BSL Corpus Annotation Guidelines

Kearsy Cormier and Jordan Fenlon

Deafness Cognition and Language (DCAL) Research Centre, University College London
49 Gordon Square, London WC1H 0PD

Contact: bslcorpusproject@ucl.ac.uk

Updated 23 October 2014

1 Acknowledgments

- **UK Economic and Social Research Council (ESRC):** The research council which funded the BSL Corpus Project (http://bslcorpusproject.org, 2008-2011, RES-062-23-0825) and subsequent corpus-based projects (e.g. BSL Directional Verbs Project, 2012-2014, ES/K003364/1).
- **BSL Corpus Project**: The many team members of the original BSL Corpus Project and follow-on projects who have contributed to the building of the BSL Corpus so far (http://bslcorpusproject.org/team), particularly Adam Schembri.
- **Auslan Corpus Project**: The team members of the Auslan Corpus Project, which was the inspiration for the BSL Corpus and the basis for these current set of guidelines, particularly Trevor Johnston.
- **DCAL:** Other staff at the UCL Deafness, Cognition and Language (DCAL) Research Centre who have contributed their academic knowledge and dissemination of information.
- **BSL Corpus participants** from the British Deaf community for being involved in the BSL Corpus Project.
- **ELAN: The Language Archive of the Max Planck Institute** in continuing to develop and improve the functionality in ELAN which has supported the flexibility and ease in annotating data. We would like to thank in particular, Han Slöetjes for supporting these developments as well as always being a helpful hand for any pressing issues or queries with ELAN.

2 Introduction

This document aims to describe the annotation conventions as they have been used for the first release of the BSL Corpus annotations, made available online in September 2014. The BSL Corpus is a collection of around 125 hours of signing by deaf native and near-native BSL signers from 8 regions around the UK (Schembri, Fenlon, Rentelis, & Cormier, 2011; Schembri, Fenlon, Rentelis, Reynolds, & Cormier, 2013). It was published as a partly open-source, partly restricted-access video collection in 2011, and is hosted by UCL CAVA (Human Communication Audio-Visual Archive for UCL). The narrative and lexical elicitation data are open access, while the conversation and interview data are restricted to registered researchers only. Further information about the movies, the annotations and the restrictions can be found on the BSL Corpus web site, http://www.bslcorpusproject.org/cava/. Both CAVA and a version of this Corpus for a general audience can be found from our Data page: http://www.bslcorpusproject.org/data/.

To date, there are around 100 files that have been annotated at the lexical (ID gloss) level and that are available on CAVA: 25 each from Birmingham, Bristol, London and Manchester from the conversation data. A substantial part of this annotation work has been carried out for a lexical frequency study (Fenlon, Schembri, Rentelis, Vinson, & Cormier, 2014) with the remainder done as part of a study on directional verbs (Cormier, Fenlon, & Schembri, 2014; Fenlon, Schembri, & Cormier, 2014). The present document reflects the annotation conventions that were in use when annotation for the lexical frequency began in 2010 until the end of the directional verbs study in 2014. These guidelines are based largely on annotation guidelines for the Auslan Corpus as they existed in 2010 (for current version of the Auslan Corpus annotation guidelines, see Johnston 2014). ID gloss annotations link to BSL SignBank which initially began as a lexical database and has been developed into an online dictionary (http://bslsignbank.ucl.ac.uk). For advanced access to BSL SignBank, which provides access to ID glosses, register on SignBank as a researcher (staff or student); follow links for providing Feedback. For more on ID glosses, see section 4.3.

The preferred minimum number of annotation tiers for any sign language corpus is three: two ID gloss tiers and also a translation tier (Johnston, 2014). So far rough translations have been done for some of the BSL Corpus data; these will be made available at a later date depending on funds.

3 Using ELAN

Our work to date has used various versions of ELAN from 3.0 and 4.7.2. ELAN allows the inserting of annotations with a precise time-alignment referring to the video(s) in display, using multiple tiers to display different kinds of information about a particular sign in a particular time-length/annotation length. This makes it possible to easily match annotations with the raw video data and allows for flexibility of having different kinds of information annotated on multiple tiers in one file. We use the latest version (and do so consistently on all the computers accessing the corpus) in order to exploit and take advantage of the latest functionality tools in ELAN as well as avoid errors, file conflicts, inconsistent practices and old versioning layouts in comparison to new

versions if using several different versions. We found that by taking this approach, it enabled a smoother workflow and files within the team and also it improved some function ability in ELAN, in particular the multiple file processing. Some functions in ELAN are particularly useful for corpus data where several eaf files are produced and can be manipulated; such as multiple search, find and replace, multiple file processing, importing and exporting.

Currently with a primarily two-person team, the way we operate is that one person works on one file, whereas another person works on another and verbally communicates any updates. We are currently exploring a "versioning" system where users can work on the same file simultaneously and thus those files will be consistently updated, or create a system where a file can be 'locked' if already in use by another person. This practice will especially be very useful in a large team and being able to carefully control file usage and updates.

For more information about ELAN and its features, the manual can be downloaded here: https://tla.mpi.nl/tools/tla-tools/elan/

3.1 Opening an eaf file: Single and double video view

Figure 1. eaf file for Participant X.

Each eaf file consists of information about one participant (Participant X); using LH-IDgloss and RH-IDgloss as the two main tiers for annotating depending on which hand(s) are being used by Participant X. The single view showing only Participant X is presented first as a master media which is the first video media file to be linked to the eaf file (this is what ELAN will ask you to find first when you open the eaf, followed by the double view). Next to the single view of Participant X is the double view which includes Participant X's signing partner, Participant Y. Annotating data always follows the single video view (Participant X). The start of annotating starts with the first sign that is produced. The double view is used both as a backup source of information in case the single view production of a particular sign by Participant X is not clear, and also to allow reference to Participant Y for content of information to support annotating the data.

3.2 File naming conventions

It is important for corpus files to be named appropriately in order to enable a quick scan of information and for files to be easily located.

3.2.1 ELAN eaf filenames

ELAN eaf files are named with the following demographics in this order: Region, Participant number, Gender, Age, Ethnicity, Deaf/Hearing family, Task.

Possible values
Region: LN (London), BM (Birmingham), BL (Bristol), BF (Belfast), N (Newcastle), C (Cardiff), GW (Glasgow)
Participant number within region: A number between 01 and 36 (for most regions this is a number between 01 and 30 but there are higher numbers in a few regions)
Gender: M or F

Age: Age in years
Ethnicity: A (Asian), B (Black), W (White)
Deaf/Hearing family: D or H
Task: L (lexical elicitation), I (interview), C (conversation), N (narratives), or NC (combined narrative/conversation)

Example
BM07F35WHC

Region	Participant number	Gender	Age	Ethnicity	Deaf/Hearing Family	Task
BM	07	F	35	W	H	C

3.2.2 Video filenames

Video files are named with a slightly different convention from the eaf files as video files consist of either a view of a single participant or a view of two participants. Video filenames consist of a code for Region followed by Participant number(s) followed by Task.

Possible values
Region: L (London), BM (Birmingham), BL (Bristol), BF (Belfast), N (Newcastle), CF (Cardiff), G (Glasgow)
Participant number within region: A number between 1 and 36 (for most regions this is a number between 1 and 30 but there are higher numbers in a few regions)
Task: l (lexical elicitation), i (interview), c (conversation), n (narratives), n-c or n+c (combined narrative/conversation)

Examples
BM7c: Birmingham participant 7 (single view) in combined narrative/conversation task
G13+14l: Glasgow participants 13 and 14 (double view) in lexical elicitation task

3.3 Basic template in ELAN

The basic template for the first release of BSL Corpus annotations has two tiers: RH-ID gloss and LH-ID gloss. As of October 2014, our work-in-progress template uses 49 tiers that are arranged in a hierarchy and controlled vocabularies have been assigned to some tiers. See Appendix for a list of work-in-progress tiers.

4 RH-ID gloss tiers and LH-ID gloss tiers

These tiers represents all manual material articulated on either the right or left hand. For first release, we only annotate lexical signs on the dominant hand. For example, the two-handed DRIVE is only annotated on the tier reflecting the signer's dominant hand (whether it is the left or right hand). *Note: This will be changed in future releases.* If each hand was producing separate meaningful units (e.g. a lexical sign on one hand and a pointing sign on the other) then these were annotated on both the RH and LH tier as appropriate.

4.1 Parsing signs: deciding on the start and end points

The start point for a sign is when the hand or hands appear to start moving away from articulating the previous sign. This is signalled by a change in direction, orientation, and/or handshape. The end point for a sign is when the hand appears to start moving towards articulating the following sign. Again, this is signalled by a change in direction, orientation, and/or handshape.

A sign sequence is considered to be finished normally when the hands begin a return to a rest position (e.g., folded arms, hands on hips, laps, or some supporting surface or object, or arms resting at the side of the body). Signers may, however, maintain their hands in a signing position without actually signing in order to signal their desire for a turn, or to hold the conversational floor.

The two hands in signing do not always move in the same way and some handshapes can spread beyond a lexical sign. For this reason, the start and end point for a sign may be different if one is looking at either the right or left hand. Annotations conducted on the RH and LH tiers are therefore done independently from the other hand.

There are small gaps (2 frames) between each annotation partly for historical reasons (in previous versions of ELAN, annotations which touch each other were problematic for exporting).

4.2 Weak activity

Weak activity is not annotated. Weak activity is defined as instances where the hand is relaxed and/or partially forming the handshape or partially copying the movement and where there is no discernible addition to meaning or communicative intent.

4.3 ID glossing of core lexical signs

- All lexical signs are annotated using an identifying gloss (*ID gloss*) from BSL SignBank if one exists (if one does not, see 6.4). An ID gloss is an English gloss (always in upper case, e.g., SISTER) that is consistently used with a unique sign (or 'lemma') to represent the sign whether in citation form or any phonological or morphological variant. If a sign needs more than one distinct English word to gloss it, hyphens are placed between the words (spaces are not used), e.g. PULL-APART not PULL APART OR PULLAPART. The ID gloss for each sign (in citation form) can be found in the BSL SignBank and is usually the same as one of the keywords associated with the sign. It is important that BSL SignBank is consulted to ensure that the right ID gloss is used at all times.

- As work on BSL SignBank is ongoing, signs are frequently encountered that have not yet been added to this resource. These signs are annotated as 'ADD-TO-SIGNBANK' as a placeholder followed by a suggested ID gloss in parentheses (e.g., ADD-TO-SIGNBANK(RED3)). These ADD-TO-SIGNBANK tokens are regularly discussed by the team to resolve any potential lemmatisation issues and then, when an appropriate ID gloss is agreed, these are added to SignBank and the placeholder replaced in the ELAN files. Further information about ID glossing and lemmatisation can be found in Fenlon et al. (under review).

- If a sign appears to be a lexical sign but is not known by the annotator, this is glossed as ADD-TO-SIGNBANK(UNKNOWN).

- Lexical variants have the same or similar/related meanings but (unlike phonological variants) they generally differ in two parameters or more from each other. Lexical variants are distinguished using a numeral tag (e.g., BROWN, BROWN2 and BROWN3). Note that the first lexical variant is not indicated with a number (e.g., BROWN not BROWN1). (This is simply to aid in the speed of annotation, eliminating the need to find/replace all tokens of e.g. BROWN with BROWN1 in ELAN and the same change made in BSL SignBank, when BROWN2 is encountered.)

- Manual negative incorporation is glossed using a negative suffix: KNOW-NOT not DON'T-KNOW. These signs will have a separate entry listed in BSL SignBank. If a sign signals negation using a non-manual headshake and with no modification to the sign, there is no need for a negative suffix. Instead, this information will be provided by the translations and will be annotated on the non-manual tiers in future.

- If a sign is repeated then each instance is annotated separately, as noted in the example below (courtesy of Johnston 2014). Note this is only done if the repetition is of the entire sign rather than for repetition of movement within a single sign for phonological or morphological purposes which are not included in the ID gloss tiers.

    - ID-gloss tiers: BOY SHOUT WOLF WOLF WOLF
      Free trans tier: *The boy cried "wolf, wolf, wolf".*

- Most compounds are found with distinct ID-glosses in BSL SignBank, e.g., the compound combining MOTHER and FATHER is a standard compound PARENTS. If a pairing of signs cannot be found in BSL SignBank as a compound, the two signs are tentatively annotated as one sign with two ID-glosses

but separated with a caret (e.g. FS:G-GRAPHIC^ART for 'graphic designer'). Each of these tokens will be returned to later and will either be given its own ID-gloss if it is found to be in widespread use or treated as two signs to be annotated and glossed separately. In some cases, these tokens may be best regarded as a collocation. (A collocation could be a potential compound if the overall meaning is not predictable from the two signs that are paired. A collocation is unlikely to be a compound if it is possible to insert another sign between the two signs.)

4.4 Numbers

- If a signer uses a number to refer to anything, it is glossed using words and not digits (e.g. FOUR, FOUR2, etc *not* 4). All unique number signs are listed in BSL SignBank. *Note: Some glosses for numbers have a 'N:' prefix, but others do not. This is a known inconsistency that will be corrected in the next release.*

- If a number is incorporated into a sign, the number is added to the end of the ID gloss after a hyphen, again using words and not digits: e.g., AGE-FOURTEEN *not* AGE-14.

- For a number incorporated into a lexical sign such as RANKING2, the gloss is e.g. RANKING2-THREE.

- If there is a number sequence, this is glossed as one unit in ELAN with carats (^) to separate the numbers. This is generally used when there is a sequence of signs that is difficult to segment further (e.g., suspected compounds or number sequences - NINETEEN^EIGHT^NINE not NINETEEN-EIGHTY-NINE(NINETEEN^EIGHT^NINE)).

4.5 Sign Names

NOTE: It is likely that the annotation of sign names has not been fully consistent with these guidelines in the current release. For future releases we will aim to make these more consistent.

- Signs name are entered with the prefix *SN:* followed by the proper name. The sign name for a person called *Peter* would be written as follows: SN:PETER, unless the sign name is identical in form to a lexical sign (see below).

- If the sign name is identical in form to a lexical sign, the ID gloss for the relevant sign may be identified after the name in brackets: e.g., SN:MISS-JENKINS(HAIR-BUN), SN:WEMBLEY(STADIUM) or SN:OSAMA-BIN-LADEN(BEARD).

- If the sign name uses a lexical sign that is in BSL SignBank but the annotator is unable to determine the name of the referent in this instance then the gloss UNKNOWN is used (e.g., SN:UNKNOWN(WOLF)).

- If the sign name is based on fingerspelling, the form is entered in brackets after the name and follows the conventions for fingerspelling as outlined below: e.g., SN:PETER(FS:PETER(PR)), (SN:PETER(FS:P-PETER), or SN:ALEX-FERGUSON(FS:A-ALEX^FS:F-FERGUSON)).

- If the sign name represents a sequence of both fingerspelling and a lexical sign, the whole sequence is entered as one sign name. The fingerspelled element and lexical element are included in brackets separated by a caret (e.g., SN:JOHN-KING(FS:J-JOHN^KING)).

- It can be difficult to determine when a fingerspelled sequence is in fact a sign name. Generally, we assume that fingerspelled sequences that use the initial letter of the name or fingerspelled sequences that are reduced so that they appear like lexical signs are sign names that have some degree of conventionalisation. Therefore, fully fingerspelled sequences (e.g., FS:BARRY where each letter is articulated clearly) are typically entered as fingerspelled sequences and not sign names (i.e., they do not have the prefix SN attached to them).

- Sign names are often for people but may be for e.g. places, organisations, etc. Some signs names however have been judged to be institutionalised and consistently used (some more than others) across the British Deaf community and thus are included in BSL SignBank as lexical signs (e.g., LONDON, BRISTOL, SEE-HEAR, DEAFINTELY-THEATRE).

- If the annotator cannot determine what the signed sequence represents, the glosses INDECIPHERABLE or unknown may be used (e.g., SN:INDECIPHERABLE(FS:INDECIPHERABLE(H)), SN:UNKNOWN(UNKNOWN)).

4.6 Signs from foreign sign languages / other sign systems

- Occasionally, signers use a sign that appears to be borrowed from a foreign sign language. These signs are also assigned an ID gloss and occur in BSL SignBank, along with doubtful lexeme tag (i.e., to indicate it is not a sign that belongs to BSL's core lexicon) and a Note to indicate that it is a possible borrowing. A similar approach is taken with signs that are borrowings from other signed systems (e.g., Paget-Gorman Sign System).

4.7 Buoys

Buoys are configurations on the non-dominant hand while the dominant hand continues to sign. There are several types of buoys that can be expected in signed discourse. The table below details some of the buoys that have been observed and annotated to date – these are based on Liddell (2003).

ID Gloss	Definition
LBUOY	List buoy
PBUOY	Pointer buoy
FBUOY	Fragment buoy
TBUOY	Theme buoy

4.7.1 List buoys

- When producing a list buoy, a certain number of fingers are held stretched out on the non-dominant hand, each one referring to an entity or idea, that are all somehow related, often sequentially. For example, an index finger can be held up to indicate the first of a series of items. The list buoy is annotated as LBUOY in each instance.

- If the signer points to a list buoy with the dominant hand, the point is annotated as PT:LBUOY.

4.7.2 Pointer buoys

- With a pointer buoy, the non-dominant hand points to a location associated with an important element in the discourse while the dominant hand continues signing. This is annotated as PBUOY. *Note: Some signs annotated as pointer buoys could instead be a type of pointing sign (see below) or theme buoys.*

4.7.3 Fragment buoys

- With a fragment buoy, the non-dominant hand is held from a preceding sign, it is intended, and it carries some meaning. Typically this meaning is conveyed by the signer pointing to or looking at or directing attention to the fragment buoy in some way. This differs from perseveration whereby the non-dominant hand is held from a preceding sign but may or may not be intended to carry any meaning.

This is annotated as FBUOY. *Note: There are some known mistakes in annotation of fragment buoys in the first release; some of them are actually LBUOYs.*

- If the signer points to a fragment buoy with the dominant hand, the point is annotated as PT:FBUOY.

4.7.4 Theme buoys

- With a theme buoy, the non-dominant hand (normally in the form of a vertical extended index finger) represents an important theme in the discourse while the dominant hand continues signing. This is annotated as TBUOY. *Note: Some signs annotated as TBUOYs could instead be PBUOYs or classifier constructions (particularly CLL).*

4.8 Conventions for non-core native and non-native signs

This section includes pointing signs, gestures, classifier signs, and fingerspelling.

4.8.1 Pointing Signs

- All pointing signs are prefixed with 'PT' so that they can be retrieved quickly across multiple files via the search function within ELAN. Each token is then further categorised according to its function. Below is a table detailing the different functions of pointing signs identified thus far in the BSL Corpus and the glossing conventions associated with each type.

ID Gloss	Function
PT:PRO1SG	1st person singular
PT:PRO2SG	2nd person singular
PT:PRO3SG	3rd person singular
PT:PRO1PL	1st person plural
PT:PRO2PL	2nd person plural
PT:PRO3PL	3rd person plural
PT:DET	Determiner
PT:LOC	Locative
PT:POSS1SG	1st person possessive
PT:POSS2SG	2nd person possessive
PT:POSS3SG	3rd person possessive
PT:POSS1PL	1st person plural possessive
PT:POSS2PL	2nd person plural possessive
PT:POSS3PL	3rd person plural possessive
PT:BODY	Point to a body part
PT:LBUOY	Point to a list buoy
PT:FBUOY	Point to a fragment buoy
PT:BUOY	Point to a buoy (of unspecified type)*
PT:	Ambiguous point

**Note: It is possible that some signs annotated as PT:BUOY are actually buoys (see previous section) and not points to buoys. This will be checked in later releases.*

- First person pronouns are typically unambiguous (as are second person pronouns), at least in terms of the referent they are pointing to (not necessarily the referent intended, depending on e.g. presence of constructed action). However, third person pronouns and locative points (i.e., adverbials) are often difficult to distinguish - there is some discussion in the literature about whether strictly pronominal functions for pointing can be distinguished from other uses of pointing in sign languages or indeed in gesture (Cormier, Schembri, & Woll, 2013; Johnston, 2013a, 2013b).

- Third person pronouns are typically identified as a point to the peripheral signing space, away from the conversational partner, and serving as a referent for another person in the discourse. Locative points are identified as a point to a location (e.g., associated with a place-name mentioned in the discourse) or the location at which a topic being discussed was situated. Ambiguous points may be ambiguous between a referent and the location of that referent (e.g., a person or the location where that person is standing), making it difficult to determine whether they are primarily locative or pronominal pointing signs.

- If two possible functions exist for a given pointing sign and it is difficult to decide between the two, then both possibilities are entered separated by a forward slash (e.g., PT:LOC/PT:PRO3SG). More than two possibilities may also be entered each separated by a forward slash.

- If there are more than two possible functions for a given pointing sign, then just the prefix is entered 'PT:'.

- Locative points can function as adverbials. They may point to a location in space or they may be used to refer to a specific point in time (e.g., points outwards from the signer may refer to the future, downwards points may refer to the present, and points directed behind may refer to the past).

- Pointing signs functioning as determiners are identified by their syntactic position adjacent to nominal signs and by prosody. If the point (whether it occurred before or after a noun) could be grouped with a noun as a single cohesive prosodic unit, then it is classed as a determiner.

    - Demonstratives (e.g., English equivalent of 'that' or 'this') are glossed as either PT:PRO3SG or PT:DET depending on its syntactic position. For example, the signed equivalent of 'that house' would be glossed as PT:DET and the signed equivalent of 'where do you want **this**?' would be glossed as PT:PRO3SG.

  NOTE: It is likely there is some inconsistency in annotating locative points when they function as determiners – e.g., HOUSE PT:LOC vs. HOUSE PT:DET.

- Points to the body are also prefixed with 'PT' followed by BODY (e.g., PT:BODY). We also provide additional information as to which body part the pointing sign is directed towards (e.g., PT:BODY-LEG).

  NOTE: It is likely there is inconsistency in use of PT:BODY.

- If a pointing sign points to the same location several times (e.g., is repeated for emphasis), this is annotated with one gloss. That is, the full annotation includes the sequence of pointing signs made to the same location.

- Plural points are determined based on function (i.e. meaning) rather than form. This is so we can determine whether there is a specific form associated with plural marking (e.g., an arc movement) at a later date.

  NOTE: It is possible that annotation of plurality has not always been based on function over form.

- Plural points may also incorporate number signs. If a number sign has been incorporated into a plural point, then the number it represents is suffixed to the end of the gloss after a hyphen (e.g., PT:PRO1PL-2, PT:PRO2PL-3, PT:PRO3PL-4). *Note: Number incorporated signs are subject to lexical variation just as number signs are. Therefore it is likely that in future releases we will change this system to e.g. PT:PRO3PL-FOUR2, such that the final suffix is an ID gloss for the number sign rather than the numeral.*

4.8.2 Classifier signs

- Classifier signs are annotated with the prefix *CL* and an additional letter specifying the type of classifier sign (note that we have followed widespread practice of referring to these signs as classifier signs, even though CLs have very little in common with classifier constructions in spoken languages).

Prefix	Name	Explanation
CLL	Classifier sign: Location	Depicts the location of entities, often by a short movement at a location or a hold
CLM	Classifier sign: Movement or displacement	Depicts the movement of entities
CLH	Classifier sign: Handling	Depicts the handling of an object
CLSS	Classifier sign: showing size and shape	Depicts the size and shape of entities, most often with a tracing movement but also sometimes with a hold

- The prefixing matter is followed, after a colon, by a description of the general meaning of the sign (e.g., *PERSON-MOVE or VEHICLE-MOVE*, rather than *THE-PERSON-ON-THE-RIGHT-WITH-LONG-HAIR-MOVES-SLOWLY-DIAGONALLY-TO-THE-LEFT-OUT-THE-DOOR-IN-ANGER*).

- Classifier sign sub-type categorisation is usually made easier by looking at the immediate linguistic environment or context-of-utterance rather than simply at the form of the sign alone. For example, in the following two strings the same form on the dominant hand is given handling status in one but verb of location status in the other, as a result of considering the type of sign that immediately precedes each instance (pronominal in the first, verbal in the second):

RH-IDgloss	PRO1SG	CLH:PUT-CUP-ON-FLAT-SURFACE
LH-IDgloss		CLL:FLAT-SURFACE
RH-IDgloss	HAVE	CLL:CUP-ON-FLAT-SURFACE
LH-IDgloss		CLL:FLAT-SURFACE

Following on from this, it will be evident that even though many classifier signs use both hands, often only a single entity or action is depicted. However, each hand usually carries its own semantic load in that depiction, so the annotator may categorise each hand differently, e.g., the dominant as CLH and the subordinate as CLL as above.

NOTE: Conventions for annotating classifier constructions will change in future releases.

4.8.3 Gesture

- All gesture annotations begin with 'G:' followed by a brief description of its meaning (not form – i.e. G:HOW-STUPID-OF-ME and not G:HIT-PALM-ON-FOREHEAD).

- The gesture with upturned hands (also known as the 'palm-up gesture') is annotated as G:WELL. This is the second most frequent token in the BSL Corpus to date.

- Some emblems are lexicalised and glossed as lexical signs without the gesture prefix (e.g., GOOD) although this is not always the case (e.g., G:FUCK-OFF). Whether with the gesture prefix or not, emblems have been added and are being added to BSL SignBank. As with G:WELL, the form associated with G:FUCK-OFF is consistently recognised using the same gloss (also provided in BSL SignBank). This is in recognition of the fact that they appear to have consistent form/meaning mappings even though their lexical status is unclear.

- Tokens of constructed action are also recognised as instances of gestural activity. Such instances are marked with the prefix 'G:CA:'. As with classifier signs and other types of gesture, this prefix is followed by a brief description of the token's meaning (e.g., G:CA:HOLD-HANDS-UP-IN-FRIGHT). CA tokens are not lexicalised and thus are not included in BSL SignBank.

- Sequences of constructed action can appear to be difficult to separate from the category of handling classifier signs. Within the BSL Corpus, all instances where the handshape mimics the actual handshape used to carry out the activity in real life are annotated as constructed action sequences first by default, following Cormier et al. (2012). Additional evidence is required to label the token as a handling classifier sign. This may include instances of modification that appear to be typical of lexical verbs (e.g., aspectual modification).

4.8.4 Fingerspelling

Fingerspelled forms in BSL represent a sequence of hand configurations that have a one-to-one correspondence with the letters of the English alphabet. Fingerspelled forms often violate phonological constraints associated with core native signs and are said to belong to what is known as the 'non-native lexicon' (Brentari & Padden, 2001). Below are the conventions used for fingerspelled sequences whilst annotating the BSL Corpus.

- Fingerspelling is annotated with the fingerspelled word prefixed with *FS* for 'fingerspelling' followed by a colon and then the word spelled, e.g., FS:WORD.

- If not all the letters of a word are spelled, but it is clear what word the signer is attempting to fingerspell, the full spelling of the intended word is entered (not the misspelling – e.g. FS:WORD(WRD) and not FS:WRD).

- If not all the letters of a word are spelled, and it is *not* clear what word the signer is attempting to fingerspell, the gloss INDECIPHERABLE is used followed by the actual letters produced in brackets (e.g., FS:INDECIPHERABLE(GTH)) or FS:B-INDECIPHERABLE).

- If the fingerspelling is for multiple words, *a new annotation* per word is begun even if it is one continuous act of fingerspelling (e.g., FS:MISS FS:JENKINS and not FS:MISSJENKINS).

- If the form is a single fingerspelled letter (or single fingerspelled letter repeated), the letter and the word it stands for are included in the annotation. In this case, the fingerspelled letter precedes the word it represents (e.g., FS:F-FORTUNE, FS:C-CONTRIBUTION). These sequences are also known as single manual letter signs (SMLS signs). Even if the single manual letter is repeated, only one English letter is included in the annotation (e.g., FS:F-FORTUNE not FS:FF-FORTUNE).

- Some SMLS or fingerspelled sequences are actually lexical signs in their own right. These include MOTHER, YEAR, YELLOW, DAUGHTER and CLUB. Although they are based on fingerspelling, these signs do not have the prefix 'FS:' because we do not use this prefix when the sign in question is a fully lexical sign. These signs are in BSL SignBank.

- The distinction between fingerspelled forms and lexicalised fingerspelled signs is often difficult to maintain given that many fingerspelled forms can appear partly nativised (i.e., may be in the process of becoming a fully lexical sign). For example, the sign SATURDAY3 is a fingerspelled loan (based on S-A-T) that is considered partly nativised for it does not follow constraints imposed on fully lexical signs - e.g., it violates the selected fingers constraint (Brentari, 1998). To categorise these tokens in a principled way, we use guidelines based on Cormier et al. (2008). Fingerspelled loans (with remnants of 2 or more letters) are accepted as lexical signs if there is evidence of phonological restructuring to make them more native-like. Additionally, independent agreement from native signers can also be sought to be certain that this form is consistently used for that meaning (e.g., SATURDAY3 was accorded lexical status following these criteria).

4.9 Errors, mistakes, uncertainties, false-starts, unknown signs

- If a sign appears to be a lexical sign but is not known by the annotator, this is glossed as ADD-TO-SIGNBANK(UNKNOWN).

- If two possibilities exist for a single token and it is difficult to decide between the two, then both ID glosses are entered with a forward slash between each token (e.g., LOOK/THINK). More than two ID

glosses are entered, each separated by a forward slash, if more than two possibilities exist (e.g., LOOK/THINK/SEE). This convention is also used to demonstrate ambiguity between sign types - for example, when one is uncertain if a given token is a lexical sign or a sequence of constructed action (e.g., TEA/G:CA:DRINK-TEA).

- If the identity of a sign is uncertain, but a possibility exists, then the ID gloss is entered prefixed by a '?' (e.g., ?HOME). This may be used because the token looks like it could be a phonological variant of this lemma or it may be a separate (new) lemma.

- If a sign cannot be identified (e.g., it is poorly articulated or has not been completed) and it is impossible to say with any certainty what the sign is, then the sign is glossed as INDECIPHERABLE. *Note: There are some possible inconsistencies in annotation of INDECIPHERABLE versus ADD-TO-SIGNBANK(UNKNOWN) that we hope to fix in later releases.*

- If a sign is not finished but is easily identified, the corresponding ID gloss is entered followed by FALSE-START in brackets (e.g., DOG(FALSE-START)). This convention is also extended to signs that are indecipherable because the signer did not finish what they were going to sign (INDECIPHERABLE(FALSE-START)).

5 Appendix: Work-in-progress basic template in ELAN as of Oct 2014

Below is a table showing the full set of tiers of the basic template applied to all ELAN files when initially created, for work-in-progress. The ordering of tiers follows the hierarchy of tiers within the basic template. Some of these tiers are study specific. The tiers listed in yellow denote parent tiers. Labels that have been indented represent child tiers. Rows in a lighter shade are child tiers that are, in turn, also parent tiers. Some of these tiers have been annotated for some files; these will be made available online at a later date when they have been more systematically annotated and checked.

Tier	In use?	Study specific?	CV available?	Definition (unless self-explanatory)
RH-IDgloss	YES	NO		This tier represent all manual material articulated on the right hand
RH-Handshape	YES	NO	YES	
RH-Location	NO	-		
RH-Movement	NO	-		
RH-OtherPhon	NO	-		
RH-Orientation	NO	-		
RH-Grammatical Category	YES	NO	YES	
LH-IDgloss	YES	NO		This tier represent all manual material articulated on the right hand
LH-Handshape	NO	NO	YES	
LH-Location	NO	-		
LH-Movement	NO	-		
LH-OtherPhon	NO	-		
LH-Orientation	NO	-		
LH-Grammatical Category	YES	NO	YES	
Clause	YES	NO		Potential clause-like unit
RH-Argument	YES	NO	YES	Arguments produced on the right hand
RH-Animacy	YES	NO	YES	
RH-Coreference	YES	NO	YES	
RH-Person	YES	NO	YES	
RH-Number	YES	NO	YES	
LH-Argument	YES	NO	YES	Arguments produced on the left hand
LH-Animacy	YES	NO	YES	
LH-Coreference	YES	NO	YES	
LH-Person	YES	NO	YES	
LH-Number	YES	NO	YES	
Actor non-present	YES	NO	YES	

Actor-Animacy	YES	NO	YES	
Actor-Coreference	YES	NO	YES	
Actor-Person	YES	NO	YES	
Actor-Number	YES	NO	YES	
Undergoer non-present	YES	NO	YES	
Undergoer-Animacy	YES	NO	YES	
Undergoer-Coreference	YES	NO	YES	
Undergoer-Person	YES	NO	YES	
Undergoer-Number	YES	NO	YES	
Actor-modified	YES	YES	YES	Has the verb been modified for its actor argument? (Directional verbs study)
Undergoer-modified	YES	YES	YES	Has the verb been modified for its undergoer argument? (Directional verbs study)
Direction and placement	YES	YES	YES	In what direction does the verb move? (Directional verbs study)
CA-with verbs	YES	YES	YES	Does constructed action occur with the verb in question? (Directional verbs study)
Translation	YES	NO		
Comments	YES	NO		
IPoss Study Tier	YES	YES	YES	Ongoing study on inalienable possession
Position	YES	YES	YES	
Identity	YES	YES	YES	
Utterance	YES	NO		
Clause composite	YES	NO	YES	
Turn	YES	YES		
Non-manual back channel	YES	YES		
Manual back channel	YES	YES		

6 References

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

Brentari, D., & Padden, C. A. (2001). Native and foreign vocabulary in American Sign Language: A lexicon with multiple origins. In D. Brentari (Ed.), *Foreign vocabulary: A cross-linguistic investigation of word formation* (pp. 87-119). Mahwah, NJ: Lawrence Erlbaum Associates.

Cormier, K., Fenlon, J., & Schembri, A. (2014, 2-5 January 2014). *Directionality in British Sign Language is not obligatory: The importance of corpus data when considering "agreement"*. Paper presented at the 88th Annual Meeting of the Linguistic Society of America, Minneapolis, Minnesota.

Cormier, K., Quinto-Pozos, D., Sevcikova, Z., & Schembri, A. (2012). Lexicalisation and de-lexicalisation processes in sign languages: Comparing depicting constructions and viewpoint gestures. *Language and Communication, 32*(4), 329-348.

Cormier, K., Schembri, A., & Tyrone, M. E. (2008). One hand or two? Nativisation of fingerspelling in ASL and BANZSL. *Sign Language and Linguistics, 11*(1), 3-44.

Cormier, K., Schembri, A., & Woll, B. (2013). Pronouns and pointing in sign languages. *Lingua, 137*, 230-247. doi: 10.1016/j.lingua.2013.09.010

Fenlon, J., Cormier, K., & Schembri, A. (under review). Building BSL SignBank: The lemma dilemma revisited.

Fenlon, J., Schembri, A., & Cormier, K. (2014, 8-11 July 2014). *The role of gesture in directional verbs in British Sign Language: a corpus-based study.* Paper presented at the Sixth conference of the International Society for Gesture Studies, San Diego, CA.

Fenlon, J., Schembri, A., Rentelis, R., Vinson, D., & Cormier, K. (2014). Using conversational data to determine lexical frequency in British Sign Language: The influence of text type. *Lingua, 143*, 187-202.

Johnston, T. (2013a). Functional and formational characteristics of pointing signs in a corpus of Auslan (Australian sign language): are the data sufficient to posit a grammatical class of 'pronouns' in Auslan? *Corpus Linguistics and Linguistic Theory, 9*(1), 109-159.

Johnston, T. (2013b). Towards a comparative semiotics of pointing actions in signed and spoken languages. *Gesture, 13*(2), 109-142.

Johnston, T. (2014). *Auslan Corpus Annotation Guidelines,* http://media.auslan.org.au/attachments/Johnston_AuslanCorpusAnnotationGuidelines_14June2014.pdf. Sydney, Australia: Macquarie University.

Liddell, S. K. (2003). *Grammar, gesture and meaning in American Sign Language*. Cambridge: Cambridge University Press.

Schembri, A., Fenlon, J., Rentelis, R., & Cormier, K. (2011). *British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008-2011 (First Edition)*. London: University College London. http://www.bslcorpusproject.org.

Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S., & Cormier, K. (2013). Building the British Sign Language Corpus. *Language Documentation and Conservation, 7*, 136-154.