A New Web Interface to Facilitate Access to Corpora: Development of the ASLLRP Data Access Interface (DAI)

Carol Neidle*, Christian Vogler**
*Boston University, **Gallaudet University
*Boston University Linguistics Program, 621 Commonwealth Avenue, Boston, MA 02215
**Gallaudet University, Technology Access Program, 800 Florida Ave NE, Washington DC 20002
E-mail: carol@bu.edu, christian.vogler@gallaudet.edu

Abstract

A significant obstacle to broad utilization of corpora is the difficulty in gaining access to the specific subsets of data and annotations that may be relevant for particular types of research. With that in mind, we have developed a web-based Data Access Interface (DAI), to provide access to the expanding datasets of the American Sign Language Linguistic Research Project (ASLLRP). The DAI facilitates browsing the corpora, viewing videos and annotations, searching for phenomena of interest, and downloading selected materials from the website. The web interface, compared to providing videos and annotation files off-line, also greatly increases access by people that have no prior experience in working with linguistic annotation tools, and it opens the door to integrating the data with third-party applications on the desktop and in the mobile space. In this paper we give an overview of the available videos, annotations, and search functionality of the DAI, as well as plans for future enhancements. We also summarize best practices and key lessons learned that are crucial to the success of similar projects.

Keywords: web interfaces, access to corpora, corpus management

1. Introduction

Linguistically annotated video corpora for signed languages can be enormously valuable for research in linguistics and computer-based sign language recognition, with many other potential applications, including education. Construction of such corpora is time-consuming, and linguistically controlled data collection yielding high-quality video files requires resources and interdisciplinary collaboration. The substantial investment in corpus development will have greatest benefit if corpora can be shared widely.

A significant obstacle, however, to broad utilization is the difficulty in gaining access to the specific subsets of data and annotations that may be relevant for particular types of research. With that in mind, we have developed a web-based Data Access Interface (DAI), to provide access to the expanding datasets of the American Sign Language Linguistic Research Project (ASLLRP), available at http://secrets.rutgers.edu/dai/queryPages/. The DAI facilitates browsing the corpora, viewing videos and annotations, searching for phenomena of interest, and downloading selected materials from the website. We have also found these same tools invaluable for verifying the consistency of our annotations.

Here we give an overview of available video files and linguistic annotations, summarize current functionalities of the DAI, and discuss directions for ongoing development. We also offer some lessons learned that might be of interest to others engaged in corpus management.

2. Available data sets

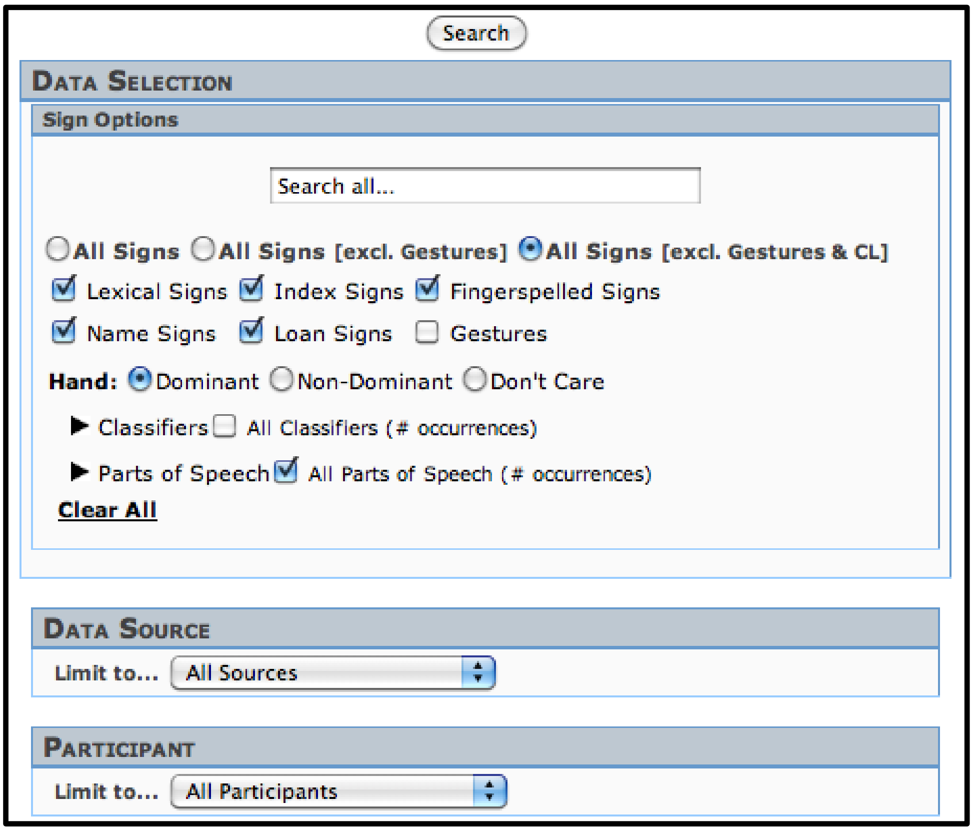
The DAI now allows access to the National Center for Sign Language and Gesture Resources (NCSLGR) Corpus, ASL videos collected and linguistically annotated at Boston University. Synchronized video files, available in compressed and uncompressed formats, show the signing from the front and side and include a close-up view of the face. Linguistic annotations of manual and nonmanual components of the signing have been carried out using SignStream® (Neidle 2002b) and are available in XML format. Manual signs are represented by unique gloss labels. Annotation conventions are documented (Neidle 2002a, 2007).

Annotations are available for 19 short narratives (1002 utterances) plus 885 additional elicited utterances, all from Deaf native signers of ASL (with most of these data coming from four signers). This constitutes a total of 1,888 linguistically annotated utterances, including 1,920 distinct canonical signs (grouping together close variants) and 11,861 total sign tokens.

Linguistic annotations include the start and endpoints of each sign, identified by a unique gloss label, part of speech, and start and end points of a range of non-manual behaviors (e.g., raised/lowered eyebrows, head position and periodic head movements, expressions of the nose and mouth) also labeled with respect to the linguistic information that they convey (serving to mark, e.g., different sentence types, topics, negation, etc.). The annotations are available via an XML format. For the DTD and documentation of the XML format, see http://www.bu.edu/asllrp/ncslgr-for-download/download-info.html.

3. Functionalities of the interface

As shown in Figure 1, the DAI user can search for specific text in gloss fields and can narrow the search to specific classes of signs or search for particular types of classifiers or parts of speech. Figure 2 displays a small section of the alphabetical listing of all signs (based on the selection in Figure 1), with sign variants grouped together, enabling the user to select a particular sign or variant (and potentially a specific signer). Selecting FINISH from this



Search

DATA SELECTION

Sign Options

Search all...

○ All Signs ○ All Signs [excl. Gestures] ● All Signs [excl. Gestures & CL]

☑ Lexical Signs ☑ Index Signs ☑ Fingerspelled Signs

☑ Name Signs ☑ Loan Signs □ Gestures

Hand: ● Dominant ○ Non-Dominant ○ Don't Care

▶ Classifiers □ All Classifiers (# occurrences)

▶ Parts of Speech ☑ All Parts of Speech (# occurrences)

Clear All

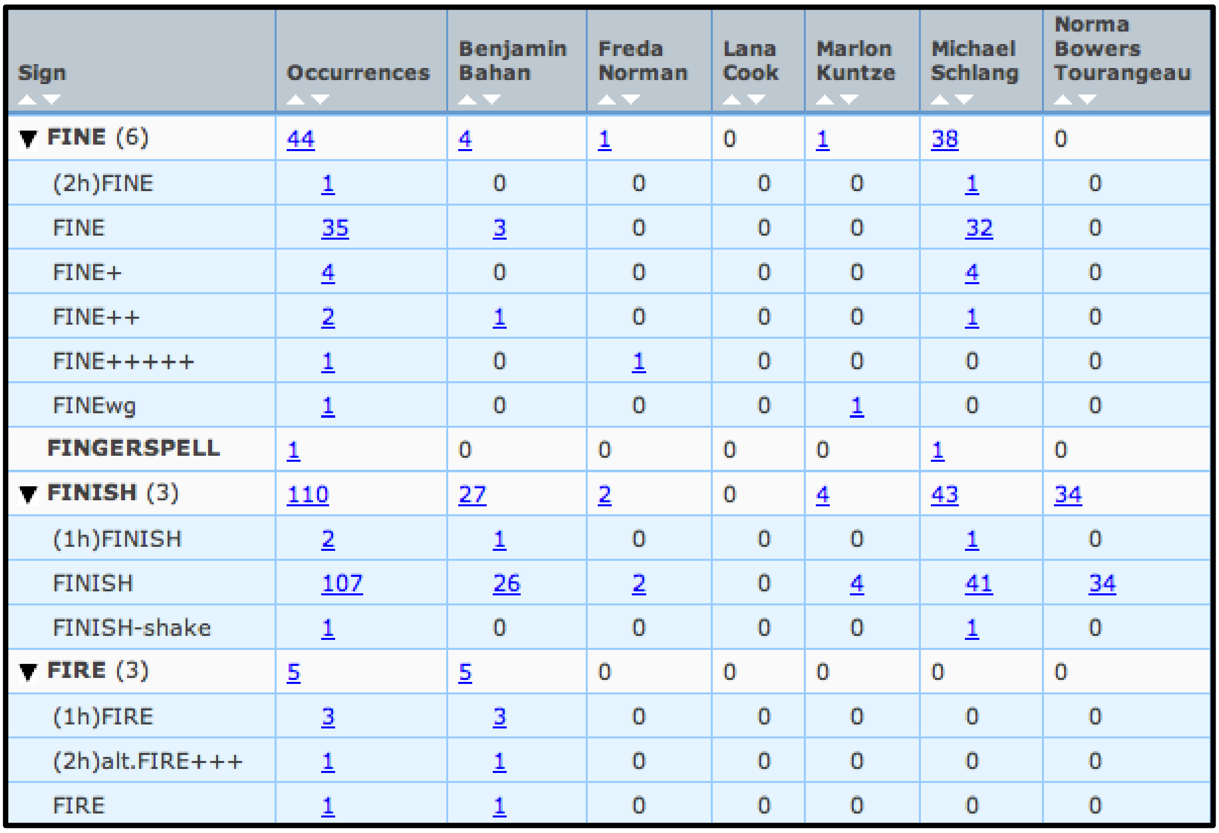
DATA SOURCE

Limit to... All Sources

PARTICIPANT

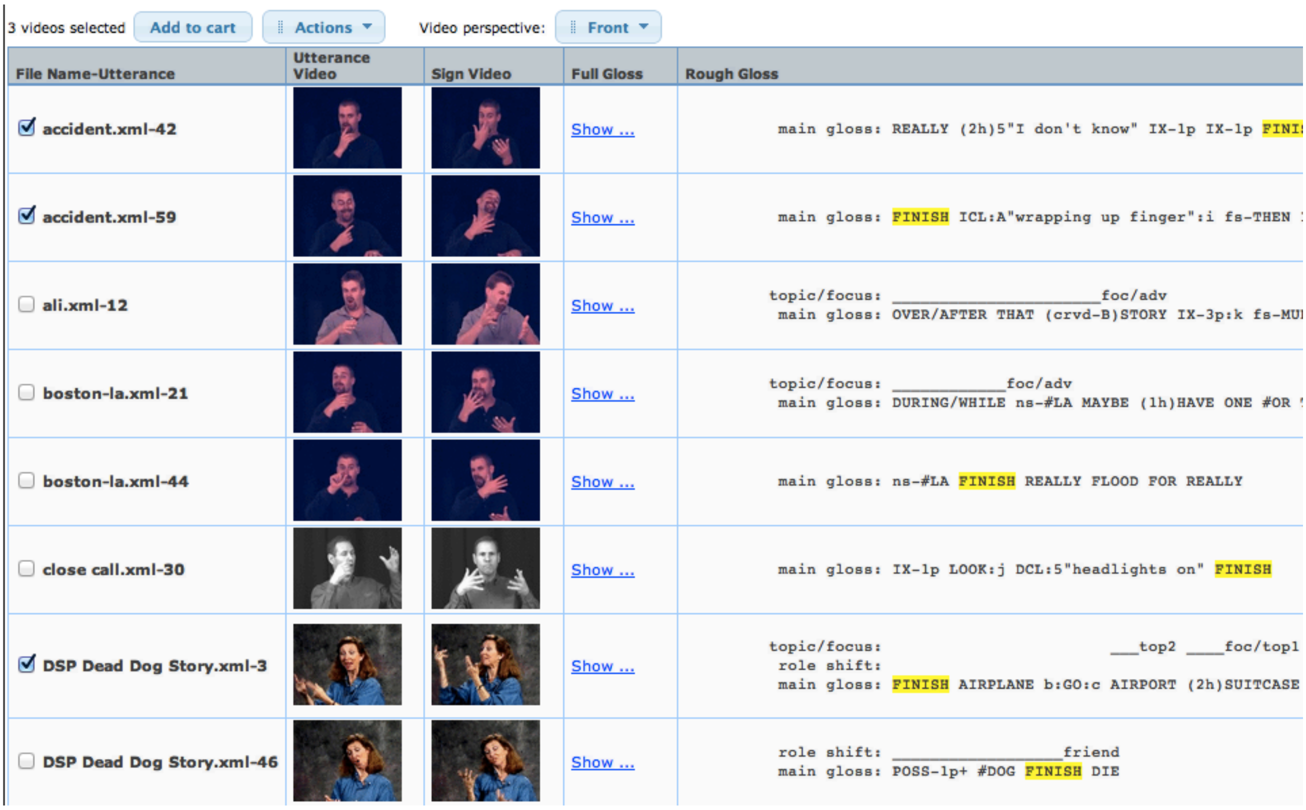
Limit to... All Participants

Figure 1: DAI Screenshot showing sample search query



Sign	Occurrences	Benjamin Bahan	Freda Norman	Lana Cook	Marlon Kuntze	Michael Schlang	Norma Bowers Tourangeau
▼ **FINE** (6)	44	4	1	0	1	38	0
(2h)FINE	1	0	0	0	0	1	0
FINE	35	3	0	0	0	32	0
FINE+	4	0	0	0	0	4	0
FINE++	2	1	0	0	0	1	0
FINE+++++	1	0	1	0	0	0	0
FINEwg	1	0	0	0	1	0	0
FINGERSPELL	1	0	0	0	0	1	0
▼ **FINISH** (3)	110	27	2	0	4	43	34
(1h)FINISH	2	1	0	0	0	1	0
FINISH	107	26	2	0	4	41	34
FINISH-shake	1	0	0	0	0	1	0
▼ **FIRE** (3)	5	5	0	0	0	0	0
(1h)FIRE	3	3	0	0	0	0	0
(2h)alt.FIRE+++	1	1	0	0	0	0	0
FIRE	1	1	0	0	0	0	0

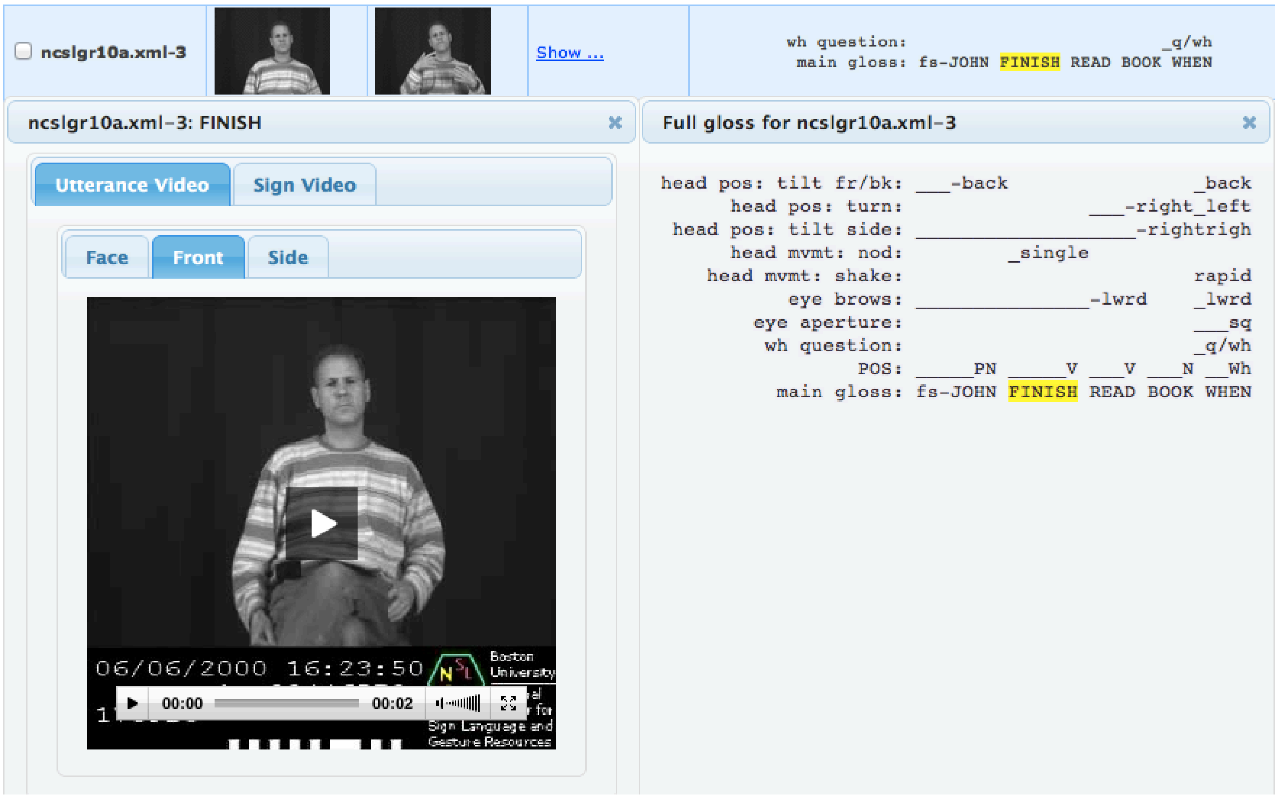
Figure 2: DAI Screenshots showing subset of results



3 videos selected | Add to cart | Actions ▾ | Video perspective: | Front ▾

File Name-Utterance	Utterance Video	Sign Video	Full Gloss	Rough Gloss
☑ accident.xml-42			Show ...	main gloss: REALLY (2h)5"I don't know" IX-1p IX-1p FINI
☑ accident.xml-59			Show ...	main gloss: FINISH ICL:A"wrapping up finger":i fs-THEN
□ ali.xml-12			Show ...	topic/focus: ______________________foc/adv main gloss: OVER/AFTER THAT (crvd-B)STORY IX-3p:k fs-MU
□ boston-la.xml-21			Show ...	topic/focus: ____________foc/adv main gloss: DURING/WHILE ns-#LA MAYBE (1h)HAVE ONE #OR
□ boston-la.xml-44			Show ...	main gloss: ns-#LA FINISH REALLY FLOOD FOR REALLY
□ close call.xml-30			Show ...	main gloss: IX-1p LOOK:j DCL:5"headlights on" FINISH
☑ DSP Dead Dog Story.xml-3			Show ...	topic/focus: ___top2 ____foc/top1 role shift: main gloss: FINISH AIRPLANE b:GO:c AIRPORT (2h)SUITCASE
□ DSP Dead Dog Story.xml-46			Show ...	role shift: __________________friend main gloss: POSS-1p+ #DOG FINISH DIE

Figure 3: Display of sentences with FINISH, with annotations selected for later download.



□ ncslgr10a.xml-3 | Show ...

```
 wh question:                        _q/wh
main gloss: fs-JOHN FINISH READ BOOK WHEN
```

ncslgr10a.xml-3: FINISH ✖

Utterance Video | Sign Video

Face | Front | Side

06/06/2000 16:23:50

00:00 00:02

Full gloss for ncslgr10a.xml-3 ✖

```
head pos: tilt fr/bk: ___-back                     _back
      head pos: turn:                    ____-right_left
 head pos: tilt side: ___________________-rightrigh
      head mvmt: nod:         _single
    head mvmt: shake:                          rapid
           eye brows: _______________-lwrd     _lwrd
        eye aperture:                          ___sq
         wh question:                          _q/wh
                 POS: _____PN _____V ___V ___N __Wh
          main gloss: fs-JOHN FINISH READ BOOK WHEN
```

Figure 4: User can view the detailed gloss, and play movies for the sign and the utterance in which it occurs from multiple viewpoints

chart would bring up the display in Figure 3, which includes still images of the relevant material; here the user can switch among the available camera perspectives for each annotation (frontal and face, in some cases also side and stereo camera view), and mark annotations for later download. It is also possible to play back online the videos corresponding to just the sign FINISH or to the utterance containing it, and to display a more complete transcription (Figure 4).

After annotations have been marked for later download, users can call up the download tool. This tool allows them to select which specific video files to download for the selected annotations, where the available choices are sign only, the utterance containing the sign, or the entire story video in which the sign occurs, or any combination of these (along with the linguistic annotations in the XML format). This greatly increases the utility of the DAI, as it is possible to focus on specific signs or linguistic phenomena, and easily obtain a collection of all available videos that exhibit them.

4. Best practices and lessons learned

Managing large corpora and making them available to the community entails a unique set of challenges. We present some key lessons that we believe are essential to the success of any similar project.

4.1. Presentation of data

Presenting signs as still images saves users and annotators time and effort. If the start and end frames of each annotation are presented as thumbnail images, users may be able to detect at a glance whether an annotation is of interest. As compared with having access only to videos (which are time-consuming to watch), availability of still images also greatly speeds up validation and consistency checks – if an annotation is inconsistent with the other ones in the same category, it is likely to manifest in a difference in still frames.

4.2. Resource Management

Keep metadata separate from file names and assets. Enforcing a consistent coding scheme across thousands of file names and file headers is nearly impossible. It is much easier to keep metadata consistent and up-to-date if it is encoded in a centralized spreadsheet or database. Note that although it may seem to be useful to have some indication of the file's contents in the file naming convention, the downside is that if any of the metadata changes or is corrected later, the file name also would have to be updated to reflect the change, which can break existing external links to the asset.

Designate only one asset as the authoritative source on metadata, and auto-generate other assets from there. Having metadata available in multiple formats is often unavoidable; for example, it may need to be present in the database tables, in a spreadsheet for easy manipulation by the team maintaining the corpus, as a web page, and in a textual format for easy distribution to third parties. Unfortunately, there is a high risk of ending up with conflicting metadata for assets, which would result in having to sort out the conflict manually in a laborious process. It follows that only one of these formats can be updated with new and corrected information, and it must be very clear throughout the lifecycle of the project which one it is to be. Moreover, all other metadata assets need to be automatically (i.e. programmatically) generated from the authoritative source, so as to avoid introducing inconsistencies due to human error. Automating this process also makes it more likely that the information is always kept up-to-date across all formats.

Separate file location and names. Files can move, as systems are upgraded, or redundancy is built in. If the location is encoded separately, only this part needs to be updated, rather than every link to a file. The DAI uses a two-part schema of the form:

<url prefix> <path to file>

where URL prefix points to a location on the server that hosts a collection of related content, such as all XML annotation files, or all videos from a specific camera. Moving the collection to a different location entails updating only a single row in the table that contains the affected URL prefix.

Be mindful of cross-platform issues. Different operating systems have different restrictions on file names; for instance, colons are not allowed on Windows. This can cause problems both for users who want to download the data sets, and for copying the data across hard drives with different file systems (such as copying from HFS+ to NTFS and vice versa). In a large corpus that has thousands or even tens of thousands of assets, running into these problems can result in significant delays and expenses. Choosing the intersection of all the restrictions on Windows, Mac OS X and Unix variants – or even restricting file names to alphanumeric characters – is the safest way to proceed, and should be planned and done before any of the data are collected.

4.3. Development Processes

Plan for continuity. In an academic environment, the design and development must be managed by a project lead, who can commit long-term, and who has the skills to review other contributors' designs and code. Leaving students, who can drop out at any moment or graduate, in charge of the project will induce significant expenses and delays. The project lead, in particular, must understand the overall design of the project, so as to hold hands with new members while they get up to speed on the design and code.

Use version control on all source files *and* third-party libraries. The time *will* come when a bug is introduced that can be triaged only by investigating an earlier project revision. Any third-party dependencies must be included in those revisions to guard against the possibility of newer versions of a library being incompatible with the older version of the source code, or the case where a third-party library introduces a bug. Having version control also allows for easy separation of development and release branches, and makes it easy to fix bugs on the release branch, without having to wait for the development branch to get into a releasable state. Modern distributed version control systems, such as Git and Mercurial, make this mode of development especially simple and painless for the developers.

4.4. Database Design

Think queries, not data format. The types of queries that need to be supported drive the design of the database and the tables. They inform every decision that pertains to the tables, the relationships between tables, database views, and choice of indices, and can result in a representation of the annotations that is markedly different from the one chosen for the annotation file format. Doing the design off the annotation file format is a sure way to run into data management and performance problems down the road. For a concrete example, consider the organization of information in tiers in the annotation file formats and in the program chosen to carry out the annotations. If two tiers are tightly linked – such as in the case of tagging a gloss with the part of speech and the canonical form of the sign –, queries are much more efficient if this linkage is made clear in an explicit relationship in the database tables, rather than using a generic tier model.

Use a collection of tags. Standardized tagging of annotations (e.g., is a sign fingerspelled? plural? does it use a non-standard location? etc.) provides a powerful and efficient way to search for specific linguistic phenomena. In fact, in the DAI the distinctions among lexical signs, loan signs, classifiers, name signs, fingerspelled signs, indexed signs, and gestures are implemented in this manner (see also Figure 1). In the annotation file formats some of this information may come from separate tiers or be implicit in the naming conventions of glosses. In the DAI database population process, however, this information is extracted and put in an explicit relationship with the annotations, as explained in the previous point on queries.

5. Plans for future development and integration of additional data types

Planned enhancements to the DAI include:

1) Integration of other types of corpora;
2) Functionalities to enable additional types of searches;
3) Providing annotations in additional formats;
4) Display of various kinds of statistical information;
5) Integration of new technologies, as they become available.

5.1. Integration of additional types of corpora

The interface will be modified to allow integration of other types of corpora, including the American Sign Language Lexicon Video Dataset (ASLLVD), a corpus containing over 3,000 citation forms of lexical signs, each produced by between 1 and 6 native signers, resulting in a total of about 9,000 tokens, which have been annotated for start and end handshapes, among other things (Neidle et al. 2012). Design decisions will have to be made about how best to allow users to move easily among the different types of data sets, e.g., to look up a sign to see variations in production of citation forms by different signers, and to see the sign in context in examples from our corpus of continuous signing.

We would also like to allow access, through this interface, to portions of the *Deaf Studies Digital Journal* (DSDJ) http://dsdj.gallaudet.edu/, edited by Ben Bahan and Dirksen Bauman. (See also http://www.gallaudet.edu/News/Pioneering_digital_journal_to_launch_November_4.html.)

5.2. Additional search functionalities

It will, before long, be possible to search for

- Grammatical constructions, such as questions (of various types), negations, conditionals, relative clauses (correlatives), topics;
- Nonmanual signals, such as eye aperture, head tilt, raised/lowered eyebrows, body lean;
- Words in the English translation field.

Searches for text based on Sign ID (represented by a unique English-based gloss label) corresponding to a specific ASL sign will include the ability to restrict text searches to whole word (by default) or to search for text strings. We will incorporate searches based on:

- Video properties: e.g.: types of available viewpoints (frontal, side, stereo, face); availability of color video
- Availability of calibration data
- Subject wearing long/short sleeves
- Subject wearing glasses.

We will also explore the possibilities of allowing searches to be based on:

- Frequency (making it possible to search for items that have a minimum number of tokens);
- Sign duration;
- Number of subjects (making it possible to search for productions of a minimum number of signers);
- Specific signers (making it possible to view the set of productions of one or several individuals);
- Characteristics of signers (particularly as the corpus grows), such as gender or age range.

Furthermore, since the new corpora will include other types of annotations, we will also need to extend the search functionalities to enable appropriate searches of those data sets. This will include the ability to limit searches to specific sign types (lexical signs, index signs, classifier constructions of different types, fingerspelled signs, loan signs, name signs). We will also provide a way to search for the initial and/or final hand shape for the sign, or other phonological properties of the sign (e.g., signs containing a particular hand shape or movement type, or signs articulated with one or two hands).

We welcome suggestions about features that might be useful for different communities of potential users. This web-based interface could be especially useful for those who use ASL as a primary language and for those learning the language. We will be working with prospective users from these groups to design tools to facilitate the kinds of access that might be anticipated.

5.3. Annotations in other formats

Although the annotations currently are available in an easy-to-parse SignStream-specific XML format, we realize that researchers have their own preferences with respect to what annotation software they use. We plan to make the annotations, at a minimum, available in the ELAN EAF format and welcome suggestions as to what other formats should be supported.

5.4. Display of statistical information about the corpora

We plan to add functionality to view statistics about common metrics for measuring the size of the available corpus, including number of utterances, signs, length of the videos, size in MB, and so on. These numbers will make it possible to compare the key characteristics of the corpus to related work at a glance.

5.5. Integration of new technologies for display of, and access to, data, as they become available

Our future plans include the display of information on annotations in new ways. One of these ways consists of integrating data that can be measured by a computer as opposed to humans, such as graphs and numbers showing changes in eyebrow height and head movement for large samples of the corpus.

We also plan to facilitate the integration of the data with third-party and mobile applications. The biggest promise of having the DAI available on the web, as opposed to distributing it off-line, lies in making it available as an online service, such that cloud-based applications can take advantage of it. For example, a sign language dictionary available on mobile devices would be able to search for and retrieve concrete usage examples for the sign in question, which can be an invaluable tool for second language learners. Taking this approach will also enable other creative ways to use a corpus that we have not yet even envisioned.

6. Resources

- Database access interface: http://secrets.rutgers.edu/dai/queryPages/
- XML file format and DTD: http://www.bu.edu/asllrp/ncslgr-for-download/download-info.html.

7. Acknowledgments

The production of the NCSLGR corpus has involved significant contributions by many people. The Computer Science personnel responsible for the technical aspects of the video capture and management of video files included Stan Sclaroff and Vassilis Athitsos. The linguistic aspects of the data elicitation and annotation were carried out by Carol Neidle, Dawn MacLaughlin, Robert G. Lee, Benjamin Bahan, Michael Schlang, and many other ASLLRP participants. We are indebted to David Greenfield, the programmer for SignStream® through version 2, and to Iryna Zhuravlova who developing the Java reimplementation with additional features. Preliminary versions of the interface were developed by Ben Waber, Eugene Khafizov, and others. Thanks also to Richard Goodrow for his assistance with current development efforts. We are grateful for support from the National Science Foundation (NSF) that made this research possible (CNS-1059218 and CNS-1059221). The work reported here on development of the interface has also been supported in part by grants from NSF (CNS-0958442 and CNS-0964385).

7. References

Neidle, Carol. 2002a. *SignStream™ Annotation: Conventions used for the American Sign Language Linguistic Research Project.* American Sign Language Linguistic Research Project Report No. 11, Boston University.

—. 2002b. SignStream™: A Database Tool for Research on Visual-Gestural Language. *Journal of Sign Language and Linguistics 4*.203-14.

—. 2007. *SignStream™ Annotation: Addendum to Conventions used for the American Sign Language Linguistic Research Project.* American Sign Language Linguistic Research Project Report No. 13, Boston University.

Neidle, Carol, Ashwin Thangali & Stan Sclaroff. 2012. Challenges in Development of the American Sign Language Lexicon Video Dataset (ASLLVD) Corpus. Paper presented to the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC 2012, Istanbul, Turkey, 2012.