Digging into Signs:
Emerging Annotation Standards for Sign Language Corpora

Kearsy Cormier[1], Onno Crasborn[2], Richard Bank[2]

[1]Deafness Cognition and Language Research Centre, University College London
49 Gordon Square, London, WC1H 0PD, United Kingdom
[2]Centre for Language Studies, Radboud University Nijmegen
PO Box 9103, NL-6500 HD Nijmegen, The Netherlands
E-mail: k.cormier@ucl.ac.uk; {o.crasborn, r.bank}@let.ru.nl

Abstract

This paper describes the creation of annotation standards for glossing sign language corpora as part of the Digging into Signs project (2014-2015, http://www.ru.nl/sign-lang/projects/digging-signs/). This project was based on the annotation of two major sign language corpora, the BSL Corpus (British Sign Language) and the Corpus NGT (Sign Language of the Netherlands). The focus of the gloss annotations in these data sets was in line with the starting point of most sign language corpora: to make general corpus annotation maximally useful regardless of the particular research focus. Therefore, the joint annotation guidelines that were the output of the project focus on basic annotation of hand activity, aiming to ensure that annotations can be made in a consistent way irrespective of the particular sign language. The annotation standard provides annotators with the means to create consistent annotations for various types of signs that in turn will facilitate cross-linguistic research. At the same time, the standard includes alternative strategies for some types of signs. In this paper we outline the key features of the joint annotation conventions arising from this project, describe the arguments around providing alternative strategies in a standard, as well as discuss reliability measures and improvement to annotation tools.

Keywords: Annotation standards, Glossing, Corpora, Lexical database, Signbank, ELAN annotation software

1. Introduction

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. Describing these videos, however, still has to be done manually and is extremely time consuming. Partly due to the fact that sign languages lack a commonly used writing system, annotation of lexical signs involves assigning a unique gloss to each sign: the ID-gloss (Johnston, 2008) As Johnston (2014a) emphasises, there are good arguments for prioritising annotation over transcription. These ID-glosses are stored in a 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 manual signs (or manual articulations more generally) are lexicalized, such as classifier constructions.

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, 2008a; Johnston, 2014b; Cormier & Fenlon, 2014; Mesch & Wallin, 2015), there is no general agreement on annotation standards. Recent arguments for standardising sign language corpus annotation have been made by Johnston (2008) and Schembri & Crasborn (2010). More agreement on how to gloss not only lexical but also partly-lexical and non-lexical manual actions will facilitate the access to corpus data by other researchers, and will thus contribute to the empirical study of sign languages in general and to comparative analyses in particular.

The current paper describes some aspects of our proposal for such annotation standards for glossing sign language corpora. They are the results of the Digging into Signs project (2014-2015, http://www.ru.nl/sign-lang/projects/digging-signs/). Our proposal includes a universal standard for some aspects of glosses, while offering alternatives for others. We will therefore also outline some motivations for offering alternatives when needed. The full proposal is published as a PDF document online (Crasborn, Bank & Cormier, 2015).

The focus of the project was in line with the starting point of most sign language corpora: to make general corpus annotation maximally useful regardless of the particular research focus. Therefore the joint annotation guidelines that were the output of the project focus on basic annotation of hand activity, and ensure that annotations can be made in a consistent way for all sign language corpora, providing annotators with the means to create consistent annotations for various types of signs that in turn will facilitate cross-linguistic research.

The aforementioned project not only aimed at setting a standard for the field of sign language studies throughout the world, but also to make significant advances toward two of the world's largest machine-readable datasets for sign languages – specifically the BSL Corpus (British Sign Language, http://bslcorpusproject.org) and the Corpus NGT (Sign Language of the Netherlands, http://www.ru.nl/corpusngt). We start by describing these corpora in section 2, then discussing some aspects of the annotation standard and the related issue of reliability in sections 3 and 4, respectively. Section 5 briefly characterises our efforts thus far to promote the standard. Finally, section 6 describes some related new functionalities in the annotation tool that is used for creating and exploiting these corpora: ELAN. Section 7 provides a brief conclusion.

2. Data Description

We will briefly outline the form and contents of the NGT and BSL corpora, and the previous annotation practices for both datasets.

2.1. The Corpus NGT

The Corpus NGT (Crasborn, Zwitserlood & Ros, 2008) was recorded between 2006 and 2008. It contains 72 hours of dialogues by 92 signers of various age groups from 5 regions in the Netherlands (Crasborn & Zwitserlood, 2008b), and includes both elicited narratives (fables) and free conversation. The great majority of video and annotation files are publicly available at The Language Archive (TLA) of the Max Planck Institute for Psycholinguistics (MPI) in Nijmegen (see https://hdl.handle.net/1839/00-0000-0000-0004-DF8E-6@view). The recent third public release of the annotation files (June 2015) includes tiers for gloss annotations, mouth action annotations, sentence level translations, and a tier for examples referred to in publications. About 20% of the corpus is annotated for the hands, less so for the mouth. The corpus specific annotation guidelines (that can be found on the TLA website as well: https://hdl.handle.net/1839/00-0000-0000-0020-B7CA-4@view) cover all aspects of annotation of the Corpus NGT, not just the publicly available part.

2.2. The BSL Corpus

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, 2014; 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 on the BSL Corpus Data page: http://www.bslcorpusproject.org/data/.

As of 2016, there are around 100 files that have been annotated primarily for manual activity at the lexical level (on right hand and left hand tiers) 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, 2015; Fenlon, Schembri, & Cormier, under review). Additionally, under the Digging into Signs project, an additional 50 files have been annotated at the lexical (ID gloss) level: 25 each from Belfast and Glasgow from the narrative data. Annotation guidelines for manual activity used for all of these files can be found on the BSL Corpus website (http://www.bslcorpusproject.org/cava/).

3. Annotation Standards

Some core features of the gloss annotation guidelines for these two corpora are shared with most researchers in the field: glosses are written words in the standard orthography for a spoken language that uniquely identify a sign form (that is, they function as formal identifiers rather than as translations), they are written in capital letters, and when multiple words are needed for the ID they are separated by hyphens. Moreover, the language of the glosses is trivial in a sense: while it makes most sense to use the spoken language best known to the signers and annotators, it has also been argued that the glosses should match the language of the publication in the case of the citation of examples (Frishberg, Hoiting & Slobin, 2012). As ELAN allows for multilingual controlled vocabularies (Crasborn & Sloetjes, 2014), gloss annotations can be added in one language and displayed in another to other users. The Corpus NGT glosses were created in Dutch, but can also be displayed in their English form.

For most other aspects of glosses, however, a lot of variation can be observed. Our main goal was to develop annotation standards for glosses of signs in sign language corpora, particularly for partly-lexical or non-lexical material. A comprehensive description of the annotation guidelines that were the output of the Digging into Signs project can be found in Crasborn, Bank & Cormier (2015). To summarise, we identified 22 categories (see Table 1) and extensively compared and adapted our (former) annotation practices for both the NGT and BSL corpora (Crasborn, Bank, Zwitserlood, Van der Kooij, De Meijer, & Sáfár, 2015, and Cormier, Fenlon, Gulamani, & Smith, 2015, respectively). This was achieved by several rounds of pilot annotation of small amounts of data from both corpora.

1	Basic gloss	12	Number incorporation
2	Two-handed signs	13	Ordinal numbers
3	Buoys	14	Sign names
4	Lexical variants	15	Fingerspelling
5	Repetition	16	Pointing signs
6	Compounds	17	Classifier/depicting signs
7	Manual negative incorporation	18	Type-like classifier/depicting signs
8	Directional verbs	19	Shape constructions
9	Plurality	20	Gestures
10	Numbers	21	Palm up
11	Number sequences	22	Manual constructed action

Table 1: 22 categories on which agreement was sought

On most of these items we reached agreement, although there are some minor notational issues. For example, as prefixes to indicate certain items, the NGT team prefers to use special characters for brevity, whereas the BSL team prefers abbreviations to provide some memory support, such as in Table 2.

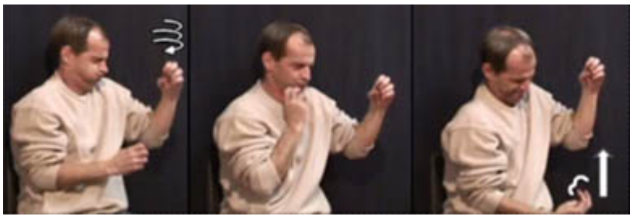
	BSL prefix	NGT prefix
Fingerspelling	FS:	#
Sign name	SN:	*
Gestures	G:	%

Table 2: examples of notational differences

When doing cross-linguistic comparisons, however, these differences are easily overcome by a simple search and replace in the annotation software or a text editor. Similarly trivial notation differences can be found for the glossing of (cardinal and ordinal) numbers and number sequences.

We realised, however, that some aspects of the NGT and BSL corpora are (and will remain) different. These include not only annotation conventions, but also file and tier naming conventions. Also, for specific projects with particular research questions, additional tiers will be needed in order to describe different phenomena related to the manual articulators. As these will contain properties or classifications particular to certain research questions and are more likely to reflect specific theoretical perspectives, standardisation will be more of a challenge. However, it is important to be aware that even general glossing conventions also come with linguistic assumptions (Cormier, Crasborn, & Bank, 2016).

A point in case is the annotation of some types of buoys (meaningful perseveration) vs. meaningless perseveration, and this was a reason to suggest two alternative approaches to glossing instances of non-dominant hand spreading. The notion of buoy as first proposed by Liddell (2003) characterises spreading of the non-dominant hand that fulfils the discourse function of highlighting information. Depending on the sign that is held in its final position, different types of buoys are distinguished: theme, pointer or fragment buoys (list buoys behave differently and are not included here). For both the time alignment and the content of the annotation, different options are proposed. The spreading behaviour of a sign can either be annotated by adding a separate gloss annotation for the hold part of the sign, or the length of the annotation for the sign can be so long as to include both the movement part and the long hold at the end. For the content of the gloss annotation, one can opt for only the gloss of the source sign, include an explicit labelling of that sign as functioning as a buoy, or add a categorisation of the type of buoy. Depending on the amount of linguistic analysis one wants to include in the gloss tier (and thus require from annotators), either a more phonetic or a more functional approach will be attractive. In some cases, it may be possible to use the corpus annotations to test which approach works best (Cormier, Crasborn, & Bank, 2016).



1a:

GlossL	MOVE+O	BE+O-------------------------------------	
GlossR		CAT	MOVE+2

1b:

GlossL	MOVE+O	FBUOY	FBUOY
GlossR		CAT	MOVE+2

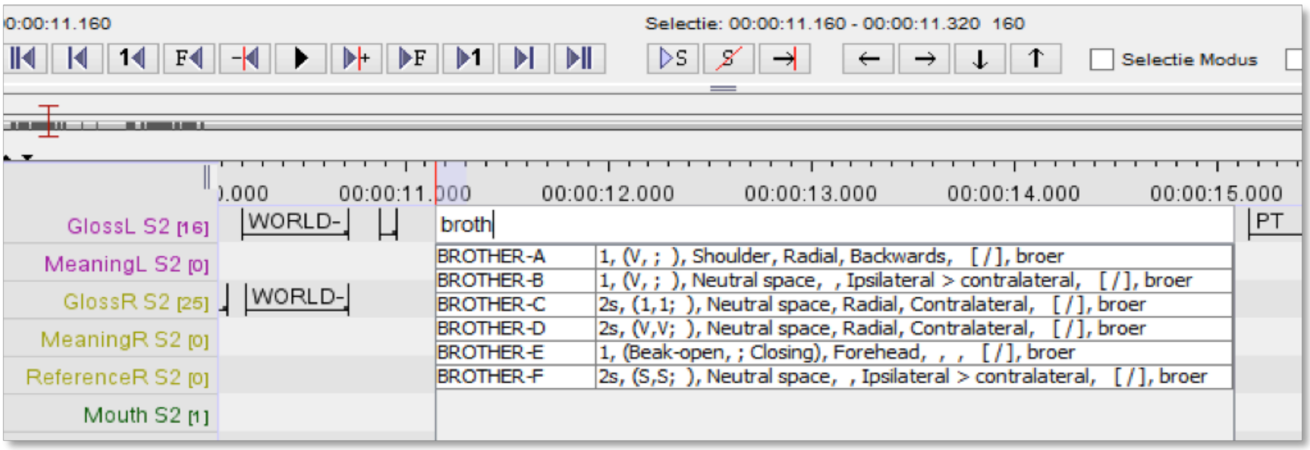
Figure 1: Example from NGT with non-dominant hand annotated as perseveration (1a) versus as a fragment buoy (1b). MOVE+O and MOVE+2 are depicting constructions which include movement with an O-handshape and 2-handshape respectively. BE+O is a depicting construction with no movement and an O-handshape.

4. Reliability and Validity

Near the end of the project, in order to address an additional aim of testing reliability of these annotation standards, we also conducted a small reliability study of each corpus, with 2 annotators independently annotating a sample of BSL data and 3 annotators independently annotating a sample of NGT data. (Cross-linguistic reliability was not possible because none of the annotators knew both sign languages.) Reliability of the BSL data (around 200 annotations, content of annotations only) was 75% across the 2 annotators. Reliability of the NGT data (around 150 annotations, content of annotations only) was an average of 71% across the 3 pairs of annotators. A content analysis of the present annotation data is taking place at the time of writing. We further plan to develop and apply more detailed measures of reliability in the near future. This will include measurements on alignment of annotations, which was outside the scope of the Digging into Signs project.

5. Improvement of Annotation Software

One of the aims of the project was to improve software for sign language corpus annotation. This project exploited the most widely used multimedia annotation tool in sign language research: ELAN (tla.mpi.nl/tools/tla-tools/elan), developed by the Max Planck Institute for Psycholinguistics (Wittenburg, Brugman, Russel, Klassmann & Sloetjes, 2006). MPI tools are open source software which are well documented and supported. The multilingual user interface of ELAN (like that of other annotation tools) allows access to the software for research assistants with limited knowledge of English, like some of the deaf annotators in the Dutch team. Version 4.9.0 of ELAN was released in May 2015, and included an improvement in the use of External



broth

BROTHER-A	1, (V, ;), Shoulder, Radial, Backwards, [/], broer
BROTHER-B	1, (V, ;), Neutral space, , Ipsilateral > contralateral, [/], broer
BROTHER-C	2s, (1,1;), Neutral space, Radial, Contralateral, [/], broer
BROTHER-D	2s, (V,V;), Neutral space, Radial, Contralateral, [/], broer
BROTHER-E	1, (Beak-open, ; Closing), Forehead, , , [/], broer
BROTHER-F	2s, (S,S;), Neutral space, , Ipsilateral > contralateral, [/], broer

Figure 2: creating a new annotation from an ECV. The user is presented with a drop-down list with possible choices of ID gloss based on the first few letters that are typed in, along with basic phonological information about each to help with identification.

Controlled Vocabularies (ECVs). An ECV provides the annotator with a list of choices (Figure 2) that are based on a lexicon (see below). Working with a lexicon-based ECV eliminates spelling errors, and greatly reduces the number of choices our annotators have to make. The ECV file is no different in format than a regular controlled vocabulary, but because of its size is stored externally, on a server, rather than in each annotation file. The ECV file is an XML file that stores a value and a description for each lexical entry in one or more languages, just like the inline vocabularies in ELAN documents. The list of glosses taken from a lexicon (the values) can thus include extra information (in the description field), which can contain for instance phonological information of the citation form, or information about the semantics.

With release 4.9.0 of ELAN, this description field can be shown at the time of selecting a new gloss. In Figure 1, a screenshot of this drop-down list is shown, with in the second column phonological information on the all glosses in the lexicon that start with ‘broth-’. The format is as follows: handedness, (strong hand, weak hand, handshape change), location, absolute orientation: movement, movement direction, movement shape, [number of occurrences / number of signers], keywords/translation equivalents. By displaying phonological information about an ID-gloss at the time of creating a new annotation, annotators can assure themselves that indeed they are selecting the right ID-gloss for the right form, without necessarily having to look up the gloss and video in the lexicon itself every time.

Additionally, a Tier Set function has been created (in beta testing at the time of writing), by which a different selections of tiers can each be assigned a name, after which the user can quickly hide and show groups of tiers in the timeline viewer and other menus. With the large number of tiers that are created for many corpora, it is a challenge to present all and only the desired information at any given time. The Tier Set function allows users to quickly display a specific (pre-defined) set of tiers for a specific purpose, for instance in order to make a quick annotation on a tier that a user is not normally working on. Annotators that are normally focussing on the gloss and mouth tiers can thus quickly show the handshape tiers to annotate a deviant handshape and then hide it again, or quickly hide or show translation tiers depending on the annotator’s needs. This results in an uncluttered workspace with easier access to relevant tiers.

The lexicons that form the basis of these lists of glosses are the NGT Signbank (http://signbank.science.ru.nl) and BSL Signbank (http://bslsignbank.ucl.ac.uk/), forks of the original Auslan Signbank (http://www.auslan.org.au/). In future it is possible that ECVs within ELAN could be adapted to work with lexical databases unrelated to these.

6. Conclusion

In summary, the Digging into Signs project provided some much needed improvement to sign language annotation software tools and also brought the field of sign language corpus research one step closer to achieving cross-linguistic annotation standards for sign language data.

However, several challenges remain. Changing existing annotations in a corpus to conform with changed annotation standards is a lot of work, and unfortunately we haven’t yet been able to implement all proposed standards into our existing annotations. However, all annotations added to our corpora in current projects make use of the new standards, and older annotations will follow in due time. Also, as annotation standards are implemented and evaluated, it is possible that some changes may be needed, resulting in a need to revisit and change the standards. Open access, sharing and transparency across annotators and projects will help ensure these issues can be addressed and resolved as this field of corpus sign linguistics moves forward.

7. Acknowledgements

This research was possible thanks to the Digging into Data Challenge which for this project included joint funding from Joint Informations Systems Committee (JISC), Arts and Humanities Research Council (AHRC) and Economic and Social Research Council (ESRC) in the UK (grant 3152) and the NWO in the Netherlands (grant 640.006.011). We thank our project teams: Jordan Fenlon, Sannah Gulamani, Sandra Smith in the UK and Merel van Zuilen and Anique Schüller in the Netherlands.

8. References

Cormier, K., Crasborn, O., & Bank, R. (2016). *The role of theory in sign language corpus annotation.* Paper presented at the 12th Conference on Theoretical Issues in Sign Language Research, Melbourne, Australia, , 4-7 January 2016.

Cormier, K. & Fenlon, J. (2014). BSL Corpus Annotation Guidelines, v. 1. Deafness, Cognition and Language Research Centre, University College London. http://www.bslcorpusproject.org/wp-content/uploads/BSLCorpusAnnotationGuidelines_23October2014.pdf.

Cormier, K., Fenlon, J., Gulamani, S., & Smith, S. (2015). BSL Corpus Annotation Conventions, v. 2.1. Deafness, Cognition and Language Research Centre, University College London. http://www.bslcorpusproject.org/wp-content/uploads/BSLCorpus_AnnotationConventions_v2.1_July2015.pdf

Cormier, K., Fenlon, J., & Schembri, A. (2015). Indicating verbs in British Sign Language favour motivated use of space. *Open Linguistics, 1*(1), 684-707. doi:10.1515/opli-2015-0025

Crasborn, O., Bank, R., & Cormier, K. (2015). Digging into Signs: Towards a gloss annotation standard for sign language corpora. Nijmegen: Radboud University & London: University College London. http://www.ru.nl/publish/pages/723853/dis_annotation_guidelines_4may2015.pdf. DOI: 10.13140/RG.2.1.2468.5840.

Crasborn, O., Bank, R., Zwitserlood, I., Van der Kooij, E., De Meijer, A., & Sáfár, A. (2015). Annotation conventions for the Corpus NGT. Version 3, June 2015. Radboud Universiteit Nijmegen. https://hdl.handle.net/1839/00-0000-0000-0020-B7CA-4@view

Crasborn, O., Mesch, J., Waters, D., Nonhebel, A., Van der Kooij, E., Woll, B., & Bergman, B. (2007). Sharing sign language data online: Experiences from the ECHO project. *International Journal of Corpus Linguistics*, *12*(4), 535–562.

Crasborn, Onno, & Slöetjes, Han. (2014). Improving the exploitation of linguistic annotations in ELAN. In N. Calzolari, K. Choukri, & et al. (Eds.), *Language Resources and Evaluation 2014*. Paris: ELRA. http://www.lrec-conf.org/proceedings/lrec2014/pdf/567_Paper.pdf

Crasborn, O., & Zwitserlood, I. (2008a). Annotation of the video data in the Corpus NGT. http://www.ru.nl/publish/pages/527859/corpusngt_annotationconventions.pdf

Crasborn, O., & Zwitserlood, I. (2008b). The Corpus NGT: an online corpus for professionals and laymen. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood, & E. Thoutenhoofd (Eds.), *Construction and exploitation of sign language corpora. Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages (LREC)* (pp. 44–49). Paris: ELRA. http://www.lrec-conf.org/proceedings/lrec2008/workshops/W25_Proceedings.pdf

Crasborn, O., Zwitserlood, I., & Ros, J. (2008). Corpus NGT. An open access digital corpus of movies with annotations of Sign Language of the Netherlands (Video corpus). Centre for Language Studies, Radboud University Nijmegen.

Fenlon, J., Schembri, A., & Cormier, K. (under review). Modification of indicating verbs in British Sign Language: A corpus-based study.

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. doi:10.1016/j.lingua.2014.02.003

Frishberg, N., Hoiting, N., & Slobin, D. I. (2012). Transcription. In R. Pfau, M. Steinbach, & B. Woll (Eds.), *Sign language: An international handbook* (pp. 1045-1075). Berlin: Mouton de Gruyter.

Johnston, T. (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)* (pp. 82–87). Paris: ELRA. http://www.lrec-conf.org/proceedings/lrec2008/workshops/W25_Proceedings.pdf.

Johnston, T. (2014a). The reluctant oracle: using strategic annotations to add value to, and extract value from, a signed language corpus. *Corpora*, *9*(2), 155–189.

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

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

Mesch, J., & Wallin, L. (2015). Gloss annotations in the Swedish Sign Language Corpus. *International Journal of Corpus Linguistics*, *20*(1), 103–121.

Schembri, A., & Crasborn, O. (2010). Issues in creating annotation standards for sign language description. In P. Dreuw, E. Efthimiou, T. Hanke, T. Johnston, G. Martinéz Ruiz, & A. Schembri (Eds.) *Corpora and Sign Language Technologies. Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages. Language Resources and Evaluation Conference (LREC)* (pp. 212–216). http://www.sign-lang.uni-hamburg.de/lrec2010/lrec_cslt_01.pdf

Schembri, A., Fenlon, J., Rentelis, R., & Cormier, K. (2014). British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008-2014. London: Deafness, Cognition and Language Research Centre, University College London.

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

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, D. Tapias (Eds.), *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 1556–1559). Genoa. http://www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf