Evaluating American Sign Language Generation Through the Participation of Native ASL Signers

Matt Huenerfauth
Computer Science Department, CUNY Queens College
The City University of New York
65-30 Kissena Blvd, Flushing, NY 11375 USA
1-718-997-3264
matt@cs.qc.cuny.edu

Liming Zhao, Erdan Gu, Jan Allbeck
Center for Human Modeling & Simulation
University of Pennsylvania
3401 Walnut St., Philadelphia, PA 19104 USA
1-215-573-9463
{liming,erdan,allbeck}@seas.upenn.edu

ABSTRACT
We discuss important factors in the design of evaluation studies for systems that generate animations of American Sign Language (ASL) sentences. In particular, we outline how some cultural and linguistic characteristics of members of the American Deaf community must be taken into account so as to ensure the accuracy of evaluations involving these users. Finally, we describe our implementation and user-based evaluation (by native ASL signers) of a prototype ASL generator to produce sentences containing classifier predicates, frequent and complex spatial phenomena that previous ASL generators have not produced.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *language generation, machine translation*; K.4.2 [**Computers and Society**]: Social Issues – *assistive technologies for persons with disabilities*.

General Terms
Design, Experimentation.

Keywords
American Sign Language, Animation, Natural Language Generation, Evaluation, Accessibility Technology for the Deaf.

1. Background and Motivations
American Sign Language (ASL) is a full natural language – with a linguistic structure distinct from English [10] [13] – used as a primary means of communication for approximately one half million people in the United States [11]. A majority of deaf 18-year-olds in the United States have an English reading level below average 10-year-old hearing students [5], and so machine translation software that could translate English text into ASL animations could significantly improve these individuals' access to information, communication, and services. Previous English-to-ASL machine translation projects [17] [18] have been unable to generate classifier predicates, a type of ASL phrase used to indicate the spatial location, size, shape, and movement of objects. Because classifier predicates are frequent in ASL and are necessary for conveying many spatial concepts in the language, we have developed a classifier predicate generator that can be incorporated into an English-to-ASL machine translation system.

During a classifier predicate, signers use their hands to position, move, trace, or re-orient imaginary objects in the space in front of them to indicate the location, movement, shape, contour, physical dimension, or some other property of corresponding real world entities under discussion. Classifier predicates consist of a semantically meaningful handshape and a 3D hand movement path [10]. A handshape is chosen from a closed set based on characteristics of the entity described (whether it is a vehicle, human, animal, etc.) and what aspect of the entity the signer is describing (surface, position, motion, etc).

For example, the sentence "the man walks between the tent and the frog" can be expressed in ASL using three classifier predicates. (Figure 1.) First, a signer performs the ASL sign TENT while raising her eyebrows (to introduce a new entity as a topic). Then, she moves her hand in a "Spread C" handshape (fingers curved like loosely holding a ball) forward and slightly downward to a point in space where an imaginary miniature tent could be envisioned. Next, the signer performs the sign FROG with eyebrows raised and makes a similar motion with a "Hooked V" handshape (index and middle fingers extended and bent slightly) to a location where a frog is imagined. Finally, she performs the sign MAN (with eyebrows raised) and uses a "Number 1" handshape (index finger extended upward) to trace a motion path between the locations of the 'tent' and the 'frog.' "Spread C" handshapes are typically used for bulky objects, "Hooked V" for animals, and "Number 1" for upright humans.

As part of our research into English-to-ASL machine translation systems, we have created a prototype system for generating ASL sentences that contain classifier predicates. We will discuss some of the implementation details of the system later in this paper, but first we will consider some important issues in the design of evaluation studies for ASL animation generators.

2. Selecting an Evaluation Method for ASL
Since there have been few sign language generation or machine translation systems developed, there has been correspondingly little work on how to best evaluate such systems. Broadly speaking, evaluations of natural language generation software fall into two major categories: automated and user-based. We will discuss how the special linguistic properties of American Sign

ASSETS'07, October 15–17, 2007, Tempe, Arizona, USA.

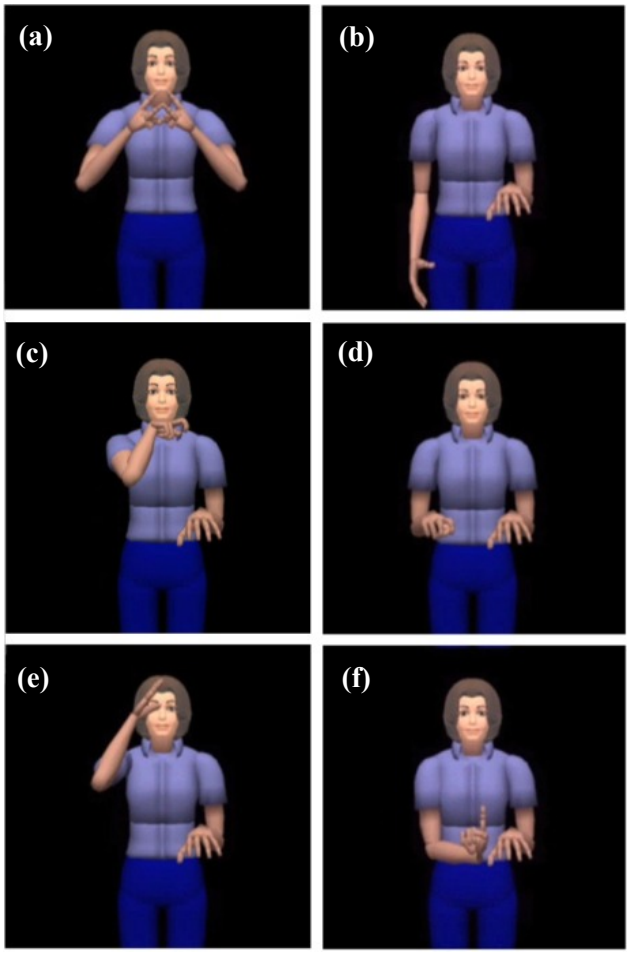

Figure 1: Images from our system's animation of classifier predicates for "the man walks between the tent and the frog." (a) ASL sign TENT, eyes at audience, brows raised; (b) Spread C handshape and eye gaze jumps to tent location; (c) ASL sign FROG, eyes at audience, brows raised; (d) Hooked V handshape and eye gaze to frog location; (e) ASL sign MAN, eyes at audience, brows raised; (f) Number 1 handshape (for the man) moves forward between the 'tent' and 'frog' while the signer's eye gaze tracks the movement path of the man.

Language make automated evaluation approaches difficult to employ, and we will describe how various linguistic and cultural factors affect the design of user-based studies of American Sign Language generation software.

2.1 Automated Evaluation of NL Generation

Automated approaches for evaluating the output of natural language generation systems have primarily been designed for evaluating systems that produce text output in some written language. These evaluation methods are "automated" in the sense that they compare the output of the system to a list of possible "correct answers" of desirable output strings that have been provided. These human-produced text strings are a set of possible correct output sentences the system should produce – often called "gold-standard" strings. To measure the performance of a natural language generation system, its output string is compared to the gold-standard string, and the degree of similarity between them (often calculated as the number of sub-phrases of various lengths that the strings have in common) is calculated [3] [15].

A major advantage of an automated evaluation method is that the string-distance metric is repeatable, and you can therefore reliably track the progress of a system under development over time against a constant set of gold-standard strings. There is also a significant cost-savings over a user-based evaluation design – the time and expense in recruiting users and conducting a study can make automatic techniques (when possible) an attractive approach to evaluating a natural language generation system.

In some cases, a set of humans are asked to write correct output sentences in the written language specifically for the purposes of evaluation. In such studies, the human participants will look at the same data that the system uses as the input to its natural language generation process – perhaps a knowledge base of semantic information to be conveyed – and they construct one or more grammatically correct sentences in some written language to express that information. In other cases, a source of gold-standard strings can be harvested from some naturally-arising source. For natural language generation systems that serve as the output component of a machine translation system, it is often possible to obtain a large sample of text with a corresponding sentence-by-sentence human-produced translation into another language (known by linguists as a "parallel corpus"). Such corpora often occur when government agencies provide records in multiple official languages or when news agencies provide translations of their articles. To evaluate the machine translation system from a source language to a destination language, the source-language version of each sentence in the corpus can be used as the input to the machine translation system, and the system's output can be compared to the destination-language version of that same sentence in the parallel corpus.

Since there can often be more than one correct answer for the output of a system producing natural language – e.g. there can be multiple correct translations for a given sentence – some metrics compare the output of the system to a list of possible gold-standard strings. The system may decide which string is closest, and then calculate the distance from that gold-standard or it may consider the similar features of the system's output to all of the sentences in the set [15].

2.2 Automatic Evaluation for ASL Systems

Several factors make the use of automatic evaluation approaches difficult for ASL. Sign languages typically lack standard written forms that are commonly used by signers. While we could use some artificial ASL writing system for the generator to produce as output (for evaluation purposes only), we have no source of gold-standard strings for the evaluation. Without a writing system in common use, it is not possible to "harvest" some naturally arising source of parallel English-ASL written corpora – no such corpora exist. It is also unclear whether human ASL signers could accurately or consistently produce written forms of ASL sentences to serve as gold-standards for such an evaluation. Further, the actual end users of an ASL generation system would never be shown artificial written ASL; they would instead see ASL animation output. Thus, evaluations based on strings would not test the full process – including the final synthesis of the "string" into an animation – a step during which errors may arise.

Even if we were to build a large corpus of ASL in some written form, the linguistic properties of ASL may confound the use of string-based evaluation metrics. An ASL performance

consists of the coordinated movement of several parts of the body in parallel (i.e. face, eyes, head, hands), and so a simple string that lists the signs performed during a sentence would be a very lossy representation of the original performance [6]. The string would likely not encode the non-manual parts of the sentence, and so any string-based metric operating on this string would fail to consider those important aspects of the performance.

Discourse considerations (e.g. topicalization) can also result in movement phenomena in ASL that may change the order of signs in a sentence without substantially changing its semantics; such movements would affect the score returned by a string-based metric while the meaning may change little. The use of head-tilt and eye-gaze during the performance of ASL verb signs may also license the dropping of entire constituents, e.g. the noun phrase subject or direct object of the sentence [13]. Since the entities discussed during an ASL conversation are often associated with locations in space around the signer at which head-tilt or eye-gaze is aimed during the verb sign, the entity is still expressed (via the head/eyes) though no manual signs are performed for it. An automatic metric may penalize such a sentence (for missing a constituent) while the information is still being conveyed.

Finally, ASL classifier predicates convey a lot of information in a single complex 'sign' (handshape indicating semantic category and movement showing 3D path/rotation), and it is unclear how we could "write" the 3D data of a classifier predicate in a string-based encoding or how to calculate an edit-distance between a gold-standard classifier predicate and a generated one.

Some researchers have empirically evaluated several automatic string-based evaluation metrics for sign language and have shown that string-based metrics do a poor job of identifying the best quality sign language translations [12]. These researchers propose building large parallel written/sign corpora that contain more information than just the input (English) sentence and the output (sign language) sentence for each pair. If the corpus were also annotated with additional syntactic and semantic information, then the more sophisticated evaluation metrics they propose could be enabled. To build such detailed corpora of sufficient size for a large-scale evaluation would be an extremely time-consuming and expensive prospect.

2.3 User-Based Evaluation of NL Generation

User-based evaluations of natural language generation systems have several advantages over automatic evaluation. Automatic metrics merely consider whether the gold-standard strings bear superficial similarity to the string generated by the system. The true meaning, tone, style, and other subtleties of the system's output are not explicitly considered. For instance, the system might generate an output sentence which is a perfectly good output but which through some oversight was not one of the alternatives included in the set of gold-standard strings. Even if the meaning of the sentence is the same as the meaning of a gold-standard string (but their exact wording differs), the system may be penalized. Further, the inclusion of a single word (such as "not") in the output of the evaluated system may not make a large difference in the superficial similarity of the strings, but it may have a major impact on the meaning of the sentence (and its correctness). Such subtleties are lost on an automatic evaluation metric, and only human judges who are asked to look at the output of a system and score its success can consider such factors.

Another advantage of user-based evaluation is that the output of the system does not have to be in the form of a written string. Human judges can listen to speech output or view animations – as is needed for sign language output. Because assigning a score to a sentence to rate its grammaticality or quality can be somewhat subjective, user-based evaluations often give the judges some form of objective task that they must accomplish to demonstrate their degree of comprehension of the sentence being evaluated.

2.4 User-Based Evaluation of ASL Systems

For the reasons above, we have selected a user-based design for the evaluation of our ASL classifier predicate generation system. For a user-based study of an ASL generator, some cultural and linguistic characteristics of the anticipated users of the system, members of the American Deaf community who use ASL, must be addressed to ensure the success and accuracy of the evaluation.

2.4.1 Identifying Native ASL Signers

When conducting a study in which human subjects evaluate the output of a natural language generation system, it is important for the subjects to be native speakers of that language. There are some subtleties that only a native user of a language can discern. Many adult users of American Sign Language learned ASL later in life as a second language – some did not experience hearing impairment until after childhood and others did not have access to sign language early in life (either due to family circumstances or placement in a lip-reading/speech focused educational program). An ideal native user of ASL is someone who learned ASL in early childhood through interaction with deaf family members at home or through experiences at a residential school for the deaf. Improper screening of subjects for an evaluation study can lead to the recruitment of judges who may not be sufficiently critical of the system's language output [13].

During the screening process, asking questions such as "How well do you sign?", "Are you a native signer?", or "Is ASL your first language?" can be ineffective and culturally insensitive. Many (though certainly not all) deaf people in the U.S. feel that usage of ASL is a central element of Deaf Culture and a requirement for membership in the Deaf community [14]. Responses to the above questions may be motivated by an individual's cultural beliefs and sense of community affiliation – rather than a consideration of linguistic skills. There is also a potential for some individuals to be offended at having their skill at ASL challenged – especially if done so by a hearing researcher. Such questions could be interpreted as challenging whether the individual is "deaf enough" or "culturally Deaf." A better alternative is to ask questions during screening that target whether the potential subject has had life experiences typical of a native signer: "Did you grow up using ASL as a child?", "Did your parents use ASL at home?", "Did you attend a residential school where you used ASL?", etc.

2.4.2 Creating a Comfortable ASL-Signing Setting

When seeking grammaticality judgments from ASL signers, it is important to minimize characteristics of the experimental environment that could prompt the signer to code-switch to a more English-like form of signing or accept such signing as being grammatically correct [13]. Many ASL signers are accustomed to switching to such signing when interacting with hearing individuals – especially those with basic levels of signing skill. To avoid this, the subject should be surrounded by ASL and not exposed to non-native English-like signing. During the study, instructions should be given to participants in ASL – preferably by another native signer. If possible, participants should be engaged in conversation in ASL before the experiment to help

Table 1: ASL sentences that were included in the evaluation study (with English glosses of each).

Transcript of ASL Sentence with Classifier Predicates (CPs)	English Gloss of the Sentence
ASL sign TENT; CP tent location; sign FROG; CP frog location; sign MAN; CP man path.	The man walks between the tent and the frog.
ASL sign TENT; CP tent location; sign TREE; CP tree location (near tent).	The tree is near the tent.
ASL sign TABLE; CP table location; sign LIGHT; CP lamp location (atop table).	The lamp is on the table.
ASL sign TABLE; CP table location; sign WOMAN; CP woman location (next to table).	The woman stands next to the table.
ASL sign WOMAN; CP woman location; sign MAN; CP man path (alongside woman).	The man walks next to the woman.
ASL sign WOMAN; CP woman location; sign MAN; CP man path (away from woman).	The man walks away from the woman.
ASL sign WOMAN; CP woman location; sign MAN; CP man path (up to woman).	The man walks up to the woman.
ASL sign CAR; CP car path (turning left).	The car turns left.
ASL sign HOUSE; CP house location; sign CAR; CP car path (toward house).	The car drives up to the house.
ASL sign HOUSE; CP house location; sign CAT; CP cat location; sign CAR; CP car path.	The car drives between the house and the cat.

produce an ASL-immersive environment. The employment of ASL interpreters does not necessarily guarantee an ASL environment will be created; interpreters often use a variety of signing communication systems – depending on the circumstance and the deaf client. Interpreters for this kind of study should have near-native ASL fluency, and they should be asked before the experimental session to use ASL (and not Signed English, etc.).

Many deaf people in U.S. have experienced educational or clinical settings in which use of English/speech has been valued more highly than use of ASL/signing. When asking ASL signers for their opinions on the grammaticality of various forms of signing, some ASL linguists have been careful to surround the experimental subject with other native ASL signers in a conversational setting [13]. This can prevent the creation of a clinical- or official-feeling environment in which signers may be more prone to use English-like signing or feel that English-like signing that they see is grammatically acceptable.

As with any study, it is important that users feel comfortable criticizing the system being evaluated. In this context, it is important that the subject does not feel like someone who built the system is sitting with them while they critique it – or else they may not feel as comfortable offering negative opinions about the system. When conducting user-based evaluations of written-language generation software, this is somewhat less of an issue since the study participants are shown a piece of text and asked to evaluate it – there may be little presumption that a computer programmed by that researcher wrote the text. With an ASL animation system, it is more obvious that a computer produced the output, and there could be a presumption that the researcher who is in the room helped to write the software.

3. The Design of Our ASL Generation System

We have built a prototype generation module that produces ASL sentences that contain classifier predicates. Such a component can be incorporated into a full English-to-ASL machine translation system to enable it to translate English sentences that discuss the movement of people and objects. Classifier predicates are the way such spatial information is conveyed in ASL. This paper focuses on our evaluation of a prototype implementation of this system, and the implementation details of the system are outlined briefly below. The system's development is ongoing, and additional technical details can be found here [6] [7] [8] [9].

Our prototype can translate a limited range of English input sentences (discussing the locations/movements of a set of people or objects) into animations of ASL performance in which an onscreen human-like character performs a set of classifier predicates to convey the locations and movements of the entities in the English text. Table 1 includes shorthand transcripts of some ASL animations produced by the system; the first sentence corresponds to the classifier predicate animation in Figure 1.

To be used in an English-to-ASL machine translation system, our system assumes the use of software that can calculate a 3D set of positions for a set of objects discussed using spatial language in English. Various such systems have been developed [2] [4]. When given a 3D model of the arrangement of a set of objects whose location and movement should be described in ASL, our system produces an animation of ASL sentences containing classifier predicates to describe the scene. The software overlays a 3D model of invisible placeholders for each object onto the volume of space surrounding the signing character; these placeholders are used to select the hand locations/movements to use during the classifier predicates representing each object.

To build a complete ASL performance containing multiple classifier predicates with accompanying referring expressions – as in Figure 1 – our system uses a planning-based architecture. Templates of classifier predicate performance are stored in the system's library; each stores a set of animation movements that are modified based on the 3D location of the object being shown (where the signer's hand needs to reach in 3D space). The output of the planning process is a structure that represents how the various parts of the animated signing character's body should move in parallel and in sequence over time [6].

This animation specification is performed by an animated human character in the Virtual Human Testbed [1], and the head of the character is controlled using the Greta facial animation software [16]. In addition to the movements of the signing character's arms, the character raises its eye-brows to indicate topicalization of noun phrases in the performance, aims its eye-gaze at points in space as required during classifier predicates, and tilt its head to accommodate natural eye-gaze movements [9].

4. The Design of Our Evaluation Study

We designed and performed a user-based evaluation of our ASL classifier predicate generator; the study took into account the need to identify native ASL signers and the importance of preventing code-switching to English, as discussed in Section 2.4 above.

4.1 Questions Asked in Our Study

Native ASL signers were shown the output of our ASL generator and were asked to rate each animation on ten-point scales for understandability, naturalness of movement, and grammatical correctness. These three categories were chosen because we believe that the understandability of the animation is a key criterion (since our goal is to make more information and services accessible to low-literacy deaf users) and that the grammatical correctness and naturalness are factors that can contribute to the

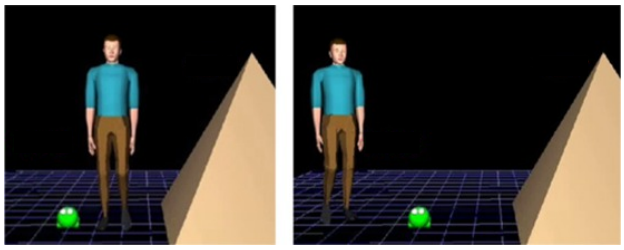

Figure 2: Still image from correct the visualization animation (left) and one of the confusables (right) for the ASL sentence in the study glossed as "the man walks between the tent and the frog" (Table 1).

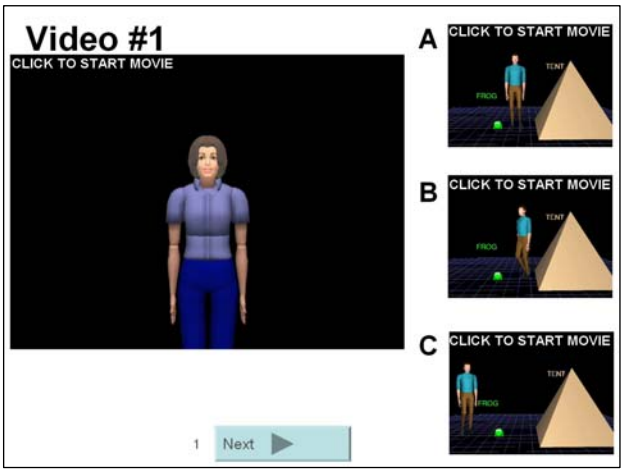

Figure 3: Screenshot from evaluation program.

understandability. Asking about the grammaticality of the output can help to identify problems in the linguistic planning of the output sentences, while asking about the naturalness of movement can identify problems in the animation portion of the system.

To make the evaluation less subjective and to better evaluate whether the animation conveyed the proper meaning, signers were also asked to complete a matching task. After viewing a classifier predicate animation produced by the system, signers were shown three short animations showing the movement or location of the set of objects that were described by the classifier predicate. The movement of the objects in each animation was slightly different, and signers were asked to select which of the three animations depicted the scene that was described by the classifier predicate. One animation was an accurate visualization of the location and movement of the objects, and the other two animations were "confusables" – showing orientations/movements for the objects that did not match the classifier predicate (Figure 2). To focus our evaluation on the classifier predicates (and not the referring expressions), the objects appearing in all three visualizations for a sentence was the same. Thus, it was the movement and orientation information conveyed by the classifier predicate (and not the object identity conveyed by the referring expression) that would distinguish the correct visualization from the confusables. For example, the three visualizations were created for the sentence "the man walks between the tent and the frog" (the frog and tent remain in the same location in each): (1) a man walks on a path between a tent and a frog, (2) a man stands in between a tent and a frog, and (3) a man starts at a location not between a tent and a frog and walks on a path never crossing between them.

Finally, signers were asked to comment on ways to improve the ASL animations. This feedback will be used to help direct future development efforts toward those portions of the system that native ASL signers felt required the most improvement.

4.2 A Lower Baseline: Signed English

Since this prototype is the first generator to produce animations of ASL classifier predicates, there are no other systems to compare it to in our study. The results are more meaningful if compared to a lower-bound on the system's performance. For a lower baseline, we wanted animations that reflected the current state of the art in broad-coverage English-to-sign translation. Since there not yet any broad-coverage English-to-*ASL* MT systems, we used Signed English transliterations as our lower baseline. Signed English is a form of communication in which each word of an English sentence is replaced with a corresponding sign, and the sentence is presented in original English word order without accompanying ASL linguistic features such as meaningful facial expressions or eye-gaze. Such animations have limited accessibility benefit for low-English-literacy deaf users since the English sentence is not translated into ASL grammatical structure – it retains its English structure. Having an environment with minimal English influence is important during the study so that when subjects view these Signed English animations, they have not already been primed to accept English-like signing as being grammatically correct ASL.

Simply having a lower baseline does not provide a numerical definition of success – since our prototype is the first ASL generator evaluated against it. Obtaining scores relative to a baseline does make it easier for future researchers to compare the performance of their ASL classifier predicate generators to ours.

4.3 Sentences Evaluated in the Study

Ten ASL animations from our generator were selected for this study based on several criteria. Sentences consist of classifier predicates of movement and location – the focus of our research. The categories of objects discussed in the sentences require a variety of ASL handshapes. Some sentences describe the location of objects, and others describe movement. Sentences describe from one to three objects in a scene, and some pairs of sentences actually discuss the same set of objects, but moving in different ways. Since the referring expression generator was not a focus of our prototype, all referring expressions are just an ASL noun phrase consisting of a single sign (phrases like "FROG" before a classifier predicate) – some one-handed and some two-handed.

To create the Signed English animations for each sentence, some additional signs were added to the generator's library of signs. (ASL does not traditionally use signs such as "THE" that are used in Signed English.) A sequence of signs for each Signed English transliteration was concatenated, and smooth transitional movements for the arms and hands between each sign were added. The English glosses in Table 1 correspond to the Signed English sentence animations presented to subjects in the study.

4.4 Survey Questionnaire and User-Interface

A simple on-screen user-interface was created that displayed one signing animation and three alternative object-visualizations on the screen at a time (Figure 3). After creating a slide for each of the 20 animations (10 ASL, 10 Signed English), the slides were placed in random order in a presentation (in a different order for each user). A user could re-play the animations as many times as desired before pressing the "Next" button at the bottom of the slide to go to the next signing animation. Subjects recorded their

Good ASL grammar? (10=Perfect, 1=Bad):
10 9 8 7 6 5 4 3 2 1
Easy to understand? (10=Clear, 1=Confusing):
10 9 8 7 6 5 4 3 2 1
Natural? (10=Moves like person, 1=Like robot):
10 9 8 7 6 5 4 3 2 1
Which picture/movie on the right matches? A B C

Figure 4: Sample question from the survey form.

responses by circling a choice on a paper survey form (Figure 4). Subjects rated each of the animations on a 1-to-10-point scale for ASL grammatical correctness, understandability, and naturalness of movement. Subjects were also asked to select which of the three animated visualizations (choice A, B, or C) matched the scene as described in the animated sentence that was performed.

While there was English text on the paper-based questionnaire, it was designed such that signers were presented with an identical set of four multiple-choice questions for each of the 20 animations. The questionnaire was not full English sentences but rather short cue phrases for each question (Figure 3). Although English text appeared on the survey form, this was considered better than having a researcher sit with each participant asking questions in ASL and taking notes on their responses. Even if the researcher was a native ASL signer (to avoid exposing the participant to English-like signing), interaction with a researcher during the entire session may have affected subjects' responses by producing a very clinical setting or one in which subjects felt that the person conducting the experiment was involved in the creation of the system being evaluated.

After these 20 slides, 3 more slides appeared containing animations from our generator (repeats of animations used in the main part of the study.) These three slides only showed the "correct" animated visualization for that sentence. Subjects were asked to comment on the animation's speed, colors/lighting, hand visibility, correctness of hand movement, facial expression, and eye-gaze. Signers were also asked to offer any comments they had about how the animation should be improved.

4.5 Recruitment & Interaction with Subjects

Personal contacts in the local deaf community in Philadelphia helped to recruit friends, family, and other associates who met the screening criteria. To participate, an individual needed to be a native ASL signer (as discussed in Section 2.4). Subjects were preferred who had learned ASL since birth, had deaf parents that used ASL at home, and/or attended a residential school for the deaf as a child (where they were immersed in an ASL-signing community). Of our 15 subjects, 8 met all three criteria, 2 met two criteria, and 5 met only one criterion (1 grew up with ASL-signing deaf parents and 4 attended a residential school for the deaf from an early age). As an informal check on participants' level of ASL fluency, a native ASL signer was present during 13 of the 15 sessions to converse in ASL with each participant.

During the study, instructions were given to participants in ASL, and a native signer was present during 13 of the 15 sessions to answer questions or explain experimental procedures. This signer engaged the participants in conversation in ASL before the session to help produce an ASL-immersive environment. Participants were given instructions in ASL about how to respond to each of the survey items. For grammaticality, they were told that "perfect ASL grammar" would be a 10, but "mixed-up" or

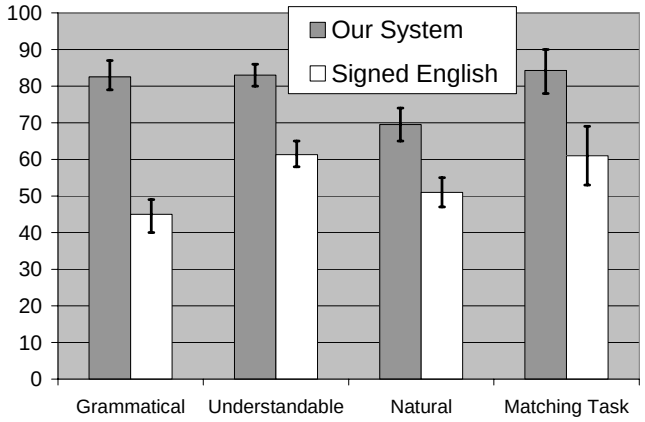

Figure 5: Grammaticality, understandability, naturalness, and matching-task success scores for our system vs. Signed English.

"English-like" grammar should be a 1. For understandability, they were told that "easy to understand" sentences should be a 10, but "confusing" sentences should be a 1. For naturalness, they were told that animations in which the signer moved "smoothly, like a real person" should be a 10, but animations in which the signer moved in a "choppy" manner "like a robot" should be a 1.

5. Results of the Evaluation Study

There were two groups of animations tested: ASL classifier predicate animations produced by our system and the Signed English lower-baseline animations. There were ten sentences included in the study, and so 20 animations were evaluated: 2 groups × 10 sentences. (While we calculated scores for individual animations in this study to identify issues with particular sentences, only per-group results are discussed in this paper due to space limitations.) Since there were 15 participants in the study, we collected a total of 300 responses. Each response consisted of a score in four categories: grammaticality (1 to 10), understandability (1 to 10), naturalness of movement (1 to 10), and whether the signer identified the visualization that correctly matched the animation (1=correct, 0=incorrect). Figure 5 shows scores for grammaticality, understandability, naturalness, and match-success percentage for each group. The classifier predicate generator's higher scores are significant ($\alpha = 0.05$, pairwise Mann-Whitney U tests with Bonferonni-corrected p-values).

5.1 Perceived vs. Actual Understanding

Among the 300 responses, there are significant ($\alpha = 0.05$) pairwise correlations between grammaticality, understandability, naturalness, and match-success (Table 2 contains Pearson's R-values, those values that are significant are shown with an asterisk*). Grammaticality, naturalness, and understandability

Table 2: Pairwise correlation coefficients (Pearson's R values) between categories on the 300 responses in the study (150 responses for animations from our system and 150 for the Signed English animations). Values that meet the significance level ($\alpha = 0.05$) are marked with an asterisk*.

	Grammaticality	Understandability	Naturalness
Understandability	0.63*		
Naturalness	0.64*	0.68*	
Match-Success	0.09	0.15*	0.06

were moderately correlated – not surprising since grammaticality and naturalness of an animation could affect understandability. A surprising result was the weak correlation between match-success and understandability. We would have expected that the respondent's perception of the understandability to be more closely correlated with her actual success at selecting the right visualization. The understandability score that a respondent selected was not a strong indicator of whether she would select the proper visualization (i.e. whether or not she understood the spatial information conveyed by the sentence). There appears to be a difference between a respondent's *perceived* understanding and her *actual* understanding of an animation. Thus, reported understandability scores are no substitute for the visualization matching data. Without collecting the match-success values, we may not have been able to determine whether respondents actually understood each animation.

5.2 Qualitative Feedback from Participants

During the last three slides in the study, subjects were asked to comment on the animation speed, visibility of the signer's hands, color and lighting, correctness of hand movements, correctness of facial expressions, and correctness of eye-gaze. They were also invited to recommend ways to improve the animations. Of the 15 subjects, eight said that some animations were a little slow, and one felt they were very slow. Eight subjects mentioned that the animations should have more facial expressions, and 4 of these specifically mentioned missing nose and mouth movements. Two subjects wanted the signer to show more emotion. Four subjects said the signer's body should seem more loose/relaxed or that it should move more. Two subjects felt that eye-brows should go higher when raised, and three felt there should be more eye-gaze movements. Two subjects felt the blue color of the signer's shirt was a little too bright, and one disliked the black background.

A few subjects commented on certain ASL signs that they felt were performed incorrectly. For example, three subjects discussed the sign "FROG": one felt it should be performed a little more to the right of its current location, and another felt that the hand should be oriented with the fingers aimed to the front.

Some participants commented on the classifier predicate portions of the performance, which were the focus of our system. For example, in the sentence "the man walks between the tent and the frog," one subject felt that it would be better to actually perform the sign TENT in the 3D location at which the tent is imagined in the signing space – instead of using the "Spread C" handshape to show the tent's location. Certain ASL signs can be signed in alternate locations to set up objects in space in this way.

6. Attempted Motion-Capture Upper Baseline

To add an upper-baseline for the evaluation, the participants could be asked to compare the animations from our system to videotapes of human signers. However, we didn't want subjects to focus on the superficial differences between the video and the animations, we wanted them to focus on any grammatical or movement errors that our animated signer might make. We therefore experimented with recording a native ASL signer (using a motion-capture suit and datagloves) performing classifier predicates. We were hoping that we could use the motion-capture data collected to animate a virtual human character superficially identical to the one used by our system. We hoped this character controlled by human movements could serve as an upper-baseline in the study. Thus, we could compare classifier predicate animations from our system to classifier predicate animations created from human movements while controlling for variation in the visual appearance.

A motion-capture room with an Ascension Technologies ReActor II system and a pair of wireless CyberGloves from Immersion Corporation was used to record a human ASL signer. In this system, 30 infrared-emitting markers are attached (via Velcro to a spandex suit) at key locations on the body, and sensors around the room triangulate the position of each marker at a rate of 30Hz. The gloves record 22 joint-angle measurements for each hand using resistive band-sensors and transmit the data to a host computer via Bluetooth. The motion-capture room initially contained many curtains and equipment; to make the room more comfortable/accessible for a deaf participant, the dim lighting was increased and various curtains and equipment moved to create a line of sight between all of the people who needed to be present to collect the motion-capture data. Once in the suit, the signer reported that the suit and gloves were comfortable and did not feel as if they were impeding her movement. The gloves are made of spandex with thin sensor strips on the back of the joints; the fingers had sufficient freedom of movement that the ASL interpreters present during the study could still understand the signer's fingerspelling and signing through the gloves.

To ensure that we were recording fluent ASL, it was important that our contributor was a native signer – someone who learned ASL in early childhood through interaction with deaf family members or experiences at a residential school for the deaf. We also needed to minimize the English influence in the environment so that the signer would not be prone to code-switch to English-like signing [13]. To elicit the ASL classifier predicate sentences, the signer was shown the "correct" animated visualization for each of the ten sentences – showing a set of objects moving in a 3D scene. The signer was asked to use ASL to describe the arrangement of the objects in the scene as she might to another ASL signer that she was having a conversation with. Asking a signer to imagine conversing with another signer can help prevent code-switching to English-like signing [13].

Data from a motion-capture session usually requires manual editing after collection. If the placement of markers on the body suit is not perfect, if the real human's body proportions differ from the model, or if there is incorrect location-triangulation when markers are occluded, then the software may not correctly calculate the skeleton joint angles. Unfortunately, the data we collected contained sufficient errors that despite post-processing clean-up, the resulting animations contained enough movement inaccuracies that native ASL signers who viewed them felt they were actually *less* understandable than our system's animations – they were not an upper-baseline. In future work, we will explore alternative upper-baselines to compare our system's animations to: animation from alternative motion-capture techniques (that require less post-collection animation "clean-up" work), hand-coded animations based on a human signer's performance, or simply a video of a human signer performing ASL sentences.

7. Conclusions

The user-based evaluation of our system differs from evaluations of "broad-coverage" natural language generation systems, ones that are able to accommodate a wide variety of inputs and are closer to being used by actual users. In an evaluation of a broad-coverage NLG system, we would obtain performance statistics for the system as it carries out a linguistic task on a large corpus or

"test set." This paper has described an evaluation of a prototype system; so, we were not measuring the linguistic coverage of the system but rather its functionality. Did signers agree that the animation output: (1) is actually a grammatically-correct and understandable classifier predicate and (2) conveys the information about the movement of objects in the 3D scene being described? We expected to find animation details that could be improved in future work; however, since there are currently no other systems capable of generating ASL classifier predicate animations, any system receiving an answer of "yes" to questions (1) and (2) above is an improvement to the state of the art.

This evaluation also served as a kind of pilot study to help determine how to best evaluate sign language generation systems. One interesting outcome was the low correlation we observed between the understandability score that a participant gave to an animation and her actual success at selecting the proper visualization for that animation – thus calling into question whether "perceived understandability" scores represent "actual understanding." Future evaluation studies of ASL systems should therefore continue to include a comprehension task (whether it be a matching task or some other kind of activity). The use of perceived understandability scores is no substitute for this data.

Subjects were comfortable critiquing ASL animations, and most suggested specific (and often subtle) elements of the animation to improve. This feedback suggested new modifications we can make to the system (and then evaluate again in future studies). Because the comments subjects made during the study (in the non-numeric portion of the evaluation) were of such high quality, future studies should continue to elicit such feedback.

Informally, we observed that the subjects in our study tended to have fairly strong English skills – this was not part of our screening criteria, nor was it a factor controlled for in this study. In future work, we may prefer subjects to better reflect the future user base of ASL generation software – users with low English literacy. Recruiting such a specialized group of users would be even more difficult, and the benefit must be weighed against the additional resources required. (In any case, we would expect that our system would do even better relative to the Signed English lower-baseline when judged by subjects with lower English literacy, since such users would likely derive even less value from the Signed English transliterations of the English sentences.)

A final contribution of this work is that we believe our evaluation method (with its comprehension task and use of a baseline) may be useful for evaluating other animations, such as gesture generation for embodied conversational agents.

8. ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (Award #0520798 "SGER: Generating Animations of American Sign Language Classifier Predicates," Universal Access Program, 2005.) We are grateful for software donated by Siemens UGS Tecnomatix and Autodesk. We also thank Mitch Marcus, Martha Palmer, and Norman Badler for their guidance and support.

9. REFERENCES

[1] N. Badler, J. Allbeck, S.J. Lee, R. Rabbitz, T. Broderick, & K. Mulkern. 2005. New behavioral paradigms for virtual human models. SAE Digital Human Modeling.

[2] N. Badler, R. Bindiganavale, J. Allbeck, W. Schuler, L. Zhao, S. Lee, H. Shin, & M. Palmer. 2000. Parameterized action representation & natural language instructions for dynamic behavior modification of embodied agents. AAAI Spring Symposium.

[3] S. Bangalore, O. Rambow, S. Whittaker. 2000. Evaluation Metrics for Generation. Proceedings of the First International Conference on Natural Language Generation.

[4] R. Coyne and R. Sproat. 2001. WordsEye: an automatic text-to-scene conversion system. ACM SIGGRAPH.

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

[6] M. Huenerfauth. 2006. Representing Coordination and Non-Coordination in American Sign Language Animations. *Behaviour & Information Technology*, 25:4.

[7] M. Huenerfauth. 2006. Generating American Sign Language Classifier Predicates for English-to-ASL Machine Translation. Dissertation, U. Pennsylvania.

[8] M. Huenerfauth. In Press. Representing ASL Classifier Predicates Using Spatially Parameterized Planning Templates. In M.T. Banich and D. Caccamise (Eds.), *Generalization of Knowledge.* Mahwah, NJ: Erlbaum.

[9] M. Huenerfauth, L. Zhao, E. Gu, & J. Allbeck. 2007. Design and Evaluation of an American Sign Language Generator. 45th Annual Meeting of the Association for Computational Linguistics. Workshop on Embodied Language Processing. Prague, Check Republic.

[10] S. Liddell. 2003. Grammar, Gesture, and Meaning in American Sign Language. UK: Cambridge U. Press.

[11] R.E. Mitchell, T.A. Young, B. Bachleda, & M.A. Karchmer. 2006. How Many People Use ASL in the United States? Why Estimates Need Updating. *Sign Language Studies*, Vol. 6, Number 3.

[12] S. Morrissey & A. Way. 2006. Lost in Translation: The Problems of Using Mainstream MT Evaluation Metrics for Sign Language Translation. *5th SALTMIL Workshop on Minority Languages, LREC-2006*.

[13] C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan, & R.G. Lee. 2000. *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure.* Cambridge, MA: The MIT Press.

[14] C. Padden & T. Humphries. 2005. Inside Deaf Culture. Cambridge, MA: Harvard University Press.

[15] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation, 40th Annual Meeting of Assoc. for Computational Linguistics.

[16] S. Pasquariello & C. Pelachaud. 2001. Greta: A simple facial animation engine. In 6th Online World Conference on Soft Computing in Industrial Applications.

[17] L. Zhao, K. Kipper, W. Schuler, C. Vogler, N.I. Badler, & M. Palmer. 2000. Machine Translation System from English to American Sign Language. Assoc. for MT in the Americas.

[18] É . Sáfár & I. Marshall. 2001. The architecture of an English-text-to-Sign-Languages translation system. Recent Advances in Natural Language Processing.