American Sign Language Generation: Multimodal NLG with Multiple Linguistic Channels

Matt Huenerfauth
Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104 USA
matt@huenerfauth.com

Abstract

Software to translate English text into American Sign Language (ASL) animation can improve information accessibility for the majority of deaf adults with limited English literacy. ASL natural language generation (NLG) is a special form of multimodal NLG that uses multiple linguistic output channels. ASL NLG technology has applications for the generation of gesture animation and other communication signals that are not easily encoded as text strings.

1 Introduction and Motivations

American Sign Language (ASL) is a full natural language – with a linguistic structure distinct from English – used as the primary means of communication for approximately one half million deaf people in the United States (Neidle et al., 2000, Liddell, 2003; Mitchell, 2004). Without aural exposure to English during childhood, a majority of deaf U.S. high school graduates (age 18) have only a fourth-grade (age 10) English reading level (Holt, 1991). Technology for the deaf rarely addresses this literacy issue; so, many deaf people find it difficult to read text on electronic devices. Software for translating English text into animations of a computer-generated character performing ASL can make a variety of English text sources accessible to the deaf, including: TV closed captioning, teletype telephones, and computer user-interfaces (Huenerfauth, 2005). Machine translation (MT) can also be used in educational software for deaf children to help them improve their English literacy skills.

This paper describes the design of our English-to-ASL MT system (Huenerfauth, 2004a, 2004b, 2005), focusing on ASL generation. This overview illustrates important correspondences between the problem of ASL natural language generation (NLG) and related research in Multimodal NLG.

1.1 ASL Linguistic Issues

In ASL, several parts of the body convey meaning in parallel: hands (location, orientation, shape), eye gaze, mouth shape, facial expression, head-tilt, and shoulder-tilt. Signers may also interleave lexical signing (LS) with classifier predicates (CP) during a performance. During LS, a signer builds ASL sentences by syntactically combining ASL lexical items (arranging individual signs into sentences). The signer may also associate entities under discussion with locations in space around their body; these locations are used in pronominal reference (pointing to a location) or verb agreement (aiming the motion path of a verb sign to/from a location).

During CPs, signers' hands draw a 3D scene in the space in front of their torso. One could imagine invisible placeholders floating in front of a signer representing real-world objects in a scene. To represent each object, the signer places his/her hand in a special handshape (used specifically for objects of that semantic type: moving vehicles, seated animals, upright humans, etc.). The hand is moved to show a 3D location, movement path, or surface contour of the object being described. For example, to convey the English sentence "the car parked next to the house," signers would indicate a location in space to represent the house using a special handshape for 'bulky objects.' Next, they would use a 'moving vehicle' handshape to trace a 3D path for the car which stops next to the house.

Proceedings of the ACL Student Research Workshop, pages 37–42,
Ann Arbor, Michigan, June 2005. ©2005 Association for Computational Linguistics

1.2 Previous ASL MT Systems

There have been some previous English-to-ASL MT projects – see survey in (Huenerfauth, 2003). Amid other limitations, none of these systems address how to produce the 3D locations and motion paths needed for CPs. A fluent, useful English-to-ASL MT system cannot ignore CPs. ASL sign-frequency studies show that signers produce a CP from 1 to 17 times per minute, depending on genre (Morford and MacFarlane, 2003). Further, it is those English sentences whose ASL translation uses a CP that a deaf user with low English literacy would need an MT system to translate. These English sentences look structurally different than their ASL CP counterpart – often making the English sentence difficult to read for a deaf user.

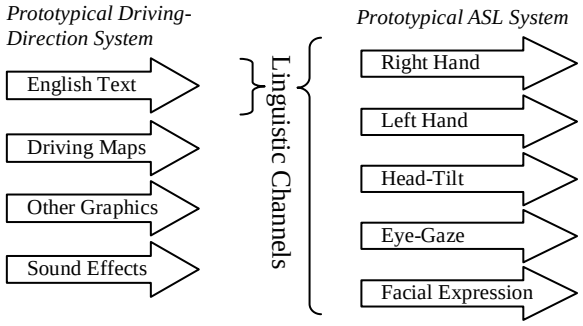
2 ASL NLG: A Form of Multimodal NLG

NLG researchers think of communication signals in a variety of ways: some as a written text, other as speech audio (with prosody, timing, volume, and intonation), and those working in Multimodal NLG as text/speech with coordinated graphics (maps, charts, diagrams, etc). Some Multimodal NLG focuses on "embodied conversational agents" (ECAs), computer-generated animated characters that communicate with users using speech, eye gaze, facial expression, body posture, and gestures (Cassell et al., 2000; Kopp et al., 2004).

The output of any NLG system could be represented as a stream of values (or features) that change over time during a communication signal; some NLG systems specify more values than others. Because the English writing system does not record a speaker's prosody, facial expression or gesture[1], a text-based NLG system specifies fewer communication stream values in its output than does a speech-based or ECA system. A text-based NLG system requires *literate* users, to whom it can transfer some of the processing burden; they must mentally reconstruct more of the language performance than do users of speech or ECA systems.

Since most writing systems are based on strings, text-based NLG systems can easily encode their output as a single stream, namely a sequence of words/characters. To generate more complex signals, multimodal systems decompose their output into several sub-streams – we'll refer to these as "channels." Dividing a communication signal into channels can make it easier to represent the various choices the generator must make; generally, a different processing component of the system will govern the output of each channel. The trade-off is that these channels must be coordinated over time.

Figure 1: Linguistic Channels in Multimodal Systems



Instead of thinking of channels as dividing a communication signal, we can think of them as groupings of individual values in the data stream that are related in some way. The channels of a multimodal NLG system generally correspond to natural perceptual/conceptual groupings called "modalities." Coarsely, audio and visual parts of the output are thought of as separate modalities. When parts of the output appear on different portions of the display, then they are also generally considered separate modalities. For instance, a multimodal NLG system for automobile driving directions may have separate processing channels for text, maps, other graphics, and sound effects. An ECA system may have separate channels for eye gaze, facial expression, manual gestures, and speech audio of the animated character.

When a language has no commonly-known writing system – as is the case for ASL – then it's not possible to build a text-based NLG system. We must produce an animation of a character (like an ECA) performing ASL; so, we must specify how the hands, eye gaze, mouth shape, facial expression, head-tilt, and shoulder-tilt are coordinated over time. With no conventional string-encoding of ASL (that would compress the signal into a single stream), an ASL signal is spread over multiple channels of the output – a departure from most Multimodal NLG systems, which have a single linguistic channel/modality that is coordinated with other non-linguistic resources (Figure 1).

[1] Some punctuation marks loosely correspond to intonation or pauses, but most prosodic information is lost. Facial expression and gesture is generally not conveyed in writing, except perhaps for the occasional use of "emoticons." ;-)

Of course, we could invent a string-based notation for ASL so that we could use traditional text-based NLG technology. (Since ASL has no writing system, we would have to invent an artificial notation.) Unfortunately, since the users of the system wouldn't be trained in this new writing system, it could not be used as output; we would still need to generate a multimodal animation output. An artificial writing system could only be used for internal representation and processing, However, flattening a naturally multichannel signal into a single-channel string (prior to generating a multichannel output) can introduce its own complications to the ASL system's design. For this reason, this project has been exploring ways to represent the hierarchical linguistic structure of information on multiple channels of ASL performance (and how these structures are coordinated or uncoordinated across channels over time).

Some multimodal systems have explored using linguistic structures to control (to some degree) the output of multiple channels. Research on generating animations of a speaking ECA character that performs meaningful gestures (Kopp et al., 2004) has similarities to this ASL project. First of all, the channels in the signal are basically the same; an animated human-like character is shown onscreen with information about eye, face, and arm movements being generated. However, an ASL system has no audio speech channel but potentially more fine-grained channels of detailed body movement.

The less superficial similarity is that (Kopp et. al, 2004) have attempted to represent the semantic meaning of some of the character's gestures and to synchronize them with the speech output. This means that, like an ASL NLG system, several channels of the signal are being governed by the linguistic mechanisms of a natural language. Unlike ASL, the gesture system uses the speech audio channel to convey nearly all of the meaning to the user; the other channels are generally used to convey additional/redundant information. Further, the internal structure of the gestures is not generally encoded in the system; they are typically atomic/lexical gesture events which are synchronized to co-occur with portions of speech output. A final difference is that gestures which co-occur with English speech (although meaningful) can be somewhat vague and are certainly less systematic and conventional than ASL body movements. So, while both systems may have multiple linguistic channels, the gesture system still has one primary linguistic channel (audio speech) and a few channels controlled in only a partially linguistic way.

3 This English-to-ASL MT Design

The linguistic and multimodal issues discussed above have had important consequences on the design of our English-to-ASL MT system. There are several unique features of this system caused by: (1) ASL having multiple linguistic channels that must be coordinated during generation, (2) ASL having both an LS and a CP form of signing, (3) CP signing visually conveying 3D spatial relationships in front of the signer's torso, and (4) ASL lacking a conventional written form. While ASL-particular factors influenced this design, section 5 will discuss how this design has implications for NLG of traditional written/spoken languages.

3.1 Coordinating Linguistic Channels

Section 2 mentioned that this project is developing multichannel (non-string) encodings of ASL animation; these encodings must coordinate multiple channels of the signal as they are generated by the linguistic structures and rules of ASL. Kopp et al. (2004) have explored how to coordinate meaningful gestures with speech signal during generation; however, their domain is somewhat simpler. Their gestures are atomic events without internal hierarchical structure. Our project is currently developing grammar-like coordination formalisms that allow complex linguistic signals on multiple channels to be conveniently represented.[2]

3.2 ASL Computational Linguistic Models

This project uses representations of discourse, semantics, syntax, and (sign) phonology tailored to ASL generation (Huenerfauth, 2004b). In particular, since this MT system will generate animations of classifier predicates (CPs), the system consults a 3D model of real-world scenes under discussion. Further, since multimodal NLG requires a form of scheduling (events on multiple channels are coordinated over a performance timeline), all of the linguistic models consulted and modified during ASL generation are time-indexed according to a timeline of the ASL performance being produced.

[2] Details of this work will be described in future publication.

Previous ASL phonological models were designed to represent non-CP ASL, but CPs use a reduced set of handshapes, standard eye-gaze and head-tilt patterns, and more complex orientations and motion paths. The phonological model developed for this system makes it easier to specify CPs.

Because ASL signers can use the space in front of their body to visually convey information, it is possible during CPs to show the exact 3D layout of objects being discussed. (The use of channels representing the hands means that we can now indicate 3D visual information – not possible with speech or text.) To represent this 3D detailed form of meaning, this system has an unusual *semantic* model for generating CPs. We populate the volume of space around the signer's torso with invisible 3D objects representing entities discussed by CPs being generated (Huenerfauth, 2004b). The semantic model is the set of placeholders around the signer (augmented with the CP handshape used for each). Thus, the semantics of the "car parked next to the house" example (section 1.1) is that a 'bulky' object occupies a particular 3D location and a 'vehicle' object moves toward it and stops.

Of course, the system will also need more traditional semantic representations of the information to be conveyed during generation, but this 3D model helps the system select the proper 3D motion paths for the signers' hands when "drawing" the 3D scenes during CPs. The work of (Kopp et al., 2004) studies gestures to convey spatial information during an English speech performance, but unlike this system, they use a logical-predicate-based semantics to represent information about objects referred to by gesture. Because ASL CPs indicate 3D layout in a linguistically conventional and detailed way, we use an actual 3D model of the objects being discussed. Such a 3D model may also be useful for ECA systems that wish to generate more detailed 3D spatial gesture animations.

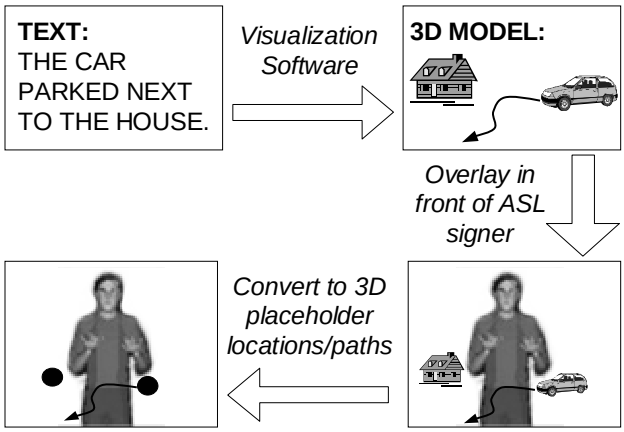
The discourse model in this ASL system records features not found in other NLG systems. It tracks whether a 3D location has been assigned to each discourse entity, where that location is around the signer, and whether the latest location of the entity has been indicated by a CP. The discourse model is not only relevant during CP performance; since ASL LS performance also assigns 3D locations to objects under discussion (for pronouns and verbal agreement), this model is also used for LS.

3.3 Generating 3D Classifier Predicates

An essential step in producing an animation of an ASL CP is the selection of 3D motion paths for the computer-generated signer's hands, eye gaze, and head tilt. The motion paths of objects in the 3D model described above are used to select corresponding motion paths for these parts of the signer's body during CPs. To build the 3D placeholder model, this system uses preexisting scene-visualization software to analyze an English text describing the motion of real-world objects and build a 3D model of how the objects mentioned in text are arranged and move (Huenerfauth, 2004b). This model is "overlaid" onto the volume in front of the ASL signer (Figure 2). For each object in the scene, a corresponding invisible placeholder is positioned in front of the signer; the layout of placeholders mimics the layout of objects in the 3D scene. In the "car parked next to the house" example, a miniature invisible object representing a 'house' is positioned in front of the signer's torso, and another object (with a motion path terminating next to the 'house') is added to represent the 'car.'

The locations and orientations of the placeholders are later used by the system to select the locations and orientations for the signer's hands while performing CPs about them. So, the motion path calculated for the car will be the basis for the 3D motion path of the signer's hand during the classifier predicate describing the car's motion. Given the information in the discourse/semantic models, the system generates the hand motions, head-tilt, and eye-gaze for a CP. It stores a library containing templates representing a prototypical form of each CP the system can produce. The templates

Figure 2: Converting English Text to 3D Placeholder



are planning operators (with logical pre-conditions, monitored termination conditions, and effects), allowing the system to "trigger" other elements of ASL signing performance that may be required during a CP. A planning-based NLG approach, described in (Huenerfauth, 2004b), is used to select a template, fill in its missing parameters, and build a schedule of the animation events on multiple channels needed to produce a sequence of CPs.

3.4 A Multi-Path Architecture

A multimodal NLG system may have several presentation styles it could use to convey information to its user; these styles may take advantage of the various output channels to different degrees. In ASL, there are multiple channels in the linguistic portion of the signal, and not surprisingly, the language has multiple sub-systems of signing that take advantage of the visual modality in different ways. ASL signers can select whether to convey information using lexical signing (LS) or classifier predicates (CPs) during an ASL performance (section 1.1). These two sub-systems use the space around the signer differently; during CPs, locations in space associated with objects under discussion must be laid out in a 3D manner corresponding to the topological layout of the real-world scene under discussion. Locations associated with objects during LS (used for pronouns and verb agreement) have no topological requirement. The layout of the 3D locations during LS may be arbitrary.

The CP generation approach in section 3.3 is computationally expensive; so, we would only like to use this processing pathway when necessary. English input sentences not producing classifier predicates would not need to be processed by the visualization software; in fact, most of these sentences could be handled using the more traditional MT technologies of previous systems. For this reason, our English-to-ASL MT system has multiple processing pathways (Huenerfauth, 2004a). The pathway for handling English input sentences that produce CPs includes the scene visualization software, while other input sentences undergo less sophisticated processing using a traditional MT approach (that is easier to implement). In this way, our CP generation component can actually be layered on top of a pre-existing English-to-ASL MT system to give it the ability to produce CPs. This multi-path design is equally applicable to the architecture of written-language MT systems. The design allows an MT system to combine a resource-intensive deep-processing MT method for difficult (or important) inputs and a resource-light broad-coverage MT method for other inputs.

3.5 Evaluation of Multichannel NLG

The lack of an ASL writing system and the multichannel nature of ASL can make NLG or MT systems which produce ASL animation output difficult to evaluate using traditional automatic techniques. Many such approaches compare a string produced by a system to some human-produced 'gold-standard' string. While we could invent an artificial ASL writing system for the system to produce as output, it's not clear that human ASL signers could accurately or consistently produce written forms of ASL sentences to serve as 'gold standards' for such an evaluation. And of course, real users of the system would never be shown artificial "written ASL"; they would see full animations instead. User-based studies (where ASL signers evaluate animation output directly) may be a more meaningful measure of an ASL system.

We are planning such an evaluation of a prototype CP-generation module of the system during the summer/fall of 2005. Members of the deaf community who are native ASL signers will view animations of classifier predicates produced by the system. As a control, they will also be shown animations of CPs produced using 3D motion capture technology to digitally record the performance of CPs by other native ASL signers. Their evaluation of animations from both sources will be compared to measure the system's performance. The multichannel nature of the signal also makes other interesting experiments possible. To study the system's ability to animate the signer's hands only, motion-captured ASL could be used to animate the head/body of the animated character, and the NLG system can be used to control only the hands of the character. Thus, channels of the NLG system can be isolated for evaluation – an experimental design only available to a multichannel NLG system.

4 Unique Design Features for ASL NLG

The design portion of this English-to-ASL project is nearly complete, and the implementation of the system is ongoing. Evaluations of the system will

be available after the user-based study discussed above; however, the design itself has highlighted interesting issues about the requirements of NLG software for sign languages like ASL.

The multichannel nature of ASL has led this project to study mechanisms for coordinating the values of the linguistic models used during generation (including the output animation specification itself). The need to handle both the LS and CP subsystems of the language has motivated: a multi-path MT architecture, a discourse model that stores data relevant to both subsystems, a model of the space around the signer capable of storing both LS and CP placeholders, and a phonological model whose values can be specified by either subsystem.

Since this English-to-ASL MT system is the first to address ASL classifier predicates, designing an NLG process capable of producing the 3D locations and paths in a CP animation has been a major design focus for this project. These issues have been addressed by the system's use of a 3D model of placeholders produced by scene-visualization software and a planning-based NLG process operating on templates of prototypical CP performance.

5 Applications Beyond Sign Language

Sign language NLG requires 3D spatial representations and multichannel coordinated output, but it's not unique in this requirement. In fact, generation of a communication signal for any language may require these capabilities (even for spoken languages like English). We have mentioned throughout this paper how gesture/speech ECA researchers may be interested in NLG technologies for ASL – especially if they wish to produce gestures that are more linguistically conventional, internally complex, or 3D-topologically precise.

Many other computational linguistic applications could benefit from an NLG design with multiple linguistic channels (and indirectly benefit from ASL NLG technology). For instance, NLG systems producing speech output could encode prosody, timing, volume, intonation, or other vocal data as multiple linguistically-determined channels of the output (in addition to a channel for the string of words being generated). And so, ASL NLG research not only has exciting accessibility benefits for deaf users, but it also serves as a research vehicle for NLG technology to produce a variety of richer-than-text linguistic communication signals.

Acknowledgments

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

References

Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (Eds.). 2000. Embodied Conversational Agents. Cambridge, MA: MIT Press.

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

Huenerfauth, M. 2003. Survey and Critique of ASL Natural Language Generation and Machine Translation Systems. Technical Report MS-CIS-03-32, Computer and Information Science, University of Pennsylvania.

Huenerfauth, M. 2004a. A Multi-Path Architecture for Machine Translation of English Text into American Sign Language Animation. In Proceedings of the Student Workshop of the Human Language Technologies conference / North American chapter of the Association for Computational Linguistics annual meeting: HLT/NAACL 2004, Boston, MA, USA.

Huenerfauth, M. 2004b. Spatial and Planning Models of ASL Classifier Predicates for Machine Translation. 10th International Conference on Theoretical and Methodological Issues in Machine Translation: TMI 2004, Baltimore, MD.

Huenerfauth, M. 2005. American Sign Language Spatial Representations for an Accessible User-Interface. In 3rd International Conference on Universal Access in Human-Computer Interaction. Las Vegas, NV, USA.

Kopp, S., Tepper, P., and Cassell, J. 2004. Towards Integrated Microplanning of Language and Iconic Gesture for Multimodal Output. Int'l Conference on Multimodal Interfaces, State College, PA, USA.

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

Mitchell, R. 2004. How many deaf people are there in the United States. Gallaudet Research Institute, Grad School & Prof. Progs. Gallaudet U. June 28, 2004. http://gri.gallaudet.edu/Demographics/deaf-US.php

Morford, J., and MacFarlane, J. 2003. Frequency Characteristics of ASL. Sign Language Studies, 3:2.

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