Synthetic and acquired corpora: meeting at the annotation

Rosalee Wolfe[*], John C McDonald[*], Jerry Schnepp[*], Jorge Toro[**]
[*]DePaul University, Chicago, IL
{wolfe, jmcdonald, jschnepp}@cs.depaul.edu

[**]Worchester Polytechnic Institute, Worcester, MA
jatoro@wpi.edu

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 recorded exemplars.

Synthetic corpora serve multiple disciplines. Because they contain the geometric data necessary for intelligent visual-detection algorithms, they can aid in the automatic recognition of signed languages (Zahedi, Keysers, & Ney, 2005) (Farhadi, Forsyth, & White, 2007). They can also provide visual depictions of linguistic processes and act as a verification tool for data integrity and hypothesis testing (Hanke & Storz, 2008).

Synthesized signs can also 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, et al., 2006; Wolfe, 2007).

A current challenge in using synthetic corpora for this purpose is producing animations that are legible, linguistically acceptable, and convincingly humanlike. This involves the development of as many inferences as possible about sign production from annotations of recorded corpora. Creating poses corresponding to handshape and articulator positioning is a start, but new strategies are needed to facilitate the production of many adjectival and adverbial modifiers.

Some of these strategies will involve human kinematics, but a significant portion will depend on establishing new standards of annotation. Annotation is thus the meeting ground of conventional and synthetic corpora. In conventional corpora, annotation identifies linguistic phenomena and supports the testing of hypotheses (Neidle, Kegl, Maclaughlin, Bahan, & Lee, 2000) (Athitsos, et al., 2010); in synthetic corpora, annotation drives the creation and synthesis of animation.

Developing new standards will require considering some difficult issues such as establishing gloss tags (Alkoby, Bernath, Hochgesang, Mirus, & Pascual, 2010) and labeling processes that occur on the face. Should an annotation simply label the pose of a human feature, such as “brows up”? Should the label incorporate whether the signer’s intent is linguistic (“Yes/no question marker”) or extralinguistic (“happy”)? Additional challenges arise when processes co-occur, as when signing a yes-no question in an angry fashion. Should an annotation system be able to represent intensity? We posit that a system of optional, ancillary information that can address these issues.

Another deep issue is flexibility. Should an annotation system force a researcher into decisions about labeling that may make it difficult to create new queries in the future? As recording technology continues to improve, researchers will be able to identify new subtleties (Hochgesang & Witworth, 2010). What provisions would give an annotation system the flexibility to incorporate future discoveries?

Expanding conventional annotation standards will have many potential benefits, including the creation of corpora large enough to facilitate corpus-based machine translation. Standard annotation would enable research groups to benefit from shared resources.

Bibliography

Alkoby, K., Bernath, J., Hochgesang, J., Mirus, G., & Pascual, P. (2010). Construction of an ID gloss database for American Sign Language. *Theoretical Issues of Sign Language Research 10: Contributions fo Corpus and Applied Linguistics.* West Lafayette, IN.

Athitsos, V., Neidle, C., Sclaroff, S., Nash, J., Stefan, a., Thangali, A., et al. (2010). Large lexicon project: Amercain Sign Language video corpus and sign language indexing/retrieval algorithms. *Proceedings of the Workshop on the representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora.* Valetta, Malta.

Farhadi, A., Forsyth, D., & White, R. (2007). Transfer learning in sign language. *Conference on Computer Vision and Pattern Recognition*, (pp. 1-8). Minneapolis, MN.

Hanke, T., & Storz, J. (2008). iLex: A database tool for integrating sign language corpus linguistics and sign language lexicography. *Sixth International Conference on Language Resources and Evaluation. (LREC 2008) Workshop W25. 3rd Workshop on the Representation and Processing of Sign Languages* (pp. 64-67). Marrakech, Morocco: ELRA.

Hochgesang, J. A., & Witworth, C. (2010). Phonetics, phonology and transcription practices in Aerican Sign Language. *Theoretical Issues of Sign Language Research 10: Contributions fo Corpus and Applied Linguistics.* West Lafayette, IN.

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

Wolfe, R., Alba, N., Billups, S., Davidson, M. J., Dwyer, C., Jamrozik, D. G., et al. (2006). An improved tool for fingerspelling recognition. *Technology and Persons with Disabilities.* Los Angeles, CA.

Wolfe, R., Davidson, M. J., & McDonald, J. (2006). Using an animation-based technology to support reading curricula for deaf elementary schoolchildren. *Twenty-second Annual International Technology & Persons with Disabilities Conference.* Losa Angeles, CA.

Zahedi, M., Keysers, D., & Ney, H. (2005). Appearance-based recognition of words in American Sign Language. *Second Iberian Conference on Pattern Recognition and Image Analysis [IbPRIA 2005]*, (pp. 511-519). Estoril, Portugal.