Gallaudet University Documentation of ASL (GUDA) - Whither a corpus for ASL?
by Julie A. Hochgesang, Jennifer Willow, Rafael Treviño, Emily Shaw

In this presentation, we discuss the challenges of creating a nation-wide representative corpus for American Sign Language and describe our preliminary efforts in sourcing primary data from existing video collections at Gallaudet University to create a language documentation of ASL. We propose that our developing language documentation, the Gallaudet University Documentation of ASL (GUDA), has the potential to become a monitor-style corpus (McEnery & Hardie, 2011). Although not designed as a corpus from the start, GUDA aims to pull together video resources already existing somewhere on Gallaudet from the early 1900s to today showcasing the use of ASL across users, discourse genres and time. GUDA will hold digital centralization, accessibility, cross-disciplinary benefit, community stewardship and collaboration within its core vision (Berez-Kroeker, Gawne, Kung, Kelly, Heston, Holton, Pulsifer, Beaver, Chelliah, Dubinsky, Meier, Thieberger, Rice, & Woodbury, 2018).

One of the challenges to creating a representative and nation-wide ASL corpus is the size of North America and the large diverse communities of Deaf ASL users. The geographical distribution of Deaf Americans is unique compared to signing populations represented by other sign language corpus projects. This need to account for representation of multiple communities within the larger Deaf community in North America presents challenges at every step of corpus development (from collecting the films, to storing/archiving the data, to annotation and analyses). Given that Gallaudet has brought people together from all over North America, it is well-positioned to represent this significant regional variation and the language as a whole over time. GUDA will work to centralize diverse video datasets of ASL use that could be representative of a wide range of ASL usage, language register, settings, and content.

For any corpus to be a "lasting multipurpose record of a language" (Himmelman, 2006), it must include cross-disciplinary cooperation, lasting stakeholder involvement and sustainable resources. Gallaudet University, established over 150 years ago, is uniquely situated to provide all these. GUDA is able to build its digital infrastructure using a web platform supported by Gallaudet University (which is advantageously situated to keep up with modern technological needs and the evolving nature of web accessibility). The digital landing site will act as a point of access for those interested in the data as it is enriched over time even prior to becoming a full corpus, including Deaf community members and researchers both on and off Gallaudet campus. This infrastructure (see Figure 1) will both house data and point to stable sources of data.

Figure 1. An Overview of the GUDA Infrastructure

For existing datasets (primary data only or comprehensive datasets) that already have a stable URL, GUDA will point to these sources and offer searchability through its infrastructure. For other data sources without stable URLs, GUDA will house the data using current video hosting

services by Gallaudet (Kaltura). The data sources will be organized and searchable along with their metadata, annotation files (using the SLAAASh data annotation protocols and ASL Signbank (Hochgesang, Lillo-Martin, Crasborn, 2018)), and terms of use through the GUDA landing site. This site is in development and features include graded levels of access and protocols for participant re-consent. This infrastructure is ideal for others to link their research, further enriching GUDA as a monitor corpus. Not only does this provide them the benefit of sharing their data in a platform that they may not have the resources to construct themselves, but it can also fulfil ethical responsibilities by making materials available to their stakeholders. For the community, these collections contain retellings of their histories, stories, culture, and ways of being.

Having been given access to the Deaf communities' histories and languages, linguists reciprocate as stewards. By using already existing ASL video sources, we are salvaging the "digital detritus" (Bird & Simons, 2003) of Gallaudet ASL video collections and hopefully creating a representative ASL corpus that will become a resource. During this massive endeavour, care must be taken to complete and standardize the metadata within collections for more comprehensive searchability resulting in fuller cross-discipline benefit. We also must take special care to document sources accurately and ethically, especially participant consent which will require re-consent measures (Chen Pichler, Hochgesang, Simons, & Lillo-Martin, 2016). Although it may take some extra work that's unprecedented for current best practices for sign language corpora (Fenlon, Schembri, Johnston & Cormier, 2015), it's well worth the effort given that GUDA can become a vital resource to test claims that have been made about ASL in the literature based on a small number of signers and their grammaticality judgments as well as a lasting resource for the creation of new research.

References

Bird, S. & G. Simons. 2003. Seven dimensions of portability for language documentation and description. *Language*, 79(3). 557-582.

Berez-Kroeker, A., Gawne, L., Kung, S., Kelly, B., Heston, T., Holton, G., Pulsifer, P., Beaver, D., Chelliah, S., Dubinsky, S., Meier, R., Thieberger, N., Rice, K., & Woodbury, A., (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics,* 56*(1).* 1-18.

Chen Pichler, D., J. Hochgesang, D. Simons & D. Lillo-Martin. (2016). Community Input on Re-consenting for Data Sharing. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J. Hochgesang, J. Kristoffersen, & J. Mesch (Eds.), *Workshop Proceedings: 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining / Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016) (29-34). Paris: ELRA. (ELRA).

Himmelmann, Nikolaus. (2006). Language documentation: What is it and what is it good for? In J. Gippert, N.P. Himmelmann, U. Mosel (Eds), *Essentials of Language Documentation*. 1- 30. New York: Mouton de Gruyter.

Hochgesang, J.A., O. Crasborn, & D. Lillo-Martin. (2018). ASL Signbank. New Haven, CT: Haskins Lab, Yale University. https://aslsignbank.haskins.yale.edu/

Fenlon, J., A. Schembri, T. Johnston, & K. Cormier. (2015). Documentary and corpus approaches to sign language research. 156-172. *The Blackwell guide to Research Methods in Sign Language Studies*, ed. by E. Orfanidou, B. Woll & G. Morgan. Oxford: Blackwell.

McEnery, T. & A. Hardie. (2011). *Corpus Linguistics: Method, Theory and Practice (Cambridge Textbooks in Linguistics)*. Cambridge University Press.