Jisc Final Report

Project Information			
Project Identifier	*To be completed by Jisc*		
Project Title	Digging into signs: Developing standard annotation practices for cross-linguistic, quantitative analysis of sign language data		
Project Hashtag	#digsigns		
Start Date	1 June 2014	**End Date**	31 May 2015
Lead Institution	University College London		
Project Director	Dr. Kearsy Cormier		
Project Manager	Dr. Kearsy Cormier		
Contact email	k.cormier@ucl.ac.uk		
Partner Institutions	Radboud University Nijmegen		
Project Web URL	http://www.ru.nl/sign-lang/projects/digging-signs/		
Programme Name	Digging into Data		
Programme Manager	Chris Brown		

Document Information			
Author(s)	Dr. Kearsy Cormier		
Project Role(s)	Principal Investigator (UCL)		
Date	31 May 2015	**Filename**	JISC_Final Report Digging into Signs_2June2015.doc
URL	*If this report is on your project web site*		
Access	This report is for general dissemination		

Document History		
Version	**Date**	**Comments**
1	2 June 2015	

Table of Contents

1 ACKNOWLEDGEMENTS 3

2 PROJECT SUMMARY 3

3 MAIN BODY OF REPORT 3

3.1 PROJECT OUTPUTS AND OUTCOMES 3
3.2 HOW DID YOU GO ABOUT ACHIEVING YOUR OUTPUTS / OUTCOMES? 4
3.3 WHAT DID YOU LEARN? 4
3.4 IMMEDIATE IMPACT 6
3.5 FUTURE IMPACT 6

4 CONCLUSIONS 6

5 RECOMMENDATIONS 6

6 IMPLICATIONS FOR THE FUTURE 6

7 REFERENCES 7

8 APPENDICES (OPTIONAL) 7

1 Acknowledgements

This project was funded by the Digging into Data Challenge, phase 3. Thanks to the partner PI on this project, Dr. Onno Crasborn (Radboud University Nijmegen, Netherlands), and to the entire Dutch team.

2 Project Summary

For sign languages used by deaf communities, linguistic corpora have until recently been unavailable, due to the lack of a writing system and a written culture in these communities, and the very recent advent of digital video. Recent improvements in video and computer technology have now made larger sign language datasets possible; however, large sign language datasets that are fully machine-readable are still elusive. This is due to two challenges.

1. Inconsistencies that arise when signs are annotated by means of spoken/written language.
2. The fact that many parts of signed interaction are not necessarily fully composed of lexical signs (equivalent of words), instead consisting of constructions that are less conventionalised.

As sign language corpus building progresses, the potential for some standards in annotation is beginning to emerge. But before this project, there were no attempts to standardise these practices across corpora, which is required to be able to compare data crosslinguistically. This project thus had the following aims:

1. To develop annotation standards for glosses (lexical/word level)
2. To test their reliability and validity
3. To improve current software tools that facilitate a reliable workflow

Overall the project aimed not only to set a standard for the whole 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).

3 Main Body of Report

3.1 Project Outputs and Outcomes

Output / Outcome Type *(e.g. report, publication, software, knowledge built)*	**Brief Description and URLs (where applicable)**
Knowledge built & exchanged	Knowledge built from pilot annotation process of both corpora and knowledge exchanged particularly at the Digging into Signs international workshop held on 30-31 March: http://www.bslcorpusproject.org/events/digging-workshop/
Guidelines	Cross-corpus sign language annotation guidelines: Crasborn, Bank and Cormier (2015b) : http://www.ru.nl/sign-lang/projects/digging-signs/
Annotations	Open access corpus annotations for BSL and NGT: Planned annotations have been completed and will be uploaded to CAVA (for BSL) and TLA (for NGT), along with release notes and ELAN templates by mid-June 2015
Dictionaries	Online lexica for BSL and NGT: http://bslsignbank.ucl.ac.uk ; http://signbank.science.ru.nl
Software	Improvements to the ELAN software : version 4.9.0 released 27 May 2015 : https://tla.mpi.nl/tools/tla-tools/elan/release-notes/
Publication	White paper about cross-corpus sign language annotation guidelines (planned for November 2015)

3.2 How did you go about achieving your outputs / outcomes?

Our main goal was to develop annotation standards for glosses of signs in sign language corpora, particularly for semi-lexical or partly-lexical material. (Annotation of fully lexical signs have their own challenges but many of the issues for lexical signs have been addressed in Fenlon, Cormier & Schembri, 2015.) The project began with an extensive comparison and adaptation of (former) annotation practices for both the BSL and NGT sign language corpora (Crasborn et al., 2015a; Cormier et al., 2015). This process of adaptation was achieved by several rounds of pilot annotation of small amounts of data from both corpora. In the end we achieved 5000 new annotations of each language, and revised annotations of existing 15,000 annotations within each language, as initially planned. 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 average 71% across the 3 pairs of annotators. More detailed measures of reliability will be done for the white paper publication planned for submission by November 2015.

In addition to several rounds of pilot annotation and reliabilty, an important part of the project was engaging with our most important stakeholders at a workshop that we hosted in March 2015 – the Digging into Signs Workshop (http://www.bslcorpusproject.org/events/digging-workshop/). The workshop was enormously successful, with around 80 academics attending representing 19 sign languages from all over the globe. Hosting this workshop allowed us to present an early draft of joint annotation guidelines to leaders of other sign language corpus projects to get their feedback on the DIS proposals based on their existing practice. This was extremely useful not only for us in terms of getting their feedback on our drafts but also as a much needed way of bringing sign language corpus project researchers back together again for the first time since the Sign Language Corpus Network was last funded (http://www.ru.nl/slcn/, 2009-2011). The final proposed annotation standards (Crasborn et al., 2015b) takes into account feedback from researchers at this workshop. For the most part, the categories that we had chosen to annotate were the same for most sign languages – we took this as a good measure of validity of our annotation guidelines (which was an additional aim of the project).

Another aim of the project was to improve software tools for sign language corpus annotations. We are committed to using open, shared, and documented standards. Thus 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. MPI tools are open source software and well documented and supported. Many tools like ELAN are available with multilingual user interfaces, also allowing access for research assistants with limited knowledge of English, like the deaf assistants to the Dutch team. Version 4.9.0 of ELAN was released on 27 May 2015. This version includes some features arising out of the current project, such as calculation of inter-rate reliability (which arose out of the reliability study of this project), and improvement to the use of Controlled Vocabularies, including External Controlled Vocabularies that are based on a lexicon (required for the present project which relies on external lexica in the form of SignBanks). Additionally, a Tier Set function has been created (currently in beta testing), 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. This will be useful for corpus annotation work: users can quickly display the handshape tier to annotate a deviant handshape, or quickly hide or show translation tiers depending on their needs. This function will be included in the next release of ELAN later in 2015 (probably version 4.9.1).

3.3 What did you learn?

We initially intended to have one final set of joint annotation guidelines based on both BSL and NGT corpora, to be used with all SL corpora. We have ended up with a single document that outlines the annotation standards devised for BSL and NGT - see Crasborn, Bank & Cormier 2015b. But some practices across the two corpora will remain different (as noted below) – for these, we also will have release notes to accompany each of the BSL and NGT corpus annotations when they are released soon after the project end in mid-2015.

In the process of attempting to standardise practices across the BSL and NGT corpora, some practices were already the same or were easily standardised. In some cases, the BSL practice was adopted by the NGT team (e.g. removal of grammatical information from the gloss tiers), in others the NGT practice was adopted by the BSL team (e.g. annotating two-handed signs on both the right and left hand tier), and in other cases, the two corpora converged on a new practice not previously used by either corpus (e.g. classifier/depicting signs and pointing signs). In addition, we found that some practices needed to be kept separate for the two corpora for a variety of reasons. In some cases, practices were kept different because of different linguistic motivations for how some things should be annotated (e.g. pointing signs or buoys). In other cases, practices diverged because the differences were trivial, ingrained in existing practice by each team and could if necessary be easily replaced (like for like) if necessary for later crosslinguistic study (e.g. use of symbols vs abbreviated prefixes for labelling of some categories). (For details, see Crasborn, Bank & Cormier 2015b.)

Still other differences were due to apparent differences between the structure of BSL versus NGT. For example, name signs in BSL very often involve some elements of fingerspelling and partly because of this, are often not lexicalised. Thus the guidelines needed for annotating name signs are complex and are closely related to the annotation practices for fingerspelling, typically only first names are included, and these names are not typically included in the BSL SignBank lexical database. Name signs in NGT however appears to prefer name signs that are lexicalised and are therefore included in the NGT SignBank, with first and last name of the individual to uniquely identify that person. This difference in practice has some implications not only in annotation but also in post-annotation processing – e.g. the NGT team anonymise their data before uploading it since first and last names are always given in the annotations.

One intended output of this project was archiving project annotations from both the BSL and NGT Corpora with The Language Archive at MPI, which is already the home of the Corpus NGT. The BSL Corpus data (including videos and annotation files) have been archived at UCL CAVA since 2011. Initially we had intended to create a mirror archive at TLA for the BSL Corpus. However, upon further investigation it has become clear that it makes more sense for the future updates to have a single archive for the annotations thus requiring regular updating in only one location. Instead we will be opting to archive the BSL Corpus video data only (without annotations) at TLA as a mirror site. This will help increase the visibility of the BSL Corpus and provide an additional backup to the entire video corpus, while ease of updating will be preserved since it is only the annotation files that will be regularly updated for the forseeable future (not the video files).

Another intended output of this project was improvements to the LEXUS software, also created by MPI for the purpose of publishing sign language lexicons that interface with ELAN. Unfortunately further development of LEXUS during the course of this project proved problematic for a number of reasons, and because of this, MPI has discontinued active development of LEXUS altogether. Instead, both the BSL and NGT corpus projects are using SignBank as their lexical database of choice. SignBank originated as a lexical database/dictionary for Auslan (Johnston, 2010; http://www.auslan.org.au/) and was adapted for BSL in 2014 – this adaptation was already in progress at the time of start of the Digging into Signs project and SignBank was always intended to be the primary lexicon for BSL (LEXUS would have been a secondary, mirror lexicon system). BSL SignBank launched in September 2014 and exists as both a public dictionary and a lexical database for researchers (http://bslsignbank.ucl.ac.uk; researcher access can be gained by registering on the system as a university staff member or research student). NGT SignBank (http://signbank.science.ru.nl) is still under development and presently exists only as a lexical database for researchers. Part of the Controlled Vocabulary improvements in ELAN (noted above) are to allow for a closer interfacing between ELAN and the SignBank system. We hope to seek further funding to enhance these interfaces further.

We have evaluated success of the workshop and project overall, including engagement by stakeholders by monitoring activity on social media (e.g. the #digsigns hashtag on Twitter). There has also been considerable activity on Facebook about the workshop and project as well.

3.4 Immediate Impact

The most notable, immediate impacts so far of the project at UCL include the capacity building of two deaf research assistants (Sannah Gulamani and Sandra Smith). In addition to the benefits to the two of them in terms of research career development, UCL has also benefited from this training (e.g. it is hoped that one or both of them will go on from here to do a postgraduate degree in sign language linguistics at UCL). UCL also benefited greatly from having hosted the Digging into Signs workshop in March 2015, as this brought a large number of sign language researchers from around the world to UCL. The wider community – particularly the community of sign language researchers - benefited from the workshop as well in terms of advancing data collection, annotation, management and dissemination techniques in corpus-based sign language research, building a network of co-operation and collaboration among sign language researchers in Europe and abroad, providing opportunities for future cross-linguistic work in the area and importantly, helped to provide training opportunities for junior researchers in the field (those who attended included both junior and senior researchers in the field). These benefits took the form of draft annotation guidelines leading up to the workshop, the discussion that took place about these guidelines at the workshop, and the networking opportunities that the workshop offered. Informal feedback from delegates who attended the workshop as well as more public feedback e.g. on social media confirms these benefits. This project has also helped raised awareness of sign language corpora, and of sign languages and deaf communities generally, amongst academics who are outside of this field via the Digging into Data Challenge cross project meetings and associated meetings (e.g. AHRC Digital Transformations).

3.5 Future Impact

Annotation standards for sign language corpora will enable sign language corpora to be more fully annotated than ever before. Such annotated corpora will have immediate and long-term practical application as valuable references for information about sign language usage (e.g. frequency) that can be consulted by researchers in linguistics, psychology, gesture studies, language development, neuroscience, and related fields, both in their own right and also in comparison with spoken/audio-visual corpora (as they begin/continue to emerge, since there are not yet annotation standards for these corpora either). Further corpus-based research on sign language structure will inform and improve sign language teaching materials which are in great need of an evidence base. This will, in turn, lead to the improvements in the training of sign language teachers and interpreters, and in the education of deaf children. Additionally, annotated corpora are sorely needed for work on computational modelling of sign language (e.g. signing avatars and automatic sign language recognition) to verify existing work. This will help computational linguists to move beyond the state of the art, which for sign languages is currently limited to automated synthesised production and automatic recognition of signs produced in isolation. We will track these impacts by following citations to the annotation standards, white paper and any other published outputs arising out of this project, as well as keeping track of mentions of the BSL Corpus, Corpus NGT, Digging into Signs on social media.

4 Conclusions

Overall the project met its objectives of creating the initial steps towards annotation standards for sign language data at the lexical/word level, testing their reliability and validity, and improving the multimedia annotation tool ELAN.

5 Recommendations

The annotation standards produced as part of this project (Crasborn et al., 2015b) outline the recommendations for how sign language corpus data should be annotated at the word/lexical level (for partly lexical signs).

6 Implications for the future

This project has created annotation standards for the initial steps towards annotation standards for sign language data at the lexical/word level. However, there is much left to be done in the future. Firstly, while most annotation conventions for partly lexical signs that have been agreed upon for the BSL and NGT corpora have been implemented for the full set of pilot annotations of new and existing data (i.e. 5000 new and 15,000 existing), some have not yet been fully implemented for all of the

existing annotations – full implementation is planned for the future. Also importantly, there are many other levels of corpus annotation both above and below the lexical level (e.g. phonetic/phonological, morphological, syntactic, discourse) that also need annotation standards. Work by Johnston (2014) for Australian Sign Language is emerging as potential best practice for some of these levels of linguistic analysis but this has not been extended across any other sign languages yet. Also there are no standards at all yet for corpus translation. These are all important areas for future research.

The project PIs Cormier and Crasborn will remain as long term contacts for the BSL and NGT corpora respectively. The primary user communities for both corpora are (sign) language researchers as well as sign language teachers, students, interpreters, and the wider Deaf community.

7 References

Cormier, K., Fenlon, J., Gulamani, S., & Smith, S. (2015). *BSL Corpus Annotation Conventions, v. 2.0, http://www.bslcorpusproject.org/wp-content/uploads/BSLCorpus_AnnotationConventions_v2_-Feb2015.pdf*: Deafness, Cognition and Language Research Centre, University College London.

Crasborn, O., Meijer, A. de, Bank, R., Zwitserlood, I., Kooij, E. van der, & Sáfár, A. (2014). Annotation conventions for the Corpus NGT. Version 2, November 2014. Radboud University Nijmegen. Retrieved from http://hdl.handle.net/1839/00-0000-0000-0020-1393-E@view

Crasborn, O., Bank, R., & Cormier, K. (2015a). Digging into Signs: Towards a gloss annotation standard for sign language corpora, first draft, February 2015. http://www.bslcorpusproject.org/wp-content/uploads/Digging_into_Signs_draft_annotation_standard_Feb2015_forweb.pdf.

Crasborn, O., Bank, R., & Cormier, K. (2015b). Digging into Signs: Towards a gloss annotation standard for sign language corpora, final version, May 2015. http://www.ru.nl/sign-lang/projects/digging-signs/

Fenlon, J., Cormier, K., & Schembri, A. (2015). Building BSL SignBank: The lemma dilemma revisited. *International Journal of Lexicography*. doi: 10.1093/ijl/ecv008

Johnston, T. (2010). From archive to corpus: transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics, 15*(1), 104-129.

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

8 Appendix: Glossary

BSL: British Sign Language
NGT: Sign Language of the Netherlands (in Dutch, Nederlandse Gebarentaal)
ELAN: Eudico Linguistic Annotator, multimedia annotation tool developed by TLA at MPI
LEXUS: Web-based lexicon (dictionary) tool developed by TLA at MPI
TLA: The Language Archive
MPI: Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands
CAVA: human Communication Audio-Visual Archive for UCL
UCL: University College London