of this project is focused exclusively on the translation of English into classifier predicates. To handle a variety of input sentences, this technology will be embedded within a complete English-to-ASL MT system that contains multiple processing pathways (Huenerfauth 2004a).

The pathway for English inputs producing CPs includes the scene-visualization software, but the pathway for other inputs will use more traditional MT technology (that is easier to implement for a wider variety of input sentences). Finally, since most deaf signers have some familiarity with non-ASL English-like forms of signing, this design would also include a "transliteration" pathway: a word-for-sign substitution process that produces a Signed English output. (This English-like output will only 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 "falls back" on simpler pathways as needed. This architecture is therefore able to blend *deep* 3D-processing and *broad* input-coverage in a single system (Huenerfauth 2004a).

3.11 Problem: No one-to-one mapping of sentences to CPs

The "parking" example at the start of this paper illustrated how a single English sentence ("the car parked next to the house") can produce multiple classifier predicates (one for the house, one for the car). In fact, it is common for several CPs 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 CPs 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.

3.12 Solution: Same formalism for inter-CP and intra-CP

To address all of the above concerns, this 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-CP and intra-CP 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: it doesn't need to pre-commit to a fixed number of classifier predicates at the start of the generation process (more can be added to the output as necessary to satisfy ASL-specific linguistic rules). For a full discussion of the classifier predicate output planning process and a worked-out example, please see (Huenerfauth 2004c).

4 User-Interface Applications of this Technology

Section 2.4 discussed how it is important that an English-to-ASL machine translation system can produce classifier predicates if it will be used in an accessible user-interface. Section 3.6 also suggested that this machine translation system could be used to produce classifier predicates for an on-screen character that must refer to other parts of a GUI on the screen. The system can do this because of its 3D graphical model, which is used to track the location and movement of the invisible placeholders that underlie classifier predicates. 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 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).

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 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. If this translation system were used, then 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 refers to user-interface elements can be exploited to simplify the translation process. During a 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 Honda," "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 this task of *reference resolution* described above.

5 Conclusion

This paper has illustrated how several misconceptions about the deaf experience, the linguistics of ASL[1], 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 CPs 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 CPs 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 CP placeholder objects and output phonological models.

This project has produced a detailed specification of the CP translation models (Huenerfauth 2004c), the generation approach (Huenerfauth 2004b), and the multi-pathway machine translation architecture in which it will be situated (Huenerfauth 2004a). The implementation of the project is ongoing, and a version of the CP-generation pathway of the system will be evaluated in a study during the summer of 2005. The CP animation output will be evaluated by members of the deaf community who are native ASL signers.

Acknowledgements

I would like to thank my advisors Mitch Marcus and Martha Palmer for their guidance, discussion, and revisions during the preparation of this work.

References

Bindiganavale, R., Schuler, W., Allbeck, J., Badler, N., Joshi, A., & Palmer, M. (2000). Dynamically altering agent behaviors using natural language instructions. In *Proceedings of the 4th International Conference on Autonomous Agents.*

Coulter, G. (ed.). (1993). *Phonetics and Phonology: Current Issues in American Sign Language Phonology.* New York: Academic Press.

Elliott, R., Glauert, J., Jennings, V., and Kennaway, J. (2004). An Overview of the SiGML Notation and SiGMLSigning Software System. In *Proceedings of the Workshop on the Representation and Processing of*

[1] While this paper has focussed on English and ASL, this project's 3D approach to classifier predicate generation is applicable to other international signed languages. While these languages have their own lexical signs and grammatical structures distinct from ASL, they all use of a 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.

Signed Languages, 4th International Conference on Language Resources and Evaluation: LREC 2004. Lisbon, Portugal.

Holt, J. (1991). Demographic, Stanford Achievement Test - 8th Edition for Deaf and Hard of Hearing Students: Reading Comprehension Subgroup Results.

Huenerfauth, M. (2003). 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.

Huenerfauth, M. (2004). 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.*

Huenerfauth, M. (2004). 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.*

Huenerfauth, M. (2004). 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.*

Huenerfauth, M. (2004). American Sign Language natural language generation and machine translation. In *Proceedings of the 6th International ACM SIGACCESS Conference on Computers and Accessibility: ASSETS 2004, Doctoral Consortium and Poster Session. Atlanta, Georgia, USA.*

Liddell, S. (2003). Grammar, gesture, and meaning in American Sign Language. Cambridge, UK: Cambridge University Press.

Liddell, S. (2003). Sources of meaning in ASL classifier predicates. In K. Emmorey (Ed.), *Perspectives on Classifier Constructions in Sign Languages. Workshop on Classifier Constructions, La Jolla, San Diego, California.*

Liu, Y. (2003). Interactive reach planning for animated characters using hardware acceleration. Doctoral Dissertation, Computer and Information Science, University of Pennsylvania.

Mitchell, R. (2004). *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://gri.gallaudet.edu/Demographics/deaf-US.php

Morford, J. & MacFarlane, J. (2003). Frequency characteristics of American Sign Language. *Sign Language Studies*, 3 (2): 213-225.

Neidle, C., Kegl, J., MacLaughlin, D., Bahan, B., & Lee, R. (2000). The syntax of American Sign Language: Functional categories & hierarchical structure. Cambridge, MA: The MIT Press.

Sáfár, É. & Marshall, I. (2001). The architecture of an English-Text-to-Sign-Languages translation system. In G. Angelova (Ed.), *Recent Advances in Natural Language Processing (RANLP), Tzigov Chark, Bulgaria.*

Wideman, C. & Sims M. (1998). Signing avatars. In *Proceedings of the Technology & Persons with Disabilities Conference.*

Zhao, L., Kipper, K., Schuler, W., Vogler, C., Badler, N., and Palmer, M. (2000). A machine translation system from English to American Sign Language. In *Proceedings of the Association for Machine Translation in the Americas, Lecture Notes in Artificial Intelligence*. Berlin: Springer-Verlag.