

The Sign And Gesture Archive (SAGA) 1.2

Susan Goldin-Meadow (Psychology), Diane Brentari (Linguistics), Jeffrey Tharsen (RCC)

Introduction

The Sign and Gesture Archive is a video data library for gesture, sign and spoken language data. Its holdings represent over 40 years of data collection, including tens of thousands of video files, thousands of coding and annotation files, and rich metadata for each source.

SAGA currently includes data from 7 sign languages: American Sign Language (ASL), British Sign Language (BSL), Mandarin Chinese Sign Language (CSL), French Sign Language (LSF), Hong Kong Sign Language (HKSL), Italian Sign Language (LIS), Nicaraguan Sign Language (NSL); plus homesign from the United States, China, Taiwan, Turkey and Nicaragua, and large datasets from the Language Development Project (LDP) of hearing participants performing co-speech and silent gesture.

The goal of SAGA is to make these data as widely available as possible to members of the University of Chicago community, and to the field at large. Our long-term goal is, where possible, to bring data from major international laboratories under the umbrella of this archive.



PI Diane Brentari collecting video data on site in Nicaragua of NSL Sign Language, Homesign and Silent Gesture.

Screenshots of the SAGA Basic Search Interface (above), and the built-in pop-up Video Player and Annotation Viewer (below).

Machine Learning and SAGA 1.2

Having successfully created the frameworks and configured the datasets for access via the website, the RCC is now developing a series of advanced methods and tools to give researchers the ability to more fully exploit the richness of this unique resource. The RCC is partnering with the Templeton Foundation in order to see what large-scale machine learning strategies can uncover about patterns and trends within the datasets that have not yet been detected by human researchers, while developing suites of online tools to empower research groups to take advantage of the vast amounts of data linked in SAGA, and to make it easy for researchers to incorporate new datasets into the Archive to drive large-scale comparative research projects.

The Basic and Advanced Search User Interfaces

The main focus of the SAGA website is its Basic Search user interface which allows researchers unprecedented access to the videos, coding files and metadata in the Archive. By selecting from general, specific and/or publication criteria, users can quickly find the data they are most interested in. Users can then explore these results via a pop-up window (as displayed in the center screenshot) which contains a full-featured video player, download options for the video and coding files, and can even display the data from any coding file right in the interactive interface below the video player.

As the data in the Archive is strictly controlled by IRB and privacy constraints, SAGA also contains a robust mechanism for controlling access to specific videos and coding files. If a user does not have the necessary permissions to view a file, a pop-up window allows the user to contact the PI or administrator who controls access via email and request that s/he be given permission to view and/or download the video(s) and coding file(s).

The Advanced Search User Interface

Via the Advanced Search User Interface, SAGA's users can not only use any of the criteria from the Basic Search, but can also incorporate any of the many additional data fields unique to each dataset, including transcripts of verbal utterances, the full range of coding and annotation data hand-entered over the past four decades, and any of the various descriptive markers for specific signs and gestures. The Advanced Search UI also allows for wildcards and other open-ended search methods. By pairing a robust metadata framework with a search mechanism able to handle a virtually unlimited range of source data, the RCC has developed an effective method for searching large, heterogeneous data archives, able to efficiently return a list of relevant datasets for users to explore in greater detail.