The Sign Language Analyses (SLAY) Database*

Rachael Tatman[a]

Department of Linguistics, University of Washington,
Box 352425, Seattle, WA 98195-2425, USA
rctatman@uw.edu

Summary. This paper describes the construction and input analysis of the Sign Language Analyses (SLAY) Database, both as it currently stands and as a guide to further expansion of the project. SLAY contains condensed cross-linguistic grammatical information from signed languages. It was designed so that the framework of the database can be expanded indefinitely to include investigations of new questions. It differs from similar projects (such as the World Atlas of Linguistic Structure (Haspelmath et al., 2005)) in that it focuses exclusively on signed languages and modality-specific grammatical questions.

Keywords: sign language, phonology, databases, computational methods

1 Aims of SLAY

The creation of SLAY was motivated by the fact that, while at least one cross-linguistic grammatical databases includes signed languages (Haspelmath et al., 2005; Zeshan, 2005) and there are multiple corpora of signed languages (Crasborn and Zwitserlood, 2008; Hanke et al., 2010), there has not previously been a grammatical database that focused exclusively on signed languages. As a result, it is difficult to answer questions about the cross-linguistic distribution of modality-specific grammatical features, such as parameters. This is due to the fact that databases constructed without reference to signed languages do not have the structure in place to support information on these grammatical structures. WALS, for example, does not have information on sign language parameters. SLAY fills that gap.

* This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1256082. I would like to extend my thanks to Richard Wright, Ellen Kaisse, Lorna Rozelle, Conner Kasten and the audience at the Workshop on Corpora and Databases in Linguistics for their guidance and recommendations. Any remaining oversights or errors are my own.

University of Washington Working Papers in Linguistics (UWWPL). Volume 33 (2015).

There are three main goals for the SLAY database.

1. To provide an extensible framework for looking at sign-language-specific grammatical questions cross-linguistically.
2. To provide guidelines for adding information in a standard and replicable way.
3. To make the database freely available for other researchers to use, modify and share.

The first aim was fulfilled by careful design of the database architecture, which is covered in more detail in Section 2. With the current database structure, more languages, sources and tables containing grammatical information can be added indefinitely.

The second is fulfilled by this paper. A description of the methods used to create the grammatical table already included in the database as well as a set of guidelines for adding additional tables can be found in Section 2. Information on the input analysis and guidelines for future work can be found in Section 3.

The third goal is currently the most difficult, as detailed in Section 4. Data from the database is currently publicly available via Sqlshare (Howe et al., 2012), however it lacks the full structure and is therefore difficult for other researchers to append. In addition, any additions made to a copy downloaded from Sqlshare would then need to sent back to the original reseracher and added to the database by hand, which is suboptimal.

2 Architecture

SLAY is a relational database with numeric primary keys and two main "parent" tables[1]. It was constructed using MySQL community edition (MySQL, 1995) which is an open-source MySQL edition available for Windows, Mac and Linux computing environments. MySQL was chosen for encoding because it is platform-independent, well-supported, open-source and can be obtained for free. It is also one of the mostly widely-used relational database management system so there are ample resources available for learning it and troubleshooting. All of these design choices were made with long-term growth in mind.

A relational database is made up of a number of different tables, in which data is stored. Every row of every table has a unique primary key which is used to identify that row. Further, rows may reference rows in other tables by their keys. Keys referencing other tables in this way are know as foreign keys. (Codd, 1970).

[1] Note that SLAY is not a hierarchical relational database. The "children" tables can have multiple "parents" and it is not necessary for a query to traverse from one of the "parent" tables to a "child" table. The familial terminology used here is only intended to clarify the structure of the database.

This is a particularly useful feature for this project because it allows the database to be be built modularly; it is simple to add a new table to the database looking at another facet of linguistic structure and to use foreign keys to link it with other tables already in the database.

The overall structure of the database is represented in Figure 1. Each table containing grammatical information (currently only one table containing information on parameters) can be thought of as a "child" table because it makes reference to information contained in the Languages (see Subsection 2.1) and Sources (see Subsection 2.2) tables. This reduces redundancy and ensures that all information on a specific language or source is simple to find.

Throughout the database, primary keys are numeric, even when other logical options (such as language's ISO 639 keys) were available. By using a numeric key additional rows can be added to any table in a principled way as needed without, for example, going through the process of applying for an ISO key for a newly-recorded sign language. This does have the disadvantage of making foreign keys (the primary key of a row in a different table than the one that is currently being referenced) somewhat opaque. Other researchers can re-key the tables if they find this objectionable.

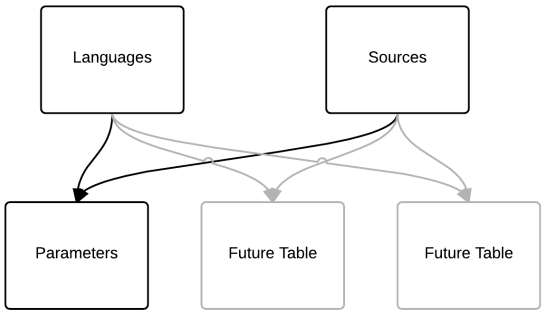

Figure 1: Current (black) and possible future (gray) structure of SLAY.

2.1 Languages Table

The languages table includes all 136 languages currently listed by Ethnologue as "deaf sign languages" (Gordon, 2005). In addition to the language name, each row in the language table includes an automatically-generated numeric primary key, the ISO 639-3 code for the language and the main country where it is used. Column names and data types can be found in Table 1. The table was populated automatically from Ethnologue's website using a Python script. It was then edited as there

appeared to be some errors. Some languages were missing (e.g. Ghardaiaest Sign Language, Caucasian Sign Language), others were dialects listed separately (e.g. Malagasy Sign Language and Norwegian Sign Language) and at least one language may be an idiolect (Rennellese Sign Language).

Column	Data Type
IdLanguages	INT (Primary key)
LanguageName	VARCHAR(45)
EthnologueId	VARCHAR(3)
Country	VARCHAR(45)

Table 1: Languages table column names and data types.

2.2 Sources Table

The reference database includes an automatically-generated numeric primary key and the title of the reference as well as optional columns for the author, year of publication, where the publication appeared, a uniform resource locator (URL) [2] and a Bibtex entry. Column names and data types can be found in Table 2. Sources were added by hand.

Column	Data Type
IdSources	INT (Primary key)
Title	MEDIUMTEXT
Author	MEDIUMTEXT
YearPublished	YEAR
AppearedIn	MEDIUMTEXT
URL	TEXT
Bibtex	LONGTEXT

Table 2: Sources table column names and data types.

2.3 Parameters Table

The parameters table, already in the database, serves as an example of the sort of grammatical information that can be added to the database. Parameters are the sub-lexical units of signed languages. They include handshape, movement, location (Stokoe, 2005), non-manuals such as facial expression (Liddell, 1978), palm orientation (Friedman, 1975) and number of hands (Bellugi and Fischer, 1972). The parameters table contains an automatically-generated numeric primary key, as well as two foreign keys: one referring to a language from the language table and

[2] My thanks to a reviewer for pointing out that the previous 150 character limit for the URL column would cut off some URLs.

another referring to a reference from the reference table. It also contains eight additional columns. The first six columns are optional and contain Boolean values[3]. These columns encoded handshape, movement, location, non-manuals, palm orientation and numbers of hands. The seventh and eights columns provided optional space for discussion of additional proposed parameters and more general notes. Each row of the table represents a unique combination of language and reference. For each of the six parameters, a Boolean with a value of 1 (or TRUE) was entered if the analysis in the reference specifically included that parameter, a 0 (or FALSE) if the analysis specially argued against that parameter and the field was left blank if it was not discussed. This allowed for a distinction between an analysis against a parameter and an analysis which simply did not include one. Data entry for this table was also by hand.

Column	Data Type
IdParameters	INT (Primary key)
LanguageParam	INT (Foreign key)
ReferenceParam	INT (Foreign key)
Handshape	BOOLEAN
Movement	BOOLEAN
Location	BOOLEAN
NonManualMarker	BOOLEAN
PalmOrientation	BOOLEAN
NumberOfHands	BOOLEAN
OtherParameters	LONGTEXT
Notes	LONGTEXT

Table 3: Parameters table column names and data types.

2.4 Adding New Tables

In keeping with the current structure of the database, new tables should have the following properties. This will allow for consistency across the tables and easier navigation and analysis as the database grows.

- Primary keys should be numeric.
- Column names for primary keys should be of the format IdTableName.
- Foreign keys for both the Languages and Sources table should be included for each observation. If multiple sources are used for a single language or *vice versa* then that should be represented by multiple rows.

[3] A Boolean is a data type which has only two possible values, commonly referred to as TRUE and FALSE, although since it was optional in this case there is also a third possibility for those cells: NULL.

- The data type should be specified for each column in the new table. Data types should as memory-efficient as possible.
- The first letter of each word in column names should be capitalized and no spaces should be used.
  - correct: NewColumnName
  - incorrect: newcolumnname, New Column Name
- The first letter of each word in the column name should be capitalized.
- Column and table names should be informative. Avoid abbreviations unless they are very common (i.e. "URL").

3 Input Analysis

Since SLAY is a meta-analytic database some analysis of the sources is required. The advantage of a database of this type is that allows quick comparison across languages for variables of interest. Without normalizing across different sources that advantage is lost and the database would more closely resemble an annotated bibliography. However, too much input analysis runs the risk of introducing biases and making the database less useful. This section discusses the input analysis for the data already in the database as well as guidelines for future work.

3.1 Input analysis for the parameters table

The first step of input analysis for the parameters table was design of the table itself. As mentioned in subsection 2.3, Boolean encoding was chosen because it could record a difference between analyses where a parameter was not discussed and those where its presence was explicitly argued against (as in Kendon (1988)). This allowed for a faithful representation of each linguist's analysis.

The second challenge was in translating–both from other languages into English and between different scholarly traditions. This is the main place where it is possible that bias was introduced to the project. For example, Engberg-Pedersen (1993) describes Danish sign language as having "place of articulation" rather than location. Based on the description of the language, this was judged to be the same as location and entered into the database as such. Another example comes from LeMaster (1997), who describes "point of articulation" and "hand configuration" in both the male and female versions of Irish Sign Language. This was entered as location and handshape. A third is Sparhawk (1978), whose description of "the moving part", "the shape of the moving parts" and "the location of any approach of contact" were analyzed as the same as movement, handshape and location. In aggregate, these judgments may influence the contents of the database. Other researchers are thus encouraged to download SLAY and make changes as they see fit;

where there was the possibility for multiple interpretations of a source's discussion of a language is has been included in the "Notes" column of the parameters table.

The third place where bias might be introduced is in the choice of sources. For this project, every attempt was made to find an academic linguistic analysis for each language surveyed. It was not always possible however; many signed language are woefully under-documented and in some cases the only resources available were dictionaries or work done in related fields. As better and more documentation becomes available, however, it will be added to the database. Due to the flexible design of the database, this process can continue indefinitely. As more sources are added they will provide more complete information on the grammatical structure of signed languages.

3.2 Input analysis for new tables

With these possible pitfalls in mind, the following guidelines should be followed when adding new data to SLAY.

- No information not included in the analyses should be added. Use a NULL value if the source does not touch on a certain grammatical feature.
- If new sources are added to the database, the following is order of preference, from most preferred to least preferred. This order was constructed under the assumption that academic linguistic analyses will have the most complete and accurate information on the grammatical structure of a language, and that the peer review process of academic work in general will catch most incorrect analyses.
  - Academic linguistic analyses
  - Academic work from related fields
  - Other sources
- If there is room for disagreement about analysis of a source, it should be noted.

4 Distribution

The database is currently publicly available via Sqlshare, a service provided by the eScience Institute at the University of Washington (Howe et al., 2012). For ease of reading, the parameter table has had its numeric foreign keys for language and reference replaced with the ISO code and Bibtex entry for the relevant language and reference, respectively. A free account is required in order to access Sqlshare. One disadvantage of the current method of distribution is that it does not maintain all the structure of the database and instead presents them as flat tables. They can be manipulated and queried directly from the web platform, however, which is a distinct advantage. The full database may be obtained by contacting the author.

References

Bellugi, Ursula and Fischer, Susan. 1972. A comparison of sign language and spoken language. *Cognition* 1(2), 173–200.

Codd, Edgar F. 1970. A relational model of data for large shared data banks. *Communications of the ACM* 13(6), 377–387.

Crasborn, Onno A and Zwitserlood, Inge. 2008. The Corpus NGT: an online corpus for professionals and laymen. In *3rd Workshop on the Representation and Processing of Sign Languages (LREC)*, pages 44–49, ELDA.

Engberg-Pedersen, Elisabeth. 1993. *Space in Danish Sign Language: The semantics and morphosyntax of the use of space in a visual language*. SIGNUM-Press.

Friedman, Lynn A. 1975. Space, time, and person reference in American Sign Language. *Language* pages 940–961.

Gordon, Raymond. 2005. Ethnologue: Languages of the world 15 th Edition. *Dallas, TX: Sil International* .

Hanke, Thomas, König, Lutz, Wagner, Sven and Matthes, Silke. 2010. DGS Corpus & Dicta-Sign: The Hamburg Studio Setup. In *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta*, pages 106–110.

Haspelmath, Martin, Bibiko, Hans-Jörg, Hagen, Jung and Schmidt, Claudia. 2005. *The world atlas of language structures*, volume 1. Oxford University Press Oxford.

Howe, B, Cole, G, Key, A, Khoussainova, N and Battle, L. 2012. Sqlshare: Database-as-a-service for long tail science. *The Cloud Computing Engagement Research Program* pages 52–56.

Kendon, Adam. 1988. *Sign languages of Aboriginal Australia: Cultural, semiotic and communicative perspectives*. Cambridge University Press.

LeMaster, Barbara. 1997. Sex differences in Irish sign language. *The Life of Language: Papers in Linguistics in Honor of William Bright* pages 67–85.

Liddell, Scott K. 1978. Nonmanual signals and relative clauses in American Sign Language. *Understanding language through sign language research* pages 59–90.

MySQL, AB. 1995. *MySQL: the world's most popular open source database*. MySQL AB.

Sparhawk, Carol M. 1978. Contrastive-identificational features of Persian gesture. *Semiotica* 24(1-2), 49–86.

Stokoe, William C. 2005. Sign language structure: An outline of the visual communication systems of the American deaf. *Journal of deaf studies and deaf education* 10(1), 3–37.

Zeshan, Ulrike. 2005. Sign languages. In *The world atlas of language structures*, pages 558–559, Oxford University Press.