Spatial Representation of Classifier Predicates for Machine Translation into American Sign Language

Matt Huenerfauth

Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
matthewh@seas.upenn.edu

Abstract

The translation of English text into American Sign Language (ASL) animation tests the limits of traditional machine translation (MT) approaches. The generation of spatially complex ASL phenomena called "classifier predicates" motivates a new representation for ASL based on virtual reality modeling software, and previous linguistic research provides constraints on the design of an English-to-Classifier-Predicate translation process operating on this representation. This translation design can be incorporated into a multi-pathway architecture to build English-to-ASL MT systems capable of producing classifier predicates.

Introduction and Motivations

Although Deaf students in the U.S. and Canada are taught written English, the challenge of acquiring a spoken language for students with hearing impairments results in the majority of Deaf U.S. high school graduates reading at a fourth-grade[1] level (Holt, 1991). Unfortunately, many strategies for making elements of the hearing world accessible to the Deaf (e.g. television closed captioning or teletype telephone services) assume that the user has strong English literacy skills. Since many Deaf people who have difficulty reading English possess stronger fluency in American Sign Language (ASL), an automated English-to-ASL machine translation (MT) system can make more information and services accessible in situations where English captioning text is at too high a reading level or a live interpreter is unavailable.

Previous English-to-ASL MT systems have used 3D graphics software to animate a virtual human character to perform ASL output. Generally, a script written in a basic animation instruction set controls the character's movement; so, MT systems must translate English text into a script directing the character to perform ASL. Previous projects have either used word-to-sign dictionaries to produce English-like manual signing output, or they have incorporated analysis grammar and transfer rules to produce ASL output (Huenerfauth, 2003; Sáfár and Marshall, 2001; Speers, 2001; Zhao et al., 2000). While most of this ASL MT work is still preliminary, there is promise that an MT system will one day be able to translate many kinds of English-to-ASL sentences; although, some particular ASL phenomena – those involving complex use of the signing space – have proven difficult for traditional MT approaches. This paper will present a design for generating these expressions.

ASL Spatial Phenomena

ASL signers use the space around them for several grammatical, discourse, and descriptive purposes. During a conversation, an entity under discussion (whether concrete or abstract) can be "positioned" at a point in the signing space. Subsequent pronominal reference to this entity can be made by pointing to this location (Neidle et al., 2000). Some verb signs will move toward or away from these points to indicate (or show agreement with) their arguments (Liddell, 2003a; Neidle et al., 2000). Generally, the locations chosen for this use of the signing space are not topologically meaningful; that is, one imaginary entity being positioned to the left of another in the signing space doesn't necessarily indicate the entity is to the left of the other in the real world.

Other ASL expressions are more complex in their use of space and position invisible objects around the signer to topologically indicate the arrangement of entities in a 3D scene being discussed. Constructions called "classifier predicates" allow signers to use their hands to position, move, trace, or re-orient an imaginary object in the space in front of them to indicate the location, movement, shape, contour, physical dimension, or some other property of a corresponding real world entity under discussion. Classifier predicates consist of a semantically meaningful handshape and a 3D hand movement path. A handshape is chosen from a closed set based on characteristics of the entity described (whether it is a vehicle, human, animal, etc.) and what aspect of the entity the signer is describing (surface, position, motion, etc).

For example, the sentence "the car drove down the bumpy road past the cat" could be expressed in ASL using two classifier predicates. First, a signer would move a hand in a "bent V" handshape (index and middle fingers extended and bent slightly) forward and slightly downward to a point in space in front of his or her torso where an imaginary miniature cat could be envisioned. Next, a hand in a "3" handshape (thumb, index, middle fingers extended with the thumb pointing upwards) could trace a path in space past the "cat" in an up-and-down fashion as if it were a car bouncing along a bumpy road. Generally, "bent V" handshapes are used for animals, and "3" handshapes, for vehicles.

Generating Classifier Predicates

As the "bumpy road" example suggests, translation involving classifier predicates is more complex than most English-to-ASL MT because of the highly productive and spatially representational nature of these signs. Previous ASL MT systems have dealt with this problem by omitting these expressions from their

[1] Students who are age eighteen and older are reading English text at a level more typical of a ten-year-old student.

linguistic coverage; however, many English concepts lack a fluent ASL translation without them. Further, these predicates are common in ASL; in many genres, signers produce a classifier predicate on average once per 100 signs (this is approximately once per minute at typical signing rates) (Morford and MacFarlane, 2003). So, systems that cannot produce classifier predicates can only produce ASL of limited fluency and are not a viable long-term solution to the English-to-ASL MT problem.

Classifier predicates challenge traditional definitions of what constitutes linguistic expression, and they oftentimes incorporate spatial metaphor and scene-visualization to such a degree that there is debate as to whether they are paralinguistic spatial gestures, non-spatial polymorphemic constructions, or compositional yet spatially-parameterized expressions (Liddell, 2003b). No matter their true nature, an ASL MT system must somehow generate classifier predicates. While MT designs are not required to follow linguistic models of human language production in order to be successful, it is worthwhile to consider linguistic models that account well for the ASL classifier predicate data but minimize the computational or representational overhead required to implement them.

Design Focus and Assumptions

This paper will focus on the generation of classifier predicates of movement and location (Supalla, 1982; Liddell, 2003a). Most of the discussion will be about generating individual classifier predicates; an approach for generating multiple interrelated predicates will be proposed toward the end of the paper.

This paper will assume that English input sentences that should be translated into ASL classifier predicates can be identified. Some of the MT designs proposed below will be specialized for the task of generating these phenomena. Since a complete MT system for English-to-ASL would need to generate more than just classifier predicates, the designs discussed below would need to be embedded within an MT system that had other processing pathways for handling non-spatial English input sentences. The design of such multi-pathway MT architectures is another focus of this research project (Huenerfauth, 2004).

These other pathways could handle most inputs by employing traditional MT technologies (like the ASL MT systems mentioned above). A sentence could be “identified” (or intercepted) for special processing in the classifier predicate pathway if it fell within the pathway’s implemented lexical (and – for some designs – spatial) resources.[2] In this way, a classifier predicate generation component could actually be built on top of an existing ASL MT system that didn't currently support classifier predicate expressions.

We will first consider a classifier predicate MT approach requiring little linguistic processing or novel ASL representations, namely a fully lexicalized approach. As engineering limitations are identified or additional linguistic analyses are considered, the design will be modified, and progressively more sophisticated representations and processing architectures will emerge.

Design 1: Lexicalize the Movement Paths

The task of selecting the appropriate handshape for a classifier predicate, while non-trivial, seems approachable with a lexicalized design. For example, by storing semantic features (e.g. +human, +vehicle, +animal, +flat-surface) in the English lexicon, possible handshapes can be identified for entities referred to by particular English nouns. Associating other features (e.g. +motion-path, +stationary-location, +relative-locations, +shape-contour) with particular verbs or prepositions in the English lexicon could help identify what kind of information the predicate must express – further narrowing the set of possible classifier handshapes. To produce the 3D movement portion of the predicate using this lexicalized approach, we could store a set of 3D coordinates in the English lexicon for each word or phrase (piece of lexicalized syntactic structure) that may be translated as a classifier predicate.

Problems with This Design

Unfortunately, the highly productive and scene-specific nature of these signs makes them potentially infinite in number. For example, while it may seem possible to simply store a 3D path with the English phrase "driving up a hill," factors like the curve of the road, steepness of hill, how far up to drive, etc. would affect the final output. So, a naïve lexicalized 3D-semantics treatment of classifier movement would not be scalable.

Design 2: Compose the Movement Paths

Since the system may need to produce innumerable possible classifier predicates, we can't merely treat the movement path as an unanalyzable whole. A more practical design would compose a 3D path based on some finite set of features or semantic elements from the English source text. This approach would need a library of basic animation components that could be combined to produce a single classifier predicate movement. Such an “animation lexicon” would contain common positions in space, relative orientations of objects in space (for concepts like above, below, across from), common motion paths, or common contours for such paths. Finally, these components would be associated with corresponding features or semantic elements of English so that the appropriate animation components can be selected and combined at translation time to produce a 3D path.

Problems with This Design

This design is analogous to the polymorphemic model of classifier predicate generation (Supalla 1978, 1982, 1986). This model describes ASL classifier predicates as categorical, and it characterizes their generation as a process of combining sets of spatially semantic morphemes. The difficulty is that every piece of spatial information we might express with a classifier predicate must be encoded as a morpheme. These phenomena can convey such a wide variety of spatial

[2] A later section of this paper describes how the decision of whether an input English sentence can be processed by the special classifier predicate translation pathway depends on whether a *motif* (introduced in that section) has been implemented for the semantic domain of that sentence.

information – especially when used in combination to describe spatial relationships or comparisons between objects in a scene – that many morphemes are required.

Liddell's analysis (2003b) of the polymorphemic model indicates that in order to generate the variety of classifier predicates seen in ASL data, the model would need a tremendously large (and possibly infinite) number of morphemes. Using a polymorphemic analysis, Liddell (2003b) decomposes a classifier predicate of one person walking up to another, and he finds over 28 morphemes, including some for: two entities facing each other, being on the same horizontal plane, being vertically oriented, being freely moving, being a particular distance apart, moving on a straight path, etc.

Liddell considers classifier predicates as being continuous and somewhat gestural in nature (2003a), and this partially explains his rejection of the model. (If there are not a finite number of possible sizes, locations, and relative orientations for objects in the scene, then the number of morphemes needed becomes infinite.) Whether classifier predicates are continuous or categorical and whether this number of morphemes is infinite or finite, the number would likely be intractably large for an MT system to process. We will see that the final classifier predicate generation design proposed in this paper will use a non-categorical approach for selecting its 3D hand locations and movements. This should not be taken as a linguistic claim about human ASL signers (who may indeed use the large numbers of morphemes required by the polymorphemic model) but rather as a tractable engineering solution to the highly productive nature of classifier predicates.

Another reason why a polymorphemic approach to classifier predicate generation would be difficult to implement in a computational system is that the complex spatial interactions and constraints of a 3D scene would be difficult to encode in a set of compositional rules. For example, consider the two classifier predicates in the "the car drove down the bumpy road past the cat" example. To produce these predicates, the signer must know how the scene is arranged including the locations of the cat, the road, and the car. A path for the car must be chosen with beginning/ending positions, and the hand must be articulated to indicate the contour of the path (e.g. bumpy, hilly, twisty). The proximity of the road to the cat, the plane of the ground, and the curve of the road must be selected. Other properties of the objects must be known: (1) cats generally sit on the ground and (2) cars generally travel along the ground on roads. The successful translation of the English sentence into these two classifier predicates involved a great deal of semantic understanding, spatial knowledge, and reasoning.

A 3D Spatial Representation for ASL MT

ASL signers using classifier predicates handle these complexities using their own spatial knowledge and reasoning and by visualizing the elements of the scene. An MT system may also benefit from a 3D representation of the scene from which it could calculate the movement paths of classifier predicates. While design 2 needed compositional rules (and associated morphemes) to cover every possible combination of object positions and spatial implications as suggested by English texts, the third and final MT design (discussed in a later section) will use virtual reality 3D scene modeling software to simulate the movement and location of entities described by an English text (and to automatically manage their interactions).

The AnimNL System

A system for producing a changing 3D virtual reality representation of a scene from an English text has already been implemented: the Natural Language Instructions for Dynamically Altering Agent Behaviors system (Schuler, 2003; Bindiganavale et al., 2000; Badler et al., 2000) (herein, "AnimNL"). The system displays a 3D animation and accepts English input text containing instructions for the characters and objects in the scene to follow. It updates the virtual reality so that objects obey the English commands. AnimNL has been used in military training and equipment repair domains and can be extended by augmenting its library of Parameterized Action Representations (PARs), to cover additional domains of English input texts.

The system's ability to interact with language and plan future actions arises from the use of PARs, which can be thought of as animation/linguistic primitives for structuring the movements in a 3D scene. PARs are feature-value structures that have slots specifying: what agent is moving, the path/manner of this motion, whether it is translational/rotational motion, the terminating conditions on the motion, any speed or timing data, etc. A single locomotion event may contain several sub-movements or sub-events, and for this reason, PARs may be defined in a hierarchical manner. A single "high-level" PAR may specify the details for the entire motion, but it may be defined in terms of several "low-level" PARs which specify the more primitive sub-movements/events.

The system stores a database of PAR templates that represent prototypical actions the agent can perform. These templates are missing particular details (some of their slots aren't filled in) about the position of the agent or other entities in the environment that would affect how the animation action should really be performed in particular situations. By parameterizing PARs on the 3D coordinates of the objects participating in the movement, the system can produce animations specific to particular scene configurations and reuse common animation code.

English lexicalized syntactic structures are associated with PARs so that the analysis of a text is used to select a PAR template and fill some of its slots. For example, there may be a PAR associated with the concept of "falling" vs. another for "jumping." While these templates must remain parameterized on the 3D location of the agent of the movement until it is known at run time, there are some properties (in this case, the direction of motion) that can be specified for each from the English semantics. During analysis of the English input text, semantic features of motion verbs are obtained from the VerbNet hierarchy (Kipper et al., 2004), and these features are also used to select and fill a particular motion template. Since VerbNet groups verbs that share common semantic/syntactic properties, AnimNL is able to link an entire set of semantically similar motion verbs to a single PAR template. Each of the verbs in the set may fill some of the slots of the motion template somewhat differently.

When a PAR template has been partially filled with information from the English text and 3D object locations, it is passed off to AnimNL's animation planner. In fact, PARs contain slots allowing them to be hierarchical planning operators: pre-conditions, effects, subplans, etc. The movements of all objects in the AnimNL system are governed by a planning process, which allows the objects in the scene to move realistically. Many spatial motions have conditions on the location, orientation, or motion state of an object and its environment before, during, and after the event. The PAR operators help the system work out the details of an animation from the limited specification of this motion provided by an English text. For example, it may determine starting and stopping locations for movement paths or select relative locations for objects in the 3D scene based on prepositions and adverbials in the English input text. The interaction and conditions of these planning operators simulate physical constraints, collision avoidance, human anatomical limitations, and other factors to produce an animation.

Using AnimNL for ASL

The MT system's classifier predicate generator can use the AnimNL software to analyze English sentences to be translated into classifier predicates. AnimNL can process this text as if it were commands for the entities mentioned in the text to follow. Based on this analysis, the AnimNL can create and maintain a 3D representation of the location and motion of these entities. Next, a miniature virtual reality animation of the objects in this representation can be overlaid on a volume of the space in front of the torso of the animated ASL-signing character. In this way, a miniature 3D virtual reality would be embedded within the original 3D space containing the standing animated virtual human. In the "bumpy road" example, a small invisible object would be positioned in space in front of the chest of the signing character to represent the cat. Next, a 3D animation path and location for the car (relative to the cat) would be chosen in front of the character's chest.

The AnimNL software can thus produce a miniature "invisible world" representing the scene described by the input text. Unlike other applications of AnimNL – where entities described by the English text would need to be rendered to the screen – in this situation, the 3D objects would be transparent. Therefore, the MT system does not care about the exact appearance of the objects being modeled. Only the location, orientation, and motion paths of these objects in some generic 3D space are important since this information will be used to produce classifier predicates for the animated ASL-signing character.

An Overly Simplistic Generation Strategy

The next section of this paper (design 3) will discuss how the "invisible world" representation can be used to generate classifier predicates. To motivate that third and final design, we will first consider an overly simplistic (and incorrect) strategy for using the virtual reality to attempt classifier predicate generation.

This simplistic "Directly Pictorial" strategy for building a classifier predicate is as follows: When a new object is introduced into the invisible world, the signing character moves its hand to a location "inside of" the transparent object. By also choosing an appropriate handshape for the character (possibly using the +animal or +vehicle features discussed above), then a classifier predicate is apparently produced that conveys the spatial information from the English text. As objects in the invisible world are moved or reoriented as AnimNL analyzes more text, the signer can express this information using additional classifier predicates by again placing its hand inside the (possibly moving) 3D object. (See Figure 1.)

Limitations of the "Directly Pictorial" Strategy

Whereas design 2 mirrored the polymorphemic model, this design is similar to that of DeMatteo (1977), who sees classifier predicates as being direct "spatial analogues" of 3D movement paths in a scene imagined by the signer (Liddell, 2003b). In this model, signers maintain a 3D mental image of a scene to be described, select appropriate handshapes to refer to entities in their model, and trace out topologically analogous location and movement paths for these entities using their hands.

Unfortunately, the model is over-generative (Liddell, 2003b). By assuming that the selection of handshapes and movements are orthogonal and that movement paths are directly representative [3] of the paths of entities in space, this analysis predicts many ASL classifier constructions that never appear in the data (containing imaginable but ungrammatical combinations of handshape, orientation, and movement) (Liddell, 2003b). Finally, the model cannot consider discourse and non-spatial semantic features that can influence classifier predicate production in ASL.

Design 3: Lexicon of Classifier Predicates

The "Directly Pictorial" strategy was just one way to use the 3D information in the invisible world representation to generate classifier predicates. This section will introduce the MT approach advocated by this paper: design 3. This design uses the invisible world but avoids the limitations of the previous strategy by considering additional sources of information during translation. Whereas previous sections of this paper have used comparisons to linguistic models to critique an MT design, this section will use a linguistic model for inspiration.

Lexicon of Classifier Predicate Templates

Liddell (2003a, 2003b) proposed that ASL classifier predicates are stored as large numbers of abstract templates in a lexicon. They are "abstract" in the sense that each is a template parameterized on 3D coordinates of whatever object is being described, and each can therefore be instantiated into many possible

[3] To illustrate how classifier predicate movements can be conventional and not visually representative, Liddell (2003b) uses the example of an upright figure walking leisurely being expressed as a classifier predicate with D handshape slightly bouncing as it moves along a path. While the hand bounces, the meaning is not that a human is bouncing but that he or she is walking leisurely.

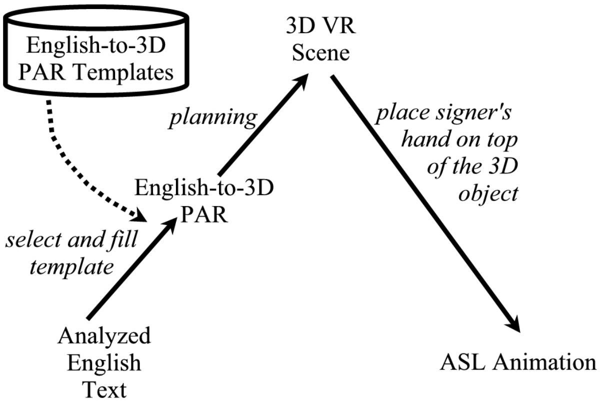

Figure 1: "Directly Pictorial" Generation Strategy (argued against in this paper). Solid lines depict transformation processes between representations, and dotted lines, information flow into a process.

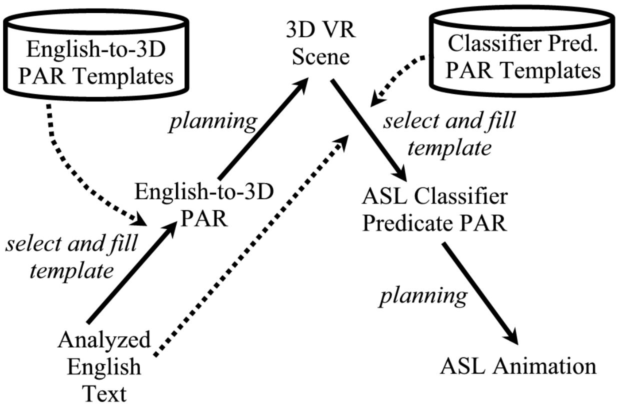

Figure 2: The Design 3 Architecture. Notice the new selection/filling process for a Classifier Predicate PAR based on: a PAR template, the 3D scene data, and English text features.

classifier predicate outputs. For example, there may be one template for classifier predicates expressing that a car is parked at a point in space; when this template is turned into an actual classifier predicate, then the 3D coordinate of the car would be filled in.

Each lexical entry stores the semantic content of a particular classifier predicate and most of the handshape and movement specification for its performance. A signer selects a template based on how well its spatial and non-spatial semantics convey the desired content. When a signer generates a classifier predicate from this template, then the locations, orientations, and specific movement paths of objects in a 3D mental spatial representation are used to fill the remaining parameters of the template and produce a full specification of how to perform the classifier predicate.

Although the previous paragraph refers to this approach as "lexical," it differs from design 1 (which augmented the English lexicon with 3D movement data) because it creates a distinct ASL lexicon of classifier predicates, and the movement information in these entries is parameterized on the data in the 3D scene. While these templates may also resemble the compositional morphemes of the polymorphemic model (the "animation lexicon" of design 2) since they both link semantics to 3D movement, these templates have more pre-compiled structure. While the morphemes required complex processing by compositional rules, the templates just need to be selected and to have their 3D parameters set.

Liddell (2003b) explains that this model avoids the under-generation of (Supalla, 1978, 1982, 1986) by incorporating a 3D spatial representation to select locations and movement paths, but it also avoids the over-generation of (DeMatteo, 1977) by restricting the possible combinations of handshapes and movement paths. Impossible combinations are explained as lexical gaps; ungrammatical classifier predicate feature combinations are simply not entries in the lexicon (Liddell, 2003b).

Classifier Predicate Templates for MT

To implement this linguistic model as an MT design, we will need: (1) a 3D scene representation, (2) a list of templates for producing the signing character's arm movements, (3) a way to link the semantics of English sentences to specific templates, and (4) a method for turning a filled template into an animation of the signer's arm. Requirement 1 is satisfied by the invisible world representation produced by the AnimNL software.

While the AnimNL software used one database of PAR templates to produce the 3D animation of objects in the invisible world, this design can fulfill requirement 2 by adding a second database, whose PAR templates will describe the animated movement of the signing character's arm as it performs a classifier predicate. (This first set will be called "invisible world" PARs, and the second, "classifier predicate" PARs.) Compared to the invisible world PARs, the classifier predicate PARs will be very simple: they will store instructions for the signing character's hand to be in a particular shape and for it move between two or more 3D coordinates in the signing space – possibly along a programmed contour.

The re-use of PAR templates suggests a method for linking the semantics of the English text to arm movement templates (requirement 3). Just as the AnimNL software used features of lexical syntactic structures to trigger invisible world PARs, design 3 can use these features to link the semantics of English sentences to classifier predicate PARs. These features can help select a template and fill some of its non-spatial information slots. Finally, data from the invisible world representation can fill the spatial parameters of the classifier predicate PAR.

Since arm movements are represented as PARs, this design can use a planning process (like that of the AnimNL software) to transform these PARs into a 3D animation script (requirement 4). While the AnimNL's planning process turned invisible world PARs into animations of invisible objects, this planning process will turn classifier predicate PARs into an animation script controlling the movement of the signing character's arm as it produces a classifier predicate. (See Figure 2.)

Generating Multiple Classifier Predicates

Up until now, this paper has focused on generating a single classifier predicate from a single

English sentence, but in fact, the actual English-to-ASL translation problem is more complex. New challenges arise when generating several interrelated classifier predicates to describe a single scene. While specifying a system to generate a single predicate has been a natural starting point (and a first priority), it is important to consider how this architecture would need to be enhanced to handle the production of multiple classifier predicates. If these issues are not considered early in the development process, then software design decisions may be made that would make the MT system difficult to extend.

While the earlier sections of this paper may have suggested that there is always a correspondence between a single English input sentence and a single ASL classifier predicate output, in fact, several classifier predicates may be needed to convey the semantics of one English sentence (or vice versa). Even when the mapping is one-to-one, the classifier predicates may need to be rearranged during translation to reflect the scene organization or ASL conventions on how these predicates are sequenced or combined. For instance, when describing the arrangement of furniture in a room, signers generally sequence their description starting with items to one side of the doorway and then circling across the room back to the doorway again. An English description of a room may be significantly less spatially systematic in its ordering.

Multiple classifier predicates used to describe a single scene may also interact with and constrain one another. The selection of scale, perspective, and orientation of a scene chosen for the first classifier predicate will affect those that follow it. If decisions about the representation of the virtual reality scene are made without considering the requirements of the later classifier predicates, then output may be produced which arranges the elements of the scene in a non-fluent manner. Often the first English sentence describing a 3D scene may not contain enough detail to make all of the choices about the scene layout or perspective. A generation approach that considers the spatial information in adjacent (later) English input sentences prior to making such decisions could produce higher quality ASL output.

Another motivation for making generation decisions for groups of related classifier predicates is that the semantics of multiple classifier predicates may interact to produce emergent meaning. For example, one way to convey that an object is between two others in a scene is to use three classifier predicates: two to locate the elements on each side and then one for the entity in the middle. In isolation, these classifier predicates do not convey any idea of a spatial relationship, but in coordinated combination, this semantic effect is achieved.

Classifier Predicate Motifs

An MT system could handle the translation complexities discussed above by using sets of multi-classifier templates called *motifs*. Instead of immediately triggering one ASL classifier as each sentence of an English text is encountered, now the system will represent collections of multiple interrelated classifier predicate templates that can be used together to describe a scene. These collective structures would allow generation decisions to be made at the scene-level, thus decoupling individual English sentences from individual classifier predicates. The motif structure could decide how many classifiers must be used to communicate some block of spatial information and how to coordinate and arrange them.

A motif would serve as a set of deep generation rules or patterns for constructing a series of ASL classifier predicates in a specific semantic genre – e.g. movement of vehicles, giving directions, furniture arrangement, movements of walking people, etc. While this paper focuses on movement and location predicates, motifs can be imagined for size and shape specifiers (e.g. stripes or spots on clothing), instrument classifiers (e.g. using handtools), and others. Each motif would contain conditional rules for determining when it should be employed, that is, whether a particular English input text is within its genre. Just like the classifier predicate PAR templates in design 3, motifs could be triggered by features of the analyzed English text.[4]

Motifs would use planning rules to select and sequence their component predicates and to choose the best viewpoint, orientation, and scale for the entire scene. Having a separate motif for each genre would allow these planning rules to be specialized for how interrelated classifier predicates communicate spatial semantic information in a particular domain – possibly using genre-specific conventions as in the "furniture arrangement" example. Each motif could translate an English sentence according to its own guidelines; so, the system could translate the same input sentence differently based on the motif genre in which it occurred.

Implementation Issues

We can extend design 3 to generate multiple classifier predicates by adding a database of motif representations to be used in the PAR-planning process. In fact, these multi-predicate motifs could be represented as additional higher-level PAR templates. In the same way that a classifier predicate PAR can be hierarchically decomposed into sub-movements of the signer's arm (each represented by a lower-level PAR), analogously, a PAR representing a multi-predicate motif can be decomposed into PARs for individual classifier predicates. In design 3, English text features immediately triggered a single classifier predicate PAR; now, English features will trigger a PAR representing a motif. During planning, the motif PAR can use English text features and 3D invisible world data to decide how to expand its sub-actions – how to select and arrange the classifier predicates to express it.

Motifs are quite domain-specific in their implementation; so, questions can be raised as to what degree of linguistic coverage this design could achieve. This MT approach is certainly not meant to cover all English input sentences – only those that should be translated as classifier predicates. While domain-specificity can sometimes make an MT approach impractical to use, this design is meant to be embedded within a complete (possibly existing) MT system for English-to-ASL that uses traditional MT technologies to handle the majority of English inputs. Because these

[4] A stochastic motif genre-identifier could also be induced from statistical analyses of English texts known to produce a certain type of classifier predicate translation.

other MT processing pathways would be available, this design can focus on linguistic depth, rather than breadth.

With the linguistic coverage of the initial system as a baseline, the addition of this design would improve the coverage incrementally by bringing additional genres (domains) of classifier predicate expressions into the system's ASL repertoire as new motifs are implemented. The non-classifier translation pathways of the MT system would handle those spatial sentences still outside of the motif coverage. The other pathways would likely produce an overly English-like form of signing for these spatial sentences: a less desirable but somewhat useful result.

Relating Motifs to ASL Linguistic Models

The previously discussed linguistic models did not include a level of representation analogous to a motif because these models were focusing on a different part of the classifier predicate generation problem. Only after a signer has decided what spatial information to communicate (content selection) and how to sequence its presentation (propositional ordering) do these models describe how to build an individual classifier predicate (surface generation). They account for how humans produce single classifier predicate expressions – not how they plan the elements of an entire scene.

Linguistic models that do explain how human signers conceptualize 3D scenes also do not use a motif-analogous representation. Here, the reason may be that the generation task for a human is significantly different than the translation task for a computer. For example, Liddell (2003a) discusses how signers could plan a 3D scene and use multiple interrelated classifier predicates to describe it, but his model relies on the human ASL signers' rich mental visualization of objects in a 3D space and their ability to map (or "blend") these locations to the physical signing space. In a translation setting, the mental 3D visualization of the English speaker is not available; the English text is the only source of information about the scene. Because English generally includes less spatial detail than ASL when describing 3D space, both MT systems and human ASL interpreters are faced with the problem of understanding the English description and reconstructing the scene when producing classifier predicates.[5] Although not as robust as a human ASL interpreter, the AnimNL software can help this MT system create a 3D representation from the English text. But we are still left with the task of interpreting the English text for semantic and discourse cues to help guide our selection of classifier predicates to express this 3D scene. Therefore, motifs are triggered and informed by features from the analysis of the English text.

As a final linguistic concern, it is useful to consider whether the addition of motifs (that use 3D data) to design 3 has placed this system in further conflict with the polymorphemic model (Supalla, 1978, 1982, 1986). While this may initially appear to be the case, the addition of motifs is actually neutral with respect to this model. The model claims that an individual classifier predicate is composed from discrete morphemes, but it does not preclude the human signer from using mental 3D visualization of the scene during the deeper generation processes (those which overlap with the work of motifs). So, the point where the model diverges with this approach is the same as where it diverged from the original design 3 – when 3D data is used to fill the parameters of the classifier predicate PAR. This surface generation stage produces the non-categorical movements and locations of the classifier predicate output.

Discussion

Advantages of Virtual Reality

The 3D representation in this design allows it to consider spatial information when making generation decisions. Not only does this help make the generation of individual classifier predicates possible, but it also allows the system to potentially consider factors like spatial layout or visual salience when making deeper generation choices inside motifs – something a system without a 3D representation could never do.

This virtual reality representation for the space used by ASL classifier predicates may also be a basis for transcribing or recording these ASL phenomena electronically. A listing of the 3D objects currently in the invisible world with their properties/coordinates and a fully specified/planned arm movement PAR could be used to record a classifier predicate performance of a human signer. This approach would record more movement detail than classifier predicate glosses used in the linguistic literature, which merely describe the motion in English words and the handshape used. It would also be more informative than a simple movement annotation since it could store its non-spatial semantics (the semantic features that triggered the movement template), its spatial semantics (the locations of the 3D objects in the scene which it is describing), and the identities of those objects (what discourse entities are they representing). This additional information would likely be of interest to researchers studying these phenomena or building MT systems to handle them.

The 3D representation also allows this system to address ASL phenomena aside from classifier predicates in novel and richer ways. One example is the non-topological use of the ASL signing space to store locations for pronominal reference or agreement (Neidle et al., 2000). These locations could be modeled as special objects in the invisible world. The layout, management, and manipulation of these pronominal reference locations (or "tokens") is a non-trivial problem (Liddell, 2003a), which would benefit from the rich space provided by the virtual reality representation. If an ASL discourse model were managing a list of entities under discussion, then it could rely on the virtual reality representation to handle the graphical and spatial details of where these "tokens" are located and how to produce the "pointing" arm movements to refer to them.

The virtual reality representation could also facilitate the production of pronominal reference to entities that are "present" around the signing character. For instance, the character may be embedded in an application where it needed to refer to "visible" objects around it in the 3D virtual reality space or to computer screen elements on a surrounding user-interface. To make pronominal reference to an object in the visible 3D virtual

[5] And neither is perfect at this task.

reality space, a copy of this object could be made inside of the signing character's invisible world model. Then this invisible world copy could be treated like a "token" by the generation system, and pronominal references to this location could be made in the same way as for the "non-present" objects above. If the 3D object changed location during the signing performance, then its invisible world "token" counterpart can be repositioned correspondingly.

The AnimNL software makes use of sophisticated human characters that can be part of the scenes being controlled by the English text. These virtual humans possess many skills that would make them excellent ASL signers for this project: they can gaze in specific directions, make facial expressions useful for ASL grammatical features, point at objects in their surroundings, and move their hand to locations in space in a fluid and anatomically natural manner (Badler et al., 2000; Bindiganavale et al., 2000). When passed a minimal number of parameters, they can plan the animation and movement details needed to perform these linguistically useful actions. If one of these virtual humans served as the signing character, as one did for (Zhao et al., 2000), then the same graphics software would control both the invisible world representation and the ASL-signing character, thus simplifying the implementation of the MT system.

Current Work

Currently, this project is finishing the specification of both the classifier predicate generation design and a multi-pathway machine translation architecture in which it could be situated (Huenerfauth, 2004). Other research topics include: defining evaluation metrics for an MT system that produces ASL animation containing classifier predicates, developing PAR-compatible ASL syntactic representations that can record non-manual signals, and specifying ASL morphological or phonological representations that can be integrated with the PAR-based animation framework.

Acknowledgements

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

References

Bindiganavale, R., Schuler, W., Allbeck, J., Badler, N., Joshi, A., and Palmer, M. (2000). Dynamically Altering Agent Behaviors Using Natural Language Instructions. 4th International Conference on Autonomous Agents.

Badler, N., Bindiganavale, R., Allbeck, J., Schuler, W., Zhao, L., Lee, S., Shin, H., and Palmer, M. (2000). Parameterized Action Representation and Natural Language Instructions for Dynamic Behavior Modification of Embodied Agents. AAAI Spring Symposium.

DeMatteo, A. (1977). Visual Analogy and the Visual Analogues in American Sign Language. In Lynn Friedman (ed.). *On the Other Hand: New Perspectives on American Sign Language.* (pp 109-136). New York: Academic Press.

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

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

Huenerfauth, M. (2004). A Multi-Path Architecture for Machine Translation of English Text into American Sign Language Animation. In the 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.

Kipper, K., Snyder, B., Palmer, M. (2004). "Extending a Verb-lexicon Using a Semantically Annotated Corpus," In the Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-04).

Liddell, S. (2003a). *Grammar, Gesture, and Meaning in American Sign Language.* UK: Cambridge University Press.

Liddell, S. (2003b). Sources of Meaning in ASL Classifier Predicates. In Karen Emmorey (ed.). *Perspectives on Classifier Constructions in Sign Languages.* Workshop on Classifier Constructions, La Jolla, San Diego, California.

Morford, J., and MacFarlane, J. (2003). "Frequency Characteristics of American Sign Language." Sign Language Studies, 3:2.

Neidle, C., Kegl, D., 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.

Sáfár, É., and Marshall, I. (2001). The Architecture of an English-Text-to-Sign-Languages Translation System. In G. Angelova (ed.) *Recent Advances in Natural Language Processing (RANLP),* (pp. 223-228). Tzigov Chark, Bulgaria.

Schuler, W. (2003). Using model-theoretic semantic interpretation to guide statistical parsing and word recognition in a spoken language interface. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL'03), Sapporo, Japan.

Speers, d'A. (2001). Representation of American Sign Language for Machine Translation. PhD Dissertation, Department of Linguistics, Georgetown University.

Supalla, T. (1978). Morphology of Verbs of Motion and Location. In F. Caccamise and D. Hicks (eds). *Proceedings of the Second National Symposium on Sign Language Research and Teaching.* (pp. 27-45). Silver Spring, MD: National Association for the Deaf.

Supalla, T. (1982). Structure and Acquisition of Verbs of Motion and Location in American Sign Language. Ph.D. Dissertation, University of California, San Diego.

Supalla, T. (1986). The Classifier System in American Sign Language. In C. Craig (ed.) *Noun Phrases and Categorization, Typological Studies in Language, 7.* (pp. 181-214). Philadelphia: John Benjamins.

Zhao, L., Kipper, K., Schuler, W., Vogler, C., Badler, N., and Palmer, M. (2000). A Machine Translation System from English to American Sign Language. Association for Machine Translation in the Americas.