Building the ASL Signbank: Lemmatization Principles for ASL

Julie A. Hochgesang[1], Onno Crasborn[2], Diane Lillo-Martin[3]
[1]Gallaudet University, [2]Centre for Language Studies, Radboud University,
[3]University of Connecticut & Haskins Laboratories
800 Florida Ave NE, SLCC 3206, Washington DC 20002, USA; PO Box 9103, 6500HD Nijmegen, Netherlands;
Department of Linguistics, 365 Fairfield Way, Unit 1145, Storrs, CT 06269-1145, USA
julie.hochgesang@gallaudet.edu, o.crasborn@let.ru.nl, diane.lillo-martin@uconn.edu

Abstract

Following the example of other sign language researchers, we are creating a Signbank, a usage-based lexical database, to maintain consistent and systematic annotation information for American Sign Language (ASL). This tool, which will be available to the public, is currently being used in conjunction with an on-going effort to prepare corpora of sign language acquisition to share with the research community. This paper will briefly report on the development of the ASL Signbank, focusing on the adopted lemmatization principles. Lemmatization of ASL signs has never been done on a scale like this before - one that has been continually refreshed by actual usage data.

Keywords: lexical database, lemmatization, ASL, signbank

1. Introduction

Signbanks, usage-based lexical databases, have been created for several signed languages (Auslan, Johnston 2001; British Sign Language, Fenlon et al. 2014; Sign Language of the Netherlands, Crasborn et al. 2016; Finnish Sign Language, Salonen et al. 2016). Given the lack of conventionalized writing systems for signed languages, a best practice for annotating is to use ID glosses (Johnston 2001), unique gloss identifiers of signs. To keep an organized database of ID glosses and the signs they represent, they are added to the Signbank as the signs are observed in the primary data while annotating. Subsequently, the signs are organized in the database using lemmatization principles. We are creating a Signbank for American Sign Language (ASL), currently used in conjunction with SLAAASh, an ongoing effort to prepare corpora of sign language acquisition to share with the research community. The Signbank itself is to be made available to the public for general use. This paper will briefly report on the development of the ASL Signbank (which now has 2600+ entries), focusing on the lemmatization principles adopted from Fenlon et al. (2015), as well as on the workflow we have implemented for creation and maintenance of ID glosses We also touch upon how we use ASL Signbank for research.

2. SLAAASh and ASL Signbank

The Sign Language Acquisition, Annotation, Archiving and Sharing (SLAAASh) project is working with a digitized video corpus of Deaf children's use of ASL, collected as spontaneous production data from 4 Deaf children of Deaf parents, ages 1;04-4;01 (Lillo-Martin & Chen Pichler 2008). We are currently annotating the primary data systematically using our ID glosses and annotation conventions. We also are engaging re-consenting protocols (Chen Pichler et al. 2016) to document permission from the children (who are now adults) and others in the recordings to share their data for research purposes. And lastly, we are also working with others to develop a web-based platform for sign language data sharing. Taken together these activities create an annotation, archiving and sharing infrastructure that can be used by other research projects also studying ASL, exponentially increasing the availability of usage-based observations of signs for research.

The entries in the ASL Signbank are produced and coded by our Signbank team headquartered at Gallaudet University. This system is allowing us to organize the ID glosses we have been developing over many years throughout the various incarnations of our projects (Chen Pichler et al 2010; Chen Pichler et al 2015). At earlier stages of this process we used homegrown efforts to organize ID glosses (single folder on a single user's computer, shared Google Drive account, shared Dropbox account). Soon we discovered that these attempts were inefficient and that we needed to turn to a lexical database solution like a Signbank.

The ASL Signbank software is modelled on the NGT Signbank, which in turn is based on the Auslan Signbank software (Cassidy et al. 2018). The software is available for developers under a public license at http://github.com/Signbank/Global-Signbank/. The ASL Signbank infrastructure has been developed and maintained by Radboud University, but it is hosted by Haskins Laboratories and Yale University in the US. In addition to organization and access, an advantage of the ASL Signbank is the availability of direct linking to the ELAN annotation software (Crasborn et al. 2016), as part of version 5.0 (released October 2017).

2.1 ASL-LEX

The ASL Signbank team is collaborating with the team building ASL-LEX (Caselli et al 2016), a publicly-available database which includes subjective frequency and iconicity judgments as well as phonological information on 1,000 signs (and more to come). Our collaboration involves sharing ID glosses, so that signs that are common across the databases can be easily accessed, as well as phonological information, so that the signs have consistent coding.

We are building up both projects simultaneously, coordinating new entries as possible while accommodating the distinct project requirements. The goals of the ASL Signbank and ASL-LEX are somewhat different: the Signbank is based on usage data (e.g., ID glosses for signs are created as they occur in our child acquisition data, as well as in the data from other research projects that use our ID glosses), while the ASL-LEX project was designed to

include elicited signs in order to represent the full range from high to low frequency and high to low iconicity, for use in psycholinguistic experiments. Despite these different goals, the projects are mutually reinforcing. Our projects are linked together by the alignment of glosses (we use the same lemmas, although we may differ in annotation ID glosses, described in sections 3 and 4); shared phonological coding (using a simplified version of the Prosodic Model (Brentari 1998)); shared iconicity ratings (subjective iconicity ratings as well as iconicity categorization); and shared lexical properties (e.g., identification of lexical class). Eventually, the actual frequency data from our corpora (in child signing and child-directed signing) will help to tie the projects even closer together.

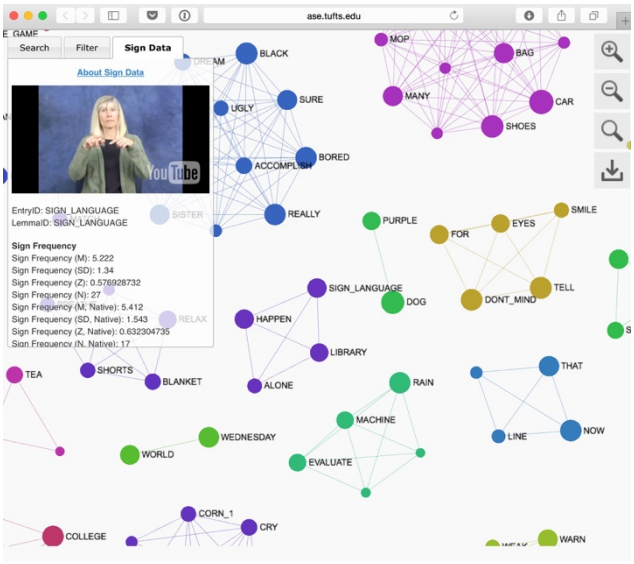


Figure 1. ASL-LEX, visualization view of phonological neighborhoods of some ASL signs

3. Overall description of signs

As with other Signbanks, our goal is to create an open-access lexical database of ASL signs with their ID glosses (Johnston 2001) to facilitate consistent and systematic annotation of sign usage in multiple data sets. Along with a movie and image of the sign and its ID gloss, each entry has information about the sign's formational components, its grammatical characteristics, and usage information. The categories of information about each sign used in the ASL Signbank have been derived from all prior Signbanks and prior annotation conventions (Chen Pichler et al 2015). As discussed briefly in 2.1, we also consider the data categories used by ASL-LEX. To illustrate, Figure 2 shows a record from the ASL Signbank for AGAIN.

Using our lemmatization principles (outlined in section 4), we have a lemma ID gloss for the sign as well as an annotation ID gloss which will be slightly different if the sign has phonological variants (which occurs when forms share all phonological features except for one or two). We enter "translation equivalents" (keywords) to facilitate the search for each sign and to represent the meaning of the sign. These can also be used in ELAN when the ASL Signbank is used as an external controlled vocabulary (ECV). This reduces the need for annotators to memorize ID glosses. The dialect field allows us specify any US region. The field "semantic field" is used to group together sets of signs that refer to the same subject. We inherited the fields from the NGT Signbank for the *morphology* section, allowing us to describe the make-up of compounds (See Crasborn et al. 2016). In the *phonology* section (coded in conjunction with ASL-LEX), we identify handedness, major location, minor location (beginning and final), dominant hand selected fingers and flexion as well as any abduction or flexion change, nondominant handshape and path movement. For the *morphosyntax* section, we identify the word or lexical class of the sign as well as its derivation history (lexicalized through fingerspelling, compounding, borrowing, et cetera), and type of iconicity. *Relations to other signs* allows us to connect ID glosses with homonyms, synonyms, variants, antonyms, hyponyms, hypernyms, and so on. *Relations to foreign signs* tracks any known connection with borrowed signs from other signed languages. *Frequency* will mark how many times the sign occurs in our corpus as well as the number of signers in the corpus who use that sign. *Publication status* and *notes* allow us to add metadata about the entry itself, especially useful for maintenance of ID glosses if they need to be changed.

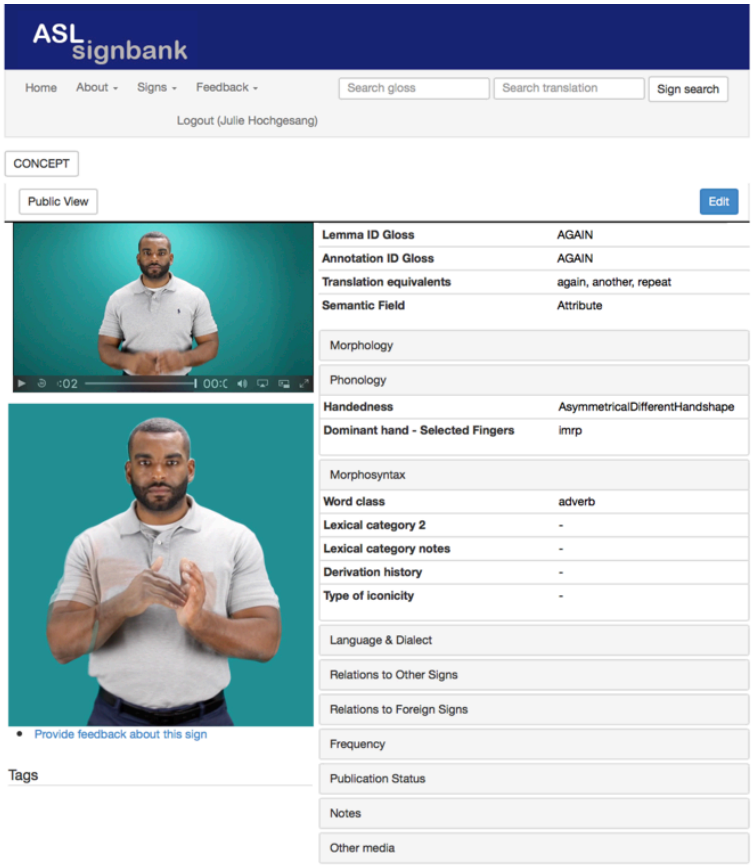


Figure 2. ASL Signbank entry for AGAIN

4. Lemmatization of ASL signs

The data in the SLAAASh corpus comes from four Deaf children, their Deaf parents, and others interacting with them. Clearly, this cannot be considered a representative selection of ASL signers or even of ASL acquirers. In this way, the SLAAASh corpus-building may be seen as different from current sign language corpora projects like the BSL Corpus project. Nonetheless, our treatment of

signs is the same – each citation form gets its own gloss. We did not pre-determine a list of signs to be included (as dictionaries might do), but assign ID glosses as we come across the signs in the primary video data. Initial assignment of ID glosses did not follow lemmatization principles by determining which forms are related to which lexemes; the only real rule we had was to give each sign form a different ID gloss. After enough entries began to accumulate, we were able to modify the organization and assignment of ID glosses on the basis of lemmatization principles described in Fenlon et al (2015). To our advantage, the connection between Signbanks and ELAN made possible in recent releases (including an external controlled vocabulary generated by a Signbank, and a Signbank Lexicon Service in ELAN) permits changes in annotation glosses recorded in ASL Signbank to be promulgated throughout the annotations in our corpus.

We generally follow the same principles as laid out in Fenlon et al (2015):

> *…we consider the citation form to be the lemma (i.e. the unmodified form of a given sign is used here as the headword of a lexeme)…The ID gloss is a unique English-based translation used primarily as an annotation tag in the corpus for all occurrences of that lexeme regardless of how it might be modified. It is important to note that the choice of the English word as an ID gloss for a particular lexeme is not meant to indicate the sign's core meaning or grammatical function. It is merely a label to uniquely identify each lexeme, to be used in annotation of sign language data, in lieu of any standardised orthography for the language. [However] (for) the purposes of annotation… it is much more useful to use ID glosses that have some meaningful connection to the lexeme, e.g. via one of the translation equivalents, since annotation is done by typing in the ID gloss" (176).*

The lemmatization principles are simple at first glance: if the meaning and the form of two entries are different, they are different lemmas and get different ID glosses; if the meaning is similar and there are one or two phonological differences, they are under the same lemma and get the same core annotation ID glosses, with the difference indicated by lowercase tags that identify the particular formational aspect responsible. For example, the ASL signs for "believe" and "precious" (Figure 3) are clearly different forms. One is two-handed, the other is one-handed.

Figure 3: ASL signs for "believe" and "precious" with their ID glosses

There are different locations, handshapes, path, and other formational features characterizing each sign. Also the meanings are, of course, quite different. So with that, they are deemed different lemmas and accordingly get unique glosses.

In Figure 4, it can be seen that two forms for 'believe' are similar but clearly different in at least one aspect, specifically the initial handshape for the dominant or strong hand (for the first and for the second). Their meaning, however, is the same. Given that, we treat them as phonological variants linked to the same lemma but with unique annotation ID glosses (marked by the lowercase tags, i.e., "b" and "ix").

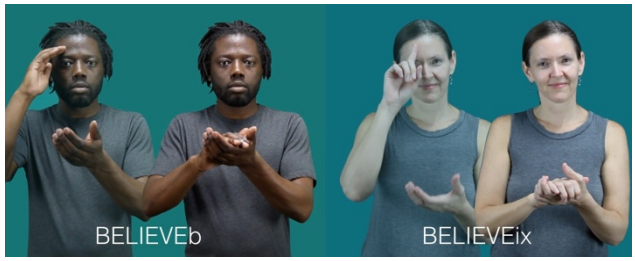


Figure 4: ASL variants for "believe" with their ID glosses

So, signs with phonological variants (e.g., BELIEVEb and BELIEVEix) are annotated with distinct ID glosses in our ELAN files, which are linked to ASL Signbank (and connected under the same lemma). The signs in Figure 5 are examples of signs under different lemmas. They are synonyms but have clearly distinct phonological forms.

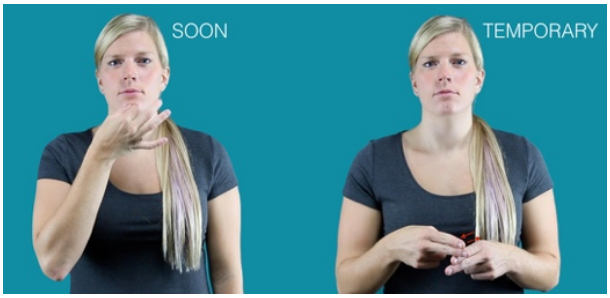


Figure 5: ASL signs for "soon" and "temporary" with their ID glosses

The signs for "soon" and "temporary" receive unique ID glosses but are marked as "related" in the ASL Signbank.

In practice, this adherence to lemmatization has been remarkably difficult but instructive. For instance, the ASL sign for "equal" (Figure 6) can have a single set of movements (bent hands move together to touch) or repeated sets.


Figure 6: ASL sign for "equal"

The "equal" forms will vary based on syntactic placement as well as intended meaning. Are they of a single lemma with one of the forms a modification of the lemma? Or are they two separate lemmas with conventionalized separate meanings? We have tentatively left this example as a single lemma with one annotation ID gloss. We will revisit it once we have a sufficient number of examples in the corpora that use the ASL Signbank.

On the other hand, the ASL signs for "show" and "example" are produced similarly but one ("show") has one set of movements and the other ("example") has shorter and repeated movements (Figure 7). With this slightly different phonological form and their conventionalized different meanings, they are separate lemmas.

Figure 7: ASL signs for "show" and "example".

These lemmatization decisions are made both by observing how these signs behave in the dataset and how they are understood by the researchers. Regular lab meetings are held to discuss sign lemmas. Frequently the discussion involves producing the sign in various modifications. For example, if the two signs are verbs, they can be modified to reflect grammatical aspect. If they are changed in the same way, they are deemed the same lemma. For example, these two forms in Figure 8 appear to be the same at first glance but they are used in different contexts and cannot be changed in the same way when modified for aspect.

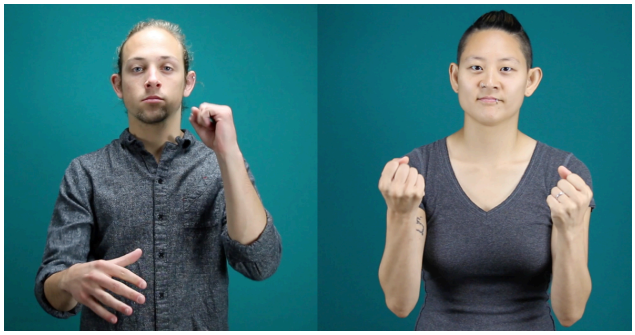


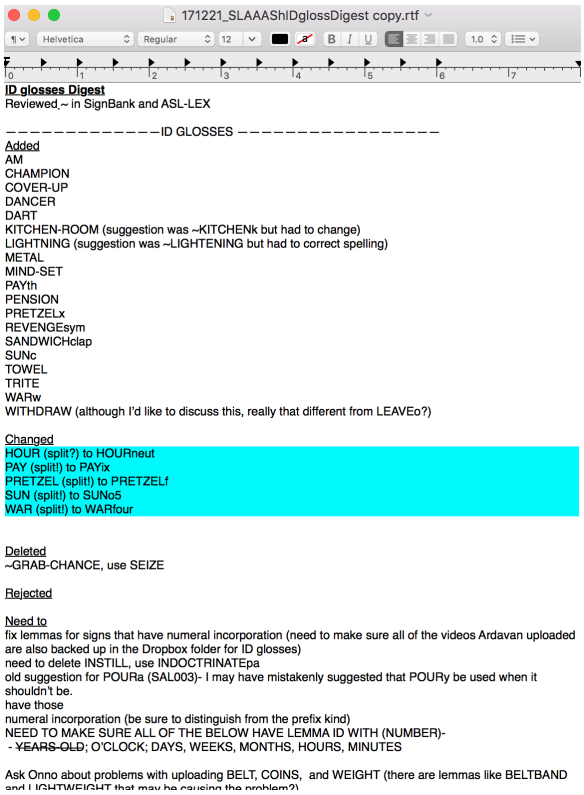
Figure 8. ASL forms basically meaning "to sign" but are different lemmas because they are used differently

For signs that may not be as lexically fixed but are specific types of signs, we use regular codes to annotate them, e.g., DS for "depicting sign", NS for "name sign", IX for "index" or pointing sign. Specific referents are added to related tiers in the annotation files. (See SLAAASh annotation conventions for more, (Hochgesang (2015)).

5. Workflow of creation and maintenance of ID glosses

The maintenance of the ID glosses is under the supervision of one person, currently Julie Hochgesang, on the SLAAASh research team. All of the annotators for SLAAASh and other research projects who use the ASL Signbank are required to follow a specific protocol for suggesting an ID gloss when the sign they need to annotate is not in the Signbank. After ensuring that they have exhausted all possibilities by searching translation equivalents on ASL Signbank, they then propose an ID gloss using their understanding of the ASL Signbank lemmatization principles. Since the SLAAASh eafs are linked to the ASL Signbank ECV, they must force the annotation field to escape the list in order to enter new entries. They prefix their suggestion with ~, e.g., ~PROPOSED-NEW-IDGLOSS. They then take a video of themselves producing the sign and upload this to the ASL Signbank. They click "proposed new sign" and add the tag "proposed new ID gloss needs review". These three steps are a triple safeguard against errors in the annotation files and accidentally adding them as permanent additions to the ASL Signbank.

The ID gloss supervisor then reviews the suggestions and ensures that the additions are not duplicates of already existing signs. Lemmatization principles as outlined in section 4 are applied. The sign is then marked as approved and tagged to be refilmed. Native/early signers are hired to produce signs which are then published in ASL Signbank. ID gloss digests (shown in Figure 9) are reports of additions, deletions, changes, ongoing issues and are shared with the entire SLAAASh project team as well as others who are registered users of the ASL Signbank.



171221_SLAAAShIDglossDigest copy.rtf

Helvetica Regular 12 1.0

ID glosses Digest

Reviewed,~ in SignBank and ASL-LEX

----------ID GLOSSES ----------

Added

AM
CHAMPION
COVER-UP
DANCER
DART
KITCHEN-ROOM (suggestion was ~KITCHENk but had to change)
LIGHTNING (suggestion was ~LIGHTENING but had to correct spelling)
METAL
MIND-SET
PAYth
PENSION
PRETZELx
REVENGEsym
SANDWICHclap
SUNc
TOWEL
TRITE
WARw
WITHDRAW (although I'd like to discuss this, really that different from LEAVEo?)

Changed

HOUR (split?) to HOURneut
PAY (split!) to PAYix
PRETZEL (split!) to PRETZELf
SUN (split!) to SUNo5
WAR (split!) to WARfour

Deleted

~GRAB-CHANCE, use SEIZE

Rejected

Need to

fix lemmas for signs that have numeral incorporation (need to make sure all of the videos Ardavan uploaded are also backed up in the Dropbox folder for ID glosses)
need to delete INSTILL, use INDOCTRINATEpa
old suggestion for POURa (SAL003)- I may have mistakenly suggested that POURy be used when it shouldn't be.
have those
numeral incorporation (be sure to distinguish from the prefix kind)
NEED TO MAKE SURE ALL OF THE BELOW HAVE LEMMA ID WITH (NUMBER)-
- ~~YEARS-OLD~~; O'CLOCK; DAYS, WEEKS, MONTHS, HOURS, MINUTES

Ask Onno about problems with uploading BELT, COINS, and WEIGHT (there are lemmas like BELTBAND and LIGHTWEIGHT that may be causing the problem?)

Figure 9: Screenshot of ID Gloss Digest

6. Use of ASL Signbank for Research

The ASL Signbank contributes to research in two broad ways. First, there are research projects that make use of the data in the Signbank itself, along with the connections to ASL-LEX. Second, there is research that is enabled by the use of Signbank in annotation of sign language data such as the SLAAASh project.

Because the Signbank includes information about the form of each sign as well as morpho-syntactic and other information, it is possible to conduct analyses based on the signs in the Signbank that would typically exceed the number of examples based on other methods. For example, analyses of the frequency of occurrence of specific phonological elements (e.g., selected fingers) can readily be made to test previous claims about markedness. Another example is the occurrence of forms that violate Battison's (1974) symmetry and dominance constraints. Although our original coding system assumed these constraints would hold, we discovered a (relatively small) number of signs that violate the constraints, and adjusted the phonological coding options accordingly.

In combination with data in ASL-LEX, it will be possible to test hypotheses about a number of questions, including extending some that have already been examined using ASL-LEX alone. For example, Caselli & Pyers (2017) used ASL-LEX data to examine the iconicity and phonological neighborhood density of signing children's vocabulary development. With child-produced and child-directed frequency information to be made available in ASL Signbank, this kind of study can be extended.

As the ASL Signbank is used in annotating primary sign language data such as the SLAAASh corpus, it will make further research possible. Using multiple file searching functions of ELAN, it is possible to identify all instances in the corpus of signs of interest, which have been uniformly annotated because of the Signbank. As a further tool, we anticipate using lexical category information in Signbank to automatically tag SLAAASh data, which can be further tested in various ways. One only need consider the vast amount of research that has been made possible by the CHILDES database (https://childes.talkbank.org/) to anticipate the range of possible studies that will be forthcoming.

7. Conclusion

The actual usage of each sign in the corpora informs the ASL Signbank, from the most basic questions (which signs to include) to refinement of the postulated linguistic features. As our data set grows, our ability to answer these kinds of questions will improve. Lemmatization of ASL signs has never been done on a scale like this before - one that has been continually refreshed by actual usage data.

8. Acknowledgements

The research reported here was supported in part by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under award number R01DC013578 and award number R01DC000183. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The work reported here is not possible without the entire SLAAASh research team, including Amelia Becker (who also provided feedback on drafts of this paper), Donovan Catt, Anna Lim Franck, Ardavan Guity, Carmelina Kennedy, Laura Mahan, Matthew Nardozza, Lettie Nazloo, Deborah Peterson, Lee Prunier, Doreen Simons, Phoebe Tay, Jacob Veeder, and all of the ASL SignBank actors. slla.lab.uconn.edu

The ASL Signbank was developed at Radboud University by Onno Crasborn, Wessel Stoop, Micha Hulsbosch, and Susan Even.

9. Bibliographical References

Battison, R. (1974). Phonological deletion in American Sign Language. *Sign Language Studies, 5*, 5-19.

Brentari, D. (1998). A prosodic model of sign language phonology. Cambridge, MA: MIT Press.

Caselli, N., Sevcikova, Z., Cohen-Goldberg, A., Emmorey, K. (2016). ASL-Lex: A Lexical Database for ASL. Behavior Research Methods. doi:10.3758/s13428-016-0742-0

Caselli, N. & J. Pyers. (2017). The road to language learning is not entirely iconic: Iconicity, neighborhood density, and frequency facilitate acquisition of sign language. Psychological Science. Vol 28, Issue 7, pp. 979 – 987.

Cassidy, S., Crasborn, O., Nieminen, H., Stoop, W., Hulsbosch, M., Even, S., Komen, E., Johnston, T. (2018). *Signbank: Software to Support Web Based Dictionaries of Sign Language*. Paper to be presented at the LREC 2018, Miyazaki, Japan.

Chen Pichler, D., Hochgesang, J.A., Lillo-Martin, D. (2015, March). BiBiBi Project ASL Annotation Conventions. Poster presented at "Digging into Signs Workshop: Developing Annotation Standards for Signed Language Corpora". University College London, London, United Kingdom (March 30-31, 2015).

Chen Pichler, D., J.A. Hochgesang, D. Lillo-Martin, & R. Quadros. (2010). Conventions for sign and speech transcription in child bimodal bilingual corpora. Languages, Interaction and Acquisition, 1(1), 11-40.

Chen Pichler, D., J.A. Hochgesang, D. Simons & D. Lillo-Martin. (2016). Community Input on Re-consenting for Data Sharing. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J.A. 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: European Language Resources Association (ELRA).

Crasborn, O. & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood & E. Thoutenhoofd (Eds.), Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the

Representation and Processing of Sign Languages. ELRA, Paris, pp. 39-43.
Crasborn, O., Bank, R., Zwitserlood, I., Kooij, E. van der, Schüller, A., Ormel, E., Nauta, E., Zuilen, M. van, Winsum, F. van., & Ros, J. (2016). *Linking Lexical and Corpus Data for Sign Languages: NGT Signbank and the Corpus NGT*. Paper presented at the The 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining, Portorož, Slovenia
Crasborn O., R. Bank, I. Zwitserlood, E. Kooij, E. Ormel, J. Ros, A. Schüller, A. de Meijer, M. van Zuilen, Y. Nauta, F. van Winsum & M. Vonk (2017). NGT Signbank. Nijmegen: Radboud University, Centre for Language Studies. ISLRN: 976-021-358-388-6, DOI: 10.13140/RG.2.1.2839.1446.
Fenlon, J., Cormier, K, Schembri, A. (2015). Building BSL Signbank: The Lemma Dilemma Revisited. *International Journal of Lexicography*, 28(2), 169-206.
Johnston, T. (2001). The lexical database of Auslan (Australian Sign Language). *Sign Language & Linguistics, 4*(1/2), 145-169.
Fenlon, J., K. Cormier, R. Rentelis, A. Schembri, K. Rowley, R. Adam, & B. Woll. (2014). BSL SignBank: A lexical database of British Sign Language (First Edition). London: Deafness, Cognition and Language Research Centre, University College London.
Fisher, J.N., J.A. Hochgesang, & M. Tamminga. (2016). Examining Variation in the Absence of a 'Main' ASL Corpus: The Case of the Philadelphia Signs Project. Proceedings of the 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining. (bit.ly/PhiladelphiaSigns) 75-80. LREC, Portorož, Slovenia, May 28 2016.
Hochgesang, J.A. (2015, updated 2016) SLAAASh ID Glossing Principles and Annotation Conventions. Ms., Gallaudet University and Haskins Laboratories. http://bit.ly/2jGbPfU
Johnston, T. (2001). The lexical database of Auslan (Australian Sign Language). Sign Language & Linguistics, 4(1/2), 145-169.
Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. International Journal of Corpus Linguistics, 15(1), 106-131.
Lillo-Martin, Diane & Chen Pichler, Deborah (2008). Development of Sign Language Acquisition Corpora. In O. Crasborn, E. Efthimiou, T. Hanke, E.D. Thoutenhoofd, & I. Zwitserlood, (Eds.), Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora; 6th Language Resources and Evaluation Conference, 129-133.
Salonen, J., Takkinen, R., Puupponen, A., Nieminen, H., & Pippuri, O. (2016). Creating Corpora of Finland's Sign Languages. 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) (pp. 179-184). Paris: European Language Resources Association (ELRA).

10. Language Resource References

ASL Signbank (2018). Lexical database for American Sign Language. URL: http://aslSignbank.haskins.yale.edu.