A Proposal for Making Corpora More Accessible for Synthesis: A Case Study Involving Pointing and Agreement Verbs

Rosalee Wolfe[1], John C. McDonald[1], Jorge Toro[2], Jerry Schnepp[1]
[1]DePaul University Chicago, IL
[2]Worchester Polytechnic Institute Worcester, MA
E-mail: {wolfe,jmcdonald,jschnepp}@cs.depaul.edu, jatoro@wpi.edu

Abstract

Sign language corpora serve many purposes, including linguistic analysis, curation of endangered languages, and evaluation of linguistic theories. They also have the potential to serve as an invaluable resource for improving sign language synthesis. Making corpora more accessible for synthesis requires geometric as well as linguistic data. We explore alternate approaches and analyze the tradeoffs for the case of synthesizing indexing and agreement verbs. We conclude with a series of questions exploring the feasibility of utilizing corpora for synthesis.

Keywords: annotation standards, corpora, sign language synthesis

1. Introduction

Sign language corpora provide a means for developing new insights into sign language (Crasborn, 2008), for supporting the documentation and curation of endangered languages (Johnston & Schembri, 2006), and for enabling alternative methods for evaluating theories, such as those describing patterns in language acquisition (Lillo-Martin & Pichler, 2008). By design, corpora are meant to support future as well as current research.

Sign language corpora could also be a valuable resource for the development of better sign language synthesizers. A synthesizer is an essential component of an automatic translation system between spoken and signed language. It can also serve as a verification tool for transcribing lexical items and can serve as a powerful basis for building flexible educational tools.

Current sign synthesizers excel at recalling items from a lexicon and concatenating them to create sentences. However, much work still needs to be done to expand the flexibility of synthesizers if they are to fulfill their promise, and corpora have the potential of serving a key role in this development.

2. Using Corpora for Improving Synthesis

The output of a synthesizer is only as good as the data used to create it. Without access to corpora, researchers miss important cases that synthesis algorithms need to model. New models must be rigorously tested with as many examples as possible. Access to corpora opens the door for thorough testing.

Corpora gathered for analysis provide large, rich collections of exemplars which are useful for algorithm development. They have three advantages over those gathered by synthesis researchers. The first is the level of quality of the recorded data, the second is the general purpose of the recorded data, and the third is the annotations accompanying the recorded data.

2.1 High-Quality Recording

Through years of experience, linguists have developed consistent methodologies for elicitation, and have established state-of-the-art recording facilities, designed specifically for capturing sign language. The results are high-quality recordings that preserve as much information as possible.

2.2 Generality

The second advantage of corpora gathered for analysis is the general nature of the data. We have found that our own elicitation techniques can become too specific when we are interested in representing a particular language construct for synthesis. As with movie directors, there the overwhelming desire to give such directions such as “now point to the red square.” For example, when informants knew we were interested in the placement of indices, it overly influenced how the informants signed the story.

2.3 Annotations

If sign language corpora were simply a collection of recordings, their usefulness for synthesis would be limited due to the time investment required to manually search the videos for the desired exemplars. The addition of annotations facilitates time-effective machine searching. Searchable annotations also provide the potential to identify exceptional cases that do not fit standard models. Synthesis algorithms need to incorporate these in order to exhibit the full range of expressiveness of natural signing.

While annotation data desired by synthesis researchers and linguistic scholars share many similarities, they differ somewhat in several key areas. To better understand these similarities and differences, the following section describes the organization of our sign synthesis system and lays the groundwork for a possible approach to utilize corpora originally intended for analysis.

3. Motivation for treatment of numeric and linguistic data

Ultimately any sign synthesis system must have access to numeric data for creating the postures and timing of animation. Our sign synthesis system combines rules, linguistic labels and numeric data. It has four major components – a handshape editor, a sign transcriber, an expression builder and a sentence generator. The first three provide user interfaces to record and store numeric, phonemic and lexical data, as shown in Figure 1. The fourth combines these data to form complete sentences.

Our earliest component was the handshape editor. It stores the information not in terms of joint rotations but as the handshape features of bend, spread and hook. A mathematical model (McDonald, et al., 2001) converts linguistic features to joint rotations, generating the numeric data required by the underlying animation engine.

The sign transcriber uses handshapes as a basis for creating signs. It then allows for the designation of an articulator, place of articulation and palm orientation. These correspond to the phonemic parameters of handshape, location and palm orientation.

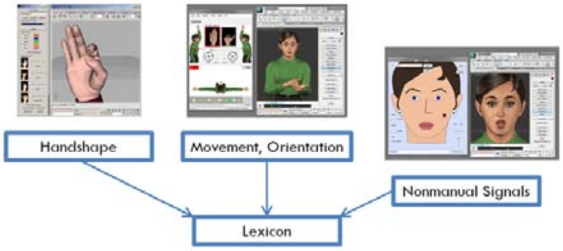
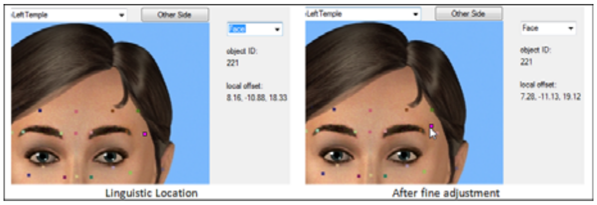


Figure 1: Handshape editor, sign transcriber, expression builder

Initial testing with members of the Deaf community indicated that more flexibility should be incorporated into this approach. Reviewers indicated that signs were awkward, and would demonstrate that sometimes a different location may be preferable to that used in the synthesized animations. As seen in Figure 2, we added a method to control positioning at a very fine level of detail. Although it still carries the linguistic tag "Left Temple", the actual geometric position of the location has changed.



LeftTemple
Face
object ID:
221
local offset:
8.16, -10.88, 18.33

Linguistic Location

LeftTemple
Face
object ID:
221
local offset:
7.28, -11.13, 19.12

After fine adjustment

Figure 2: Fine adjustment to phonemic parameter of location

The linguistic parameter of motion caused the most difficulties. We found that the rates of change are not the same for all parameters as shown in Figure 3. In this example of the sign INFORM (National Technical Institute for the Deaf, 2000), the initial handshape (label A) transitions to the final handshape (label B) in half the time required to transition from the initial location (label A) to the final location (label C). For this reason, the sign transcriber includes facilities to designate internal timing within a sign as seen in Figure 4.

Some lexical signs require the inclusion of a facial nonmanual signal. To address this requirement, the expression builder provides access to facial elements used in the formation of nonmanual signals (Schnepp, Wolfe, & McDonald, 2010).

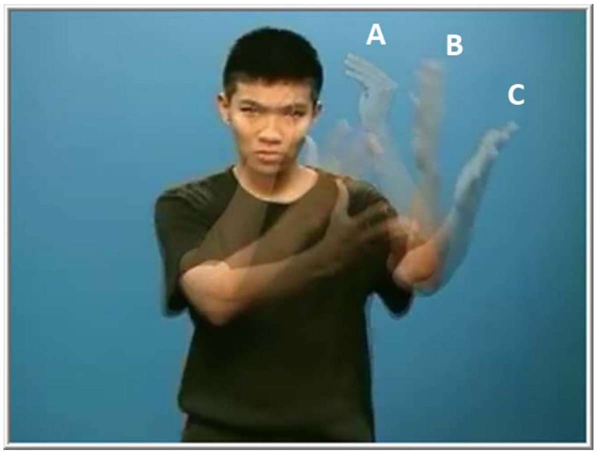


Figure 3: Varying rates of change: Handshape transition is complete before final location is achieved.

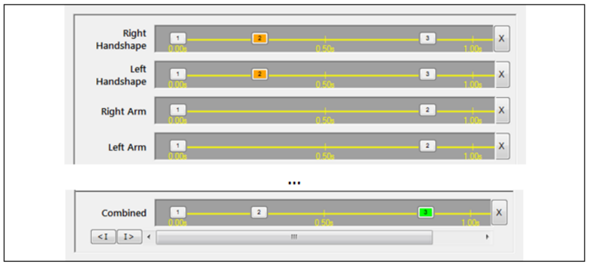


Figure 4: Timing interface

Additionally, the sign transcriber requires information about a sign's part of speech (POS). The data collected depends on the POS category. For example, agreement verbs require information about the type of agreement (object only, both subject and object), direction (backwards or forwards), and orientation agreement (Toro, 2004). The citation form is stored as data and the conjugated form is created dynamically when sentences are synthesized.

Finally the sentence generator uses a stream of text tokens as input to combine lexical items and grammar rules to generate complete sentences. The tokens can be glosses, fingerspelled words or indices. The sentence generator looks up the sign stem in the lexicon. Depending on the POS, rules modify the sign stem and may require additional information. For example, if the sentence includes an agreement verb, the user needs to specify the subject and object by designating the relevant indices. To synthesize the utterance, the sentence generator applies its grammar rules to modify the animation data, and renders the animation. Figure 5 shows the flow of data through the entire system. The signs, handshapes and nonmanual signals are represented as data, while the sentence generator is rule-based.

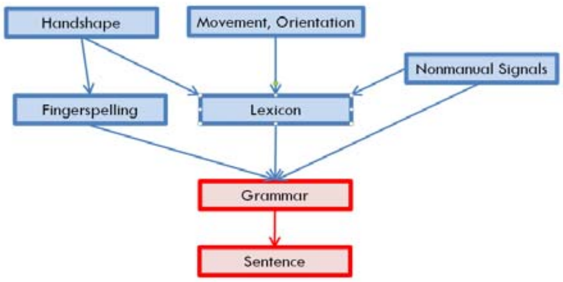


Figure 5: Flow of data towards the sentence generator. Blue indicates a data-based representation; red indicates a rule-based representation.

4. Synthesis methodology

Introducing a new language construct into this system is a five-step process:

1. Research the linguistic literature for descriptions and characterizations of the construct under study. Linguistic theory provides the guidance and inspiration for the algorithmic representation.
2. Observe examples of the construct in context. Study multiple signers in multiple contexts. This is similar to the motion studies animators create when planning an animation.
3. Use or modify the software to implement the construct. Synthesize signed sentences.
4. Conduct user tests with representatives of the Deaf community to gather feedback on questions such as
    a. Is the sentence comprehensible?
    b. Does the avatar sign it the way you would sign it?
    c. If it's not right, would you show us how it should be signed?
5. Analyze the feedback, propose refinements, and repeat the process.

It is at step two where corpora would be most useful. The next section discusses how different tiers in a corpus can benefit the synthesis process.

5. Corpora from a synthesis standpoint

Currently, we are studying processes involving indices and agreement verbs, and have been looking at the types of tiers that could support effective synthesis of sentences with these processes. Having a gloss tier is essential and ID-glosses are optimal for searching our lexicon. The start and end times for a sign are also critically useful, because they provide the average and range for a sign's duration. However, a gloss tier does not carry enough information to correctly surmise agreement verb conjugation.

Most corpora include more than a gloss tier. The following paragraphs analyze tier types and combination of tier types with respect to supporting synthesis.

5.1 Gloss and phonemic tiers

This approach focuses on descriptive annotation; where phonemic information is labeled, but syntactic designations are omitted to be as theory-neutral as possible. This set of tiers is useful for supporting verb conjugation because it contains specific information about the starting and ending location of the verb form. However, it is difficult to infer the identity of the referents from these locations. Per Padden (1990), the locus for a referent is not a precise geometric position. Further different verbs (SHOW vs. TELL) will assume different geometric positions while still indicating the same referent. Additionally, this approach requires a direct geometric interpretation of phonemes, which does not facilitate any fine-tuning required for naturally-flowing synthesis.

5.2 Gloss and syntactic tiers

In this approach, corpora contain not only glosses, but labels for POS and referents for agreement verb conjugation. A synthesizer can utilize the syntactic information to apply its rules for modifying signs. With this approach, the synthesizer makes some assumptions, placing the referents at "best guess" locations, and adjusting the verbs and nonmanual signals accordingly. Unfortunately, the synthesizer may not always make good guesses, particularly when there are more than two referents, resulting in awkward sentences.

5.3 Gloss, syntactic and phonemic tiers

Having access to both syntactic information about a referent as well as the phonemic information pertaining to its location gives a synthesizer everything it needs to create well-formed grammatical sentences that flow naturally. However, the prospect of tagging for syntax (which might need to be revised) and recording the detail of phonemic data is a nontrivial challenge.

5.4 Gloss, syntactic, and selected phonemic tiers

One possibility might be to record syntactic tags, and only a small subset of phonemic information. A synthesizer needs to know the location of a referent when it is established in the sign space, so the referent only needs to be tagged for location once in the annotation. According to Padden (1986), a location remains associated with a referent during discourse until the signer explicitly associates a new referent with the location. Since the only location data required is the first appearance of a referent, a corpus that already includes syntactic tagging would require minimal additional phonemic information.

6. Benefits of Standardized Tiers for Synthesis

Having a standardized set of tiers for synthesis would add flexibility. It facilitates the possibility of interchanging signing avatars or animation software and provides a test-bed for different approaches to synthesis such as mocap, procedural or manual animation. It also leaves open the possibility for changing avatars to accommodate different audiences (adults vs. children, addressing cultural sensitivities) or applications (real-time vs. higher fidelity rendering).

Maintaining the separation between the phonemic and syntactic representations of sign language makes it possible to create and modify movement algorithms for sign production without requiring re-annotation. Results of lexicographic research from projects such as iLex (Hanke, Storz, & Wagner, 2010) could be used to improve models of movement, resulting in more natural and believable sign synthesis. This approach could potentially accommodate the incorporation of signing styles (Heloir & Gibet, 2009) or to aid in the development of more natural variability in a signer's movements yielding a less robotic signing style.

7. Work-in-progress

We have created new algorithms for synthesizing indexing and agreements verbs based on a corpus study. For resources, we relied on the SignStream corpora (Neidle, 2002), videos from NTID (National Technical Institute for the Deaf, 2000), and our own elicited examples. The animations are viewable at http://asl.cs.depaul.edu/LREC2012. Feedback and comments are welcome.

8. More Questions than Answers

The considerations mentioned in this paper are only a beginning. The following are open questions:

- **Is it too soon to think about standardization for synthesis?**

  With standardization comes the potential benefit of increased collaboration and the possibility of sharing resources. However, premature standardization can omit important features that are then difficult and expensive to add.
- **What other information is necessary to synthesize other language constructs?**

  Although it has been posited that only a small amount of phonemic information needs to be annotated to create correct utterances involving agreement verbs, perhaps additional data is required for other cases. What other cases should be studied?
- **How can the impact of recording additional information be minimized?**

  The process of annotation is expensive, and additional tagging to support synthesis will only exacerbate the situation. Are there cases where more information can be inferred from extant data?

It is hoped that this discussion will help open a dialog to consider the alternatives and ramifications for a standardization of annotation to support synthesis.

9. References

Crasborn, O. (2008). Open Access to Sign Language Corpora. LREC 2008 6th International Conference on Language Resources and Evaluation. Workshop Proceedings. W25. 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora (pp. 33-38). Paris: ELRA.

Hanke, T., Storz, J., & Wagner, S. (2010). iLex: Handling Multi-Camera Recordings. LREC 2010. 7th International Conference on Language Resources and Evaluation. Workshop Proceedings. W13. 4th Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies. (pp. 110-111). Paris: ELRA.

Heloir, A., & Gibet, S. (2009). A Qualitative and Quantitative Characterisation of Style in Sign Language Gestures. In M. S. Dias, S. Gibet, M. M. Wanderley, & R. Bastos (Eds.), Gesture-Based Human-Computer Interaction and Simulation: 7th International Gesture Workshop, GW 2007, Lisbon, Portugal, May 23-25, 2007, Revised Selected Papers (pp. 122-133). Berlin, Heidelberg: Springer-Verlag.

Johnston, T., & Schembri, A. (2006). Issues in the Creation of a Digital Archive of a Signed Language. In L. Barwick, & L. Thieberger (Ed.), Sustainable Data from Digital Fieldwork (pp. 7-16). Sydney, Australia: Sydney University Press.

Lillo-Martin, D., & Pichler, D. C. (2008). Development of Sign Language Acquisition Corpora. LREC 2008 6th International Conference on Language Resources and Evaluation. Workshop Proceedings. W25. 3 rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora (pp. 129-133). Paris: ELRA.

McDonald, J., Alkoby, K., Carter, R., Christopher, J., Davidson, M., Ethridge, D., et al. (2001, May). An improved articulated model of the human hand. The Visual Computer, 17(3), 158-166.

National Technical Institute for the Deaf. (2000). American Sign Language Video Dictionary and Inflection Guide. Rochester, New York, USA.

Padden, C. (1986). Verbs and Role-shifting in ASL. Proceeings of the 4th National Symposium on Sign Language Research and Teaching. Las Vegas, Nevada.

Padden, C. (1990). The Relation between Space and Grammar in ASL Morphology. In C. Lucas (Ed.), Sign Language Research: Theoretical Issues (pp. 118-132). Washington, DC: Gallaudet University Press.

Schnepp, J., Wolfe, R., & McDonald, J. (2010). Synthetic Corpora: A Synergy of Linguistics and Computer Animation . Fourth Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies LREC 2010. Valetta, Malta: ELRA.

Toro, J. (2004). Automated 3D Animation System to Inflect Agreement Verbs. Sixth High Desert Linguistics Conference. Albuquerque, New Mexico.