5-21-2011

Synthetic and Acquired Corpora: Meeting at the Annotation

Rosalee Wolfe

John C. McDonald

Jerry C. Schnepp

Bowling Green State University - Main Campus, schnepp@bgsu.edu

Jorge Toro

Follow this and additional works at: https://scholarworks.bgsu.edu/vcte_pub

Repository Citation

Wolfe, Rosalee; McDonald, John C.; Schnepp, Jerry C.; and Toro, Jorge, "Synthetic and Acquired Corpora: Meeting at the Annotation" (2011). Visual Communication and Technology Education Faculty Publications. 36.
https://scholarworks.bgsu.edu/vcte_pub/36

This Presentation is brought to you for free and open access by the Visual Communication at ScholarWorks@BGSU. It has been accepted for inclusion in Visual Communication and Technology Education Faculty Publications by an authorized administrator of ScholarWorks@BGSU.
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

**Worcester 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


