Data Sets available through DAI 2

Addendum to C. Neidle & A. Opoku, **User's Guide to the American Sign Language Linguistic Research Project (ASLLRP) Data Access Interface (DAI) 2 — Version 2.** *ASLLRP Report No. 18,* Boston University, Boston, MA.
http://www.bu.edu/asllrp/rpt18/asllrp18.pdf

Most recent update: March 2022

The American Sign Language Linguistic Research Project (ASLLRP) provides Web access to a wide range of different types of ASL data, including:

- Isolated signs (**23,452** sign videos from **33** different signers)
- Continuous signing corpora (**2,651** utterances, containing a total of **20,560** sign examples available as video clips segmented from those utterances and in their utterance context, from **19** different signers)

The corpora shared by the ASLLRP incorporate data collected at Boston University and at the Rochester Institute for Technology (under the supervision of Matt Huenerfauth), as well as videos shared by DawnSignPress. Including both the citation-form signs and continuous signing corpora, we have a total of **44,012** sign tokens corresponding to **3,542** distinct signs (not including fingerspelled signs, classifiers, and gestures).

We have enforced, to the best of our ability, consistency in labeling throughout our corpora (see [5, 8]). Sign-level annotations include text-based gloss labels, annotations of sign type (lexical, loan, fingerspelled, classifier, number, and name signs, as well as gestures and compounds), and phonological properties (e.g., information about hand configurations on the 2 hands). The utterances also include sentence-level information about such things as non-manual behaviors and grammatical markings, translations, etc. Annotation conventions are documented here: [2, 3].

Organization of the website

The DAI 2 website consists of two sections, one providing access to several **ASLLRP continuous signing corpora**, and the other providing access to the **ASLLRP Sign Bank**, as follows:

[1] ASLLRP Continuous Signing Corpora: https://dai.cs.rutgers.edu/dai/s/dai

The data from the following datasets can be browsed or searched on the Web based on various sign-level and utterance-level properties:

(A) The **ASLLRP SignStream® 3 Corpus** https://dai.cs.rutgers.edu/dai/s/dai, an expanding collection of data annotated with SignStream® 3, consisting of **65** SignStream® files:

- **37** SignStream® collections from videos recorded at Boston University in 2011-2013.

- **10** SignStream® files, with annotations of a video article by **Ben Bahan**: Bahan, B. (2009/2010). Sensory Orientation. *Deaf Studies Digital Journal, Issue 1,* Fall.
- **18** SignStream® files with annotations of sentences shared by **DawnSignPress** (**https://www.dawnsign.com**), which has granted permission for users to view these data on this site, but ***not*** to download or use or redistribute these data.

(B) The **NCSLGR SignStream® 2 Corpus** **https://dai.cs.rutgers.edu/dai/s/daioriginal**, an older corpus released in 2007 (previously available on CD-ROM, **http://www.bu.edu/asllrp/cslgr/pages/** and more recently available from a different website, **http://secrets.rutgers.edu/dai/queryPages/** **[9]**). This includes **38** files (of which **19** are short narratives) that had been annotated with SignStream® 2; see the p. 4 for differences in the annotations for the older data.

Download options are available for all of the above, excluding the DawnSignPress data. Linguistic annotations for the signs and utterances in the data that can be downloaded are available in XML format. These utterances from (A) above can also be viewed and further analyzed and annotated within SignStream® **http://www.bu.edu/asllrp/SignStream/3/**, an application we have developed for analysis of visual language data shared on the Web.

[2] The ASLLRP Sign Bank: https://dai.cs.rutgers.edu/dai/s/signbank

There is also an online ASLLRP Sign Bank publicly available [4, 6, 7]. The Web interface makes it possible to search the dataset based on various criteria and to view, for specific signs, both examples from our citation-form sign datasets and segmented signs from our continuous signing corpora (viewable either individually or in their sentential context). It is now possible to view subsets of the data by sign type (lexical signs, loan signs, fingerspelled signs, number signs, classifiers, gestures, compounds).

The Sign Bank incorporates the data from (A) above, plus additional datasets containing citation-form signs:

- The **ASLLVD** data set: see **http://www.bu.edu/asllrp/av/dai-asllvd.html** **[1, 8]**
- Isolated signs collected by researchers at **Rochester Institute of Technology** under the supervision of **Matt Huenerfauth** as part of NSF grant IIS-1763569. Thanks to Abraham Glasser, Ben Leyer, Saad Hassan, and Sarah Morgenthal for their roles in helping with data collection and annotation.
- Citation-form signs provided by **DawnSignPress** (**https://www.dawnsign.com**)

New: ASLLRP Sign Bank Download Options

It is currently possible to download the ASLLRP Sign Bank citation-form sign datasets and videos from our website for use in sign recognition research, with the ability to download segmented Sign Bank examples from our continuous signing corpora to be provided from the same site in the near future.

Datasets currently available for download from **https://dai.cs.rutgers.edu/dai/s/signbank,** with accompanying annotations and a document explaining the gloss labeling [5]:

Boston University American Sign Language Lexicon Video Dataset (ASLLVD)

- **9,748** sign tokens; **6** signers

Rochester Institute of Technology (RIT) Dataset

- **11,801** sign tokens; **12** signers

DawnSignPress (DSP) Dataset

- **1,903** sign tokens; **15** signers

See also **acknowledgments and credits**:

http://www.bu.edu/asllrp/dai-credits.html
http://www.bu.edu/asllrp/data-credits.html# credits

Statistics for these data sets are available here:

https://dai.cs.rutgers.edu/dai/s/runningstats
https://dai.cs.rutgers.edu/dai/s/runningstatsdai1

Please pay careful attention to the **terms of use**:

For **Continuous signing data from the DAI**: **http://www.bu.edu/asllrp/dai-terms.html**
For **ASLLRP Sign Bank data**: **http://www.bu.edu/asllrp/signbank-terms.pdf**

Differences in the way data are represented in these online data sets

In addition to our most recent datasets (our expanding **ASLLRP Corpus**) annotated with **SignStream® version 3 <http://www.bu.edu/asllrp/SignStream/3/>** and displayed through our improved Data Access Interface, **DAI 2.** We are also providing access to older datasets annotated with **SignStream® version 2** (our **NCSLGR Corpus: http://www.bu.edu/asllrp/ncslgr-for-download/download-info.html**) and previously displayed through an older Web interface (our **DAI** [9], updated to work around the expiration of Adobe's Flash Player as of 12/31/2020).

Representation of gloss information on the two hands has changed significantly between the older and the newer data, because of enhancements to the SignStream® annotation tool. Previously, gloss labels were inserted solely on the upper tier of the 2-hand display, with annotation on the second tier only in case the non-dominant hand was doing something unexpected, or something independent of what was being produced by the dominant hand. So, for a regular two-handed sign, we would only have included a gloss label in the main gloss tier, with nothing annotated for the second gloss tier. In SignStream® version 3, however, labels appear in the non-dominant tier for all signs that involve use of the non-dominant hand; thus all 2-handed signs include gloss information (plus start and end handshapes, also newly added in version 3) on both of the gloss tiers.

References

[1] Athitsos, V., C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, and A. Thangali. (2008) The American Sign Language Lexicon Video Dataset. IEEE Workshop on Computer Vision and Pattern Recognition for Human Communicative Behavior Analysis. **http://vlm1.uta.edu/~athitsos/publications/athitsos_cvpr4hb2008.pdf**

[2] Neidle, C. (2002) SignStream™ Annotation: Conventions used for the American Sign Language Linguistic Research Project, In American Sign Language Linguistic Research Project Report No. 11, Boston University, Boston, MA. **http://www.bu.edu/asllrp/asllrpr11.pdf**

[3] Neidle, C. (2007) SignStream™ Annotation: Addendum to Conventions used for the American Sign Language Linguistic Research Project, In American Sign Language Linguistic Research Project Report No. 13, Boston University, Boston, MA. **http://www.bu.edu/asllrp/asllrpr13.pdf**

[4] Neidle, C., and A. Opoku. (2020) A User's Guide to the American Sign Language Linguistic Research Project (ASLLRP) Data Access Interface (DAI) 2 — Version 2, In ASLLRP Project Report No. 18, Boston University, Boston, MA. **http://www.bu.edu/asllrp/rpt18/asllrp18.pdf**

[5] Neidle, C., and A. Opoku. (2022) Documentation for Download of ASLLRP Sign Bank Citation-Form Sign Datasets, Boston University, ASLLRP Project Report No. 20, Boston, MA. **http://www.bu.edu/asllrp/rpt20/asllrp20.pdf**

[6] Neidle, C., A. Opoku, G. Dimitriadis, and D. Metaxas. (2018) NEW Shared & Interconnected ASL Resources: SignStream® 3 Software; DAI 2 for Web Access to Linguistically Annotated Video Corpora; and a Sign Bank. 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community. LREC 2018, May 2018; Miyagawa, Japan. **https://open.bu.edu/handle/2144/30047**

[7] Neidle, C., A. Opoku, and D. Metaxas. (2022) ASL Video Corpora & Sign Bank: Resources Available through the American Sign Language Linguistic Research Project (ASLLRP), *arXiv:2201.07899*. **https://arxiv.org/abs/2201.07899**

[8] Neidle, C., A. Thangali, and S. Sclaroff. (2012) Challenges in Development of the American Sign Language Lexicon Video Dataset (ASLLVD) Corpus. 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. LREC 2012, Istanbul, Turkey. May 2012. **https://open.bu.edu/handle/2144/31899**

[9] Neidle, C., and C. Vogler. (2012) A New Web Interface to Facilitate Access to Corpora: Development of the ASLLRP Data Access Interface (DAI) Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, the Language Resources and Evaluation Conference, LREC 2012; Istanbul, Turkey. **https://open.bu.edu/handle/2144/31886**