LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 7 (2013), pp. 136-154
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/4592

Building the British Sign Language Corpus

Adam Schembri[(1)], Jordan Fenlon[(2)], Ramas Rentelis[(2)],
Sally Reynolds[(2)] and Kearsy Cormier[(2)]
(1) La Trobe University, (2) University College London

This paper presents an overview of the British Sign Language Corpus Project—the first endeavor 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. This means, however, that participants must consent to having the video data of their sign language use made public. This puts at risk the authenticity of the linguistic data collected, as signers may monitor their production more carefully than usual. We discuss our attempt to minimize this problem by creating a dual-access archive.

1. INTRODUCTION.[1] 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, 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. Thus, unlike previous large-scale sign language sociolinguistic projects, the dataset has been given metadata descriptions, and has been archived and made accessible online. Since June 2011, video data have been available online and searchable via key metadata.[2] 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),

[1] 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)).

[2] Available at http://www.bslcorpusproject.org/data/.

E-ISSN 1934-5275

we expect that it will become a standard reference and core data source for all researchers investigating BSL structure and use.

2. SIGN LANGUAGE DOCUMENTATION AND CORPUS LINGUISTICS. Over a decade ago, McEnery & Ostler (2000) argued that corpus linguistics needed to expand and include a wider set of languages in corpus-building activities, suggesting that the intellectual and moral imperative for doing so were clear. They recognized, however, that without conventional writing systems, sign language corpus building in particular remained a challenge. Since that time, a number of social and technological changes have made the emergence of sociolinguistically informed, corpus-based sign language documentation projects both possible and increasingly desirable. As we will show here, these changes are similar to those that have led to the rise of spoken language documentary linguistics outlined by Woodbury (2003).

First, there have been major advances in the technology of digital video data representation and maintenance. Developments in video camera technology, manipulation of digital video data, and computer storage capacity have led to digital video becoming more widely used by researchers in linguistics. Video annotation software, such as ELAN (Wittenburg et al. 2006.), has become available so that digital video files containing sign language data may be time-aligned with associated annotations. In a field in which widespread consensus on sign language notation and transcription systems has long been an issue and is still yet to be achieved (Johnston 1991, Miller 2001 van der Hulst & Channon 2010), the use of digital video annotation software has transformed research practices, and made open-access online archives of annotated sign language data possible for the first time.

Second, following many years in which much research efforts were focused on evidence for universal grammar, linguistics has begun to pay more attention to language diversity (e.g., Evans & Levinson 2009). Sign language can and should play a larger role in our understanding of language diversity, as this group of languages represents an interesting test case for linguistic theory (Cormier, Schembri & Woll 2010). Currently, there appears to be a premature consensus in the wider linguistics literature that many of the phenomena in sign language grammar can be fully understood within existing frameworks of grammatical description (e.g., Fromkin et al. 2006). As pointed out by Evans & Levinson (2009), the issue here is that some of the generalizations across signed and spoken languages require such a degree of abstraction away from the data that a clear understanding of unique features of linguistic diversity is lost. Work on 'classifier' constructions (Schembri 2003), 'agreement' verbs (Schembri & Cormier 2009) and pronominal forms (Cormier, Schembri & Woll in press) has raised questions about how much these aspects of sign language grammar share with equivalent spoken language systems. In order to advance our understanding of these, and other key aspects of sign language structure and use, more data is needed. Corpora of sign languages – machine-readable datasets of language recordings collected from large samples of signers – provides spontaneous, naturalistic data against which existing and future claims about the structure and use of specific sign languages can be tested.

Third, there is a growing recognition that many sign languages are endangered (Johnston 2004, Nonaka 2004). In developed nations, medical advances (such as vaccination for rubella), the closure of centralized schools for deaf children and improving hearing-aid and cochlear implant technology have changed both the demographics of deafness

and the transmission of sign languages (Johnston 2004, Knoors & Marschark 2012). Sign languages exist in unique sociolinguistic situations, as only 5-10% of the deaf community acquires them as first languages from deaf parents, with the majority of signers having acquired them (sometimes as delayed first languages) from deaf peers in schools for deaf children or in social networks in early adulthood. Therefore, whilst the BSL community in the UK continues to be strong and thriving with deaf parents continuing to transmit their language to their deaf and hearing children, this situation is likely to change in the near future as the impact of these social and technological developments are felt. Thus, sign language documentation projects provide an important means of recording sign languages as they are used today for future generations.

Finally, there is an increasing awareness that sign language documentation projects are vital to address concerns in deaf communities about the need for more language description to support sign language teaching, sign language interpreter training and the education of deaf children, as we discuss in the next section.

3. BSLCP RESEARCH QUESTIONS. In order to exemplify the kinds of research questions that can be explored using a corpus-based approach to the study of BSL, the BSL Corpus Project included two studies on sociolinguistic variation (specifically phonological and lexical variation and change) and one on lexical frequency (Cormier, Fenlon, Rentelis & Schembri 2011; Fenlon et al. 2013; Stamp, Schembri, Fenlon, Rentelis & Cormier 2011). In the phonological variation study, the aim was to investigate how variation in hand configuration and palm orientation in signs produced with the 1 handshape correlates with linguistic factors (e.g., the preceding or following segment in the case of phonological variable) and social factors (e.g., the signer's region, gender, age, language background, socio-economic class and ethnicity). In the lexical variation and change study, we aimed to document the extensive variation in the BSL lexicon, and correlate this variation with social factors similar to the set used in the phonological variation study.

The motivation for the focus on sociolinguistic variation in the BSL Corpus Project is that it responds to one of the main issues raised above – namely, that linguists need to work more closely with communities and speakers of languages that are the focus of documentation projects because, in the case of endangered languages, the communities are naturally concerned about the future of their languages. This is true for BSL (currently undergoing a period of rapid change and increasing contact with speakers of English) as much as any sign language (e.g., Buxton 2011).

Like many Western sign languages, BSL is a national sign language, used by a comparatively large signing community (or 'macro-community'; see Schembri 2010), but it appears to have emerged historically out of a collection of micro-community sign languages. In the nineteenth century, there were 22 residential schools for deaf children in different parts of the UK (Kyle & Woll 1985) but, given the lack of social mobility at the time, deaf people did not travel and had few opportunities to meet deaf people from other regions. The UK is a relatively small country, but the result of this social situation was the emergence of quite distinct varieties of BSL in each of these local deaf micro-communities which developed around the residential schools. Only since the late twentieth century have we seen an emergence of a national deaf identity and a sign language macro-community in the UK. Some signers in Bristol, for example, reported difficulty understanding sign-

ers from other parts of the country as recently as 1980 (Kyle & Allsop 1982). Significant regional lexical differences across the country still exist, but some variants appear only to be used by older generations of British signers. Many in the British deaf community are very concerned about documenting and preserving the rich variation before it disappears, and recent research does provide some support for the idea that there appears to be dialect leveling occurring (Stamp et al. under review).

The motivation for studying lexical frequency in BSL was similarly to address concerns in deaf communities about the need for more language description to support stakeholders such as sign language teachers. Spoken language corpora have long been used to generate word frequency lists, which are, in turn, used to design learner dictionaries and inform teaching curricula (e.g., Collins COBUILD dictionary, Longman Dictionary of Contemporary English). For signed languages, frequency studies are rare with only three studies having been conducted: one on American Sign Language (Morford & MacFarlane 2003), another on New Zealand Sign Language (McKee & Kennedy 1999, 2006) and the most recent on Auslan (Australian Sign Language) (Johnston 2011). Of these three studies, only the Auslan frequency study is based on a machine-readable corpus. Before the BSL Corpus Project, no frequency data was available for BSL. Therefore, the BSL lexical frequency study (Cormier et al. 2011) demonstrates an enormous benefit of the corpus to the BSL teaching community as well as the wider research community.[3] With sign language corpora being created around the world, it is expected that this will pave the way for further frequency studies, although they are likely to be based on smaller datasets compared to spoken language studies due to technological limitations and the lack of (for nearly all signed languages) a comprehensive lexical database (see the section on annotation below).

3.1. 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 analyze 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 et al. 2001), Auslan and NZSL (McKee, McKee & Major 2011). Drawing on a previously implemented research design also allowed us to adopt elements of these approaches that have been successful in past projects, as well as to identify potential weaknesses in the project methodology early on and address them before they became problematic.

In corpus building, the aim is generally to select a representative sample of language use from a particular language community (e.g., McEnery & Wilson 1996). Creating a representative sample of sign language users has also been the stated aim of previous large-

[3] Sign language researchers (e.g., Vinson, Cormier, Denmark, Schembri & Vigliocco 2008) have attempted to address this gap in research by collecting subjective familiarity ratings of 300 lexical signs from native signers and thus exploiting a reported correlation between subjective familiarity ratings and corpus-based frequency lists in spoken languages (e.g., Balota, Pilotti & Cortesse 2001). These ratings have then been used to inform experimental design in psycholinguistic studies. However, Cormier et al. (2011) caution against using subjective familiarity ratings as a replacement for frequency in sign language studies, because it is unclear whether familiarity and frequency are as correlated for sign languages as has been reported for spoken languages.

scale studies into sociolinguistic variation and change in sign languages (e.g., Lucas et al. 2001). Strictly speaking, however, selecting a truly representative sample of the British deaf community is not possible. Not enough is known about the population of deaf sign language users in the United Kingdom in order to recruit a representative group whose characteristics we can confidently say reflect those of the larger population. For example, published estimates as to the number of deaf individuals who use BSL vary (Schembri et al. 2010), and we know little about the distribution of deaf community members across the various towns and cities of the UK. Like previous sociolinguistic studies, we were interested in a number of demographic variables, including gender, age, region, ethnicity, socioeconomic class and age of sign language acquisition, so we employed a quota sample to ensure that an appropriate distribution of these key social factors could be found in our dataset. As explained below, participants were recruited according to these variables, with recruitment of participants with specific demographic characteristics stopping when the quota was filled.

Creating a truly representative sample of sign language situational varieties is similarly difficult. Little is known about register variation in sign languages (see, for example, Johnston & Schembri 2007). In addition, there are many constraints on filming to consider. As explained below, signers need to be seated in front of a plain backdrop facing a camera, appropriately dressed and with appropriate lighting to maximize the usability of the data collected. Not considering these external factors related to filming could easily compromise the integrity of the data. We focused our data collection on four situational varieties in BSL: personal experience anecdotes, spontaneous conversations, structured interviews, and a word list. These data types were the bare minimum required for our studies into sociolinguistic variation and change, but we hope that future work on the corpus will expand the sample to include a wider range of text types in BSL.

3.1.1 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 exhaustive documentation of the regional varieties of the language, it is not known exactly how many distinct regional dialects exist. In order to undertake our studies into sociolinguistic variation and change in BSL, we needed a sample of at least 30 individuals from each region, so that sufficient numbers of participants with a mix of demographic characteristics could be included. Given the time and resources available for the project, 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). We believed that each site included a relatively large deaf community because community centers for deaf people existed in each city, and because they represented some of the largest urban centers in the UK. The continuing or prior existence of centralized deaf schools in each of these 8 cities provided further motivation for selection. The preference for collecting data from larger British cities with thriving deaf communities was strategic, since it was easier to recruit a sufficient number of signers from a variety of backgrounds to ensure that our quota sampling targets

could be met (i.e., so that our data could be analyzed for the influence of the major social factors of region, gender, age, age of sign language acquisition, socioeconomic class and ethnicity on variation). 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.

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. Therefore, those who had recently moved to the city in question were not invited to take part regardless of the degree to which they mixed with the local deaf community since relocating there. Each participant considered the local deaf community to be an integral part of their day-to-day lives, having grown up in that community or having interacted with that community for a substantial part of their lives.

3.1.2 PARTICIPANTS. In total, 249 participants were filmed as part of the BSL Corpus Project. We aimed to recruit only participants who reported learning BSL before the age of 7, although this criterion was relaxed in some cases, because some flexibility was required when our quotas could not be filled (many people we approached did not wish to participate, most often because of a concern about being filmed). 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. We also made sure that we recruited a sufficient number of native signers (i.e., signers who learnt to sign from birth from deaf parent(s) or an elder sibling) within each region. 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). We did not recruit hearing signers (native or otherwise) and we tried to exclude deaf people who learnt to sign later than age 12. Research has shown that the age at which a child is exposed to sign language as a first language has a considerable effect on their sign language proficiency in adulthood (Cormier, Schembri et al. 2012, Emmorey 2002). 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. Due to variable patterns of language transmission within the deaf community (e.g., as mentioned above, the language is often learned from peers in schools for deaf children, or after school as a young adult in other settings), clear differences can be seen between older and younger signers (for example, in the use of lexical items, see Sutton-Spence, Woll & Allsop 1990; Stamp et al. under review). Recruitment to the BSL Corpus project was designed to reflect this variation by ensuring that participant selection was balanced across four 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. Participants in the oldest age group (65 years and older) were most likely to have been educated in residential schools for deaf children, sometimes with approaches that emphasized the use of fingerspelling (the use of specific signs representing each of the 26 letters of the alphabet to spell out English words) and/or the development of speech-reading skills. BSL may, however, have been used by school children with each other in the dormitories and in the playground. Like the older group, participants in the 51-64 years of age group would

have been educated in centralized schools for deaf children, although many would have experienced the strong shift towards the use of speech skills and residual hearing that occurred in a number of schools following the Second World War. Those in the 31-50 years of age category would have witnessed further major changes in deaf education: the greater use of hearing aids as assistive technology improved, the move towards the use of sign supported English (a form of signed communication simultaneously produced with spoken English and in which signs follow English syntactic patterns) as a language of instruction, the closure of centralized schools for deaf children and the spread of mainstreaming. Participants in the youngest group (18-30 years of age) have seen the increasing recognition of BSL as the language of the British deaf community, but most would have been educated in mainstream school settings with few, if any, deaf peers, and with sign language interpreting of classroom instruction in spoken English by hearing teachers. Some of the members of this group would have been educated in schools using BSL as the medium of instruction (some schools have introduced bilingual approaches using BSL together with English since the 1980s).

Due to the demographics of the British deaf community, it was not possible to achieve as equal a balance of various social classes or ethnicities, but a mix of ethnic groups was attempted. We classified participants into two broad social classes based on their occupation and/or educational background. Deaf individuals with a university education and/or 'white collar' professional occupations were categorized 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). It must be pointed out, however, that the emergence of a professional class in the deaf community is a relatively recent phenomenon, following the greater access to university education through increasing provision of sign language interpreting and note-takers that began in the 1980s and which was further assisted by the passing of the Disability Discrimination Act in 1995. As such, it is doubtful there has been time for clear linguistic differentiation between socioeconomic classes to emerge.

The ethnic composition of the British deaf community is unknown. 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. This proved difficult, however, as few non-White signers met some of our other recruitment criteria, and very few Black and south Asian deaf people could be found in some of the smaller cities in the UK. Out of a total of 20 participants (8% of all project participants) in the non-White category, we filmed three individuals in Birmingham, two in Bristol, two in Cardiff, three in Glasgow, six in London, three in Manchester and one in Newcastle. No non-White participants could be recruited in Belfast.

3.1.3 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, modeled on a similar approach to participant recruitment used by Lesley Milroy in her studies of British English (Milroy 1980, Lucas et al. 2001). All fieldworkers were deaf and all but one were native signers. Each had lived much or all of their lives in their

local deaf community. Each fieldworker was trained in our recruitment and data collection procedures before they started recruiting participants for the project. With the participant recruitment criteria in mind, they approached suitable members of their local deaf community and invited them to take part in the project. They also asked these participants to suggest others to approach about participation. In some cases, participants were recruited using flyers distributed at deaf social events, or by means of notices on deaf community center noticeboards. All fieldworkers were present on the day of filming and assisted in data collection together with a deaf project researcher. They also led on some tasks given to participants (i.e., the interviews and lexical elicitation tasks). As the two university researchers involved in filming were also deaf signers, in almost all cases, no hearing people were present during filming. The research team believed that, as has been demonstrated for ASL (Lucas & Valli 1992), some BSL signers may produce contact varieties of sign language (i.e., varieties reflecting relatively more English influence, or perhaps with greater code-switching between English and BSL) when in the presence of hearing signers. In fact, having the fieldworkers lead on the tasks given to participants also ensured that the two deaf university researchers never directly took part in any of the linguistic data collected and kept further language contact influences between deaf signers from different regions to a minimum. For example, the Belfast participants often mentioned that, when visiting other parts of the UK, they tend to use signs more widely recognized within the UK because the variety of signing that they typically used in Northern Ireland was often not understood by signers living in England, Scotland and Wales. As a result, this conscious monitoring had almost become second nature to them and they confessed to making such adjustments when meeting the researchers prior to filming. By having a local fieldworker lead on the tasks, this ensured language contact influences (i.e., with the London-based university researchers) were kept to a minimum. Participants were always filmed in pairs and, where possible, were filmed with another person within, or close to, their age group. We also ensured that we filmed a mixture of mixed-sex and same-sex pairs (32 male/male, 37 female/female, and 56 mixed). As participants were required to engage in conversation for 30 minutes, we tried to make sure that participants were familiar with their partner (i.e., were friends or acquaintances) so that they felt comfortable with one another and any awkwardness could be avoided. In fact, one individual expressed a desire to be filmed again, after being filmed with a person she did not know well and with whom she did not feel particularly comfortable. As a result, someone better known to her was recruited and she was filmed again with this different individual. We also generally avoided filming together individuals who were in a long-term relationship (particularly husbands and wives), because often their conversational data was not as natural as between pairs of friends. The high level of familiarity between couples, particularly retired married couples, sometimes meant that there was less to say to each other than between friends who did not see each other on a daily basis.

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 maximize our ability to code the subtleties of the sign language data such as specific hand configurations – pale colors 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 – observation suggests that people tend not to hit target locations that are higher on the body when they are resting their elbows, for example. Furthermore, all participants were required to wear plain colored clothing on their upper body. We brought plain dark t-shirts to filming sessions in the event that interviewees arrived for the filming session wearing something that did not meet our requirements. Again, colored and patterned clothing can interfere with the ability to code subtle features of signs, such as finger configurations.

FIGURE 1. Screenshots from BSL Corpus Project video data <pair view and individual view> (L36+L37c and L36c)

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 for the data collected to be subsequently available as part of the online BSL Corpus digital video archive.[4] Once their consent was obtained, filming began following completion of a metadata questionnaire. This questionnaire consisted of 39 questions designed to provide a comprehensive overview 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 & 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), 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. In many cases, however, participants were either not asked by the fieldworker to prepare a narrative, or failed to reflect on this request prior to their arrival, and thus produced a more-or-less spontaneous narrative on the spot. This task was initially intended as a kind of warm-up activity so that both participants could become accustomed to the setting in which filming took place, with background screens/curtains behind in each participant, lighting equipment and cameras. All of the data produced in this warm-up session by participants was, however, suitable for inclusion in the corpus and was thus added

[4] See http://www.bslcorpusproject.org/data

to the collection. 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. Participants were reassured that the conversational data obtained from this part of the session would not be made publicly available on the Internet, but would only be shared with other university researchers in an attempt to lessen the effects of the observer's paradox (Labov 1972) which we discuss below. This was necessary so that participants would relax and converse freely and to ensure that the data collected was as close to the vernacular variety as possible. 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 standardization. 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.[5] We made sure that most filming sessions took place in settings familiar to the participants, such as deaf social clubs and the offices of deaf organizations, to ensure that participants felt comfortable and that it was appropriate to use a relatively informal variety of BSL.

3.1.4 ANNOTATION AND TRANSLATION. All the annotations conducted under the BSL Corpus Project are specifically linked to the planned research projects on phonological and lexical variation and lexical frequency. For the phonological variation study, this included annotations of 6330 signs from 211 signers in the conversational dataset. All tokens were annotated for handshape and orientation using a simplified coding system and 2110 of these tokens contain further annotation, such as the token's grammatical class and whether it was articulated using one or two hands (see Fenlon et al. 2013 for a detailed outline of this coding system). For the lexical variation study, gloss-based annotations were completed for 7332 signs from all 249 signers from the lexical elicitation data. For the lexical frequency study, 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 (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 categorization 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 identifying, or ID, glosses when annotating a sign language corpus). Johnston (2010) notes that two types of annotations are minimally required to achieve a machine-readable corpus: ID glossing and a written free translation. An ID gloss, as mentioned previously, 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

[5] 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/.

variants. Johnston (2010) also explains that this process is made considerably simpler if a comprehensive lexical database for the sign language in question exists. Unfortunately, prior to the BSL Corpus Project, no such lexical database was available. Existing dictionaries of BSL (e.g., Brien 1992) do not follow lemmatization 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, lemmatization 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, Fenlon 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. Like the related Auslan documentation work (Johnston 2010), this will result in one of the first primarily corpus-based dictionaries for any sign language and its link with the BSL Corpus means that, as well as information on its phonological and morphological structure, each entry can also be put into a sociolinguistic context (e.g., a given lexical or phonological variant can be quantitatively associated with a specific region, age group or other social factors in the corpus).

A written free translation is also required, together with ID glosses, in the early stages of sign language corpus building (Johnston 2010). This is particularly 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', 'interested', 'interesting', 'motivated', 'eager', 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). 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 prioritized 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 inclusion in the open access component of the corpus and thus likely to reach a larger audience (as discussed below, access to the conversational 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. (2010).

All the annotation and translation work carried out under the BSL Corpus Project was conducted in ELAN 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 license (as described below) is that they will share any annotation work they carry out with the BSL Corpus Project team. Although this represents an ideal economy of effort, it is important that widely agreed annotation standards are put in place prior to annotation work and that these standards are reviewed at regular intervals. This is to ensure that any annotations carried out (at whatever level in the language) are consistent with current practices in sign language research and can be maximally exploited by future researchers. As of 2011, we have adapted the Auslan Corpus annotation standards (Johnston 2011) for our purposes, with the intention of reviewing these standards before any further annotation work is conducted.

4. OBSERVER'S PARADOX AND AUDIENCE DESIGN. Although still relatively small in terms of annotations, the BSL Corpus dataset is already one of the largest sign language corpora in the world, involving annotated data from all 249 participants, with approximately 40,000 lexical items annotated in total. As such, the corpus lends itself naturally to large-scale investigations into sociolinguistic variation and change in BSL. Unlike previous large-scale sociolinguistic projects (e.g., Schembri et al. 2009), the dataset is archived and available on-line, and is searchable via metadata. Additionally, as annotations are also made available online in the future, the dataset will then become machine-readable and searchable via content. This means, however, that participants must consent to having the video data of their sign language use made public. This seems to put at risk the authenticity of the data collected, as signers may monitor their production more carefully than might otherwise occur. As Tagliamonte (2006) explains, a specific aim of variationist sociolinguistics is to study the vernacular variety of a community's language. Labov (1972) defines this as the variety adopted by speakers when they are monitoring their style least closely. The focus on the vernacular reflects the belief among sociolinguists that it is the most systematic variety, as it is assumed to be the variety that was acquired first and is the most free from self-conscious style-shifting or hypercorrection (Tagliamonte 2006).

This situation is particularly complex in bilingual communities, such as the British deaf community, in which all members have varying degrees of fluency in two languages: BSL and English. In Figure 2 below, we present a cline of language mixing varieties used in the British deaf community. Each of these varieties represents an abstraction, but these categories are similar to widely used categories in the sign language literature (e.g., Sutton-Spence & Woll 1999, Lucas & Valli 1992). Unlike other work, however, we have added 'BSL self-conscious style' at the left end of the continuum for varieties in which deaf bilinguals consciously reduce their English code-mixing to the best of their ability, either to achieve particular linguistic effects, as in sign language poetry, or to reflect attitudes related to linguistic purism and prescriptivism (cf. Wertheim 2006 on Tartar and Russian code-mixing). Observation of sign language use in the British deaf community suggests that 'vernacular BSL' and 'contact signing' are perhaps the varieties most commonly used in most informal situations by deaf bilinguals, although 'sign supported English' in which English is clearly the matrix language also appears common. The aim of the BSL Corpus Project is to collect as much data in vernacular BSL varieties as possible; this is the reason for the methodology described above (no hearing people present, filming only native

signers and early learners of BSL, pairing up individuals who are comfortable being filmed together, and so on).

STYLE	BSL self-conscious style	Vernacular BSL	Contact signing	Sign supported English	Spoken English
CHARAC-TERISTICS	Little or no English influence	English code-mixing, but BSL is the dominant language	Code-mixing between English and BSL with neither as dominant language	Signs are produced alongside spoken or mouthed English	Little or no BSL influence

FIGURE 2. A cline of language mixing varieties in the British deaf community

Although the (ideal) aim was to observe how deaf people use BSL with each other when they are not being observed, participants were in fact filmed using three cameras, and with lighting equipment, after they had filled in consent forms that made them fully aware that the aim of the project was to create an open-access on-line BSL corpus using the data collected. Participants were not only aware that their signing would be seen by researchers, but also potentially by anyone with a computer that has access to the Internet. In this way, combining sociolinguistic methodology with an objective to build an open-access corpus creates a unique form of the observer's paradox. Wertheim (2006) has proposed that we can use the work of Allen Bell's notion of 'audience design' to better understand the challenge posed by the observer's paradox for language documentation and corpus linguistics. Bell (1984:187) claims "...at all levels of language variability, people are responding to other people. Speakers are designing their style for their audience". He identifies five 'participant roles' for any linguistic interaction: (1) the speaker (the person who is speaking), (2) addressee (the person or people addressed by the speaker), (3) auditor (the person or people present and referred to, but not addressed by the speaker), (4) overhearer (the person or people known to be listening, but neither addressed nor referred to), and (5) eavesdropper (not known to be listening, and neither addressed nor referred to). Bell (1984) proposes that each of these roles can have an effect on the design of the linguistic interaction by the participants, with (1) having the most effect on the variety produced and (5) the least effect. To adapt the notion of audience design to sign language recordings, the unusual situation of being video-recorded for language documentation purposes may make attempts to design one's language output with an unusual type of participant role (4) in mind more salient in the data. Sign language documentation project participants do not know who the end users of the collection will be, but they know that the recordings of their signing will be watched, so how this fact impacts on their sign language production is unknown.

Some sociolinguists, however, argue that the concept of the vernacular represents an abstraction, claiming that all varieties of speech vary considerably in response to situational contexts. As such, "…the concept of an entirely natural speech event (or an entirely unnatural one) is untenable" (Milroy & Gordon 2003:50). Despite this, it is important to consider how best to adapt the variety of techniques sociolinguists use to overcome the observer's paradox, or at least to reduce its effects. One of the ways that we have tried to

overcome these concerns is to exclude the 30-minute conversations from the planned open-access archive. As explained above, participants were informed before data collection that conversation data would not be part of the open-access archive, and that researchers who wish to see it would have to fill in an online registration form that includes a confidentiality agreement. Thus the intent was to have the conversational data restricted for use by academic researchers only but to make all other subsets of data (including the narrative, lexical elicitation and interview data) openly accessible to anyone.

One issue that arose after the data were already collected (which was only discovered by the researchers, who were not presenting during filming, as annotation and translation of the data began) was that in the interview task, some participants referred to specific people and groups within the BSL-using community by name. For example, one of the questions in the sociolinguistic interview attempted to elicit responses from the interviewees about their attitudes in relation to 'good' or 'beautiful' BSL usage. Some interviewees exemplified 'good' BSL usage by naming individuals they felt to be 'good' (and at times, 'bad') signers. As a consequence, we decided to restrict all of the interview data in the same way that the conversation data is restricted—i.e., such that the interviews are viewable only by registered researchers upon signing a confidentiality agreement.

While we do not yet know the degree to which our approach to data collection has been successful in avoiding style shifting towards more English-like varieties of BSL, results of our phonological variation study (Fenlon et al. 2013) suggests that our data are not too dissimilar to data collected under different circumstances by American colleagues as part of their sociolinguistic variation project on ASL (Lucas et al. 2001). This is despite the fact that deaf participants in the ASL study were filmed in larger groups of two to six individuals, and none of the data were intended to be part of an online, open-access collection. Additionally, as explained above, we filmed our participants in front of a plain, blue screen and with studio lights to ensure the quality of the video data collected. This created a considerably less informal setting than the set-up in the ASL study where no screens or lighting were used. These combined factors mean that participants in the BSL corpus might have adopted a more formal style of BSL than the ASL used in the comparable American study. However, when we compare the ASL and BSL phonological variation studies, this does not appear to be the case. Lucas et al.'s (2001) analysis of 5195 signs that were considered to have the 1 handshape as its citation form revealed that 3128 tokens (60%) were realized as a different handshape. Although this is slightly higher than the 54% (n=1125) reported in Fenlon et al. (2013), the similar rate of variation between the two studies suggests that observer effects on the data collected in the two studies are very similar despite differences in design. Both approaches produced datasets that included significant proportions of non-citation forms of the 1 handshape—variation that is associated with more informal varieties of ASL and BSL.

5. A DUAL ACCESS CORPUS. As noted above, since summer 2011, the BSL Corpus Project video data have been available online via CAVA (the University College London "Human Communication Audio-Visual Archive"), a secure system that 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 with an Internet connection. 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 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).

6. CONCLUSION. In this paper, we have outlined the BSL Corpus Project, describing the data collection methodology and the creation of a dual access archive, as well as discussing the implications of this particular research design for linguistics research. 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 the 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 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.

References

Balota, David A., Maura Pilotti & Michael J. Cortesse. 2001. Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition* 29. 639–647.

Bell, Allan. 1984. Language style as audience design. *Language in Society* 13. 145–204.

Brien, David. 1992. *Dictionary of British Sign Language/English*. London: Faber & Faber.

Buxton, David. 2011. *Empowerment - not dependency!* Plenary presentation at the Annual British Deaf Association Conference. Belfast, 25–27 November, 2011.

Cormier, Kearsy, Adam Schembri & Bencie Woll. 2010. Diversity across sign languages and spoken languages – Implications for language universals (A response to Evans & Levinson). *Lingua* 120(12). 2664–2667.

Cormier, Kearsy, Adam Schembri & Bencie Woll. in press. Pronouns and pointing in sign languages. *Lingua*.

Cormier, Kearsy, Adam Schembri, David Vinson & Eleni Orfanidou. 2012. First language acquisition differs from second language acquisition in prelingually deaf signers: Evidence from sensitivity to grammatical judgement in British Sign Language. *Cognition* 124(1). 50–65.

Cormier, Kearsy, Jordan Fenlon, Ramas Rentelis & Adam Schembri. 2011. Lexical frequency in British Sign Language conversation: A corpus-based approach. In Peter K. Austin, Oliver Bond, Lutz Marten & David Nathan (eds.), *Proceedings of the Conference on Language Documentation and Linguistic Theory 3*, 81–90. London: School of Oriental and African Studies.

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, Eleni Efthimiou, Eleni Fotinea, Thomas Hanke, Jette Kristoffersen & Johanna Mesch (eds.), *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey*, 7–12. Paris: European Language Resources Association.

Crasborn, Onno & Thomas Hanke. 2003. *Additions to the IMDI metadata set for sign language corpora*. Paper presented at the ECHO workshop. Nijmegen University, The Netherlands, 8–9 May, 2003.

Deuchar, Margaret. 1984. *British Sign Language*. London: Routledge & Kegan Paul.

Emmorey, Karen D. 2002. *Language, cognition, and the brain: Insights from sign language research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Evans, Nicholas & Stephen Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioural and Brain Sciences* 32. 429–448.

Fenlon, Jordan, Adam Schembri, Ramas Rentelis & Kearsy Cormier. 2013. Variation in handshape and orientation in British Sign Language: The case of the '1' hand configuration. *Language and Communication* 33. 69–91.

Fromkin, Victoria, Robert Rodman & Nina Hyams. 2006. *An introduction to language*. Boston, MA: Thomson Wadsworth.

Hulst, Harry van der & Rachel Channon. 2010. Notation systems. In Diane Brentari (ed.), *Sign Languages*, 151–172. Cambridge: Cambridge University Press.

Johnston, Trevor. 1991. Transcription and glossing of sign language texts: Examples from Auslan (Australian Sign Language). *International Journal of Sign Linguistics* 2(1). 3–28.

Johnston, Trevor. 2004. W(h)ither the deaf community? Population, genetics, and the future of Australian Sign Language. *American Annals of Deaf* 148(5). 358–375.

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.

Johnston, Trevor. 2011. *Auslan Corpus Annotation Guidelines*. http://www.auslan.org.au/video/upload/attachments/AuslanCorpusAnnotationGuidelines30November2011.pdf. (30 November, 2011.)

Johnston, Trevor. 2012. Lexical frequency in sign languages. *Journal of Deaf Studies and Education* 17(2). 163–193.

Johnston, Trevor & Adam Schembri. 2007. *Australian Sign Language (Auslan): An introduction to sign language linguistics*. Cambridge: Cambridge University Press.

Knoors, Harry & Marc Marschark. 2012. Language planning for the 21st Century: Revisiting bilingual language policy for deaf children. *Journal of Deaf Studies and Deaf Education* 17(3). 291–305.

Kyle, James G. & Bencie Woll. 1985. *Sign language: The study of deaf people and their language*. Cambridge: Cambridge University Press.

Kyle, James & Lorna Allsop. 1982. *Deaf people and the community*. Bristol: University of Bristol, Centre for Deaf Studies.

Labov, William. 1972. *Language in the inner city: Studies in the Black English Vernacular*. Philadelphia, PA: University of Pennsylvania Press.

Lucas, Ceil & Clayton Valli. 1992. *Language contact in the American Deaf Community*. San Diego, CA: Academic Press.

Lucas, Ceil, Robert Bayley & Clayton Valli. 2001. *Sociolinguistic variation in American Sign Language*. Washington, DC: Gallaudet University Press.

McEnery, Tony & Nicholas Ostler. 2000. A new agenda for corpus linguistics–Working with all the world's languages. *Literary and Linguistic Computing* 15(4). 403–419.

McKee, David & Graeme Kennedy. 1999. A list of 1,000 frequently-used signs in New Zealand Sign Language. In Graeme Kennedy (ed.), *New Zealand Sign Language: Distribution, Origins, Reference*, 17–25 Wellington: Occasional Publication 2, Deaf Studies Research Unit, Victoria University.

McKee, David & Graeme Kennedy. 2006. The distribution of signs in New Zealand Sign Language. *Sign Language Studies* 6(4). 372–390.

McKee, David, Rachel McKee & George Major. 2011. Numeral variation in New Zealand Sign Language. *Sign Language Studies* 11(5). 72–97.

Miller, Christopher. 2001. Some reflections on the need for a common sign notation. *Sign language and Linguistics* 4(1/2). 11–28.

Milroy, Lesley. 1980. *Language and social networks*. Oxford: Blackwell.

Milroy, Lesley & Matthew Gordon. 2003. *Sociolinguistics: Method and interpretation*. Oxford: Blackwell.

Morford, Jill & James MacFarlane. 2003. Frequency characteristics of American Sign Language. *Sign Language Studies* 3(2). 213–225.

Nonaka, Angela M. 2004. The forgotten endangered languages: Lessons on the importance of remembering from Thailand's Ban Khor Sign Language. *Language in Society* 33. 737–767.

Pollitt, Kyra, Janet Beck, Helen Dunipace, Sue Lee, Cathryn McShane, Elvire Roberts, Sherratt Rowan, Robert Skinner, Adam Schembri & Graham H. Turner. 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 (eds.), *Developing the interpreter; Developing the profession*, 29–43. Coleford, UK: Forest Books.

Schembri, Adam. 2003. Rethinking 'classifiers' in sign languages. In Karen D. Emmorey (ed.), *Perspectives on classifier constructions in sign languages*, 3–34. Mahwah, NJ: Lawrence Erlbaum Associates.

Schembri, Adam. 2010. Documenting sign languages. In Peter K. Austin (ed.), *Language documentation and description volume 7: Lectures in language documentation and description*, 105–143). London: School of Oriental and African Studies.

Schembri, Adam, David McKee, Rachel McKee, Trevor Johnston, Della Goswell & Sara Pivac. 2009. Phonological variation and change in Australian and New Zealand Sign Languages: The location variable. *Language Variation and Change* 21(2). 193–231.

Schembri, Adam & Kearsy Cormier. 2009. *No agreement on agreement: Are we missing the point?* Paper presented at the workshop 'From Gesture to Sign: Pointing in Spoken and Signed Languages'. University of Lille, France, 4–5 June 2009.

Schembri, Adam, Kearsy Cormier, Trevor Johnston, David McKee, Rachel McKee & Bencie Woll. 2010. Sociolinguistic variation in British, Australian and New Zealand sign languages. In Diane Brentari (ed.), *Sign Languages*, 479–501. Cambridge: Cambridge University Press.

Schembri, Adam & Trevor Johnston. 2007. Sociolinguistic variation in the use of fingerspelling in Australian Sign Language (Auslan): A pilot study. *Sign Language Studies* 7(3). 319–347.

Schembri, Adam, Trevor Johnston & Della Goswell. 2006. NAME dropping: Location variation in Australian Sign Language. In Ceil Lucas (ed.), *Multilingualism and Sign Languages: From the Great Plains to Australia* (Vol. 12), 121–156. Washington, DC: Gallaudet University Press.

Stamp, Rosemary, Adam Schembri, Jordan Fenlon & Ramas Rentelis. Under review. 2012. Variation and change in British Sign Language number signs. *Journal of Sociolinguistics*.

Stamp, Rosemary, Adam Schembri, Jordan Fenlon, Ramas Rentelis & Kearsy Cormier. 2011. *Lexical variation and change in British Sign Language (BSL): Evidence for dialect levelling?* Paper presented at the Sixth International Conference on Language Variation in Europe (ICLaVE). Freiburg Institute for Advanced Studies, Germany, 29 June–1 July, 2011.

Sutton-Spence, Rachel, Bencie Woll & Lorna Allsop. 1990. Variation and recent change in fingerspelling in British Sign Language. *Language Variation and Change* 2. 313–330.

Tagliamonte, Sali. 2006. *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.

Vinson, David P., Kearsy Cormier, Tanya Denmark, Adam Schembri & Gabriella Vigliocco. 2008. The British Sign Language norms for acquisition, familiarity and iconicity. *Behaviour Research Methods* 40(4). 1079–1087.

Wertheim, Suzanne. 2006. Cleaning up for company: Using participant roles to understand fieldworker effect. *Language in Society* 35. 701–727.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann & Han Sloetjes. 2006. ELAN: a Professional Framework for Multimodality Research. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, 1556–1559.

Woodbury, Tony. 2003. Defining documentary linguistics. In Peter K. Austin (ed.), Language documentation and description, vol. 1, 35–51. London: School of Oriental and African Studies.

Woodbury, Tony. 2011. Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. In David Nathan (ed.), *Proceedings of Endangered Languages Achive Workshop on Language Documentation and Archiving*, 11–20). London: School of Oriental and African Studies.

Adam Schembri
a.schembri@latrobe.edu.au