Towards a British Sign Language Corpus: A short report

Adam Schembri¹, Jordan Fenlon², Ramas Rentelis², Sally Reynolds² and Kearsy Cormier²

¹La Trobe University
²University College London

a.schembri@latrobe.edu.au

Abstract
This paper presents an overview of the British Sign Language Corpus Project - the first endeavor ever to create a machine-readable digital corpus of British Sign Language (BSL) collected from deaf signers across the United Kingdom. In the field of sign language studies, it represents a unique combination of methodology from variationist sociolinguistics and corpus linguistics. Unlike previous large-scale sign language sociolinguistic projects, the dataset is being annotated and tagged using ELAN software, given metadata descriptions, and the video data has been made accessible, with long-term efforts to make the dataset searchable on-line.

1.0 Introduction
The British Sign Language Corpus Project (BSLCP) was a project funded by the UK Economic and Social Research Council (ESRC, 2008-2011) that aimed to create a machine-readable corpus of spontaneous and elicited British Sign Language (BSL) digital video data collected from deaf native, near-native and fluent signers across the United Kingdom. Researchers at University College London led the project, working with co-investigators based at Bangor University (Wales), Heriot-Watt University (Scotland), Queens University Belfast (Northern Ireland) and the University of Bristol (England). In the field of sign language studies, the BSLCP represents a unique combination of methodology from variationist sociolinguistics, language documentation, and corpus linguistics. The project team conducted studies on sociolinguistic variation, language change and lexical frequency in BSL while simultaneously documenting BSL via creation of a corpus. Since June 2011, video data
have been available online and searchable via key metadata at http://www.bslcorpusproject.org/data/. As annotation is added to the online digital video dataset (including gloss-level and more detailed linguistic annotations as well as English translations via ELAN annotation files), we expect that it will become a standard reference and core data source for all researchers investigating BSL structure and use.

2.0 BSLCP Methodology

The studies of lexical frequency and sociolinguistic variation in BSL (and all future studies which will draw upon this corpus) required that we collect and analyse sign language data from a large and diverse sample of the British deaf community. In order to make cross-linguistic comparison possible, the methodology employed was similar to related studies undertaken on sociolinguistic variation in ASL (Lucas, Bayley, & Valli, 2001), Auslan (Schembri et al., 2009) and NZSL (McKee, McKee, & Major, 2011).

2.1 Sites

BSL is known to exhibit considerable regional variation, particularly in the lexicon (Deuchar, 1984; Kyle & Woll, 1985; Brien, 1992; Schembri et al., 2010), but in the absence of comprehensive documentation of the regional varieties of the language, it is not known exactly how many distinct regional dialects exist. We opted to collect data from eight major sites around the UK, representing each of the four countries in the union, and the major regions within England (the largest of the four countries in terms of population). The sites across England included one each from the South-East (London), South-West (Bristol), North-East (Newcastle), North-West (Manchester) and Midlands (Birmingham). These were combined with one site each in Northern Ireland (Belfast), Wales (Cardiff), and Scotland (Glasgow). Data were collected from these sites over a period of two years beginning in Birmingham in late 2008 and ending in London in early 2010.

2.2 Participants

In total, 249 deaf participants were filmed as part of the BSL Corpus Project. To ensure that each participant was representative of their region, we attempted to recruit participants who were lifelong residents, or had at least lived or worked in that region for ten years or more. We aimed to recruit only participants who reported learning BSL before the age of 7, although this criterion was relaxed in some cases. Our records show that 95% (n = 237) of participants are signers who reported learning to sign before the age of 7 and that all but one of the
remaining 12 reported that they learnt to sign by the age of 12. In total, native signers represented 31% (n = 76) of participants (we know that this is a higher proportion than in the adult deaf community as a whole, but it may or may not be representative of those who learned to sign before age 7). The participant sample was also roughly balanced for gender; of the 249 participants, 52% (n = 129) were women.

We also attempted to recruit participants from a wide range of age groups: 18-35 years (24%, n = 59), 36-50 years (25%, n = 62), 51-64 years (27%, n = 68), 65 years or older (24%, n = 60). The division of participants into these age groups is partly motivated by changes in language policy in deaf education during the twentieth century.

We classified participants into two broad social classes based on their occupation and/or educational background. Deaf individuals with a university education and/or ‘while collar’ professional occupations were categorised as middle class (38%, n = 95), whereas individuals with no university education and having traditional ‘blue-collar’ factory or trade-related occupations were classified as ‘working class’ (62%, n = 173).

As the overall British population was about 10% non-White (mostly of Afro-Caribbean and south Asian origin) according to the 2001 Census (the most recently available statistics for the UK at the time the project began), we attempted to recruit a similar proportion of non-White participants for our project. Out of a total of 20 participants (8% of all project participants) in the non-White category, we filmed 3 individuals in Birmingham, 2 in Bristol, 2 in Cardiff, 3 in Glasgow, 6 in London, 3 in Manchester and 1 in Newcastle. No non-White participants could be recruited in Belfast.

2.3 Data collection

In order to recruit sufficient numbers of deaf people in each city, we worked with a deaf fieldworker from each site who conducted participant recruitment. This methodology was first used by Ceil Lucas and colleagues in the American deaf community, modelled on a similar approach to participant recruitment used by Lesley Milroy in her studies of British English (Milroy, 1980; Lucas et al., 2001).

Participant pairs were filmed with one camera focused on each participant and a third focused on both members of the pair, as shown in Figure 1. We used two blue background screens in order to maximise our ability to code the subtleties of the sign language data such as specific hand configurations – pale colours in the background, for example, make it very difficult to see the handshapes of fair-skinned signers (the majority of our participants were White British deaf
people, as explained above). Two freestanding lights were used, one placed near each participant, but not in view of the video camera. Participants were seated in chairs without arms to prevent them from resting their elbows while signing as this interferes with sign language production, especially the production of signs on their usual locations. Furthermore, all participants were required to wear plain coloured clothing on their upper body. We brought plain dark t-shirts to filming sessions in case interviewees arrived for the filming session wearing something that did not meet our requirements. Again, coloured and patterned clothing can interfere with the ability to code subtle features of signs, such as finger configurations.

Upon arrival, participants were required to read an information sheet and sign a consent form (both explained to them in BSL by the fieldworkers) indicating their agreement to take part in the study and for the data collected to be subsequently available as part of the online BSL Corpus digital video archive (see http://www.bslcorpusproject.org/data). Once their consent was obtained, filming began following completion of a metadata questionnaire. This questionnaire consisted of 39 questions designed to provide a comprehensive background of each participant’s language experience and were designed to conform as much as possible with metadata standards for sign language corpora proposed in Crasborn and Hanke (2003). These questions aimed to elicit information ranging from language of preference, languages used at home and at school (inside and outside of the classroom) during childhood, as well as where they
lived in the UK prior to filming and the extent to which they interacted with the deaf community. Each participant was asked prior to his or her arrival to think of a short personal experience narrative to present (lasting no longer than five minutes) during the filming session. For the first task, they were instructed to retell this narrative to their partner. Following their narratives, participants were left to themselves to engage in a 30-minute conversation where they were free to talk about anything they wanted. After the conversation session, participants were then asked to participate in a 15-minute interview led by the fieldworker on language attitudes and awareness. Interview questions ranged from asking about definitions of ‘BSL’ and how it was different to English; to knowledge of variation and change related to region, age, gender, age of BSL acquisition and audiological status (i.e., hearing versus deaf signing); and to attitudes about BSL teaching, notions of ‘correct’ usage and BSL standardisation. Finally, participants took part in a lexical elicitation task in which they were asked to produce signs that they used for 102 concepts, chosen for their known or suspected high level of sociolinguistic variation.\(^1\) We made sure that each filming session took place in settings familiar to the participants, such as deaf social clubs and the offices of deaf organisations, to ensure that participants felt comfortable and that it was appropriate to use a relatively informal variety of BSL. No hearing people were involved in any of the data collection sessions.

2.4 Annotation and translation

All the annotations conducted under the BSL Corpus Project are specifically linked to research projects on phonological and lexical variation and lexical frequency. For the phonological variation study, this included annotations of 6330 signs from 211 signers from the conversation data (Fenlon et al., 2012). For the lexical variation study, gloss-based annotations were completed for 7332 signs from all 249 signers from the lexical elicitation data (Stamp et al., 2012, 2011). For the lexical frequency study (Cormier et al., 2011), a set of approximately 25,000 signs (500 signs each from 50 participants) from the conversational data was annotated with an English gloss for each individual sign: that is, one unique, identifying gloss or ‘ID gloss’, per sign (see Johnston, 2010). Although some of the annotations mentioned here were specifically tailored for the research study in question (e.g., the handshape categorisation used in the phonological variation should not be considered to reflect annotation standards at the phonological level for sign language corpora in general), we have generally adopted the approach set out in Johnston (2010) to ensure that the work conducted here takes us towards the desired goal of corpus machine
readability (namely the use of ID glossing when annotating a sign language corpus). An ID gloss is an English gloss that is consistently used with a unique sign to represent the sign in citation form along with all its phonological and morphological variants. Johnston (2010) also states that this process is made considerably easier if a comprehensive lexical database for the sign language in question exists. Unfortunately, prior to the BSL Corpus Project, no such lexical database existed. Available dictionaries for BSL (e.g., Brien 1992) do not follow lemmatisation practices (e.g., homonyms are often grouped together under a single entry and phonological variants of a lemma are often assigned separate entries). As a result, lemmatisation work, as part of the lexical frequency study, had to be carried out concurrently with annotation. This resulted in a lexical database of approximately 1800 unique signs which is being converted into an online dictionary called BSL Signbank (see Cormier et al., 2012 for an overview of this process) by the Deafness, Cognition, and Language Research Centre as part of its work plan between 2011 and 2016.

A written free translation is also required, together with ID glosses, in the early stages of sign language corpus building (Johnston, 2010). This is particular important since the very act of ID glossing eschews context-based glosses which have often been used to present examples in the sign language literature and may also be seen as a type of translation in itself. However, with ID glossing, a single English word is used to represent all instantiations of a lexeme regardless of its grammatical context (e.g., TEACH might be used whether the sign in question means ‘teacher’ or ‘teach’) or its particular meaning/sense (e.g., EXCITED will be used whether it means ‘excited’, ‘exciting’, ‘excitement’, ‘interested’, ‘interesting’, ‘motivated’ etc.). This means it is not possible to obtain a clear understanding of an utterance’s meaning by reading the ID glosses alone. Instead, one must refer to the written English translation together with the ID glosses to achieve this. An additional advantage of a written translation is that it is much faster to produce than ID glossing and will enable us to render a larger proportion of the corpus usable in a short space of time (e.g., people may search for particular topics by searching the written translation in ELAN for certain English words). As part of the BSL Corpus Project, initial translation work was completed for 23% of the entire dataset (specifically, for 70% of the personal narrative data and 36% of the interview data). We prioritised translation of the narrative and interview data over the conversational and lexical elicitation data because it would be more advantageous (i.e., the lexical elicitation data consists primarily of single signed responses to flashcards) and because these were both originally intended for open access and likely to reach a bigger audience (as discussed below, access to the conversation data is restricted
to researchers only). The process of, and issues involved in, translating the BSL Corpus is described in some detail in Pollitt et al. (2012).

All the annotation and translation work carried out under the BSL Corpus Project was conducted in ELAN (Wittenburg et al., 2006) and will be made publicly available online in the future. Future annotation work at the ID gloss level is planned but is likely to take some time to complete as the process cannot be automated because sign language recognition technology is still in its infancy and because we do not have yet have an adequate lexical database. A further possibility for annotation lies in the use of the corpus by other sign language researchers for their own purposes. In fact, one of the requirements that applicants must agree to before they are granted a user licence (as described below) is that they will share any annotation work they carry out with the BSL Corpus Project team.

3.0 A dual access corpus

As noted above, since summer 2011, the BSL Corpus Project video data have been available online via CAVA (human Communication Audio-Visual Archive), a secure system which allows viewing and downloading of the restricted corpus data (i.e., conversations and interviews) via a user license which includes a confidentiality agreement. CAVA also allows viewing and downloading of the open-access corpus data (i.e., narratives and lexical elicitation data) by anyone. The CAVA website is searchable via participant metadata and is aimed primarily at researchers (for the restricted data) and BSL teachers, students, interpreters or others (for the ability to download the open access data for closer study). In addition to the CAVA website, the BSL Corpus video data are also available to casual users via a user-friendly browse interface which allows users to select which subset of the open-access data they would like to view by narrowing their search beginning with region (any of the 8 regions may be chosen), then task (narratives or lexical elicitation may be chosen), then age group (where the age ranges 16-40, 41-65 and 65+ may be chosen). This casual interface was created separately from CAVA in order to encourage use of the corpus by the deaf community and has a simple, easy-to-use interface with instructions in both BSL and English. This follows current trends within language documentation to design language archives such that different audiences, including language communities themselves, can use, understand and enjoy them (Woodbury, 2011).

4.0 Conclusion

In this paper, we have briefly outlined the BSL Corpus Project, describing the data collection methodology and the creation of a dual access archive. Future
uses of the BSL Corpus Project are dependent on more annotation work being completed, in particular more ID glossing and translations. Much of this work will require long-term commitment and resources. We hope to include these annotations, as well as at least some of the current data with restricted accessibility, in the open-access collection in the future. Expanding the open-access archive will require that the researchers contact all participants to request additional consent for all or part of their conversation/interview data to be made public. When this has been achieved in future, a much greater proportion of the BSL Corpus video data will be available for use by researchers, as well as for BSL teachers, students, interpreters and the wider deaf community. In the meantime, the open access video data (specifically, the personal narratives and lexical elicitation data) are freely and openly available for anyone to use, including teachers, students and interpreters.

On-going work on the BSL Corpus will provide an important and enduring digital repository of contemporary BSL as a standard reference point for research into sign language structure and use, and it will also serve as a profoundly important cultural resource for the British deaf community.

Acknowledgements
We would like to thank the BSL Corpus Project co-investigators Margaret Deuchar, Frances Elton, Donall O Baoill, Rachel Sutton-Spence, Graham Turner and Bencie Woll as well as our project collaborator Trevor Johnston. This work was supported by the Economic and Social Research Council of Great Britain (RES-062-23-082, British Sign Language Corpus Project, and RES-620-28-6001/6002, Deafness, Cognition and Language Research Centre (DCAL)).

Annotation
1 For a list of the 102 concepts that were included in the lexical elicitation task and the list of questions asked during the interview, see http://www.bslcorpusproject.org/cava/activities/.

References
Towards a British Sign Language Corpus: A short report


Pollitt, Kyra et al. 2012. "Well, it's green here, but I've seen green and green, and my mother's was always green": Initial issues and insights from translating the BSL Corpus. In Jules Dickinson & Christopher Stone (ed.), Developing the Interpreter: Developing the Profession. Coleford, UK: Forest Books.


Schembri, Adam, Kearsy Cormier, Trevor Johnston, David McKee, Rachel McKee & Bencie Woll. 2010. Sociolinguistic variation in British, Australian


