Digging into Signs:
Towards a gloss annotation standard for sign language corpora

Onno Crasborn, Richard Bank, & Kearsy Cormier
First draft, February 2015

1. Introduction

1.1 Digging into Signs project

The Digging into Signs project (http://www.ru.nl/sign-lang/projects/digging-signs/), a one year project started in June 2014, is a joint effort of University College London (PI Kearsy Cormier) and Radboud University (PI Onno Crasborn). The project aims at developing standard annotation protocols for glossing sign language corpora, and includes reliability and validity tests as well as enhancements to the ELAN annotation software and the development of the Signbank lexical database system (Johnston, 2010; Cormier et al., 2012).

The relatively recent advances in computer technology and digital video have made it possible to collect and store large datasets of sign language video recordings. Partly due to the fact that sign languages lack an easily accessible writing system, annotation of lexical signs involves assigning a unique gloss to each sign: the ID-gloss (Johnston, 2008). These ID-glosses are stored in a computerized lexical database so that signs in the corpus can consistently be identified. However, this leaves many complexities to deal with in annotation as not all signs (or manual articulations more generally) are lexicalized. The aim of the Digging into Signs project is to make a first proposal for a standard in this area.

Although several sign language corpus projects have provided guidelines for annotation (e.g. Crasborn, Mesch, Waters, Nonhebel, Van der Kooij, Woll, & Bergman, 2007; Crasborn & Zwitserlood, 2008; Johnston, 2014; Cormier & Fenlon, 2014), there is no general agreement on annotation standards. Recent arguments for standardising sign language corpus annotation have been made by Johnston (2008) and Schembri and Crasborn (2010). This document is a firm first step towards providing such general annotation standards for sign language corpora. This first step includes finding the common ground in the existing corpus glossing practices for BSL and NGT, seeking areas where change is needed, and outlining the motivations for choosing between different alternatives.

1.2 Aim of this document

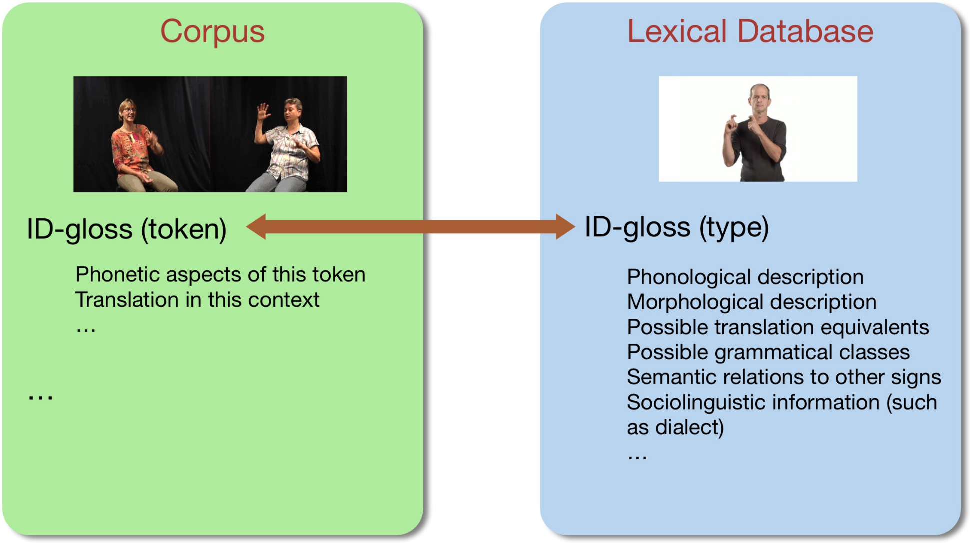
This document provides a draft of the joint annotation guidelines that will be proposed as a standard at the end of the Digging into Signs project in May 2015. It is the result of an extensive comparison and adaptation of current annotation practices in use for both the BSL and NGT sign language corpora and serves to summarize existing gloss annotation practice for these two corpora for the purpose of the Digging into Signs workshop in London on March 30-31, 2015. The final end-of-project document will take on board suggestions from the Digging workshop, and include a more detailed explanation of each category in the style of the existing separate annotation guidelines for NGT and BSL. Meanwhile, the separate guidelines for each corpus can be found in the separate BSL and NGT annotation guidelines. They can be downloaded from the workshop web page.

Some aspects of the NGT and BSL corpora are (and will remain) different: file naming conventions are different, for instance, as are tier naming conventions. For specific projects with particular research questions, different tiers will be needed in order to describe different phenomena. However, the focus for Digging is making general corpus annotation maximally useful regardless of the particular research focus. The joint annotation practices listed below focus on basic annotation of hand activity, and ensure that annotations can be made in a consistent way for both the BSL and NGT corpora, providing annotators something to go on and at the same time facilitating cross-linguistic research.

2. Annotation practice

The table below (in section 3) provides the general annotation practice agreed upon by both corpus teams (where applicable). Also, there are references to the sections in the BSL or NGT guidelines, where each topic is described in greater detail. Furthermore, BSL/NGT exceptions or specifications to general practice, if applicable, are given. First, the following will briefly outline our common annotation practices, starting with tier organization.

Gloss tiers are used for the annotation of ID-glosses. Both the BSL and NGT corpora use two tiers per signer for this: one for the left hand and one for the right hand. The choice of how to align the annotations for the two hands is partly driven by project-specific research questions or more generally the foreseen research field the corpus aims to serve. The use of ID-glosses (Johnston, 2008) is important here, because it provides a means to discriminate between the sign tokens in the corpus and the sign types in the lexicon. Further information about types and tokens should be encoded in the appropriate location, cf. the diagram below.


Corpus
ID-gloss (token)
Phonetic aspects of this token
Translation in this context
…
…
Lexical Database
ID-gloss (type)
Phonological description
Morphological description
Possible translation equivalents
Possible grammatical classes
Semantic relations to other signs
Sociolinguistic information (such as dialect)
…

Both the BSL and NGT corpus projects currently use SignBank as their lexical database of choice. SignBank originated as a lexical database/dictionary for Auslan (Johnston, 2010; http://www.auslan.org.au/), and has been / is in the process of being adapted for BSL and NGT.

Apart from the *Gloss* tiers (or rather, hand tiers), child tiers to the *Gloss* tiers may include, for example, phonological or grammatical information, or meaning. *Meaning* tiers are used to add meaning in context for each token, since the ID-gloss does not provide this information. The NGT Corpus includes a *Meaning* tier; the BSL Corpus does not but it could in the future. In addition to meaning in context of lexical signs, the NGT team uses the *Meaning* tier to describe incomplete fingerspelled words and the meaning of classifier/depicting and shape constructions. In addition, a *Reference* tier per hand is used to annotate the referents of pointing signs and shape constructions for NGT. Separate translation tiers are used for sentence-level translations in e.g. English, Dutch, etc. We refer for now to the respective annotation guidelines of the two languages for the tier setups (or templates) that are currently used for these tiers.

The Digging into Signs project focuses solely on the *Gloss* tiers (i.e. primary manual activity) for BSL and, additionally for NGT, some of their child tiers. The final output of the project will include ELAN templates to go with the proposed standard.

One of the most salient differences in annotation practice between the BSL and NGT corpora is the labels used for categories of some (mostly non-lexical) manual material: where the NGT team uses short codes for prefixes (like # for fingerspelling, or * for sign names), the BSL team uses prefixed letters (like FS: for fingerspelling, or SN: for sign name). One method results in a slightly more concise view of each annotation; with the other method it may be more straightforward to infer the meaning of the abbreviations. Annotators of the BSL and NGT corpora prefer their existing systems (respectively), so we are not planning to change these. However, if one did want to make the labelling systems consistent across languages, a simple search-and-replace is all that would be required to bring the corpora into line. The same holds for glossing manual forms of constructed action, for instance: the BSL annotations characterize the meaning in the gloss itself, whereas in the NGT annotations this is done in the *Meaning* tier. A simple script could separate or merge information as needed if a single consistent system was needed or preferred.

Other differences in possible glossing solutions are based on more principled linguistic motivations. For instance, both BSL and NGT corpora annotate list buoys in similar ways. But other types of buoys (theme buoys, pointer buoys and fragment buoys – cf. Liddell, 2003), which we refer to here as non-list buoys, are annotated differently. The NGT Corpus annotates all anticipation and perseveration of manual activity on the non-dominant hand by extending the duration of the annotation of the non-dominant hand, regardless of meaningful intent. The BSL Corpus team annotate only intentionally meaningful productions. Thus, unintended perseveration of material on the non-dominant hand is not annotated (the start and end annotation of two-handed signs in these cases follows the dominant hand). Intended perseveration on the non-dominant hand is considered to be a buoy: fragment, pointer or theme (following Liddell, 2003) and is annotated as such (following Johnston, 2014). The motivation behind these differences comes down to a difference between a preference for annotating only intentionally meaningful manual material in the BSL case versus a preference for annotating actual start and end times of signs for the purposes of prosodic analysis in the NGT case.

Another example of a difference due to linguistic motivations is in how pointing signs are annotated in the BSL and NGT corpora. In the Corpus NGT, the focus is on the form of the pointing to the extent that forms can be systematically distinguished; meaning and function is annotated on child tiers (*Reference* and *Grammatical Class*). In the BSL Corpus, focus is on the function of the pointing – e.g. pronominal signs are annotated differently from locative points and different from pointing signs that act as determiners, and this information is included in the gloss. The motivations behind these differences are partly related to the original and planned research questions to be explored via pointing in the two projects. In addition, there are different ideas about how to organize annotations: i.e. which information should be encoded within the gloss, and which should be to separate information on different tiers, for the benefit of annotators and/or anticipated end users.

3. General annotation practice by both corpus teams (where applicable).

The following table provides the general annotation practice agreed upon by both BSL and NGT corpus teams. The columns that describe the BSL/NGT exceptions or specifications also contain references to the sections in the BSL or NGT guidelines, where each topic is described in greater detail. Also, examples are given.

Topic	**General practice**	***BSL guidelines section /*** **BSL exception/specification**	***NGT guidelines section /*** **NGT exception/specification**	**Example**	
1. *Basic gloss*	All lexical signs are annotated using an identifying gloss (*ID-gloss*), written in upper case. This gloss corresponds to 'Annotation ID-gloss' in SignBank	*4.3* 'Annotation ID gloss' corresponds to lemma (citation form) only, not phonological variants	*5.3.2* 'Annotation ID gloss' may be lemma or phonological variant	BSL: SIGN SIGN02 SIGN03	NGT: SIGN-A SIGN-B SIGN-C
2. *Two-handed signs*	Two handed signs annotated on both right and left hand tiers.	*4* Start and end of two-handed signs follows dominant hand. Perseveration is not annotated. Meaningful/intentional perseveration on nondominant hand is annotated as buoy (see below).	*5.3.3.2* Start and end of two-handed signs is determined independently for each hand.	BSL/NGT: SIGN SIGN	
3. *Buoys*	Buoys get their own ID-gloss	*4.7* List, Pointer, Fragment and Theme buoys. Start and end of buoy follows start and end of co-occuring sign on dominant hand.	*5.3.7* Only List buoys (other buoys are not separately glossed, instead, non-dominant hand glosses are extended in duration; see also **2.** *Two-handed signs*, above)	BSL: LBUOY PBUOY FBUOY TBUOY	NGT: COUNTING-HAND-1 COUNTING-HAND-2
4. *Lexical variants*	Lexical variants (same or similar/related meanings but differ in two parameters or more from each other) are suffixed.	*4.3* Use a numeral tag (note that the first lexical variant is not indicated with a number. This simplifies adding variants to existing glosses)	*5.3.9.1* Use a dash and a letter (lexical and phonological variants are not distinguished)	BSL: SIGN SIGN02 SIGN03	NGT: SIGN-A SIGN-B SIGN-C
6. *Repetition*	Repeated signs are annotated seperately	*4.3*		BSL/NGT: BOY SHOUT WOLF WOLF WOLF	

Topic	**General practice**	***BSL guidelines section /*** **BSL exception/specification**	***NGT guidelines section /*** **NGT exception/specification**	**Example**	
7. *Compounds*	One ID-gloss for lexicalised compounds in SignBank, or ^ between ID-glosses for possible compounds (e.g. GRAPHIC^ART with GRAPHIC and ART in SignBank but not GRAPHIC^ART)	*4.3*	*5.3.14* No ^ for possible compounds	BSL: PARENTS GRAPHIC^ART	NGT: PARENTS GRAPHIC-ART
8. *Manual negative incorporation*	ID-Glossed using a negative suffix	*4.3*	*5.3.13*	BSL/NGT: KNOW-NOT	
9. *Directional verbs*	Directional verbs receive an ID-gloss	*4.3; 5* Grammatical/modification info on separate project-specific tiers.	*5.3.8* ID-gloss is pre- or suffixed with person number when it makes use of the 1SG location 'near body' or 'on chest'	BSL: ASK TAKE-OVER	NGT: ASK:1 1:TAKE-OVER
10. *Plurality*	Plural forms receive an ID-gloss	*4.3; 5* Grammatical info on separate project-specific tiers (see also **18.** *Points*).	*5.3.10* ID-gloss is suffixed .PL	BSL: CHILD	NGT: CHILD.PL
11. *Numbers*	Numbers receive their own ID-glosses	*4.4* Numbers are written in words	*5.3.7* Numbers are written in digits	BSL: ONE ONE2	NGT: 1-A 1-B
12. *Number sequences*	Number sequences receive one ID-gloss (see also **7.** *Compounds)*	*4.4* Carets specify the separate parts	*5.3.7* No specification of the separate parts	BSL: NINETEEN^EIGHT^NINE	NGT: 1989
13. *Number incorporation*	ID-gloss is suffixed with information about incoporated number	*4.4* ID-gloss is used for number	*5.3.7*	BSL: HOUR-FOUR02	NGT: HOUR-4
14. *Ordinal numbers*		ID-glossed as lexical signs, some of which allow number incorporation (see **13**)	*5.3.7* Suffixed -ORD	BSL: FIRST RANKING RANKING02 RANKING02-THREE	NGT: 1-ORD

Topic	**General practice**	***BSL guidelines section* / BSL exception/specification**	***NGT guidelines section* / NGT exception/specification**	**Example**	

15. *Sign names*	Prefixed	*4.5*
		Prefixed SN: followed by first name only, unless fingerspelled then fingerspelling conventions are followed (see **17.** *Fingerspelling*). If sign name is homonym with lexical sign, ID gloss for homonym is added in brackets
			5.3.17
			Prefixed * followed by both first and last name
				BSL:
				SN:FIRST
				SN:PETER(FS:P-PETER)
				SN:ALEX-FERGUSON(FS:A-ALEX^FS:F-FERGUSON))
				SN:OSAMA-BIN-LADEN(BEARD)
					NGT:
					*FIRST-LAST

17. *Fingerspelling*	Prefixed	*4.8.4* Prefixed FS: followed by word if fingerspelled fully, or by intended word followed by actual letters used if not fingerspelled fully	*5.3.16* Prefixed # followed by *perceived* letters. The presumed intended word is on the *Meaning* tier.	BSL: FS:WORD FS:WORD(WRD)	NGT: #WORD #WRD 'word'
18. *Pointing signs*	Pointing signs receive the gloss PT. with suffixes.	*4.8.1* *Function* of points including pronominal, locative or determiner functions is annotated. Points to buoys are annotated as PT: followed by type of buoy with particular fingers unspecified.	*5.3.11* *Direction* of points is annotated for spatial directions like up and down. Location of points is annotated for pointing to specific fingers of the weak hand. The referent of a pointing sign is annotated on the *Reference* child tier(s). Grammatical class distinctions may be specified on the *GrammClass* child tier.	BSL: PT:LOC PT:DET PT:PRO1SG Multiple possible functions: PT:LOC/PT:PRO3SG PT:LBUOY	NGT: PT PT:B PT:W PT:1
19. *Classifier/ depicting signs*	One of four elements for movement (MOVE, PIVOT, AT, BE), followed by classifier handshape	*4.8.2* Additionally add prefix for type of depicting sign: whole entity (DSEW), part entity (DSEP), or handling (DSH). Handshape list includes same handshapes as for NGT but some additional handshapes based on Brennan (1992) and pilot annotations.	*5.3.19* Meaning to be annotated on *Meaning* tier. More restricted handshape list than BSL.	BSL: DSEW(2)-MOVE DSEP(1)-PIVOT DSEW(2)-AT	BSL/NGT: MOVE+2 ('cat walks to and fro') PIVOT+1 ('cat's legs move around') AT+2 ('bird is here')

Topic	**General practice**	***BSL guidelines section /*** **BSL exception/specification**	***NGT guidelines section /*** **NGT exception/specification**	**Example**	
20. *Shape constructions*		*4.8.2* Annotated as type of classifier/depicting sign (DSS) but without movement	*5.3.20* Glossed as SHAPE + handshape. Meaning to be annotated on *Meaning* tier	BSL: DSS(CYL)	NGT: SHAPE+cylinder ‘drain pipe’
21. *Type-like classifier/ depciting signs*		*4.8.2* For common whole entity signs mainly depicting humans, animals and vehicles, specify handshape as well as orientation, and the referent (human, animal, entity, vehicle)	Annotate as classifier/depicting signs.	BSL: DSEW(1-VERT)-MOVE:HUMAN DSEW(FLAT-LATERAL)-AT:VEHICLE	NGT: MOVE+1 MOVE+flat
22. *Gestures*		*4.8.3* Prefixed G	*5.3.18* Generic character % for non-emblematic gesticulation. Emblematic gestures are included as lexical items in Signbank without a % prefix.	BSL: G:HOW-STUPID-OF-ME	NGT: % HEY
23. *Palm up*		*4.8.3*	*5.3.12*	BSL: G:WELL	NGT: PO
24. *Manual constructed action*		*4.8.3* Prefixed G:CA	Marked with %, meaning described on *Meaning* tier	BSL: G:CA:HOLD-HANDS-UP-IN-FRIGHT	NGT: %

2. Uncertainties

There are a couple of types of uncertainties (some of which are mutually exclusive) that are presently encoded in the following manner:

	BSL (see guidelines section 4.9)	**NGT (see guidelines sections 5.3.4 to 5.3.6, and 5.4)**
Doubt as to whether the movement is a sign or not	INDECIPHERABLE	±
Doubt about whether this gloss is chosen correctly	?GLOSS , or GLOSS1/GLOSS2/… if uncertainty is restricted	?GLOSS
First annotator doesn’t know this sign: it needs to be double-checked by colleagues	ADD-TO-SIGNBANK(UNKNOWN)	??

None of us knows this sign	ADD-TO-SIGNBANK(UNKNOWN)	???
Proposal for a new gloss, to be discussed at the weekly annotation meeting and then added to the SignBank lexicon	ADD-TO-SIGNBANK(GLOSS)	$GLOSS
New gloss needed, no proposal yet	-	$
Invisible, but likely this sign (out of video frame, behind other hand)	?GLOSS	!GLOS
Invisible, unclear or doubtful which sign it is	Options given separated by / , or INDECIPHERABLE	!
False start, but the sign is recognised as GLOSS	GLOSS(FALSE-START)	~GLOSS
False start, not clear what the sign was going to be	INDECIPHERABLE (FALSE-START)	~

3. Structural differences

Just for comparison, some structural differences between the two corpora are listed below.

Topic	**BSL**	**BSL guidelines section**	**NGT**	**NGT guidelines section**
Annotation file naming conventions	Region + Participant_number + Gender + Age + Ethnicity + Deaf/hearing_family + Task (e.g. LN01M25WDL)	3.2.1	Corpus + Number (e.g. CNGT0001) *Metadata* tiers will be created to facilitate searching for annotations + metadata	-
Video file naming conventions	Region + Participant_number(s) + task (e.g. L1l; CF5+6c)	3.2.2	Corpus + Number + Participant_number + View (e.g. CNGT0001_S003_b)	-
Video file viewing conventions	One annotation file per signer. Annotation file (e.g. LN01M25WDL) is linked to one video (e.g. L1c) showing L1 and second video showing both signers (e.g. L1+2c); annotations are for L1 only. Annotations for L2 are in a separate file, linked to videos L2c and L1+2c.	3.1	One annotation file per signer pair. Annotation file (e.g. CNGT0001) is linked to both the video of Signer 1 (e.g. CNGT0001_S003_b) and the video of Signer 2 (e.g. CNGT0001_S004_b). Annotations for both signers are in one file.	-
Parsing	Small gaps (2 frames) between annotations	4.1	Annotations are true to frames.	5.3.3.1

References

Brennan, Mary. 1992. An introduction to the visual world of BSL. In David Brien (ed.), *Dictionary of British Sign Language/English*, 1-133. London: Faber & Faber.

Cormier, Kearsy, Jordan Fenlon, Trevor Johnston, Ramas Rentelis, Adam Schembri, Katherine Rowley, Robert Adam & Bencie Woll. 2012. From corpus to lexical database to online dictionary: Issues in annotation of the BSL Corpus and the development of BSL SignBank. In Onno Crasborn et al. (eds.), *Proceedings of the 5th Workshop on the representation and processing of sign languages: Interactions between corpus and lexicon [workshop part of 8th International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey]*, 7-12. Paris: ELRA.

Crasborn, Onno, Johanna Mesch, Dafydd Waters, Annika Nonhebel, Els Van der Kooij, Bencie Woll & Brita Bergman. 2007. Sharing sign language data online: Experiences from the ECHO project. *International Journal of Corpus Linguistics* 12(4). 535–562. doi:10.1075/ijcl.12.4.06cra.

Crasborn, Onno & Inge Zwitserlood. 2008. Annotation of the video data in the Corpus NGT. http://www.ru.nl/publish/pages/527859/corpusngt_annotationconventions.pdf.

Fenlon, Jordan, Kearsy Cormier & Adam Schembri. under review. Building BSL SignBank: The lemma dilemma revisited. https://www.academia.edu/7735296/Building_BSL_SignBank_The_lemma_dilemma_revisited

Johnston, Trevor. 2008. Corpus linguistics and signed languages: No lemmata, no corpus. In O Crasborn, T Hanke, E D Thoutenhoofd, I Zwitserlood & E Efthimiou (eds.), *Construction and exploitation of sign language corpora. Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages (LREC)*, 82–87. Paris: ELRA. http://www.lrec-conf.org/proceedings/lrec2008/workshops/W25_Proceedings.pdf.

Johnston, Trevor. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics* 15(1). 106–131. doi:10.1075/ijcl.15.1.05joh.

Johnston, Trevor. 2014. Auslan Corpus Annotation Guidelines. Sydney: Macquarie University. http://media.auslan.org.au/attachments/Johnston_AuslanCorpusAnnotationGuidelines_14June2014.pdf

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

Schembri, Adam & Onno Crasborn. 2010. Issues in creating annotation standards for sign language description. *Corpora and Sign Language Technologies. Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages. Language Resources and Evaluation Conference (LREC)*, 212–216. https://www.academia.edu/5245746/Issues_in_creating_annotation_standards_for_sign_language_description.