Synthetic Corpora: A Synergy of Linguistics and Computer Animation

Jerry Schnepp, Rosalee Wolfe, John McDonald
DePaul University
243 S. Wabash Ave, Chicago IL. U.S.A.
E-mail: jschnepp@cdm.depaul.edu, rwolfe@depaul.edu, jmcdonald@cs.depaul.edu

Abstract

Synthetic corpora enable the creation of computer-generated animations depicting sign language and are the complement of corpora containing videotaped exemplars. Any design for a synthetic corpus needs to accommodate linguistic processes as well as support the generation of believable, acceptable synthesized utterances. This paper explores one possibility for representing linguistic and extralinguistic processes that involve the face and reports on the outcomes of a user test evaluating the clarity of utterances synthesized by this approach.

1. Introduction

Synthetic corpora are computer representations of linguistic phenomena. They enable the creation of computer-generated animations depicting sign languages and are the complement of corpora containing videotaped exemplars.

Synthetic corpora have the potential to serve multiple disciplines. They can aid in the automatic recognition of sign (Farhadi, et al., 2007) because they contain the geometric data required for intelligent visual detection algorithms. They can also provide visual depictions of abstract representations and act as a verification tool for data integrity and hypothesis testing (Hanke & Strorz, 2008).

Synthesized signs can be modified as they are formed. This provides the flexibility to generate an endless variety of utterances not possible with recordings and opens possibilities for automatic translation efforts. While representing sign for this purpose is still an open question, a synthetic corpus has the potential to serve in this capacity. The flexibility of synthetically-generated sign is also useful for the development of interpreter training software and self-directed learning tools for deaf children (Wolfe, 2006; Wolfe, et al., 2007)

The following describes a design for a synthetic corpus of American Sign Language. In addition to representing glosses, the corpus provides for facial nonmanual signals and extralinguistic facial communication. The paper also reports on a user evaluation of animations generated by this approach.

2. Design Goals

From an animator's perspective, utterances in sign are comprised of geometric poses and movements. Given the proper videotaped reference material, it is possible to animate any signed utterance. However, the animation does not take into account linguistic structure. Whereas the production of computer generated animation only requires timing and geometric data, the synthesis of sign requires additional information, because what is manifested physically is often the result of co-occurring linguistic and extralinguistic processes (Wilbur, 2000).

Figure 1 depicts the gloss BOOK being signed in a yes-no question with happy affect. These co-occurring functions require representation as independent entities so that they can be recombined and thus interact with each other. They have parallels to tracks used in sign annotation software (Brugman & Russell, 2004). Linguistic annotations can help animation transcribers understand the salient features of movements and poses, helping them to build far more legible animations. Thus the classification of geometric changes based on their linguistic function is mandatory for producing novel utterances.

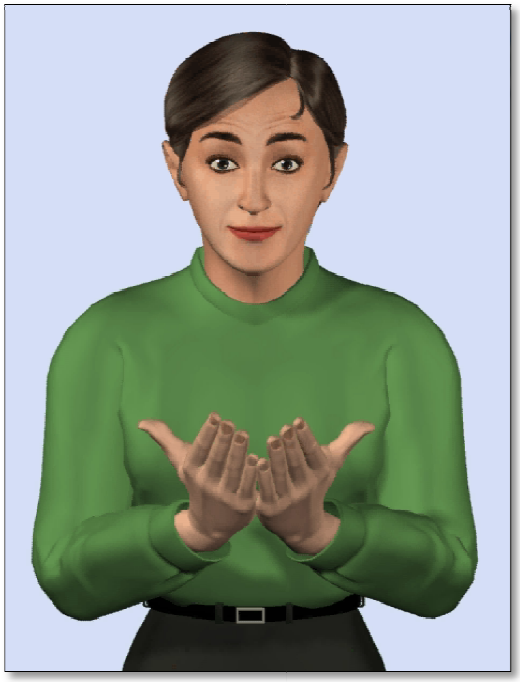

Figure 1: A happy signer asking a Yes/No question.

A desirable feature of any representation is the ability to accommodate paralinguistic and extralinguistic information. Emotional affect must be considered, as well as such phenomena as mouthing, which some populations may prefer. Researchers, however, should have the option to include or exclude this additional data when generating utterances.

To demonstrate the importance of this design goal, consider a Wh-question signed in an angry fashion, as in Figure 2. The eyebrows lower as part of producing a Wh-question. However, the emotional state of anger also involves lowering the eyebrows. The synthesis of this sentence requires that these two be depicted simultaneously.

Figure 2: An angry signer asking a Wh-question.

At first glance, the design goals of linguistics and animation would appear to be at cross purposes. Linguistic researchers often use corpora to form hypotheses through queries on linguistic features, and are interested in such abstractions as phonemes, lexical modifiers and verb agreement. In contrast, animators require extensive minute detail.

In actuality, the fields of linguistics and computer animation create a mutually beneficial synergy. Having the detailed precision required for animation can facilitate the exploration of subtle interactions among linguistic phenomena. Likewise, animators need an abstract representation to organize, combine, and synthesize complex animation data.

Regardless of the animation technique, linguistic knowledge is necessary to produce any synthetic corpus. Animators who hand-transcribe need to work closely with linguists, so that phenomena are tagged correctly. Linguistic information guides the transcription artist's efforts to produce a natural exemplar that encapsulates the essential motions of a sign.

With motion capture, the role of linguistics is no less central. Motion capture equipment generates massive amounts of data that must be cleaned to remove extraneous noise. The linguistic attributes of a sign give the cleanup artists precisely what they need to process and extract the desired motion.

3. Current Proposal

Our work uses labeled manual transcription to create detailed and accurate animations of sign. These animations require voluminous data, as they must be realistic enough to pass the scrutiny of fluent signers. However, such detail is organized using a framework that is both abstract enough to facilitate linguistic research and flexible enough to allow for the synthesis of novel utterances.

Table **1** shows the high level structure of our corpus design, which is based on abstractions used by linguists and is encoded as XML (DuCharme, 1999). High level tracks separately control the linguistic functions of gloss, syntax, and nonmanual lexical modifiers. These direct the position and timing of subordinate geometric components. Researchers have the option to add high level tracks for paralinguistic or extralinguistic functions.

High Level Tracks
- Linguistic:
  - syntax
  - gloss
  - lexical modifier
- Extralinguistic:
  - affect
  - mouthing

Syntax Block
- Label
- Start time
- End time
- Curve
- Geometry groups
  - Controllers
    - Keys

Gloss Block
- Label
- Start time
- End time
- Linguistic Component Block
  - Left Handshape
    - Label
    - Time
      - Geometry groups
        - Controllers
          - Keys
  - Right Handshape
    - Label
    - Time
      - Geometry groups
        - Controllers
          - Keys
- Geometry groups
  - Controllers
    - Keys

NM Lexical Modifier Block
- Label
- Start time
- End time
- Curve
- Viseme *(multiple possible)
  - Label
  - Time
  - Geometry groups
    - Controllers
      - Keys

Affect Block
- Label
- Start time
- End time
- Curve
- Geometry groups
  - Controllers
    - Keys

Mouthing Block
- Label
- Start time
- End time
- Curve
- Viseme * (multiple possible)
  - Label
  - Time
  - Geometry groups
    - Controllers
      - Keys

Table 1: Corpus Structure.

Each track contains blocks of time-based information. Each block has a label, a start time, an end time, as well as a collection of subordinate geometry blocks. Geometry blocks can contain animation keys or a static pose. Further, blocks can contain intensity curves that control the onset and intensity of a pose, allowing for multifarious variations.

Figure 3 demonstrates the abstraction of linguistics and the detail of animation in the case of the question "Do

you want a book?" The green curve represents the movement corresponding to the yes-no question syntactic marker. The red curve represents the influence of the affect "anger".

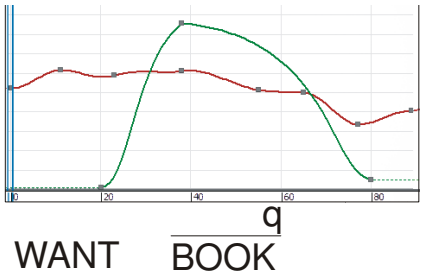

Figure 3: Intensity curves and corresponding sentence.

Although the syntactic marker co-occurs with the gloss BOOK, the green curve controlling the intensity of the corresponding pose starts before the onset of the syntactic marker and ends a significant amount of time after it. This reflects a commonly-used technique in animation whereby the action will ease-in and ease-out of a given pose (Burtnyk & Wein, 1976). Further, animation principles require that the pose not be held perfectly still at any time, thus there is no plateau in the curve.

The use of labeled poses follows common practice in animation studios where a master animator creates a dictionary of characteristic poses (Thomas & Johnston, 1981). By encapsulating minute geometric arrangements in concise groups called poses, a master animator provides an efficient mechanism for others to apply and combine poses. In a similar fashion, this corpus design allows for application and composition of linguistic processes.

4. A Case Study

To test the feasibility of this approach, we focused on the interaction of processes that take place on a signer's face. We based the design on the substantial body of literature that characterizes these processes (Grossman & Kegl, 2006; Reilly, et al., 1990; Weast, T., 2008). We also considered the feasibility of incorporating both linguistic and extralinguistic information in the design.

We conducted a study of the clarity and acceptability of the synthesized utterances. Since we aimed to represent the interactions of both linguistic and extralinguistic facial movements, we chose a set of test utterances that combined the effects of a single facial linguistic marker and a single emotive pose (See Table 2).

Twenty participants, all of whom were attending the 2009 DeafNation Expo trade show in Palatine, Illinois volunteered to participate in this study. The participants answered background questionnaires to determine their level of ASL fluency. They were informed that they could withdraw at any time during the experiment and they were naive as to its purpose. This work was reviewed and approved by the Institutional Review Board at DePaul University [JS101609CDM].

During the user test, participants viewed animations of ASL signs. During each session the participant watched short clips depicting the combination of nonmanual signals and emotional affect, as listed in Table 2. The clips are available at http://asl.depaul.edu/LREC2010. Following each clip, participants answered questions regarding its meaning and clarity.

t WHq BOOKS YOU WANT HOW-MANY (1) Happy (2) Angry CHA COFFEE TALL (3) Happy (4) Angry

Table 2: Test utterances.

The test environment comprised a PC laptop placed on a table in an exhibition booth. The test facilitator operated the laptop while the participant watched an attached monitor. The participants viewed animations full-screen on the 21" LCD monitor (resolution: 1280 x 1024 pixels). They were seated at a viewing distance of 20-40". All instructions were signed by the Deaf facilitator or the interpreter. A note-taker sat behind both the participant and facilitator while the interpreter sat across the table.

Each participant tested individually. Participants were informed that they should watch each animation carefully and that they could watch an animation as many times as they wanted.

The facilitator prefaced each animation with a short sentence establishing its context. For example, the first animation displayed "How many books do you want?" Before playing the animation the facilitator explained that the character is the owner of a book store who is taking an order from a customer.

After watching an animation, each participant answered four questions. The first question asked the participant to repeat the sentence to confirm that the animation had communicated the intended meaning. Question two presented a graphical Likert scale (Figure 4) which queried the perceived emotional state. The third question employed another Likert scale measuring the animation's clarity, from unrecognizable (1) to perfectly clear (5). The last question asked for suggestions to improve the animation.

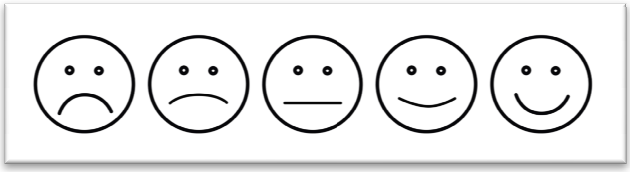

Figure 4: Likert scale measuring emotional state.

5. Results

For brevity, only responses to utterance (4) are reported here. All the results were similar and the entire data set is available at http://asl.depaul.edu/LREC2010. In

response to the first question, participants were able to replicate the utterance 100% of the time. Also, 70% rated the animation as clear or very clear (Table **3**). Each participant ascertained that the mouth shapes which characterize CHA indicate a large size. While some were confused as to the reason why the avatar appeared angry about a large cup of coffee, 95% correctly identified the intended emotional state (Table **4**). After viewing the animation, participants described her as "grumpy", "angry", "disappointed" and "negative".

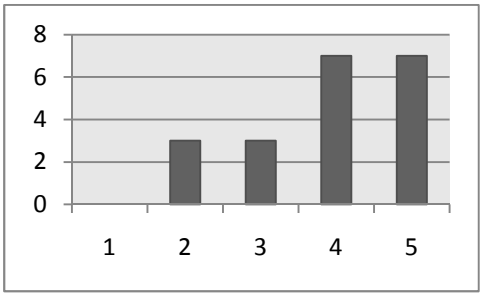

Table 3: Clarity of test utterance (4).

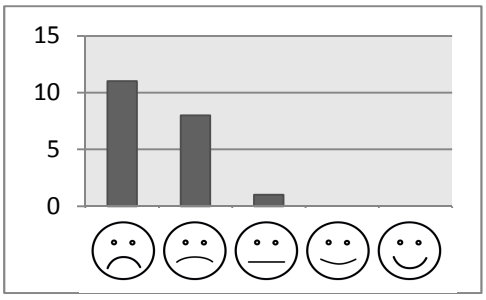

Table 4: Emotion of test utterance (4).

6. Conclusion and Future Work

The use of linguistic abstractions as a basis for animations has yielded promising results. The animations produced were well received by fluent signers and appear to communicate effectively. The data strongly suggest that the representation chosen for our corpus is flexible enough to display co-occurring facial nonmanual signals.

While this approach undoubtedly requires extension and revision, it is a step toward the automatic generation of American Sign Language. Moving forward, we plan to extend this representation to other parts of the body and test it with a wider range of utterances. We also plan to integrate the corpus structure into a more complete user interface that would facilitate the generation of ASL animations incorporating linguistic and extralinguistic features that interact on many levels including the facial nonmanual signals presented here.

7. Acknowledgements

We would like to acknowledge Nick Roessler and Brent Shiver for their help organizing and conducting user tests at DeafNation Expo, and Diana Gorman Jamrozik and Peter Cook of Columbia College Chicago for valuable discussions on nonmanual signals. We would also like to acknowledge DePaul University and The American Sign Language Project for funding.

8. References

Brugman, H. & A. Russell (2004). Annotating Multi-media / Multi-modal resources with ELAN. In: Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation. IMDI Team (2003), IMDI Metadata Elements for Session Descriptions, Version 3.0.4, MPI Nijmegen.

Burtnyk, N. & Wein, M. (1976). Interactive Skeleton Techniques for Enhancing Motion Dynamics in Key Frame Animation. Communications of the Association for Computing Machinery, Vol 19. No. 10 October 1976, 564-569.

DuCharme, B. (1999). XML: The Annotated Specification. Upper Saddle River, NJ: Prentice-Hall.

Grossman, Ruth & Judy Kegl. 2006. To capture a face: a novel technique for the analysis and quantification of facial expressions in American Sign Language. Sign Language Studies 6(3) 273–305.

Farhadi, A., Forsyth, D. & White, R. (2007). Transfer learning in sign language. In Computer Vision and Pattern Recognition, pages 1–8.

Hanke, T. & Storz, J. (2008). iLex – A database tool for integrating sign language corpus linguistics and sign language lexicography. In: Crasborn, Onno et al. (eds.): LREC 2008. 6th International Conference on Language Resources and Evaluation. Workshop Proceedings. W25. 3rd Workshop on the Representation and Processing of Sign Languages. Sunday 1st June 2008,Marrakech – Morocco. Paris: ELRA, 64-67.

Reilly, J.,McIntire, M. & Bellugi, U. (1990). Faces: The relationship between language and affect. In Virginia Volterra & Carol Erting (eds.), From Gesture to Language in Hearing and Deaf Children, New York, NY: Springer-Verlag. 128–141.

Thomas, F., and Johnston, O. (1981). The Illusion of Life: Disney Animation. New York: WaltDisney Productions.

Weast, T. (2008). Question in American Sign Language: A Quantitative analysis of raised and lowered eyebrows. PhD thesis, The University of Texas at Arlington.

Wilbur R.B. (2000). Phonological and prosodic layering of nonmanuals in American Sign Language. In Lane, H. & K. Emmorey (eds.), The signs of language revisited: Festschrift for Ursula Bellugi and Edward Klima, (pp. 213-241) Hillsdale, NJ: Lawrence Erlbaum.

Wolfe, R. (2006). An Improved Tool for Practicing Fingerspelling Recognition. Conference 2006 International Conference on Technology and Persons with Disabilities. Northridge, California, March 17-22.

Wolfe, R. McDonald, J., Davidson, M., and Frank, C. (2007) Using an Animation-based Technology to Support Reading Curricula for Deaf Elementary Schoolchildren. The 22nd Annual International Technology & Persons with Disabilities Conference. Los Angeles, CA March 21.