Collecting and Evaluating the CUNY ASL Corpus for Research on American Sign Language Animation

Pengfei Lu
Doctoral Program in Computer Science
The Graduate Center, CUNY
City University of New York
365 Fifth Ave, New York, NY 10016
pengfei.lu@qc.cuny.edu

Matt Huenerfauth
Department of Computer Science
Queens College and Graduate Center
City University of New York (CUNY)
65-30 Kissena Blvd, Flushing, NY 11367
matt@cs.qc.cuny.edu

Abstract

While there is great potential for sign language animation generation software to improve the accessibility of information for deaf individuals with low written-language literacy, the understandability of current sign language animation systems is limited. Data-driven methodologies using annotated sign language corpora encoding detailed human movement have enabled some researchers to address several key linguistic challenges in ASL generation. This article motivates and describes our current research on collecting a motion-capture corpus of American Sign Language (ASL). As an evaluation of our motion-capture configuration, calibration, and recording protocol, we have conducted several rounds of evaluation studies with native ASL signers, and we have made use of our collected data to synthesize novel animations of ASL, which have also been evaluated in experimental studies with native signers.

1 Introduction

Sign languages are natural languages conveyed by movements of the hands, arms, torso, head, face, and eyes. There are different sign languages used around the world, and sign languages in different countries are not typically mutually intelligible. Further, the sign language in a country is often distinct from the spoken/written language in use in that country; a sign language is not merely a

manual representation of the word order of the local spoken language. Typically, a sign language will emerge and develop naturally through its use among a community of signers in some region.

In the U.S., American Sign Language (ASL) is the primary means of communication for about one-half million people (Mitchell et al., 2006). ASL has a distinct word-order, syntax, and lexicon from English; it is not a representation of English using the hands. Although written-language reading is an important part of the curriculum for deaf students, lack of auditory exposure to English during the language-acquisition years of childhood leads to lower literacy for many adults. In fact, the majority of deaf high school graduates in the U.S. have only a fourth-grade (age 10) English reading level (Traxler, 2000). While these statistics have focused on ASL and the U.S., similar trends in sign language usage, literacy, and deafness are present in many countries.

Most technology used by people who are deaf does not address this literacy issue; many deaf people find it difficult to read the English text on a computer screen or on a television with text captioning. Software to present information in the form of animations of ASL could make information and services more accessible to deaf users, by displaying an animated character performing ASL, rather than English text. Research on synthesizing ASL animations will therefore benefit people who are deaf or hard-of-hearing with low literacy of English reading and use ASL as their preferred language – this is the target user group of ASL animation technologies.

1.1 Research Goals and Organization of this Article

Our research goal is to create technologies that make it easier to synthesize computer animation of ASL, to increase the accessibility of information available on websites, presented by computer software, or through future captioning technologies. This article primarily focuses on our efforts to construct and evaluate a corpus of ASL – containing 3D movement data and linguistic annotations – to support our research on ASL animation synthesis technologies.

Section 1.2 explains why animation technologies for ASL are necessary, with advantages over videos of human signers in various scenarios. Section 1.3 briefly presents some linguistic aspects of ASL that make it challenging to create synthesis technology. Section 1.4 surveys prior work on ASL animation technologies and explains how the majority of research in the field does not make use of data recorded from human signers (i.e., they are not "data-driven," unlike most modern research in the field of Natural Language Processing). Section 1.5 explains how there is a critical lack of corpora resources for ASL – in particular, corpora containing 3D movement data recorded from humans – to support research on ASL animations.

We have therefore begun a multi-year project to collect and annotate a motion-capture corpus of ASL. Prior ASL corpus-building projects are outlined in section 1.5, and section 2 describes our data collection and annotation methodology. We intend to use this corpus of human motion data and linguistic annotation to train statistical models for use in ASL animation synthesis technologies; we also anticipate that this corpus will be of interest to ASL linguists and other researchers. Our initial research focus is to model where signers tend to place spatial reference points around them in space and to discover patterns in the motion paths of indicating verbs (linguistic details in section 4). There are several important aspects of our research:

- We use a novel combination of hand, body, head, and eye motion-tracking technologies and simultaneous video recordings (details in section 2.1 and 2.2).
- We collect multi-sentence single-signer ASL discourse through various elicitation techniques designed to elicit desired linguistic phenomena (details in section 2.3).
- We annotate novel linguistic information relevant to the use of space around the signer's body to represent entities under discussion (details in section 2.4).
- We have conducted evaluations to determine whether our motion-capture equipment configuration is sufficiently sensitive and well-calibrated – such that we can record key movement details of the human performance (details in section 3).
- We involve ASL signers in the research: as evaluators of our generation software, as research assistants conducting evaluation studies, and as corpus annotators. This involvement of native signers allows us to conduct rigorous evaluation studies and add linguistically accurate annotations to our corpus. (Details in section 2.4 and 3.)

- We train mathematical models of sign language based on data extracted from our corpus. For example, we have trained models of the movement path of ASL indicating verbs based on examples of verb performances for different arrangements of the verb's subject and object in the signing space; these models produce high-quality animations (Huenerfauth and Lu, 2010b; Lu and Huenerfauth, 2011a). (Details in section 4.)

1.2 Applications of ASL Synthesis Research

While writing systems for ASL have been proposed (e.g., SignWriting, 2012), none is widely used in the Deaf community; so, websites or software must display videos or animations of ASL. A limitation of video is that if the information content on a website is frequently updated, the video would need to be frequently and largely re-recorded for each modification. For many websites or applications, there would be a benefit from enabling easier editing of an ASL message; a human author could prepare a "script" of the ASL message, which could be synthesized automatically into animation or video. While is possible to splice together videos of human signers to produce novel messages, it is difficult to produce smooth transitions between signs, subtle motion variations in performances, or proper combinations of facial expressions with signs. Animation-based synthesis technology enables more control, blending, and modulation than spliced videos.

Animation-based synthesis also supports privacy, collaboration, and customization. Because the face is used to indicate important information in ASL, videos of sign language must include the face of the human signer – thereby revealing the identity of the human who is producing the ASL video. Instead, a virtual human character in an animation would not reveal the human signer's identity. For wiki-style applications in which multiple authors are collaborating on information content, ASL videos could be distracting: the person performing each sentence may differ. A virtual human in an animation (that is collaboratively produced) would be more uniform. Animations allow ASL to be viewed at different angles, at different speeds, or performed by different virtual humans – depending on the preferences of the user. It may be necessary to adjust these vis-

ual properties to accommodate situational factors; for example, variations in the screen size, ambient lighting conditions, or other factors could affect the visual clarity of the animation.

1.3 Use of Spatial Reference in ASL: Challenge for Animation

Producing a fluent and understandable ASL animation requires researchers to address several challenging linguistic issues, as discussed in (Lu and Huenerfauth, 2010). In this article, we focus on one aspects of sign language that makes it challenging for NLP research: the way in which signers arrange invisible placeholders in the space around their body to represent objects or persons under discussion (Cormier, 2002; Janis, 1992; Liddell, 1990; 1995; 2003; Meier, 1990; Neidle et al., 2000; Padden, 1988).[1] An ASL generator must select which entities should be assigned locations (and where). For example, a signer may discuss someone named ‘‘Mary.’’ After mentioning Mary the first time, the signer may point to a location in space where a spatial reference point is created that represents her. On future occasions in which the signer wants to refer to Mary, the signer will simply point to this location. Sometimes the subject/object is not mentioned in the sentence, and the use of eye-gaze or head-tilt aimed at these locations is the only way in which the signer conveys the identity of the subject/object (Neidle et al., 2000). On other occasions, a signer may set up a spatial reference point on the left side of the signing space and one on the right; later, when discussing these entities contrastively, a signer may orient the torso at one side or another.

Unfortunately, modern sign language scripting, generation, or machine translation (MT) software does not *automatically* handle these spatial aspects of ASL discussed above. Humans authoring ASL animations using state-of-the-art scripting systems must manually add pointing signs to the performance that aim at various locations around the virtual human signer. In

[1] ASL linguists debate whether the locations in space occupy 3D locations or 1-dimensional location on an arc around the body (Liddle, 2003; McBurney, 2002; Meier, 1990); we do not seek to engage in this debate. In our corpus (section 2), we collect the motion data of how the signer’s body moves, and we mark the moments in time when a pointing sign occurs to refer to these entities in space. Researchers may analyze our data either in terms of arc-locations or 3D-locations where spatial reference points may be -- depending on how they interpret where the finger is pointing.

(Huenerfauth and Lu, 2012), we described experimental studies that measured how much the lack of spatial reference points, pointing pronouns, or contrastive role shift affects the usability of current ASL animation technologies. Native ASL signers evaluated ASL animations that did and did not include these phenomena. There were significant benefits to ASL animations in which the virtual character associates entities under discussion with spatial reference points in the signing space, uses pointing pronoun signs, and uses contrastive role shift. If future sign language animation research included better modeling of how the space around a signer is used for spatial references, we would expect better comprehension and user-satisfaction scores.

1.4 Prior Work on ASL Animation

In the past decade, there have been many initial attempts at generating sign language computer animations. While this article focuses on ASL, in fact, various sign languages have been the focus of computer animation and accessibility research, including Arabic Sign Language (Tolba et al., 1998), Australian Sign Language (Vamplew et al., 1998), British Sign Language (Cox et al., 2002), Chinese Sign Language (Wang et al., 2002), Greek Sign Language (Fotinea et al., 2008), Japanese Sign Language (Shionome et al., 2005), Korean Sign Language (Kim et al., 1996; Lee et al., 1997), and Sign Language of the Netherlands (Prillwitz et al., 1989).

Prior work on sign language computer animations can be divided into two areas: scripting and generation/translation. Scripting systems allow someone who knows sign language to "word process" an animation by assembling a sequence of signs from a lexicon and adding facial expressions. For example, the eSIGN project (2012) created tools for content developers to build sign databases and assemble scripts of signing for web pages (Kennaway et al., 2007). Sign Smith Studio (Vcom3D, 2012) is a commercial tool for scripting ASL. Both of these scripting systems enable a user to select signs from a pre-built dictionary and to arrange them on a timeline to construct complete sentences; often the user can also add facial expressions or other movements to produce

more fluent animations. While these systems do require significant work from an ASL-knowledgeable user to produce animations, when compared to the alternative method of producing sign language animations (by carefully posing all of the joint angles of a human figure using some general-purpose 3D animation software), it is clear that the scripting systems make the process of producing sign language animations more efficient and easier.

Sign language generation research focuses on techniques for further automating the planning or creation of sign language sentences; a common use of generation technologies is within a system for translating from a written-language sentence into a sign language animation. Most prior sign language generation or MT projects have been short-lived, producing few example outputs (Zhao et al., 2000; Veale et al., 1998). The English to British Sign Language system (Marshall and Safar, 2001; Safar and Marshall, 2002) for the European Union's VISICAST project used hand-built translation transfer rules from English to British Sign Language; the system allowed for human intervention during MT to fix errors before synthesizing an animation output. The later eSIGN Project (Elliott et al., 2006; eSIGN, 2012; Kennaway et al., 2007) included a human-assisted machine translation system to convert a written language text into a script for a virtual sign language character to perform. The human can intervene to modify the translation: e.g., finding an appropriate gloss for a sign with several possible meanings. The human user may also add facial expressions, body movements and pauses, or the user may alter the position or location of a sign.

While data-driven machine-learning approaches have been common for NLP researchers working on written/spoken languages in the past two decades, recently more sign language animation synthesis researchers are using data from human signers in the development of their systems. Bungeroth et al. (2006) transcribed German Sign Language sentences from sign language interpretation of television weather reports to build a small corpus (consisting of strings of transcription of the signs), in support of statistical sign language translation research. Morrissey and Way (2005)

examined how statistical example-based MT techniques could be used to translate sign language by using a small corpus (also consisting of strings of sign sequences) of Dutch Sign Language. Segouat and Braffort (2009) built a small corpus of French Sign Language (individual signs and multi-sign utterances) by asking annotators to build animations based on video recordings; they used this corpus to study coarticulation movements between signs. Also, studying French Sign Language animations, Gibet et al. (2011) captured full-body human motion-capture data, which they used as source material to splice together novel messages. All of these researchers had to construct small corpora for their work (some merely consisting of transcripts of sign sequences and some consisting of human movement data). A limiting factor in the further adoption of data-driven research methodologies for sign language animation is the lack of sufficiently detailed corpora of ASL (specifically, annotated corpora that include recordings of movement from human signers).

1.5 Prior Sign Language Corpora Resources

Sign language corpora are in short supply and are time-consuming to construct because, without a writing system in common use, it is not possible to harvest some naturally arising source of ASL "text." To create a corpus, it is necessary to record the performance of a signer (through video or motion-capture). Humans must then transcribe and annotate this data by adding time-stamped linguistic details. For ASL (Neidle et al., 2000) and European sign languages (Bungeroth et al., 2006; Crasborn et al., 2004, 2006; Efthimiou and Fotinea, 2007), signers have been videotaped and experts marked time spans when events occur – e.g. the right hand performs the sign "CAR" during time index 0-50 milliseconds, and the eyebrows are raised during time index 20-50 milliseconds. Such annotation is time-consuming to add; the largest ASL corpus has a few thousand sentences.

Even if large video-based corpora of sign language were created and annotated, they may not actually be sufficient for animation research. In order to learn how to control the movements of an animated virtual human based on a corpus, we need precise hand locations and joint angles of

the human signer's body throughout the performance. Asking humans to write down 3D angles and coordinates during an annotation process is time-consuming and inexact; instead, some researchers have used computer vision techniques to model the signers' movements – see survey in (Loeding et al., 2004). Unfortunately, the complex shape of the hands and face, the rapid speed of signing, and frequent occlusion of parts of the body during signing limit the accuracy of vision-based recognition; it is not yet a reliable way to build a 3D model of a signer for a corpus. Motion-capture technology (discussed in section 2.1) is required for this level of detail. While researchers have constructed some collections of motion-capture data of various human movements, e.g. (CMU Graphics Lab Motion Capture Database, 2012), there has not yet existed a motion-capture ASL corpus; a corpus consisting of sign language movements, with linguistic annotation, is needed for ASL animation research.

2 Collecting an ASL Corpus to Support Animation Research

Section 1.3 discussed how it is important for human signers (and for animated characters) to associate entities under discussion with locations in the signing space. Spatial modification of signs is an inherent part of the grammar of ASL, and they are present in fluent and understandable ASL. If we had a sufficiently large sample of ASL discourse, with linguistic annotation of when the signers established and referred to locations in the signing space in this manner, we could learn patterns in where and when this spatial reference occurs. Of course, we would want our corpus to contain motion-capture data recordings of where exactly the human signer was pointing in the signing space. Similarly, researchers studying other linguistic issues, e.g., coarticulation, as in the case of Segouat and Braffort (2009), would benefit from a corpus containing motion-capture recordings of human signers. Thus, we are collecting the first motion-capture corpus of ASL, and we are releasing the first portion of this corpus to the research community. This section describes the equipment, recording, annotation, and initial release of this corpus.

2.1 Our Motion-Capture Configuration

Assuming an ASL signer's pelvis bone is stationary in 3D space (the humans we record are sitting on a stool), we want to record movement data for the upper body. We are interested in the shapes of each hand; the 3D location of the hands; the 3D orientation of the palms; joint angles for the wrists, elbows, shoulders, clavicle, neck, and waist; and a vector representing the eye-gaze aim. We are using a customized configuration of several commercial motion-capture devices (as shown in Figure 1(a), worn by a human signer): a pair of motion-capture gloves, an eye tracker helmet, a head tracker system, and a motion-capture body suit.

For high quality sign language data, it is important to accurately record the subtle movements of the fingers. While there are various technologies available for digitizing the movements of a human's fingers, many of these techniques require line-of-sight between a recording camera and the fingers. The rapid movements of sign language (and the frequent occlusions caused by one hand blocking another) would make such technologies inappropriate. For our project, we ask signers to wear a pair of Immersion CyberGloves® (Figure 2(c)). Each of these flexible and lightweight spandex gloves has 22 flexible sensor strips sewn into it that record finger joint angles so that we can record the signer's handshapes. The gloves still permit comfortable movement; in fact, humans viewing someone in the gloves are able to discern ASL fingerspelling and signing.

The signer being recorded also wears an Applied Science Labs H6 eye-tracker (Figure 2(d)), a lightweight head-mounted eye-tracker. The camera on the headband aims downward, and a small clear plastic panel in front of the cheek reflects the image of the participant's eye. In order to calibrate the eye-tracker system, we place a clear plastic panel on an easel in front of the signer, with several numbered dots (with known placements) on the panel. We ask the participant to look at each dot in sequence during the calibration process. Data from an Intersense IS-900 system (Figure 2(a) and 2(b)) is used to compensate for head movement when calculating eye-gaze direction.

This acoustical/intertial motion-capture system uses a ceiling-mounted ultrasonic speaker array (Figure 2(a)) and a set of directional microphones on a small sensor (Figure 2(b)) to record the location and orientation of the signer's head. A sensor sits atop the helmet, as shown in Figure 1(a).

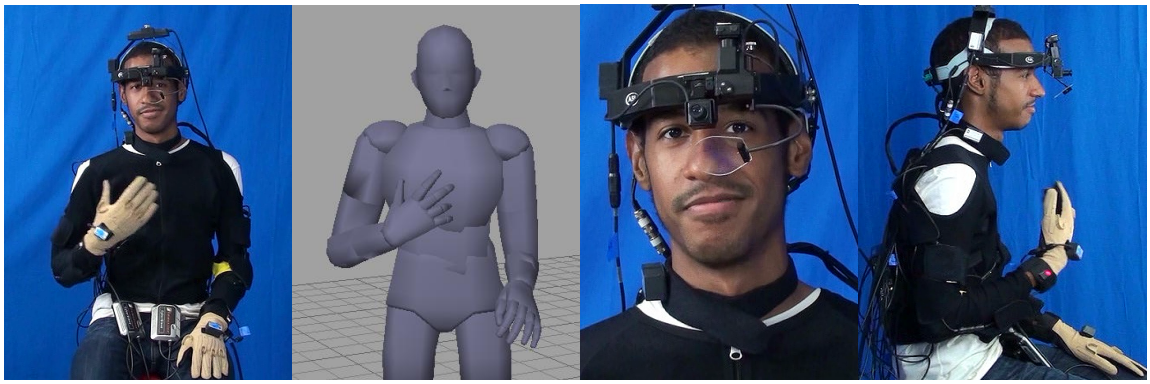

Figure 1: (a) Signer wearing motion-capture equipment (shown in evaluation study in section 3.3), (b) Animation produced from motion-capture data (shown in evaluation studies in sections 3.2 and 3.3), (c) The face view of the signer, (d) The right-side view of the signer.

Figure 2: (a) Intersense IS-900 ceiling-mounted ultrasonic speaker array, (b) Intersense IS-900 ceiling-mounted ultrasonic sensor, (c) Animazoo IGS-190 sensor on the top of one Immersion CyberGlove, (d) Applied Science Labs H6 eye-tracker.

Finally, the signer also wears an Animazoo IGS-190 bodysuit (Figure 1(a)); this system consists of a spandex suit covered with soft Velcro to which small sensors attach. A sensor placed on each segment of the human's body records inertial and magnetic information. A sensor is also placed atop the Immersion cyberglove shown in Figure 2(c). Participants wearing the suit stand facing north with their arms down at their sides at the beginning of the recording session; given this

known starting pose and direction, the system calculates joint angles for the wrists, elbows, shoulders, clavicle, neck, and waist. We do not record leg/foot information in our corpus. Prior to recording data, we photograph each participant standing in a cube-shaped rig of known size; next, we draw a human skeleton model atop this photograph and label the corners of the cube-shaped rig in the photo. This process allows us to identify bone lengths of the human participant, which are needed for the IGS-190 system to accurately calculate joint angles from the sensor data.

2.2 Video Recording and Data Synchronization

Our motion-capture recording sessions are videotaped to facilitate later linguistic analysis and annotation. Videotaping the session also facilitates the "clean up" of the motion-capture data in post-processing, during which algorithms are applied to adjust synchronization of different sensors or remove "jitter" or other noise artifacts from the recording. Three digital high-speed video cameras film front view, facial close-up, and side views of the signer (Figure 3); a similar camera placement has been used in video-based ASL-corpora-building projects (Neidle et al., 2000). The views are similar to those shown in Figure 1, but the camera image is wider than the photos in that Figure. The facial close-up view is useful when later identifying specific non-manual facial expressions during ASL performances. To facilitate synchronizing the three video files during post-processing, a strobe light is flashed once at the start of the recording session.

To facilitate synchronization of the videos and the motion capture data from the Animazoo IGS-190 body suit and Intersense IS-900 head tracker, we ask the signer in the motion-capture equipment to perform a very quick head movement (turn the head to one side) immediately after the strobe light is flashed at the start of the recording (described above), so that we can identify the moment easily when the signer's head turns in: the three videos, the data from the body suit, and the data from head tracker; this allows us to synchronize all of our data streams.

To facilitate synchronization of the videos and the motion capture data from the Applied Science Labs H6 eye-tracker, we ask the signer in the data collection session to close their eyes for at least 10 seconds, after he/she opens the eyes, the strobe light flashes (described above), and he/she performs the quick head movement (described above) for the synchronization of the videos and the body suit data. We can identify the moment in time when the eyes open in the eye-tracker data stream and in the three video recordings – thereby synchronizing these data streams.

A "blue screen" curtain hangs on the back and side walls of the motion-capture studio (Figure 3). While the background of the video recording is not particularly important for our research, future computer-vision researchers who may wish to use this corpus might benefit from having a solid color background for "chroma key" analysis. Photographic studio lighting with spectra compatible with the eye-tracking system is used to support high-quality video recording.

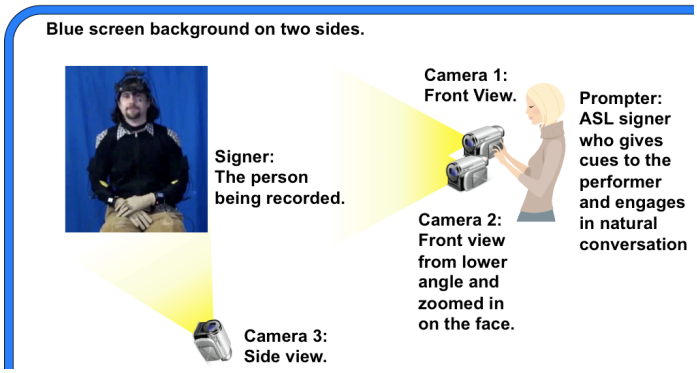

Figure 3: Diagram of an overhead view of our motion-capture studio setup.

2.3 Eliciting the ASL Corpus

During data collection, a native ASL signer (called the "prompter") sits directly behind the front-view camera to engage the participant wearing the suit (the "performer") in natural conversation (Figure 3). While the corpus we are collecting consists of unscripted *single-signer* discourse, prior ASL corpora projects have identified the importance of surrounding signers with an ASL-centric

environment during data collection. Some ASL linguists (Neidle et al., 2000) have warned other researchers about the dangers of permitting English influences in the experimental/recording environment when you want to collect video corpora of sign language. Such English influences can affect how the signer performs. English influence in the studio must be minimized to prevent signers from inadvertently code-switching to an English-like form of signing. Thus, it is important that a native signer acts as the prompter, who conversationally communicates with the deaf participants to elicit the verb, sentence, or story being recorded for the corpus.

Advertisements posted on Deaf community websites in New York City asked whether potential participants had grown up using ASL at home or whether they attended an ASL-based school as a young child. Of the 8 participants we have recorded for the corpus: 7 grew up with parents who used ASL at home (the 8th is deaf with hearing parents and started learning ASL as an infant, age 1.5), 2 were married to someone deaf/Deaf, 7 used ASL as the primary language in their home, 8 used ASL at work, and 8 had attended a college where instruction was primarily in ASL. The signers were 8 men of ages 21-34 (mean age 27.9).

We prefer to collect multi-sentence passages with a varied number of entities under discussion; we also prefer to record passages that avoid complex spatial descriptions, which are not the focus of our research. In (Huenerfauth and Lu, 2010a), we discussed details of: the genre of discourse we record, our target linguistic phenomena to capture (spatial reference points and inflected verbs), the types of linguistic annotation added to the corpus, and the effectiveness of different "prompts" used to elicit the desired type of spontaneous discourse. As described in (Huenerfauth and Lu, 2010a; Lu and Huenerfauth, 2011a), we have experimented with different prompting strategies over the years to elicit ASL signing in which signers establish different numbers of pronominal reference points in space, perform longer monologues, and other linguistic considerations. Our corpus contains passages in which signers discuss their personal histories, their recollection of

news stories or movies, their explanation of encyclopedic information, their opinion about hypothetical scenarios, their comparison of people or things, their description of a page of photos, and other categories described in (Huenerfauth and Lu, 2010a; Lu and Huenerfauth, 2011a). Table 1 lists some of the prompts we have used. While some prompts use English text, the English influence was minimized by using a delay of 30 minutes between when texts were read and when ASL was recorded. Further, the participant was asked to discuss concepts in their own words, to a native ASL signer behind the camera, with whom they had been conversing in ASL about unrelated topics during the 30 minutes. Table 2 provides statistics about the passages we have collected.

Type of Prompt	Description of This Prompting Strategy
Personal Introduction	Please introduce yourself and discuss your background, your hobbies, your family and friends, your education, your employment, etc.
Compare (people)	Compare two people you know: your parents, some friends, family members, etc.
Compare (not people)	Compare two things: e.g. Mac vs. PC, Democrats vs. Republicans, high school vs. college, Gallaudet University vs. NTID, travelling by plane vs. by car, etc.
Photo Page	Look at this page of photos (of people who are in the news recently) and then explain what is going on with them.
Opinion / Explain Topic	Please explain your opinion on this topic (given) or explain the concept as if you were teaching it to someone.
Personal Narrative	Please tell a story about an experience that you had personally.
News Story	Recount a brief news article after you read it.
Children's Book	Explain a story as you remember after you read a short children's book.
Repeat Conversation	Explain what you saw after watching a 3-minute video of an ASL conversation or of a captioned English conversation.
Wikipedia Article	Recount a 300-word Wikipedia article after you read it, e.g. "Hope Diamond"
Hypothetical Scenario	What would you do if: you were raising a deaf child? You could have dinner with two famous or historical figures?
Recount Movie/Book	Describe your favorite movie or book.

Table 1: Examples of some prompts used to elicit the corpus.

Recording session No.	Signer	Glosses in total	Length of video in total (seconds)	Average number of glosses per passage	Average video length of the passages
#1	A	730	364	56.2	28.0
#2	B	434	236	48.2	26.2
#3	C	1291	571	117.4	51.9
#4	D	1512	665	116.3	51.2
#5	E	735	310	91.9	38.8
#6	F	4516	2048	180.6	81.9
#7	G	3467	1786	102.0	52.6
#8	H	983	633	140.4	90.4
#9	H	5125	3474	119.2	80.8
#10	A	2634	1425	67.5	36.5
#11	B	2165	1178	54.1	29.4

Table 2: The properties of our corpus.

2.4 Annotating the ASL Corpus

A team of native ASL signers at our lab (including linguistics undergraduate students and local deaf high school students) use SignStream™ (Neidle et al., 2000) to annotate our corpus. After one annotator finishes a file, it is crosschecked by at least two others, and disagreements are discussed in an adjudication meeting. The annotations that we have begun to release (section 2.5) include English translations of the entire passage and sign glosses (with time alignment to the video). Figure 4 shows a transcript and its English translation of a passage collected using the "Compare (not people)" prompt. Table 3 explains the notation in the transcript, more details in (Neidle et al., 2012). Figure 5 shows how parallel timeline tracks are available for the main gloss and a non-dominant hand gloss, a row which is used when a signer uses his/her non-dominant hand to perform a sign or when the signer performs different signs simultaneously on the two hands.

fs-MAC fs-PC BOTH COMPUTER IX-1-p:1,2 BUT DO-DO fs-MAC UMM NONE/NOTHING fs-VIRUS fs-PC NONE/NOTHING HAVE fs-VIRUS POP-UP PROBLEM IX-1-s:2 NOTHING IX-1-s:2 fs-MAC EASY TO/UNTIL USE fs-PC CAN #BE COMPLICATED UMM fs-PC IX-1-s:1 CAN BUSINESS GOOD FOR BUSINESS A-LOT fs-OF fs-WORD PAPER CL"type page" fs-MAC IX-1-s:2 NONE/NOTHING NEGATIVE PLUS IX-1-s:S FAVORITE/PREFER fs-MAC IX-1-s:S LOVE fs-MAC EASY TO/UNTIL USE TO/UNTIL LESS PROBLEM FIRST-IN-LIST-2 VERY-FAST PICTURE_2 TWO CL"UPLOAD" PERFECT IX-1-s:S HAPPY WITH THAT COMPUTER	Both Mac and PC are computers; they are different. Mac doesn't have viruses while PC does have virus problems. Mac is easy to use, while PC is complicated. PC is good for business use because of a lot of word papers can be typed. Mac doesn't have any negative or positive. I love to use Mac because it is easy to use, fewer problems. It's very fast with uploading pictures. I am happy with that computer.

Figure 4: Transcript of a passage we collected using the "Compare (not people)" prompt.

Type of notation	Explanation of this notation
fs-*X*	Fingerspelled word
IX-1-p:1,2	Index sign (pointing), handshape-#1, plural, spatial reference points #1 and #2.
IX-1-s:1	Index sign (pointing), handshape-#1, singular, spatial reference point #1.
IX-1-s:2	Index sign (pointing), handshape-#1, singular, spatial reference point #2.
CL"*X*"	Classifier Predicate, meaning gloss is provided inside the quotation marks.

Table 3: The notations in the transcript in Figure 4 and 5.

Main Gloss	fs-MAC		BOTH	COMPUTER	IX-1-p:1,2	BUT	DO-DO		…
Non-Dom. Hand Gloss		fs-PC							
SRP#1 Establishment					PC				…
SRP#2 Establishment					MAC				…
SRP#1 References					*r*				…
SRP#2 References					*r*				…

Main Gloss	IX-1-s:2	fs-MAC	EASY	TO/UNTIL	USE	fs-PC	CAN	#BE	COMPLICATED	…
Non-Dom. Hand Gloss										…
SRP#1 Establishment										…
SRP#2 Establishment										…
SRP#1 References										…
SRP#2 References	*r*									…

Main Gloss	fs-PC	IX-1-s:1	CAN	BUSINESS	GOOD	FOR	BUSINESS	A-LOT	…
Non-Dom. Hand Gloss									…
SRP#1 Establishment									…
SRP#2 Establishment									…
SRP#1 References		*r*							…
SRP#2 References									…

Figure 5: Excerpts from the annotation timeline for the passage shown in Figure 4.

Our annotations include information about when spatial reference points (SRPs) are established during a passage, which discourse entity is associated with each SRP, when referring expressions later refer to an SRP, and when any verbs are spatially inflected to indicate an SRP. These SRP establishments and references are recorded on parallel timeline tracks to the glosses and other linguistic annotations. Figure 5 shows an annotation timeline for this passage; the first and second rows ("Main Gloss" and "Non-Dom. Hand Gloss") list the sequence of glosses. The signer establishes two spatial reference points: The first time that the signer points to two locations in 3D space around his body (glossed as "IX-1-p:1,2"), he establishes an SRP at one location to represent "PC", and another SRP at a second location, to represent "Mac." Those two SRPs are referred to again later in the passage when the signer performs "IX-1-s:1" and "IX-1-s:2" signs (see Table 3).

In Figure 5, the third row ("SRP#1 Establishment) and fourth row ("SRP#2 Establishment") indicate when a new spatial reference point has been created. (The numbers #1 and #2 are arbitrary identifiers, customarily we use #1 for the first SRP established in a passage and #2 for the second, etc.) When an SRP is established, then an annotation is added to the appropriate row with start- and end-times that align to the sign or phrase that established the existence of this SRP. The

label of the annotation is meant to be a brief gloss of the entity referenced by this SRP. Some of the information on the gloss rows corresponds to the numbers ("1" and "2") for these SRPs; specifically, the integer after the colon at the end of the gloss "IX-1-s:1" indicates that the pointing sign is referring to SRP #1. A pointing sign directed at SRP #2 appears as "IX-1-s:2". By assigning each SRP an index number in this manner, the gloss of each pronominal reference to an SRP is marked with this index number (following a colon at the end of a gloss in the transcription).[2]

Rows 5 and 6 of Figure 5 indicate when signs later in the passage refer again to the SRPs. A separate "SRP Reference" row is created for each SRP that is established; while this example shows a passage with two SRPs, there could be more rows if needed. Whenever the entity is referred to during the passage (including during the initial establishment of the SRP), the "SRP Reference" row receives an annotation with a label "r" (for "reference"). Figure 6 shows the average number of SRP establishments (i.e., the number of unique SRPs established per passage) and the average number of SRP references per passage for each recording session. Some signers were more loquacious, producing stories of greater length, after a prompt, leading to a higher number of SRPs.

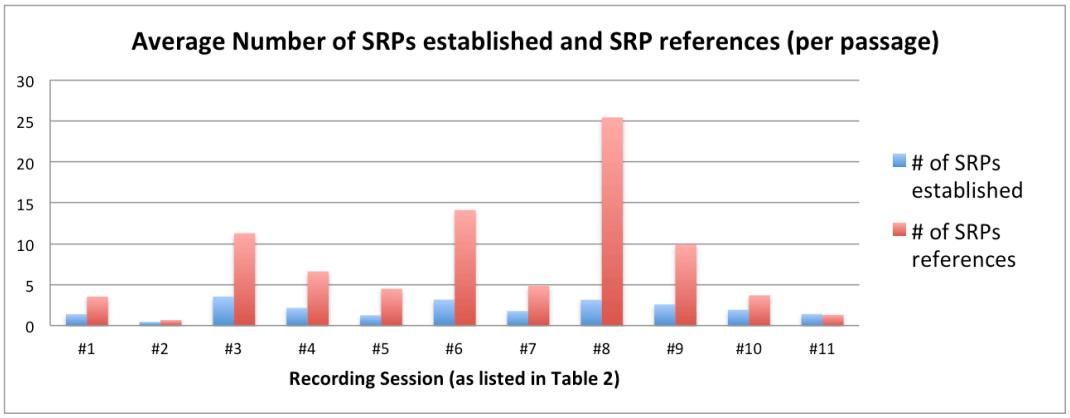

Figure 6: Average number of SRPs established/referenced (per passage) for each recording session.

[2] Verbs whose motion path is modified to indicate their subject or object location in the signing space are also marked with colons and index numbers, corresponding to the SRPs that are serving as the subject or object of the verb. Annotation of movements of the signer's torso toward locations in space is planned for future work.

2.5 First Release of the Data

We began motion-capture data collection in the summer of 2009 and have so far collected and linguistically annotated 242 ASL passages from 8 signers during 11 recording sessions (212 minutes of annotated ASL motion-capture data). We are now releasing the first sub-portion of our corpus that has been checked for quality and is ready for dissemination. The released segments of the corpus correspond to "recording sessions" number 6, 7, and 10 (signer E, G, and A), from Table 2. This release includes 98 passages performed by 3 native signers. The data includes Autodesk Motion Builder files of the motion-capture recording, BVH files (another commonly used file format for motion-capture data), high-resolution video recordings, and annotations for each passage. The annotations are in the form of plaintext files exported from SignStream™ (Neidle et al., 2000). Additional details about the corpus are available in (Lu and Huenerfauth, 2012a). Given this is our first release, we welcome advice and feedback from other researchers about how we should organize this corpus so that it is most useful. Future corpus releases may contain revisions of motion data formats, additional linguistic annotation, and additional passages (from the other recording sessions). Our lab website contains details about the steps required to obtain access our corpus: http://latlab.cs.qc.cuny.edu/corpus.

3 Evaluating Our Collected Motion Data

This section addresses the question: *Have we successfully configured and calibrated our motion-capture equipment so that we are recording good-quality data that will be useful for NLP research?* If a speech synthesis researcher were using a novel microphone technology to record human speakers to build a corpus, that researcher would want to experimentally confirm that the audio recordings were of high enough quality for research. Since the combination of motion-capture equipment we are using is novel and because there have not been prior motion-capture-

based ASL corpora projects, we must evaluate whether the data we are collecting is of sufficient quality to drive ASL animations of a virtual character. In corpus-creation projects for traditional written/spoken languages, researchers typically gather text, audio, or (sometimes) video. The quality of the gathered recordings is typically easier to verify and evaluate; for motion-capture data, a more complex experimental design is necessary (details in sections 3.1, 3.2, and 3.3). We want to measure how well we have compensated for several possible sources of error in our recordings:

- If the connection between a sensor and the recording computer is temporarily lost, then data gaps occur. We selected equipment that does not require line-of-sight connections and tried to arrange the studio to avoid frequent dropping of any wireless connections.
- As discussed previously, participants are asked to perform a quick head movement and distinctive eye blink pattern at the beginning of the recording session to facilitate our "synchronization" of the various motion-capture data streams during post-processing. If done incorrectly or inexactly, then a timing error is introduced into the data.
- Electronic and physical properties of motion-capture sensors can sometimes lead to small random errors (called "noise") in the data; we can attempt to remove some of this noise with smoothing algorithms applied to the data afterward.
- Differences between the bone lengths and other body proportions between the human and the "virtual skeleton" of the animated character being recorded could lead to "retargeting" errors; these errors manifest as body poses of the human that do not match the body poses of the virtual human character. We must be careful in the measurement of the bone lengths of the human and in the design of the virtual animation skeleton.
- To compensate for how motion-capture equipment sits on the body on different occasions or on different humans, we must set "calibration" values at the beginning of each recording session to adjust sensor sensitivity and offsets; e.g., we designed a novel protocol for calibrating gloves for ASL signers (Lu and Huenerfauth, 2009).

Section 3.3 presents a previously unpublished study in which we used motion-capture data from our corpus to synthesize animations of sign language, and we showed these animations to native ASL signers, who answered comprehension questions and subjective survey questions about the quality of the animations. In that study, we presented videos of the humans who were recorded for our corpus as an upper-baseline of comparison (some researchers prefer the term "ceiling" to the term "upper baseline"). As expected, the videos of humans received higher evaluation scores than our animations, but we believe that the scores obtained for our animations indicated that we

are collecting a corpus of reasonable quality. Before presenting the results of this study, we will summarize the results two other past studies (in sections 3.1 and 3.2). These studies are presented as a "lower baseline" for comparison – to enable the reader to better understand and interpret the results of the novel study presented in section 3.3.

Individuals who are not familiar with the process of using motion-capture data to animate virtual humans to produce sign language animations are often surprised at the challenges involved in this work. The first instinct of many researchers who are new to this field is that an easy way to make high quality animations of signing would be to directly control the movements of a virtual avatar based on the motion-capture recordings from a human. In fact, the results of such "direct puppetry" of a virtual human are often extremely difficult to understand, appear jerky, and lack accuracy in how the locations of the body align and touch.[3] In a prior evaluation study (Lu and Huenerfauth, 2010), we compared ASL animations from direct puppetry to ASL animations produced using a scripting/synthesis approach (that concatenated sign lexical items encoded as keyframe movement targets with motion-interpolation between keyframes), and we found that the synthesized animations achieved higher comprehension and subjective quality scores.

3.1 Early Study with Low Comprehension Scores for Motion-Capture Animations

In Huenerfauth (2006), we constructed a prototype system for planning and synthesizing ASL animations (containing a specific linguistic construction); the synthesizer used movement interpolation through keyframe location/orientation targets for the hands at specific moments in time. For our user-based evaluation of the resulting animations, we wanted to include "upper baseline" ani-

[3] While human animations can be produced from motion-capture data (as is done for films or video games), typically, extensive manual post-editing work is required from skilled animators to correct and adjust the movements to produce a clear and understandable result. While we are using a "direct puppetry" approach to synthesize some ASL animations for these evaluation studies -- so that we may evaluate the quality of our motion data -- we do not intend to use direct puppetry to produce ASL animations for our later research. Instead, we will use the corpus as training data to build machine-learning models of various ASL linguistic phenomena, and we intend to synthesize novel ASL animations based on these resulting models. This paradigm is exemplified in (Huenerfauth and Lu, 2010b; Lu and Huenerfauth, 2011a).

mations that would be very natural and understandable. Somewhat naïve to the complexities of using motion-capture data for direct puppetry (summarized above), we decided to use motion-capture data to produce an animation of a virtual human character. During the data collection process, we manually adjusted the settings of the cybergloves (although, we later learned that we were not sufficiently accurate in our calibration), and we adjusted the body size of the virtual human avatar to approximately match the size of the human recorded (although, we also later learned that we were not sufficiently accurate in this regard). The human performed identical sentences to those that were synthesized by our ASL system, and we produced some animations based on this data.

When we conducted our experiment to evaluate the quality of the ASL animations produced by our system, we asked participants to rate the animations they saw on Likert scales for grammatical correctness, understandability, and naturalness of movement; they also answered comprehension questions about the animation to determine how well they understood the information it conveyed. While we had expected the motion-capture-based direct-puppetry animations to be the *upper* baseline in our study (since a human was controlling the moments to produce fluent ASL sentences), we found that the motion-capture animations achieved the *lowest* scores in our study. In Figure 7, adapted from (Huenerfauth, 2006), the bar marked "MO" is the motion-capture data, the "CP" is the synthesized animation from our system, and the "SE" is an animation of a character performing signs in English word order, which was used as another baseline.

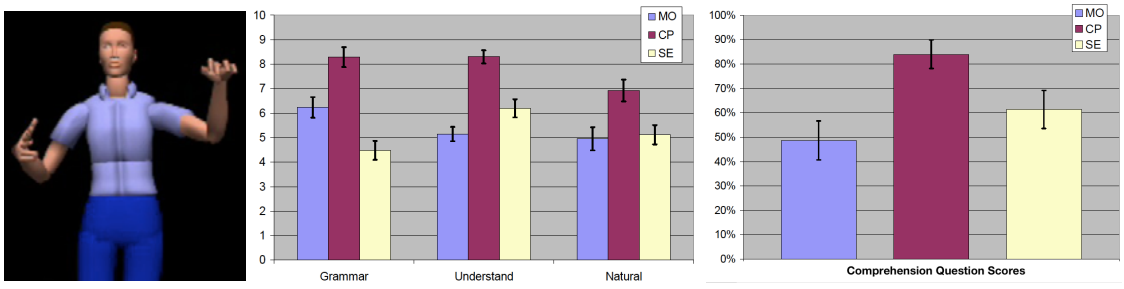

Figure 7: Screenshot of an animation produced from motion-capture data, scores for Likert-scale subjective evaluation questions, scores for comprehension questions – images from (Huenerfauth, 2006).

Despite our efforts, we had not sufficiently controlled the various sources of error in the data collection and animation process, and the resulting animations were very poorly understood. Our ASL synthesis system, which was only an early prototype system, actually achieved higher scores than the motion-capture-based animations in this case. Aside from the calibration and body-size issues mentioned above, it was also clear that the motion-capture based animations had some movement-jitter in the data that was inadequately smoothed prior to producing the animations.

3.2 Prior Study Evaluating Motion-Capture Animation from our CUNY Studio

Beginning in 2008, we established a new motion-capture recording studio at CUNY, as described in section 2.1, which consisted of some equipment that was designed to address problems we had encountered in our prior work. For instance, based on some experiences with dropped wireless signals from a pair of gloves used in 2006, we opted for a wired set of cybergloves in our new studio. Based on problems with occlusion of the hands during some signs in 2006, we opted for a non-optical based motion-capture body suit based on intertial/magnetic sensors in our new studio. We also made efforts to improve our equipment calibration, body size retargeting, and data-stream synchronization, which had been potential sources of error in 2006. These details appear in sections 2.1 and 2.2 of this article, and additional publications about our methodological improvements for ASL motion-capture appear in (Lu and Huenerfauth, 2009; Lu and Huenerfauth, 2010).

We conducted a study to evaluate the quality of the motion-capture data we were capable of collecting using our new studio and recording process. A focus of our methodological work was on how to best calibrate the cybergloves worn by ASL signers in an efficient and accurate manner. Given the 22-sensors sewn into each glove (each of which has a "gain" and "offset" setting which must be adjusted to achieve good data), the calibration of these gloves is non-trivial. Since the gloves sit differently on the hands on different occasions, the calibration must be performed for each session. Unskilled personnel may not be able to achieve good calibrations (and thus the data

collected does not produce recognizable hand shapes), and we found that even skilled personnel who managed to calibrate the gloves accurately required too much time (over 1 hour per glove). Thus, we designed a novel calibration protocol for the cybergloves designed to be efficient, accessible to deaf participants, and yielding accurate calibration results (Lu and Huenerfauth, 2009).

To evaluate our glove-calibration process, we collected motion-capture data from a native ASL signer, under two different recording conditions: (1) using a pre-existing fast glove calibration process (referred to as the "old calibration" in Figures in this section) and (2) using our newly designed glove calibration protocol. We produced ASL animations via direct puppetry using this data, and we showed these animations to native ASL signers, who evaluated the results (by answering Likert-scale subjective evaluation questions and comprehension questions about the information contained in the passages). In this way, we could compare whether our new glove calibration protocol allowed us to record better quality motion-capture data.

A native ASL signer (22-year-old male who learned ASL prior to age 2) performed a set of 10 ASL stories based on a script we provided; he wore the motion-capture equipment described in section 2.1 and underwent the calibration process and protocols described in section 2.2. As a proof-of-concept "reality check" on the quality of the motion-capture data that had collected, we wanted to use a simple animation approach (direct puppetry) to "visualize" our collected data without adding any visual embellishments. Autodesk MotionBuilder software was used to produce a virtual human whose movements were driven by the motion-capture data collected; see Figure 1(b). Figure 8 illustrates the transcript of one example story used in the experiment; the signer rehearsed and memorized each story (cue cards were available during the session).[4]

[4] It is important to note that the pre-scripted stories used in this study are not part of our CUNY ASL Corpus, described in section 2, which contains unscripted data. Thus, the pre-scripted cue-card text in Figure 8 should not be confused with the after-the-fact gloss annotation transcript shown in Figure 4, which was produced by an annotator watching a video recording of the unscripted ASL performance after-the-fact. Pre-scripted stories were used in this study because the signer needed to perform each story two times – once with each of the different glove calibrations being compared.

(a)	(b)
LAST FALL, MY AUNT #SALLY SHE PLAN #GARAGE #SALE. KNOW++ ? SET-UP TABLE OUTSIDE HOUSE. OLD THINGS DON'T-WANT, SELL. SOMETIMES, ADVERTISE IN NEWSPAPER. IF ARRIVE EARLY, CAN FIND GOOD THINGS, GOOD PRICES. CHEAP. TEND SELL: (list-of-5) (1st) OLD BOOKS (2nd) MAGAZINES (3rd) TOYS (4th) ART (5th) CLOTHING. OLD DRESSES, SHE WEAR PAST 1950s, SHE SELL MANY. ALSO, MUSIC RECORDS, MOST $1. I HELP AUNT SET-UP. FINISH. WRONG! NONE SHOW-UP.	Last fall, my Aunt Sally planned a garage sale. Do you know what that is? You set up a table outside the house, and then you can sell old things that you don't want anymore. Sometimes, you can advertise it in the newspaper. If you arrive early at one, you can often find good stuff at good prices. Stuff is cheap. People tend to sell old books, magazines, toys, art, clothing, etc. There were a bunch of old dressed that my aunt used to wear back in the 1950s; she sold a bunch of them. Also, there were records for sale, most for $1. I helped my aunt set everthing up. Unfortunately, when we were done, there was a bad surprise: no one showed up!

Figure 8: A story used in our first evaluation study: (a) a transcript of the sequence of ASL signs in the story and (b) an English translation of the story.

Our choice of a simplistic direct-puppetry animation allowed us to examine whether our motion-capture data itself is understandable, without applying any appearance enhancements to the animation. The avatar had a very simple appearance: without any face details. Since no facial expression information was collected by the motion-capture equipment, any facial expression would have been an after-the-fact embellishment from an animator. The lack of facial expression is a critical limitation of the ASL animations for this study because essential linguistic information is conveyed by facial expressions and eye movements in ASL; thus, the face-less animations produced for this study would certainly have limited understandability. Of course, we do not believe that the face-less direct-puppetry animation shown in this study could be used as an "end-product" that would appear in an application for deaf users. Simple animations were used in this study so that we could evaluate the quality of the corpus data in isolation, that is, in a "raw" form.

Using questions designed to screen for native ASL signers (Huenerfauth, 2008), we recruited 12 participants to evaluate the ASL animations. A native ASL signer conducted the studies, in which participants viewed an animation and were then asked two types of questions after each: (1) ten-point Likert-scale questions about the ASL animation's grammatical correctness, understandability, and naturalness of movement and (2) multiple-choice comprehension questions about basic facts from the story. The comprehension questions were presented in ASL, and answer

choices were presented in the form of clip-art images (so that strong English literacy was not necessary). Examples of the questions are included in (Huenerfauth, 2008).

Figure 9 shows how the 10 animations produced using the older glove calibration process ("Old Calibration") and the animations produced using our newly designed glove calibration protocol ("Mocap2010") had similar scores for the grammaticality, understandability, and naturalness of the signing movement. The Mocap2010 animations had higher comprehension question scores. Statistically significant differences are marked with an asterisk ($p<0.05$). The Likert-scale data was analyzed using Mann-Whitney pairwise comparisons with Bonferroni-corrected p-values; non-parameteric tests were selected because the Likert-scale responses were not normally distributed. The comprehension question data was analyzed using an ANOVA.

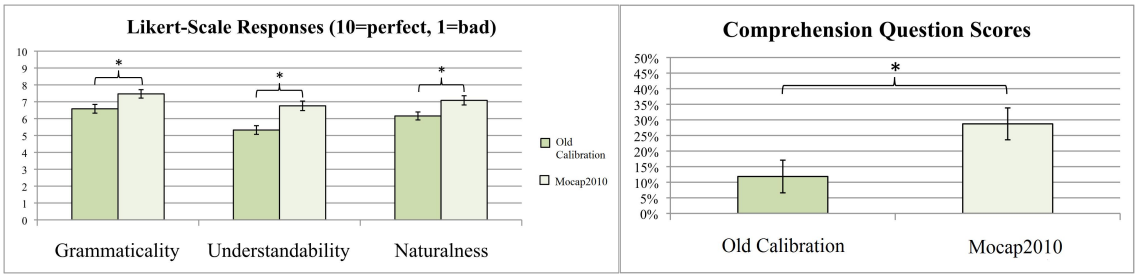

Figure 9: Likert-scale scores and comprehension scores from the evaluation study in 2010.

These results demonstrate the progress our laboratory has made in collecting motion-capture recordings of sign language performances with higher levels of quality – in this case, with a specific focus on just one aspect of our methodology, the glove calibration process. What is particularly notable is the low comprehension score for the "Old Calibration" animations. While the glove calibration in that data was lower quality for those animations, the calibration of the bodysuit and other sensors had already been carefully adjusted. We had previously developed protocols for measuring the bone lengths of human participants and matching them to the size of the virtual skeletons, synchronizing and aligning the data from the gloves with the data from the bodysuit, and other methodology issues – details in (Lu and Huenerfauth, 2010; Lu and Huenerfauth, 2012a).

Thus, it is notable how difficult to understand animations based on direct-puppetry can be – even when there is only one component of the motion-capture setup which has been poorly calibrated.

A limitation of this study is that it is not presenting an evaluation of actual data that is included in our corpus. This study used pre-scripted stories (with somewhat limited linguistic complexity), and since the signer memorized and performed the story while viewing cue-cards, it is likely that the signing is more formal, slower, and more precisely articulated than would be natural/unscripted ASL signing. Therefore, we conducted a new study (previously unpublished, described in section 3.3) with a larger number of sign language stimuli, sign language data drawn from our actual corpus (instead of "scripted" stories), and a larger number of research participants.

3.3 New Corpus Comparison Study

After collecting data for three years, we conducted another round of experiments to evaluate our motion-capture data. In this study, we use actual recordings from our corpus: specifically, we used 12 stories from the 3 native signers (average age 27) that are included in the released portion of our corpus (section 2.5). The average passage length was 97.7 signs, and topics include: personal information, news stories, explanation of encyclopedia articles, and short narratives. As in our prior study, MotionBuilder was used to produce a virtual human based on the data – see Figure 10(b).

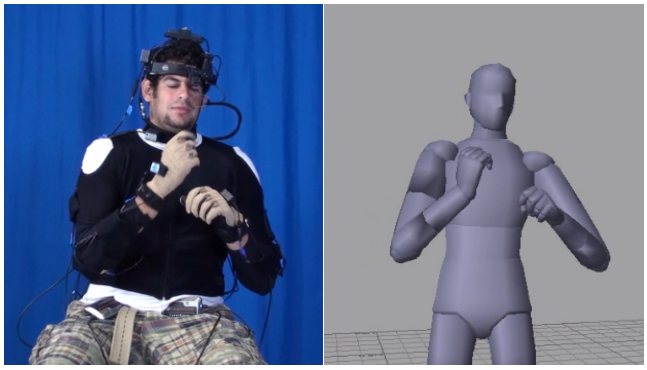

Figure 10: (a) A video recording of a human wearing the motion-capture body suit during the data collection process, (b) the recording-based virtual human; (both shown in the study in section 3.3).

A novel aspect of this study is that we chose to compare the direct-puppetry animation to the actual video recording of the human signer (while wearing the motion-capture body suit) during the data collection session (Figure 10). Since the video of the human has a much more detailed and natural appearance than our virtual human signer, we would expect higher naturalness and understandability scores for the video. Further, our motion-capture equipment does not record facial expressions (and so our virtual human in Figure 10(b) has no facial details), thus we would also expect higher naturalness and understandability scores for the video of the human signer, since this additional facial information is included in that version of each story.

We recruited 12 native ASL signers to evaluate these ASL animations and videos. A native ASL signer conducted the studies, in which participants viewed an animation and were then asked two types of questions after each: (1) ten-point Likert-scale questions about the ASL animation's grammatical correctness, understandability, and naturalness of movement and (2) multiple-choice comprehension questions about basic facts from the story. The comprehension questions were presented in the form of videos in which a native signer presented the questions in ASL, and answer choices were presented in the form of clip-art images (so strong English literacy was not necessary). Identical questions were used to evaluate the motion-capture animations and the human video. For example, the "Mac vs. PC" story shown in Figure 4 was one of the stories used in this study; one comprehension question for this story was "Which computer does the signer think is good for business?" Examples of the clip-art answer choices appear in Figure 11. This set of pictures in Figure 11 is the answer choices to the question "Which does the signer think is good for business?" The transcript of this story for this question has been shown in Figure 4.

Figure 11: Example of the answer choices presented as clip-art images to the native signers.

Before presenting the results of this study, we shall first consider how we would interpret the results of this experiment. We can consider some bases for comparison:

- As an "upper" basis of comparison, we look to some of our prior work, where we conducted evaluations of computer-synthesized ASL animations, using state-of-the-art generation models (Lu and Kacorri, 2012). These animations were the result of several years of research into how to produce high-quality ASL animations and were produced with the intervention of a human animator (they were not automatically produced); we compared the understandability of these animations to videos of human signers. The animations had comprehension questions scores that were 15% lower than human videos and Likert-scale subjective scores that were 30% lower than human videos.
- As a "lower" basis of comparison, we note that the motion-capture animations we produced in (Huenerfauth, 2006) had comprehension question scores that were approximately 40% lower than synthesized animations and subjective scores that were approximately 25% lower. As another "lower" basis of comparison, we note that the experiments in section 3.2 revealed that animations produced using our "old" equipment calibration led to approximately 50% lower comprehension scores and approximately 14% lower subjective scores, as compared to our new calibration protocols.

Further, there are reasons why we expect the Mocap2012 animations in this new study to receive lower scores than our high-quality computer-synthesized animations in prior work, lower scores than human videos of ASL, or lower scores than the simpler scripted stories in section 3.2:

- The Mocap2012 animations were produced via direct puppetry control of a virtual human without any additional clean-up or editing work by a human animator to enhance the understandability of the animation prior to its use in the study, as typically done when artists use motion-capture data to produce animations. While researchers have used motion-capture data to produce high-quality animations of sign language, e.g., (Gibet et al., 2011), they made use of sophisticated modeling and processing techniques to enhance the quality of the animation. Because we are interested in the raw, unenhanced use of our motion-capture data, we are using simplistic direct puppetry techniques in this study.
- There were no face details or eye gaze in the Mocap2012 animations (as compared to human videos or the computer-synthesized animations in our prior work). Since facial

expressions and eye-gaze are an important part of ASL, this limits the potential scores that the Mocap2012 animations may achieve.

- The Mocap2012 animations consisted of unscripted ASL passages produced spontaneously by native signers discussing unrestricted topics (though elicited by prompts). Thus, we would expect the Mocap2012 animations to contain more complex movements, diverse linguistic structures, and subtleties that might be difficult to perceive in an animation), as compared to the scripted Mocap2010 animations in our earlier study or the computer-synthesized animation in our prior work. If animations are on simple topics, with pre-scripted messages, that avoid complex linguistic structures, at slower speeds, then it is more likely that the animation will be able to convey its (simple) message successfully to a human viewer.
- The Mocap2012 animations consisted of continuous signing (with no pauses inserted between signs) at natural human speeds, unlike the computer synthesized animations in our prior work, which contained some pauses and lower speed.

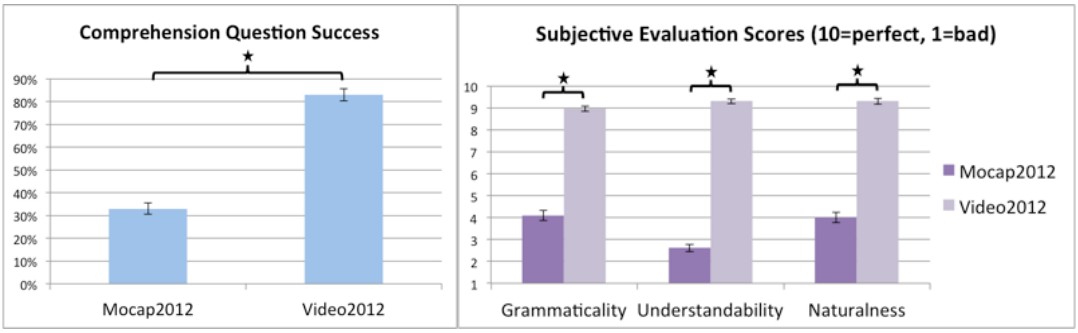

Figure 12: The comprehension-question and Likert-scale subjective scores for the motion-capture direct-puppetry virtual human animation and the video of a human signer shown in this study.

Figure 12 displays results of the Likert-scale subjective questions and comprehension-question success scores for the video recordings of a human and the recording-based motion capture virtual human evaluated in this study. Our scoring system gives +1 for correct answers and -0.25 for wrong answers. The videos (marked as "Video2012") have higher comprehension scores and subjective scores rather than the motion-capture animations (marked as "Mocap2012"). Statistically significant differences are marked with an asterisk ($p<0.05$). Likert-scale data was analyzed using Mann-Whitney pairwise comparisons with Bonferroni-corrected p-values, and comprehension question success rate data was analyzed using an ANOVA.

This is a promising result because the motion-capture data used to produce the Mocap2012 animations was taken directly from the released portion of our CUNY ASL Corpus (section 2.5).

As expected, comprehension scores for the video of a human signer (Video2012) were much higher than the scores for Mocap2012; an explanation for this result is that the video included greater visual detail, the video included a face and eyes, the video had natural coloring and shadows, etc. While the signer is wearing some motion-capture equipment in the video, the face and hands are visible in the video, and in an open feedback period at the end of the study, no participants mentioned difficulty at understanding the humans in the video (although some participants asked about why the people in the videos were wearing such equipment).

The comprehension question scores for Mocap2012 animations in the new study (Figure 12) were similar to those of the Mocap2010 animations in our prior study (Figure 9); both had comprehension question success scores near 33%. Given the similar comprehension scores, it was interesting that the Likert-scale subjective evaluation scores for Mocap2012 were lower than those for Mocap2010. One reason for the lower Likert-scale scores may be that, in the 2012 study, videos of a human signer were shown as an upper baseline for comparison in the experiments. In the 2010 study, no videos of humans were included: two forms of animation were compared. Based on this observation, we hypothesized that when viewed in comparison to a video of a real human signer (with high quality facial appearance, emotional subtlety of facial expression, real shadows and colors, etc.), an animation of a virtual human signer may seem (subjectively) to be of lower quality. In other words, perhaps an ASL animation would receive lower Likert-scale subjective evaluation scores when it is being evaluated in an experiment in which videos of human signers are shown for comparison. To investigate this hypothesis, we conducted a set of experiments to determine quantitatively how the inclusion of videos of human signers as a baseline for comparison may affect the scores collected in a user-based evaluation study of ASL animations (Lu and Kacorri, 2012). We found that the use of a video of a human signer shown as a basis of comparison in an evaluation study of ASL animation can lead to lower Likert-scale subjective scores for the ASL animations.

4 The Real Test of Quality: Doing Research with the Corpus

The research question addressed by these prior experiments was whether our motion-capture configuration and recording protocols enabled us to collect motion-data of sufficient quality for data-driven ASL animation research. Given the challenges in producing an understandable animation of a face-less virtual human from direct puppetry of motion-capture data (see section 3.3), the fact that the resulting animations were partially understandable was a positive result. Given the scoring approach used for our comprehension questions (which includes negative score penalties for incorrect answers, so as to yield scores of 0% from random guessing), we see that the comprehension question results for the Mocap2012 animations were well above chance. Given the similarity in the comprehension scores between the Mocap2012 and Mocap2010 animations, we characterize the results of this evaluation to be a successful indication that the motion-capture corpus contains good-quality data. Of course, given the lack of any other available annotated motion-capture corpus of ASL, even if the motion-capture data in the corpus were not perfect quality, this would still be a contribution to the field, for which there is a lack of available motion-capture ASL corpora.

While this suggests that our data is of good quality, a *real* test of our corpus is for it to be used in animation research. If we can build useful ASL-animation synthesis software based on analysis of this corpus, then we will know that we have good quality motion-capture data. In fact, we are already doing so: In our current research, we are using data from our corpus to train models used to synthesize novel ASL animations. Specifically, we are studying ASL verbs whose motion path depends on where in the signing space an SRP for their subject and object have been established (Lu and Huenerfauth, 2011a). Based on native signers' evaluations of the animations that result from our models, our methodology appears successful – details in (Lu and Huenerfauth, 2011b; 2012b). This research is mentioned in this article to illustrate how the collection of sign language movement data from humans (and mathematical modeling of this data) can yield a high quality ASL animation system, as measured in user-based studies. This first use of our corpus to

tackle a challenging problem in sign language generation provides further evidence that the quality of the motion-capture data we are collecting at CUNY is sufficient for supporting computational linguistic research on ASL.

5 Conclusions and Future Work

Given the accessibility benefits of providing information in the form of ASL for people who are deaf or hard-of-hearing and who use ASL, our laboratory is focused on advancing ASL animation synthesis technologies that make it easier to produce linguistically accurate and understandable animations. To address some challenging linguistic issues in ASL (e.g., the use of spatial reference, which was a focus of this article), the ASL animation research community is in need of additional human movement data with linguistic annotation. Previous researchers assembled small corpora *ad hoc*, and not all of these corpora contained human motion-capture data. This article has presented the CUNY ASL Motion-Capture Corpus, the first portion of which is now released to the research community. Our goal is for this corpus to contain sufficient quality motion-capture data and useful linguistic annotations to support research on ASL animation (and possibly for researchers studying ASL linguistics or other issues in human motion).

In addition to providing details about the equipment calibration, elicitation, data processing, and annotation process, this article has presented some prior and previously unpublished studies of the quality of the motion-capture data collected. While our research paradigm for using this corpus is to train statistical and machine learning models on portions of the corpus and to use these models in our ASL animation technologies, for the purposes of evaluation, we have synthesized animation via direct puppetry from the corpus. Given the challenges in producing understandable animation in this way (discussed in section 3), it was notable that participants in studies were able to understand some information content in these direct-puppetry animations (section 3.3) – suggesting that the data is of good quality. In current work, we are using our corpus in support of

our ASL animation synthesis research (section 4) with success – also providing evidence that the corpus should be useful for other researchers. In future work, we will continue to train models of where signers place SRPs in the signing space and how verb motion-paths change based on the arrangement of SRPs; these models will be incorporated into our ASL animation software and will be evaluated in user-based studies through the participation of ASL signers.

We will continue to release more corpus data, including additional recording sessions (3 were included in our first release). We are also adding and verifying additional layers of annotation, including: part-of-speech; syntactic bracketing (NP, VP, clause, sentence); and non-manual signals (role shift, negation, WH-word questions, yes-no questions, topicalization, conditionals, and rhetorical questions). As we verify these additional layers of annotation, we will distribute them in a future corpus release. To standardize our glosses, we are also updating them to follow the gloss names used in the American Sign Language Lexicon Video Dataset (Neidle et al., 2012). As we have just distributed our first release of the corpus, we welcome feedback from the research community on how we can make future releases of this corpus most useful for their research.

Acknowledgments

This research was supported in part by the U.S. National Science Foundation under award number 0746556 and award number 1065009, by CUNY PSC-CUNY Research Award Program, and by Visage Technologies AB through a free academic license for character animation software. Jonathan Lamberton assisted with the recruitment of participants and the conduct of experimental sessions. Kenya Bryant, Wesley Clarke, Kelsey Gallagher, Amanda Krieger, Giovanni Moriarty, Aaron Pagan, Jaime Penzellna, Raymond Ramirez, Molly Sachs, Evans Seraphin, Christine Singh, Fatimah Mohammed, and Meredith Turtletaub have contributed their ASL expertise to the project.

References

J. Bungeroth, D. Stein, P. Dreuw, M. Zahedi, H. Ney. 2006. A German Sign Language Corpus of the Domain Weather Report. Proc. LREC 2006 Workshop on Representation & Processing of Sign Languages.

Carnegie Mellon University Graphics Lab Motion Capture Database. Retrived in August 2012. http://mocap.cs.cmu.edu/

K. Cormier. 2002. Grammaticalization of Indexic Signs: How American Sign Language Expresses Numerosity. Ph.D. Dissertation, University of Texas at Austin.

S. Cox, M. Lincoln, J. Tryggvason, M. Nakisa, M. Wells, M. Tutt, S. Abbott. 2002. Tessa, a System to Aid Communication with Deaf People. In 5th International ACM Conference on Assistive Technologies, pp. 205-212. ACM Press, New York.

O. Crasborn, E. van der Kooij, D. Broeder, H. Brugman. 2004. Sharing Sign Language Corpora Online: Proposals for Transcription and Metadata Categories. Proc. LREC 2004 Workshop on Representation & Processing of Sign Languages, pp. 20-23.

O. Crasborn, H. Sloetjes, E. Auer, P. Wittenburg. 2006. Combining Video and Numeric Data in the Analysis of Sign Languages within the ELAN Annotation Software. Proc. LREC 2006 Workshop on Representation & Processing of Sign Languages, pp. 82-87.

eSIGN. Retrived in August 2012. http://www.visicast.cmp.uea.ac.uk/eSIGN/index.html

R. Elliott, J. Glauert, J. Kennaway, I. Marshall, E. Safar. 2006. Linguistic Modelling and Language-Processing Technologies for Avatar-Based Sign Language Presentation. Universal Access in the Information Society 6(4), pp. 375-391.

S.E. Fotinea, E. Efthimiou, G. Caridakis, K. Karpouzis. 2008. A Knowledge-Based Sign Synthesis Architecture. Univ. Access in Information Society 6(4):405-418.

S. Gibet, N. Courty, K. Duarte, T. Le Naour. 2011. The SignCom System for Data-driven Animation of Interactive Virtual Signers: Methodology and Evaluation. ACM Transactions on Interactive Intelligent Systems, Volume 1, Issue 1, Article No. 6.

M. Huenerfauth. 2006. Generating American Sign Language Classifier Predicates For English-To-ASL Machine Translation. Doctoral Dissertation, Computer and Information Science, University of Pennsylvania.

M. Huenerfauth. 2008. Evaluation of a Psycholinguistically Motivated Timing Model for Animations of American Sign Language. The 10th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2008), Halifax, Nova Scotia, Canada.

M. Huenerfauth. 2009. A Linguistically Motivated Model for Speed and Pausing in Animations of American Sign Language. ACM Transactions on Accessible Computing. Volume 2, Number 2, Article 9.

M. Huenerfauth, P. Lu. 2010a. Eliciting Spatial Reference for a Motion-Capture Corpus of American Sign Language Discourse. Proceedings of the 4th Workshop on the Representation and Processing of Signed Languages: Corpora and Sign Language Technologies, The 7th International Conference on Language Resources and Evaluation (LREC 2010), Valetta, Malta.

M. Huenerfauth, P. Lu. 2010b. Modeling and Synthesizing Spatially Inflected Verbs for American Sign Language Animations. In Proceedings of The 12th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2010), Orlando, Florida, USA. New York: ACM Press.

M. Huenerfauth, P. Lu and A. Rosenberg. 2011. Evaluating Importance of Facial Expression in American Sign Language and Pidgin Signed English Animations. The 13th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2011), Dundee, Scotland, United Kingdom. New York: ACM Press.

M. Huenerfauth, P. Lu. 2012. Effect of Spatial Reference and Verb Inflection on the Usability of American Sign Language Animations. Universal Access in the Information Society. Berlin/Heidelberg: Springer.

W. Janis. 1992. Morphosyntax of the ASL Verb Phrase. Doctoral dissertation, State University of New York at Buffalo.

J. Kennaway, J. Glauert, I. Zwitserlood. 2007. Providing Signed Content on Internet by Synthesized Animation. ACM Trans Comput-Hum Interact 14(3):15.

J.S. Kim, W. Jang, Z. Bien. 1996. A Dynamic Gesture Recognition System for the Korean Sign Language (KSL), IEEE Trans. Syst., Man, Cybern. B, vol. 26, pp. 354–359.

C.-S. Lee, Z. Bien, G.-T. Park, W. Jang, J.-S. Kim, S.-K. Kim. 1997. Real-time Recognition System of Korean Sign Language Based on Elementary Components. In Proceeding of 6th IEEE International Conference on Fuzzy Systems, pp. 1463–1468.

S. Liddell. 1990. Four Functions of a Locus: Reexamining the Structure of Space in ASL. In C. Lucas, ed. Sign Language Research: Theoretical Issues, 176-198. Washington D.C.: Gallaudet University Press

S. Liddell. 1995. Real, Surrogate and Token Space: Grammatical Consequences in ASL. In Emmorey, Karen & Reilly, Judy. S. eds. Language, Gesture, and Space, 19-41. Hillsdale, NJ: Lawrence Erlbaum Associates

S. Liddell. 2003. Grammar Gesture and Meaning in American Sign Language. UK: Cambridge University Press.

B. Loeding, S. Sarkar, A. Parashar, A. Karshmer. 2004. Progress in Automated Computer Recognition of Sign Language, Proc. ICCHP, pp. 1079-1087.

P. Lu, M. Huenerfauth. 2009. Accessible Motion-Capture Glove Calibration Protocol for Recording Sign Language Data from Deaf Subjects. The 11th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2009), Pittsburgh, PA, USA.

P. Lu, M. Huenerfauth. 2010. Collecting a Motion-Capture Corpus of American Sign Language for Data-Driven Generation Research. Proceedings of the First Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2010), Los Angeles, CA,

P. Lu, M. Huenerfauth. 2011a. Data-Driven Synthesis of Spatially Inflected Verbs for American Sign Language Animation. ACM Transactions on Accessible Computing (TACCESS). New York: ACM Press.

P. Lu, M. Huenerfauth. 2011b. Synthesizing American Sign Language Spatially Inflected Verbs from Motion-Capture Data. Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT), in conjunction with ASSETS 2011, Dundee, Scotland.

P. Lu, M. Huenerfauth. 2012a. CUNY American Sign Language Motion-Capture Corpus: First Release. Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, The 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey.

P. Lu, M. Huenerfauth. 2012b. Learning a Vector-Based Model of American Sign Language Inflecting Verbs from Motion-Capture Data. Proceedings of the Third Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), Human Language Technologies: The 13th Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2012), Montréal, Québec, Canada.

P. Lu, H. Kacorri. 2012. Effect of Presenting Video as a Baseline During an American Sign Language Animation User Study. In Proceedings of The 14th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2012), Boulder, Colorado.

I. Marshall, E. Safar. 2001. The Architecture of an English-Text-to-Sign-Languages Translation System. In: Recent Advances in Natural Language Processing (RANLP), September 2001, Tzigov Chark, Bulgaria.

S.L. McBurney. 2002. Pronominal Reference in Signed and Spoken Language. In R.P. Meier, K. Cormier, D. Quinto-Pozos (eds.) Modality and Structure in Signed and Spoken Languages. UK: Cambridge U. Press, 329-369.

R. Meier. 1990. Person Deixis in American Sign Language. In: S. Fischer & P. Siple (eds.), Theoretical Issues In Sign Language Research, vol. 1: Linguistics. Chicago: University of Chicago Press, pp. 175-190.

R. Mitchell, T. Young, B. Bachleda, M. Karchmer. 2006. How Many People Use ASL in the United States? Sign Language Studies 6(3):306-335.

S. Morrissey, A. Way. 2005. An Example-Based Approach to Translating Sign Language. Proc. Workshop on Example-Based Machine Translation, 109-116.

C. Neidle, D. Kegl, D. MacLaughlin, B. Bahan, R.G. Lee. 2000. The Syntax of ASL: Functional Categories And Hierarchical Structure. Cambridge: MIT Press.

C. Neidle, A. Thangali and S. Sclaroff. 2012. Challenges in Development of the American Sign Language Lexicon Video Dataset (ASLLVD) Corpus. 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. LREC 2012. Istanbul, Turkey.

C. Padden. 1988. Interaction of morphology & syntax in American Sign Language. Outstanding Dissertations in Linguistics, Series IV. New York: Garland Press.

S. Prillwitz et al. 1989. An Introductory Guide to HamNoSys Version 2.0: Hamburg Notation System for Sign Languages. International Studies on Sign Language and Communication of the Deaf. Hamburg: Signum.

E. Safar, I. Marshall. 2002. Sign Language Generation Using HPSG. In Proc. Int. Conf. on Theoretical and Methodological Issues in Machine Translation, pages 105–114, TMI Japan.

J. Segouat, A. Braffort. 2009. Toward the Study of Sign Language Coarticulation: Methodology Proposal. Proc Advances in Comput.-Human Interactions, pp. 369-374.

T. Shionome, K. Kamata, H. Yamamoto, S. Fischer. 2005. Effects of Display Size on Perception of Japanese Sign Language---Mobile Access in Signed Language. Proc. Human-Computer Interaction, 22-27.

SignWriting. Retrieved in August 2011. http://www.signwriting.org/

A.S. Tolba, A.N. Abu-Rezq. 1998. Arabic glove talk (AGT): A Communication Aid for Vocally Impaired. Pattern Analysis and Applications.

C. Traxler. 2000. The Stanford Achievement Test, Ninth Edition: National Norming and Performance Standards for Deaf and Hard-of-Hearing Students. J. Deaf Studies and Deaf Education 5(4):337-348.

Vcom3D. Retrieved in August 2012. Sign Smith Studio. http://www.vcom3d.com/signsmith.php

P. Vamplew, A. Adams. 1998. Recognition of Sign Language Gestures Using Neural Networks, Austral. J. Intell. Inform. Process. Syst. 5(2), pp. 94–102.

T. Veale, A. Conway, B. Collins. 1998. Challenges of Cross-Modal Translation: English to Sign Translation in ZARDOZ System. Machine Translation 13:81-106.

C.L. Wang, W. Gao, S.G. Shan. 2002. An Approach Based on Phonemes to Large Vocabulary Chinese Sign Language Recognition, Proc. IEEE Int'l Conf. on Automatic Face and Gesture Recognition, pp. 411–416.

L. Zhao, K. Kipper, W. Schuler, C. Vogler, N. Badler, M. Palmer. 2000. A Machine Translation System from English to American Sign Language. Proc. AMTA.