Eliciting Spatial Reference for a Motion-Capture Corpus of American Sign Language Discourse

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

Pengfei Lu
The City University of New York (CUNY)
Computer Science Program, Graduate Center
365 Fifth Ave, New York, NY 10016 USA
E-mail: pengfei.lu@qc.cuny.edu

Abstract

We seek computational models of the referential use of signing space and of spatially inflected verb forms for use in American Sign Language (ASL) animations for accessibility applications for deaf users. We describe our collection and annotation of an ASL motion-capture corpus to be analyzed for our research. We compare alternative prompting strategies for eliciting single-signer multi-sentential ASL discourse that maximizes the use of pronominal spatial reference yet minimizes the use of classifier predicates.

1. Introduction

Significant numbers of deaf adults in the U.S. have relatively low levels of written English literacy (Traxler, 2000); many have difficulty reading English text on websites or other information sources. Animations of American Sign Language (ASL) make information and services accessible for these individuals. There are two major types of ASL animation technologies: scripting and generation/translation software. Scripting software allows a human author to specify the movements of a virtual human character by arranging signs and facial expressions on a timeline to be performed, e.g. (Vcom3D, 2010; Kennaway et al., 2007). Generation/translation software automatically synthesizes ASL sentences, given an English input sentence to be translated; Huenerfauth and Hanson (2009) describe and review such systems.

Our goal is to construct computational models of ASL that could be used to partially automate the work of human authors using scripting software or to underlie generation/translation systems. Specifically, we wish to model aspects of ASL linguistics that are not handled by modern ASL scripting or generation software. Signers associate entities under discussion with 3D signing space locations, and signs whose paths or orientations depend on these locations pose a special challenge: They are time-consuming for users of scripting software to produce, and they are not included in the repertoire of most modern ASL generation/translation software.

Huenerfauth (2009) found that native signers' comprehension of ASL animations improved when the animations included: (1) association of entities with locations in the signing space and (2) the use of verbs whose motion paths were modified based on these locations. Thus, users of ASL animation software would benefit from better handling of these two phenomena.

Section 2 describes how these spatial reference phenomena are frequent in ASL signing and important to the meaning of ASL sentences. Section 3 describes our overall research goals of: (1) collecting an ASL corpus using motion-capture equipment and video, (2) annotating the use of spatial reference phenomena and other linguistic features in this corpus, and (3) analyzing the human movement data in this corpus (and its relationship to the linguistic structure) to build computational models of how ASL signers associate entities under discussion with 3D signing space locations. These computational models will be incorporated into ASL animation technologies we are developing to make the resulting animations more realistic, understandable, and ultimately more useful for deaf users in accessibility applications. Section 4 discusses our corpora collection and annotation procedure; section 5 compares alternative prompting strategies we have used during year 1 of the project to elicit signing performances of the desired form. Section 6 contains conclusions and future research plans.

2. Spatial Reference Points in ASL

As in other sign languages, users of ASL frequently associate entities under discussion with locations in the signing space involved in later pronominal reference and other purposes (Liddell, 2003; Meier, 1990; Neidle et al., 2000). Various ASL constructions can be used to establish a ***spatial reference point (SRP)*** for some entity:

- Pre-nominal determiners and some post-noun-phrase adverbs consist of a pointing sign in which the entity in that noun phrase is assigned a 3D SRP location.
- Fingerspelling or some nouns may also be signed outside their standard location to establish an SRP.

The movements of other ASL signs are parameterized on the 3D locations of previously established SRPs:

- Personal, possessive, and reflexive pronouns consist of pointing movements to SRPs' 3D locations.
- Some verbs change their motion path or orientation to indicate the 3D location of their subject, object, or both. What features are modified and whether this is optional depends on the verb. These ***inflecting verbs*** (Padden, 1988) are sometimes referred to as agreeing (Cormier, 2002) or indicating verbs (Liddell, 2003).
- During verb phrases or possessive phrases, the SRP of the subject/object or possessor/possessed may be indicated by head-tilt/eye-gaze (Neidle et al., 2000).

Thus, ASL animation software that does not model SRPs cannot generate: determiners, pronouns, many noun phrases, some verb phrases, spatially inflected verbs, or possessive phrases – all of which are based on SRPs.

3. Research Goals

We seek computational models of: (1) what locations in 3D space are commonly chosen for SRPs, (2) which entities are assigned SRPs, (3) how the motion-paths of inflecting verbs change based on the 3D location of their subject's and object's SRP. Producing the hand, eye, and head movements needed to establish and refer to SRPs is burdensome for human users of ASL scripting software – and producing accurate 3D movements of spatially inflected ASL verbs is even harder. We believe that models of these three issues above could be used to partially automate this work or used to fully automate the work of ASL-animation generation/translation software.

We will build these computational models through the collection and analysis of a motion-capture corpus of ASL multi-sentential discourse. We hypothesize that linguistic features of the discourse affect the likelihood of a signer assigning an entity an SRP (and where it will be placed); we will analyze the corpus using statistical machine learning techniques to build SRP establishment models. We also believe that mathematical functions of verbs' motion paths (parameterized on SRP locations) can be induced from the collected 3D motion data; we will use regression/model-fitting techniques to construct an animation lexicon of ASL inflecting verbs that are spatially parameterized on the 3D location of the subject and/or object (so that inflected forms can be synthesized as needed by ASL scripting or generation software).

This corpus consists of motion-capture recordings of multi-sentential discourse with annotation of SRP establishment and reference. Prior researchers collected video-based corpora, e.g. (Neidle et al., 2000; Bungeroth et al., 2006; Efthimiou & Fontinea, 2007), or short sign language recordings via motion-capture, e.g. (Brashear et al., 2003; Cox et al., 2002). Researchers have designed schemes for annotating the referential use of signing space (Lenseigne & Dalle, 2005), but no previous motion-capture corpus includes such SRP annotation.

4. Corpora Collection Procedure

For our corpus, we record handshape; hand location; palm orientation; eye-gaze vector; and joint angles for the wrists, elbows, shoulders, clavicle, neck, and waist. Our novel configuration of commercial motion-capture equipment includes: two Immersion CyberGloves®, an Applied Science Labs H6 head-mounted eye-tracker, an Intersense IS-900 inertial/acoustic tracker (for the head), and magnetic/inertial sensors on an Animazoo IGS-190 bodysuit. Three high-definition digital video cameras record front, side, and facial close-up views of the signer (referred to as the "performer"). Another native signer (the "prompter") sits behind the front-view camera to converse with the performer and elicit signing to record.

In our first year, we have recorded and annotated 58 ASL passages from 6 signers (~ 40 minutes of data). To collect natural use of SRPs, we elicit *unscripted* multi-sentential single-signer discourse. Table 1 lists different prompting strategies we tried and how many recordings we collected using each. The totals for each vary because the recording session was intentionally kept relaxed/conversational to promote more natural signing:

Type	N	Description of the Prompting Strategy
Personal Intro/Info	15	Introduce yourself, describe some of your background, hobbies, family, education…
Hypothetical Scenario	4	What would you do if: You were raising a deaf child? You could have dinner with any two famous or historical figures?
Compare (not people)	9	Compare two things: e.g. Mac vs. PC, Democrats vs. Republicans, high school vs. college, Gallaudet University vs. NTID, travelling by plane vs. travelling by car, etc.
Compare (people)	7	Compare two people you know: your parents, some friends, family members, etc.
Recount Movie/Book	7	Tell us about your favorite movie or your favorite book. What happens in it?
Tell a Story (3 Wishes)	2	Invent a story using this topic: "If I had a genie that could grant three wishes, I'd…"
Repeat Conversation	6	Watch 3-minute video of ASL or captioned conversation, then explain what you saw.
Children's Book	5	Read a short children's book, then explain the story as you remember it.
Wikipedia Article	3	Read a 300-word Wikipedia article on "The History of Racial Segregation in the United States." Explain/recount the article.

Table 1: Types of prompts used during data collection with the number of stories of each type collected (N).

the prompter used different strategies to elicit signing from the performer. Sometimes the performer was verbose in their response to a prompt, but other times, he/she could think of little or nothing to say. Further, since performers were recorded for only 1 hour (after the motion-capture equipment was set-up and calibrated), we rarely had sufficient time to try all of the different prompt-types during each performer's recording session.

After collecting each story, we synchronize our video and motion-capture streams, apply the data to a 3D skeleton, and produce video segments for each story. A team of native ASL signers (including students from deaf high schools in New York) annotates the data using the SignStream™ annotation tool (Neidle et al., 2000). We annotate some traditional information: sign glosses; part-of-speech; syntactic bracketing of NPs, VPs, clauses, sentences; and non-manual marking of role shift, negation, who/what/where/when/why/how questions, yes-no questions, topicalization, conditionals, and rhetorical questions. In support of our research goals, we also annotate: when SRPs are established, which discourse entity is associated with each, when referring expressions indicate each SRP, and when any verbs are spatially inflected to indicate each SRP. Each SRP is assigned an index number, and each pronominal or verb reference to an SRP is marked with this index. These SRP establishments and references are recorded on parallel timeline tracks to the glosses and other linguistic annotations. We also mark any ***classifier predicates (CPs)*** performed; CPs are special signs in which the signer synthesizes a movement for the hands (or sometimes the

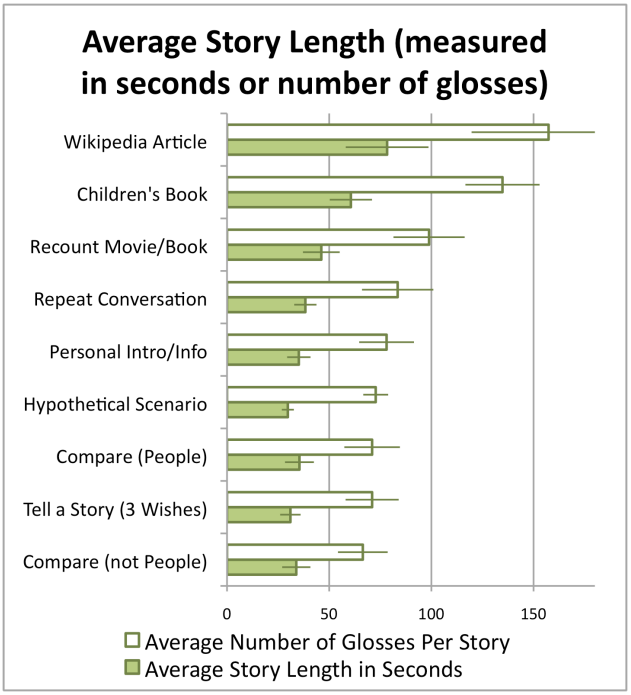

Figure 1: Length of the ASL stories collected.

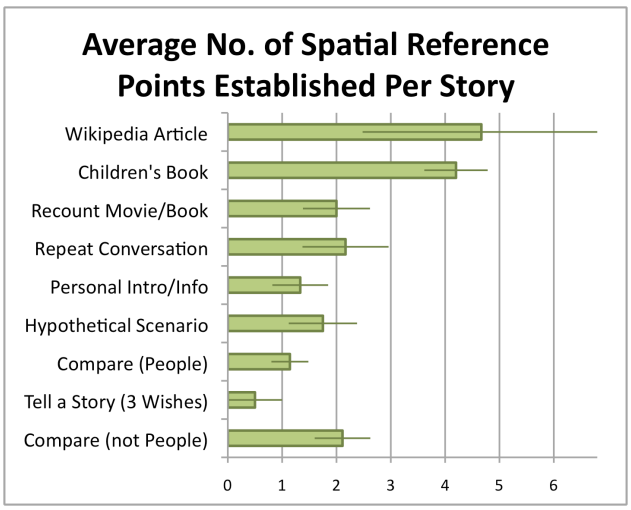

Figure 2: Spatial reference points established.

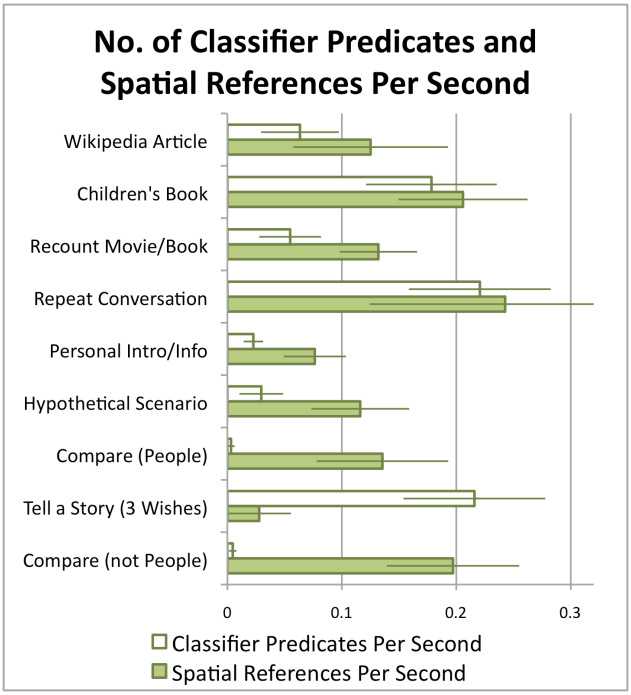

Figure 3: Number of classifier predicates and spatial references per second in each type of ASL story.

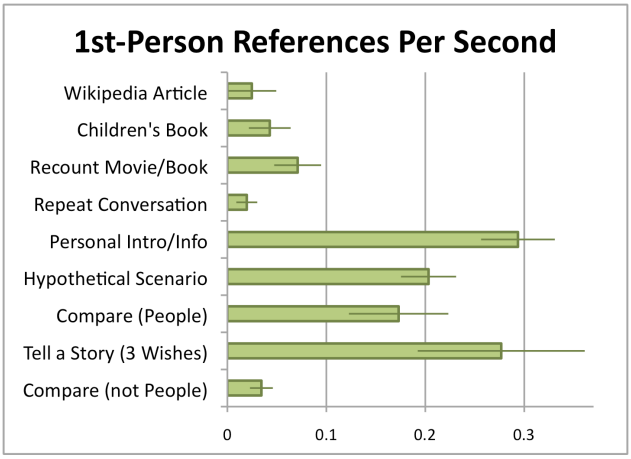

Figure 4: First-person references in the ASL stories.

body) to indicate the spatial arrangement, size, shape, or movement of people/objects in a 3D scene being described. We count CPs in order to measure the effectiveness of our prompting strategies (see section 5).

5. Comparison of Prompting Strategies

After collecting/annotating the first 58 stories, we can determine which prompting strategies were effective at collecting the desired type of ASL signing. An ideal ASL story to be collected for this corpus would:

A. Be long enough to allow for establishment of SRPs.
B. Sometimes contain multiple SRPs (perhaps 3+) to enable the study of diverse types of spatial use.
C. Contain as many pointing signs (determiners, pronouns, etc.) or inflected verbs that refer to SRPs as possible. With many examples of these ***spatial references (SRs)***, we will be able to study diverse forms of spatial use and reference in ASL signing.
D. Contain as few CPs as possible. CPs complicate how signing space is used; the interaction between CPs and SRPs is beyond the scope of our current work.

Figure 1 displays the average length of the stories collected using each prompting strategy – as measured in seconds of time or in the total number of manual signs (count of annotated glosses). Prompt types are listed in descending order based on their number of glosses; they are listed in this same order in Figures 2, 3, and 4. Error bars indicate the standard error of the mean for each value. The longest stories arose from prompts in which the performer recounted an article, book, movie, or conversation they saw recently or had seen in the past.

As listed in criterion 'B,' we'd like to collect some stories in which signers establish larger numbers of SRPs. Figure 2 displays the number of SRPs established in each story (entities assigned 3D locations for pronominal use). The longer stories generally contained more SRPs. (N.B.

If the performer referred to the prompter during the story, then the count of SRPs for that story was increased by 1. In such cases, the addressee was used as a 2nd-person referent, and thus, we counted the addressee as an SRP.)

Criteria 'C' and 'D' explain how we want to maximize the number of SRs in each story and minimize the number of CPs. Figure 3 displays the average frequency of SRs and CPs (as measured per second) in stories of each prompt-type; the values are displayed on the same graph to enable comparison of their ratio. The SRs in Figure 3 include 3rd-person and 2nd-person references, but not 1st-person (e.g. signs like "me"/"my" or inflecting verbs in which the subject/object is the signer) because these do not involve pointing to a location in the surrounding signing space. While we are not particularly interested in maximizing or minimizing the frequency of 1st-person references, we present their frequency in Figure 4 – for the sake of completeness. Unsurprisingly, the "personal intro/info," "tell a story," and "hypothetical scenario" prompts led to many 1st-person references. In some of the "compare (people)" stories, signers compared *themselves* to someone else.

6. Conclusions and Future Work

Our analysis of the different prompting strategies will guide our future data collection. Based on their high CP/SR ratio, we will no longer use the "tell a story," "children's book," and "repeat conversation" prompts. The long story lengths, high number of SRPs established, and modest CP/SR ratio of the "Wikipedia article" and "recount movie/book" prompts were promising, and we will continue to use more prompts like these in future work (selecting additional Wikipedia articles). We may further reduce the number of CPs collected by avoiding articles with spatially/visually descriptive topics. The very low CP/SR ratio of the "compare" and "personal intro/info" prompts was promising, and we will look for ways to encourage signers to elaborate further – to elicit longer stories when using these prompting strategies.

We plan on collecting/annotating approximately 200 ASL stories in total. Our experiences recording the first 58 stories have helped us to become more proficient at quickly and accurately collecting motion-capture data from signers, and we have developed new protocols for accurately and accessibly calibrating our equipment (Lu & Huenerfauth, 2009). We are also continuing to refine our annotation guide and training protocol for annotators to promote faster and more accurate annotation.

We are now beginning to analyze some collected 3D data to construct models of SRP establishment, spatial reference, and verb inflection. These models will be incorporated into ASL animation generation software we are developing to decide automatically: (1) when it should establish an SRP for an entity being discussed, (2) where it should place the SRP, and (3) how the signs later in the performance need to change based on SRP locations. In addition, we believe that our annotated ASL motion-capture corpus will be a valuable resource for future ASL linguistic researchers or computer scientists studying the synthesis of ASL animation or automatic recognition of ASL from human motion-data or video.

7. Acknowledgements

This research was supported by the U.S. National Science Foundation (award #0746556), Siemens (Go PLM Academic Grant), and Visage Technologies AB (free academic software license). Wesley Clarke, Kelsey Gallagher, Jonathan Lamberton, Amanda Krieger, and Aaron Pagan assisted with data collection/annotation.

8. References

Brashear, H, Starner, T, Lukowicz, P, Junker, H. (2003). Using multiple sensors for mobile sign language recognition. *IEEE Intl Sym Wearable Computers*, p 45.

Bungeroth, J, Stein, D, Dreuw, P, Zahedi, M, Ney, H. (2006). A German sign language corpus of the domain weather report. In C. Vettori (ed.) *2nd wkshp on represent. & processing of sign languages*, pp. 2000-3.

Cormier, K. (2002). Grammaticalization of indexic signs: how American Sign Language expresses numerosity. Ph.D. Dissertation, University of Texas at Austin.

Cox, S, Lincoln, M, Tryggvason, J, Nakisa, M, Wells, M, Tutt, M, Abbott, S. (2002). Tessa, a system to aid communication with deaf people. In *Proc. ACM Conference on Assistive Technologies*, pp. 205-212.

Efthimiou, E., Fotinea, S.E. (2007). GSLC: creation and annotation of a Greek sign language corpus for HCI. In *LNCS 4554*, Heidelberg: Springer, pp. 657-666.

Huenerfauth, M. (2009). Improving spatial reference in American Sign Language animation through data collection from native ASL signers. In *Proc. Universal Access in Human Computer Interaction*, pp 530-539.

Huenerfauth, M, Hanson, VL. (2009). Sign language in the interface: access for deaf signers. In C Stephanidis (ed.) *Universal Access Handbook*. Mahwah: Erlbaum.

Kennaway, J, Glauert, J, Zwitserlood, I. (2007). Providing signed content on Internet by synthesized animation. *ACM Transactions Computer-Human Interaction* 14(3), Article 15, pp. 1-29.

Lenseigne, B, Dalle, P. (2005). A tool for sign language analysis through signing space representation. In *Proc. sign language linguistics and application of info. technology to sign languages, Milan, Italy.*

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

Lu, P, Huenerfauth, M. (2009). Accessible motion-capture glove calibration protocol for recording sign language data from deaf subjects. In *Proc. ACM SIGACCESS Conference*, pp. 83-90.

Meier, R. (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.

Neidle, C, Kegl, J, MacLaughlin, D, Bahan, B, Lee, R. (2000). *The syntax of ASL: functional categories and hierarchical structure*. Cambridge, MA: MIT Press.

Padden, C. (1988). *Interaction of morphology and syntax in American Sign Language*. New York: Garland.

Traxler, C. (2000). The Stanford achievement test, ninth edition: national norming and performance standards for deaf and hard-of-hearing students. *J Deaf Studies & Deaf Education* 5(4), pp. 337-348.

VCom3D. 2010. Sign Smith Studio. http://www.vcom3d.com/signsmith.php. Accessed 11 March 2010.