A New Web Interface to Facilitate Access to Corpora: Development of the ASLLRP Data Access Interface*

Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus & Lexicon

Carol Neidle
carol@bu.edu
Boston University

Christian Vogler
christian.vogler@gallaudet.edu
Gallaudet University

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 who 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.

Currently available data sets

National Center for Sign Language and Gesture Resources (NCSLGR) Corpus: ASL videos collected and linguistically annotated at Boston U.

- **Synchronized video files** (compressed, uncompressed formats) show signing from front (stereo views) and side, including face close-up.
- **Linguistic annotations of manual and non-manual components** of the signing, carried out using SignStream®, are available in XML format.
  - **Start and end points of each sign, identified by unique gloss label.**
  - **Parts of speech.**
  - **Start and end points of non-manual behaviors, also labeled with respect to the linguistic information they convey (e.g., negation, question marking, etc.).**
- **Language materials: Total of 1,887 linguistically annotated utterances**, including 1,920 distinct canonical signs (grouping close variants) and 11,861 sign tokens. Contained in 19 short narratives (1,002 utterances) plus 885 additional elicited utterances from Deaf native signers of ASL (most from 4 signers).

New download capability

User can browse/search the dataset, and select materials of interest, to be added to the "download cart" for subsequent download.

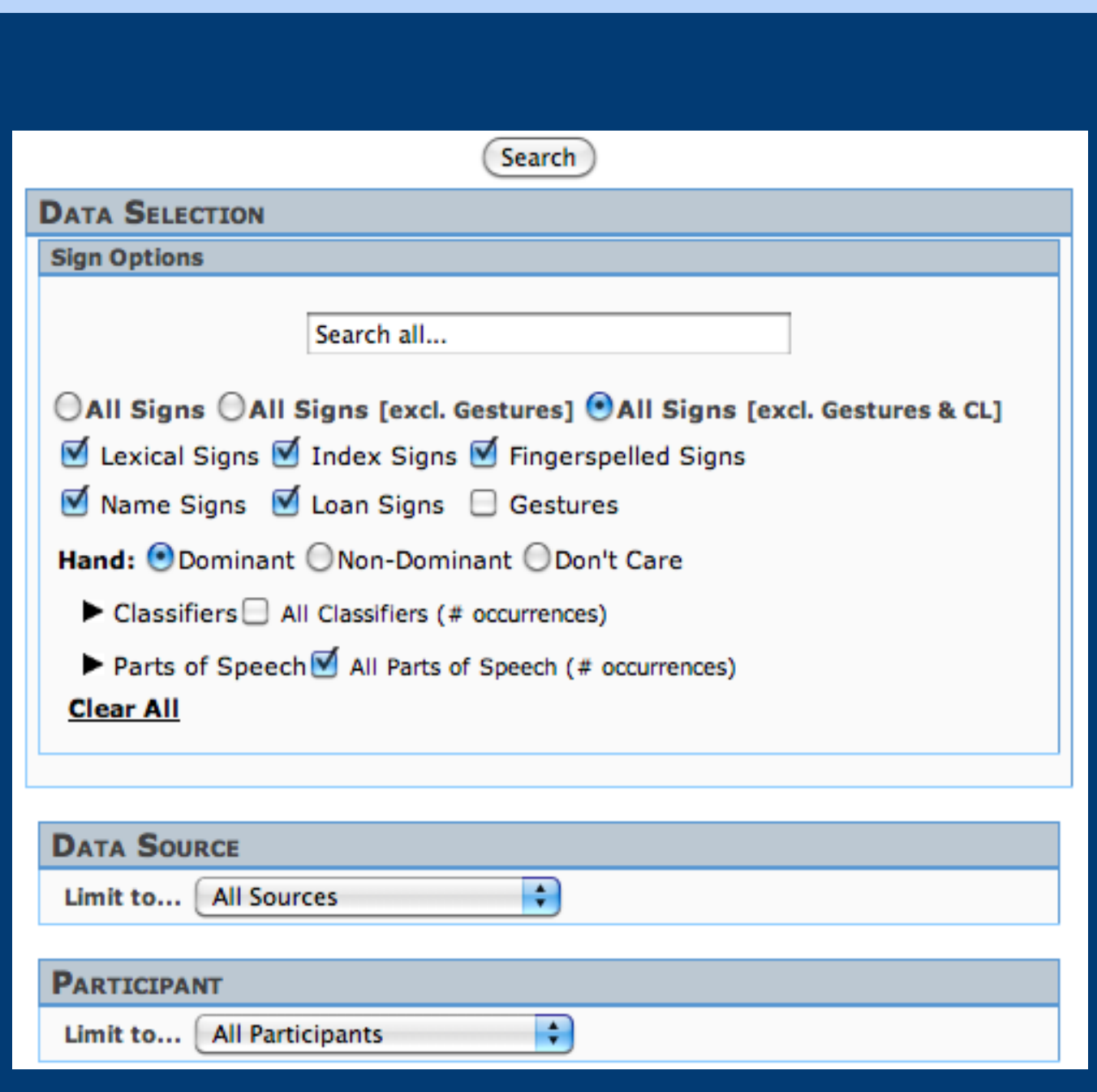
Best practices and lessons learned

- **Presentation of data**
  - **Presenting signs as still images (ideally of both start and end frames for a given sign) saves users and annotators time and effort.**
    - The user can detect at a glance whether item is of interest.
    - This facilitates consistency checks.
- **Resource management**
  - **Keep metadata separate from file names and assets.**
    It is much easier to keep metadata consistent and up-to-date if encoded in a centralized spreadsheet or database rather than encoded in filenames, as metadata can change.
  - **Designate only one asset as the authoritative source on metadata, and auto-generate other assets from there.**
  - **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 2-part schema: <url prefix> <path to file> where URL prefix points to a location on the server. 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** (e.g., restrictions on filenames).
- **Development processes**
  - **Plan for continuity.**
  - **Use version control on all source files *and* third-party libraries.**
- **Database design**
  - **Think queries, not data format.**
    The types of queries that need to be supported drive database design and tables. They inform every decision that pertains to the tables, the relationships among 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.
  - **Use a collection of tags.**
    Standardized tagging of annotations provides a powerful and efficient way to search for specific linguistic phenomena.

Resources

- **Database Access Interface (DAI):** http://secrets.rutgers.edu/dai/queryPages/
- **XML file format and DTD:**
  http://www.bu.edu/asllrp/ncslgr-for-download/download-info.html
- For **documentation of annotation conventions**:
  see reports 11 and 13 from http://www.bu.edu/asllrp/reports.html.

Functionalities of the Interface

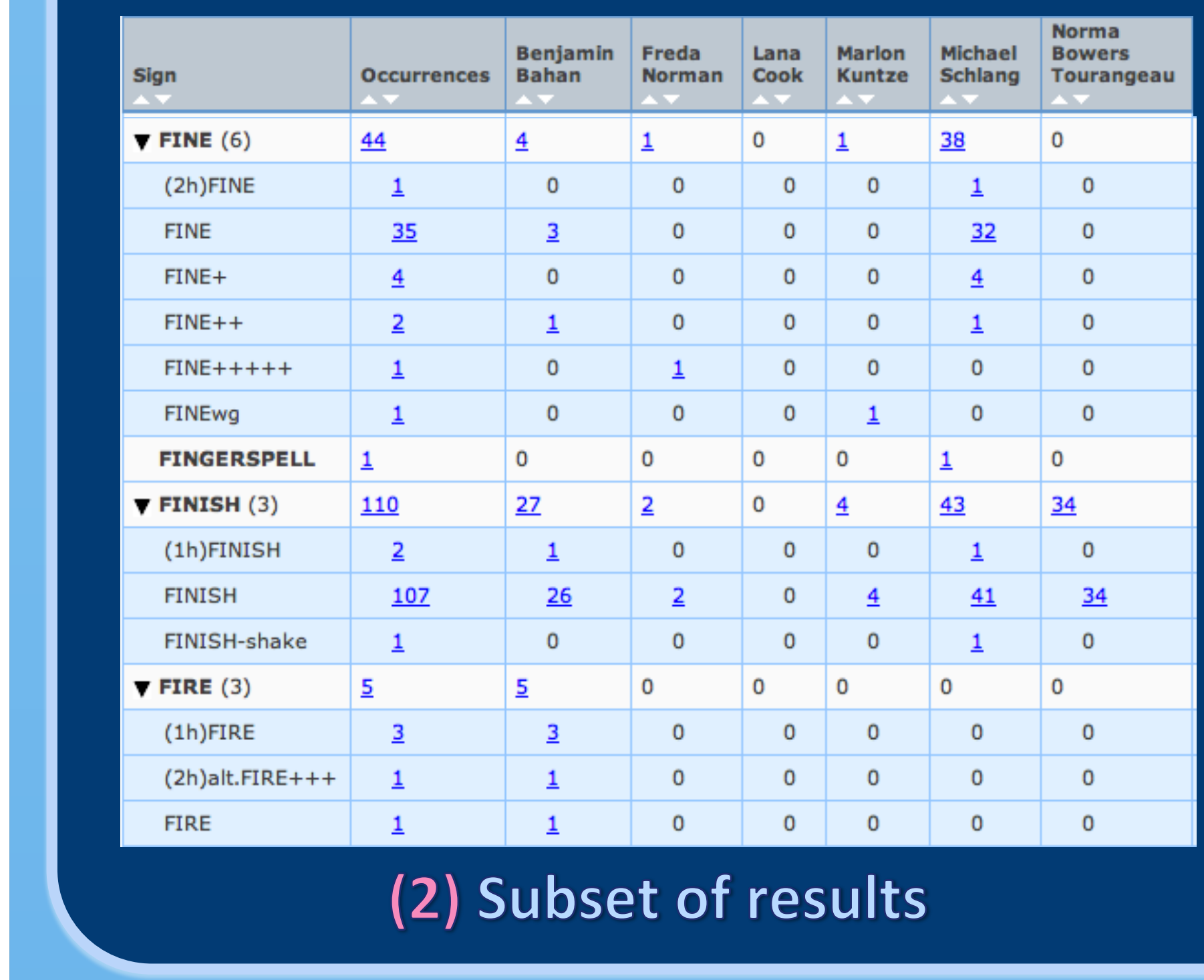


DATA SELECTION
Sign Options
Search all...
○ All Signs ○ All Signs [excl. Gestures] ● All Signs [excl. Gestures & CL]
☑ Lexical Signs ☑ Index Signs ☑ Fingerspelled Signs
☑ Name Signs ☑ Loan Signs □ Gestures
Hand: ● Dominant ○ Non-Dominant ○ Don't Care
▶ Classifiers □ All Classifiers (# occurrences)
▶ Parts of Speech ☑ All Parts of Speech (# occurrences)
Clear All

DATA SOURCE
Limit to... All Sources

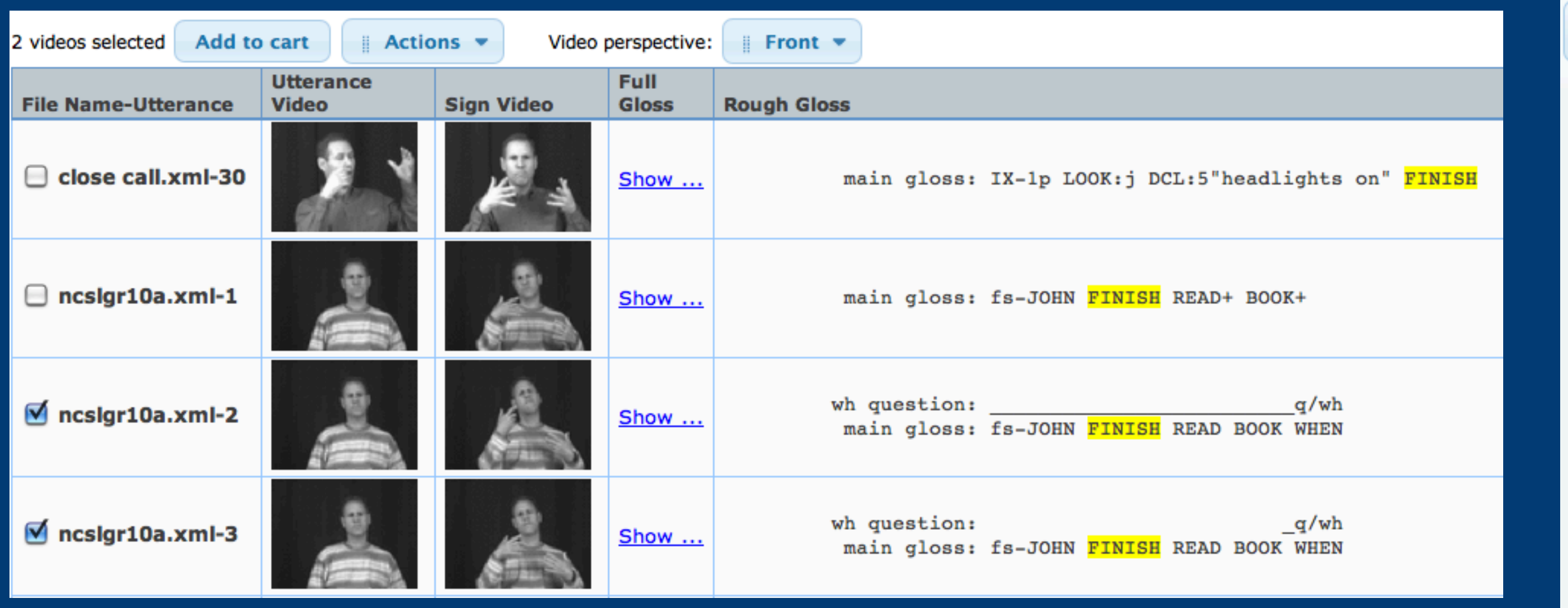
PARTICIPANT
Limit to... All Participants

(1) 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

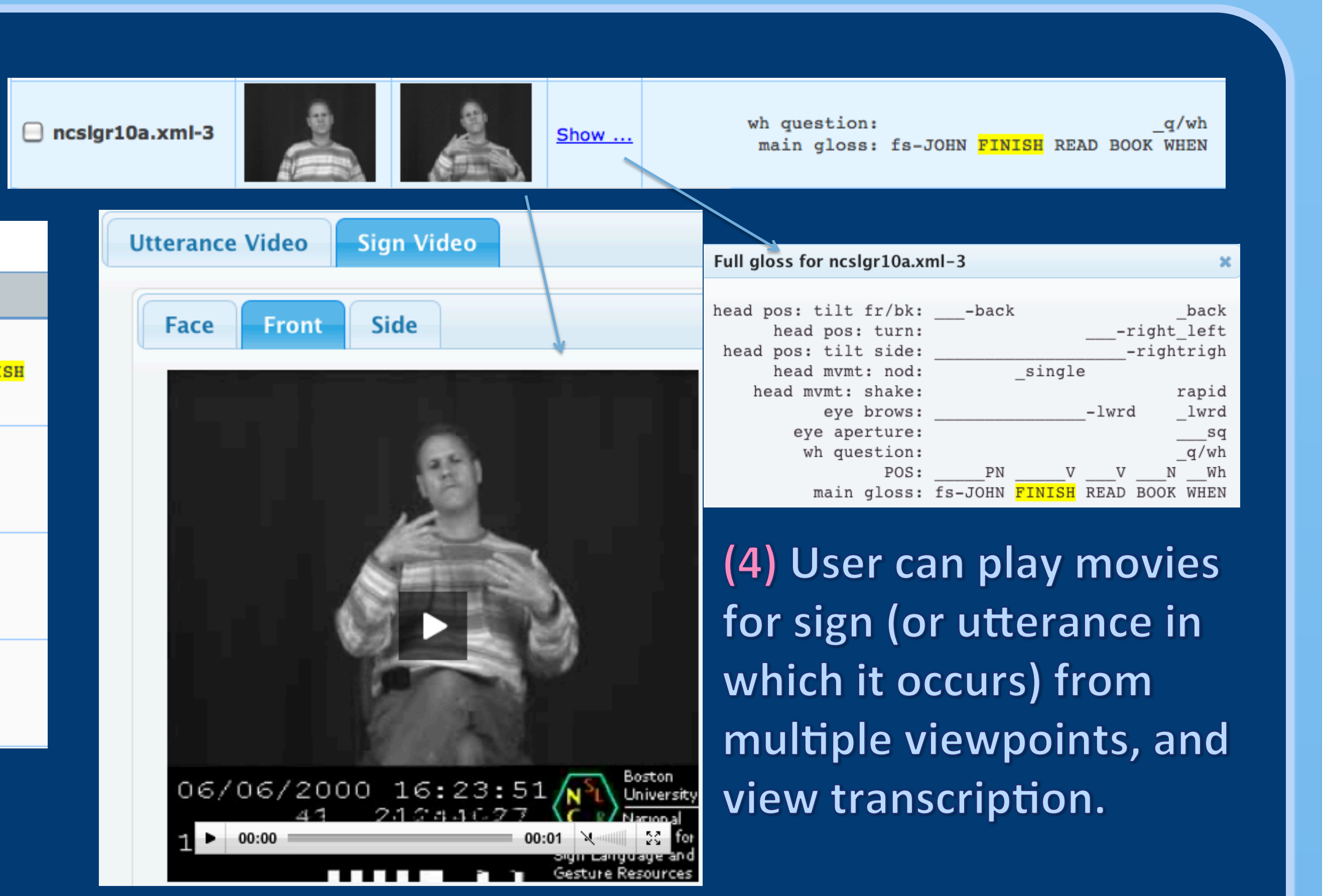
(2) Subset of results



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

File Name-Utterance	Utterance Video	Sign Video	Full Gloss	Rough Gloss
□ close call.xml-30			Show ...	main gloss: IX-1p LOOK:j DCL:5"headlights on" FINISH
□ ncslgr10a.xml-1			Show ...	main gloss: fs-JOHN FINISH READ+ BOOK+
☑ ncslgr10a.xml-2			Show ...	wh question: ____________________q/wh main gloss: fs-JOHN FINISH READ BOOK WHEN
☑ ncslgr10a.xml-3			Show ...	wh question: _q/wh main gloss: fs-JOHN FINISH READ BOOK WHEN

(3) Sentences with FINISH: selections for download

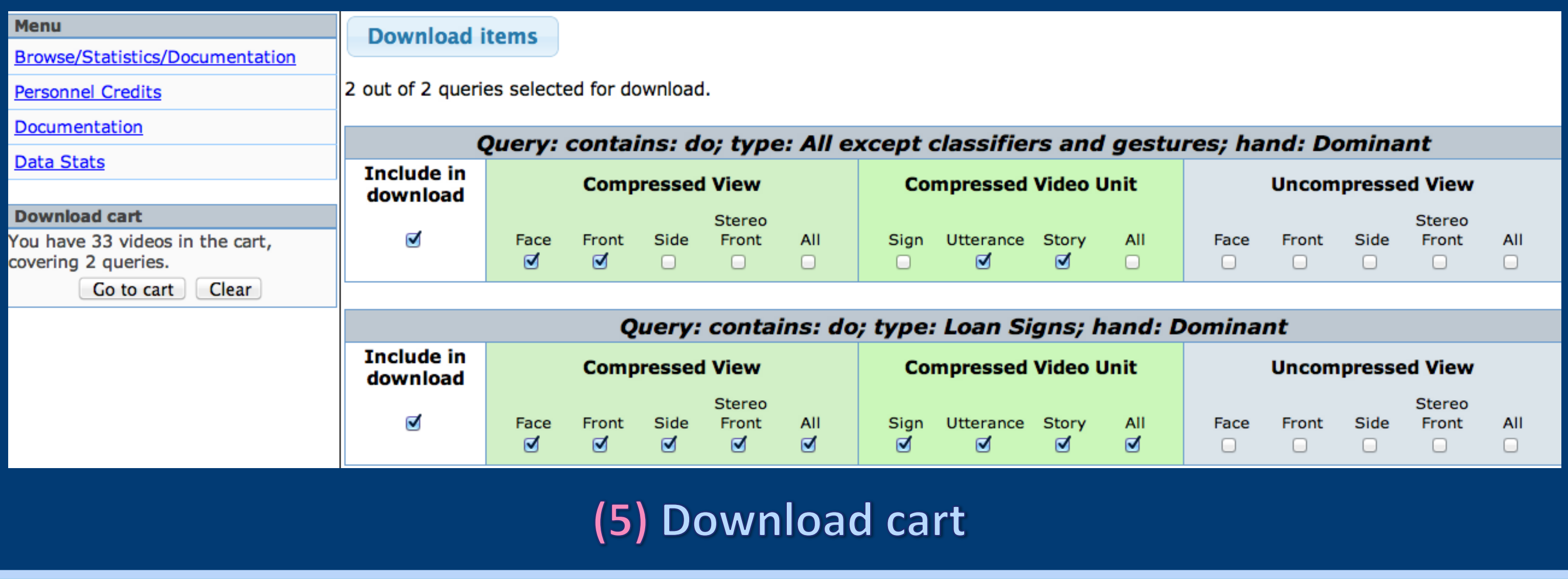


□ ncslgr10a.xml-3 | Show ... | wh question: _q/wh / main gloss: fs-JOHN FINISH READ BOOK WHEN

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
```

(4) User can play movies for sign (or utterance in which it occurs) from multiple viewpoints, and view transcription.



Menu
Browse/Statistics/Documentation
Personnel Credits
Documentation
Data Stats

Download cart
You have 33 videos in the cart, covering 2 queries.
Go to cart | Clear

Download items

2 out of 2 queries selected for download.

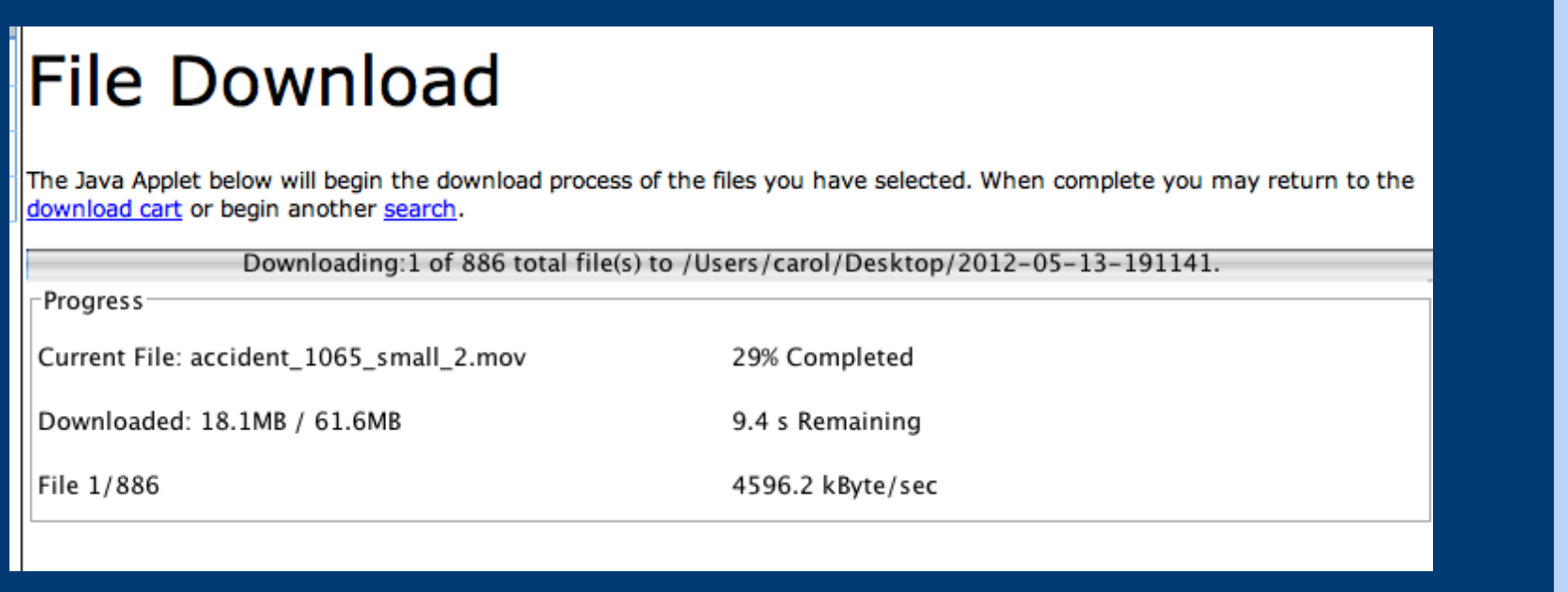
Query: contains: do; type: All except classifiers and gestures; hand: Dominant

Include in download	Compressed View					Compressed Video Unit				Uncompressed View				
	Face	Front	Side	Stereo Front	All	Sign	Utterance	Story	All	Face	Front	Side	Stereo Front	All
☑	☑	☑	□	□	□	□	☑	☑	□	□	□	□	□	□

Query: contains: do; type: Loan Signs; hand: Dominant

Include in download	Compressed View					Compressed Video Unit				Uncompressed View				
	Face	Front	Side	Stereo Front	All	Sign	Utterance	Story	All	Face	Front	Side	Stereo Front	All
☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	□	□	□	□	□

(5) Download cart



File Download

The Java Applet below will begin the download process of the files you have selected. When complete you may return to the download cart or begin another search.

Downloading:1 of 886 total file(s) to /Users/carol/Desktop/2012-05-13-191141.

Progress
Current File: accident_1065_small_2.mov — 29% Completed
Downloaded: 18.1MB / 61.6MB — 9.4 s Remaining
File 1/886 — 4596.2 kByte/sec

(6) Download in progress

Plans for future development

- Integration of other types of corpora (e.g., ASLLVD corpus — cf. Neidle, Thangali & Sclaroff, "Challenges in Development of the American Sign Language Lexicon Video Dataset (ASLLVD) Corpus," LREC 2012).
- Functionalities to enable additional types of searches – based on linguistic information, statistical frequencies, and video properties.
- Providing annotations in additional formats, e.g., ELAN EAF format.
- Display of various kinds of statistical information.
- Integration of new technologies as they become available (e.g., to provide accessibility for mobile devices).

*Acknowledgments

The production of the NCSLGR corpus has involved significant contributions by Stan Sclaroff and Vassilis Athitsos (for technical aspects of video data collection and storage); David Greenfield and Iryna Zhuravlova (for development of annotation tools and assistance with annotation file formats); and Dawn MacLaughlin, Robert G. Lee, Benjamin Bahan, Michael Schlang, among many others, for linguistic aspects of the data collection and annotation. Preliminary versions of the Web interface were developed by Ben Waber, Eugene Khafizov; thanks also to Richard Goodrow for more recent contributions. We are grateful for support from the National Science Foundation (NSF) that has made this research possible (#9528985, 9809340 and 9912573). The ongoing development effort reported here is a collaboration among BU (C. Neidle, S. Sclaroff), Gallaudet U. (B. Bahan, C. Vogler), Rutgers U. (D. Metaxas), and the U. of Texas Arlington (V. Athitsos), supported by NSF grants #0958442 and 1059218.