British Sign Language Corpus Project: Open Access Archives and the Observer's Paradox

Adam Schembri
Deafness Cognition and Language Research Centre, University College London
49 Gordon Square, London, WC1H 0PD, United Kingdom
E-mail: a.schembri@ucl.ac.uk

Abstract

The British Sign Language Corpus Project is a new three-year project (2008-2010) that aims to create a machine-readable digital corpus of spontaneous and elicited British Sign Language (BSL) collected from deaf native signers and early learners across the United Kingdom. In the field of sign language studies, it represents a unique combination of methodology from variationist sociolinguistics and corpus linguistics. The project aims to conduct a studies of sociolinguistic variation, language change and language contact simultaneously with the creation of a corpus. As such the nature of the dataset to be collected will be guided by the need to create a judgement sample of the deaf community rather than a strictly representative sample. Although the recruitment of participants will be balanced for gender and age, it will focus only on signers exposed to BSL before the age of 7 years, and adult deaf native signers will be disproportionately represented. Signers will also be filmed in 8 key regions across the United Kingdom, with a minimum of 30 participants from each region. Furthermore, participant recruitment will rely on deaf community fieldworkers in each region, using a technique of 'network sampling' in which the local community member begins by recruiting people he or she knows, and asks these individuals to recommend other individuals matching the project criteria. Moreover, the data will be limited in terms of situational varieties, focusing mainly on conversational and interview data, together with narratives and some elicitation tasks. Unlike previous large-scale sociolinguistic projects, however, the dataset will be partly annotated and tagged using ELAN software, given metadata descriptions using IMDI tools, and will be archived and made accessible and searchable on-line. As such, we hope that it will become a standard reference and core data source for all researchers investigating BSL structure and use. This means, however, that, unlike previous sociolinguistic projects on ASL and Auslan, 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 the aim of variationist sociolinguistics is to study the vernacular variety (i.e., the variety adopted by speakers/signers when they are monitoring their style least closely), open-access archives thus may not always provide the best data source. While recognising that this concept of the vernacular represents an abstraction, we discuss the possibility of overcoming this problem by making some of the conversational data password protected for use by academic researchers only, while making other parts of the corpus publicly available as part of a dual access archive of BSL.

Introduction

The British Sign Language Corpus Project (BSLCP) is a new three-year project (2008-2010) funded by the British Economic and Social Research Council that aims to create a machine-readable digital corpus of spontaneous and elicited British Sign Language (BSL) collected form deaf native signers and early childhood learners across the United Kingdom. Researchers at University College London are leading 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 and corpus linguistics. The project aims to conduct a studies of sociolinguistic variation, language change and language contact simultaneously with the creation of a corpus. Unlike previous large-scale sociolinguistic projects, however, the dataset will be partly annotated and tagged using ELAN software, given metadata descriptions using IMDI tools, and will be archived and made accessible and searchable on-line. As such, we hope that it will become a standard reference and core data source for all researchers investigating BSL structure and use.

2. 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, we will undertake five specific studies. Of these, four studies will investigate sociolinguistic variation and change in (1) a phonological variable (e.g., variation in handshape, location, movement or the presence of the non-dominant hand in a specific subset of BSL signs) (2) a grammatical variable (e.g., variation and grammaticalisation in the use of agreeing/indicating verbs or in the morphosyntactic marking of tense, aspect and modality); (3) a set of lexical variables (we will focus on fifty vocabulary items known to exhibit regional variation or undergoing language change in BSL); and (4) bilingualism and language contact (e.g., variation in the

use of English mouthing in conversational BSL or possibly also variation in the use of contact signing). Lastly, there will be a study of lexical frequency in BSL. Based on an annotated subset of 100,000 lexical items, we will investigate which are the most frequent signs in BSL conversation, adapting the methodology successfully employed for a study of lexical frequency in New Zealand Sign Language (NZSL) by McKee and Kennedy (2006).

In the sociolinguistic studies, our aim will be to investigate how the variation in three target variables correlates with linguistic (e.g., the preceding or following segment in the case of phonological variable; the degree of obligatoriness in the case of the grammatical variable) and social factors (e.g., the signer's region, gender, age, language background and possibly the strength of their social network ties, socio-economic class and ethnicity). The specific target variables will be selected after a literature review and an initial viewing of the BSL data collected (the latter is necessary because the frequency of particular linguistic variables in BSL is unknown, and we will need to focus only on those variables for which our data will yield many examples). An additional aim of our project relates to cross-linguistic differences in sociolinguistic variation and language change, with specific comparative studies planned with both a closely related sign language (Australian Sign Language, or Auslan) and an unrelated sign language (American Sign Language, or ASL).

2. BSLCP Methodology

The studies of lexical frequency, sociolinguistic variation and language change in BSL (and all future studies which will draw upon this corpus) require that we collect and analyse data from a large representative sample of the British deaf community. In order to make possible cross-linguistic comparison, the methodology is similar to related studies undertaken on ASL (Lucas, Bayley & Valli, 2001), Auslan (Schembri, Johnston & Goswell, 2006; Schembri & Johnston, 2007) and NZSL (McKee, McKee & Major, 2006). Using a previously implemented research design will also allow us to identify and propose solutions to any potential problems in the project methodology.

2.1 Site selection

To obtain a representative sample of regional variation in BSL, at least eight sites will be necessary (sites in south-east, south-west, north-east, north-west and central England as well as at one site each in Northern Ireland, Wales and Scotland). These sites will need to be carefully selected so that they are as representative of the major regional varieties of BSL as possible. In addition to this, it is most likely that the site selection will involve a focus on large urban centres where the largest number of deaf individuals is concentrated and where large residential schools for deaf children are or were previously located. This will ensure that there are sufficient numbers of deaf people to make it possible to collect a sample that is balanced for gender, age, language background, and potentially other social factors such as strength of social network ties, socio-economic class and ethnicity.

2.2 Participant selection

For this project, as has become standard in sociolinguistics research (Tagliamonte, 2006), we will recruit using a judgement sample of individuals from the British deaf community. Our aim is to recruit deaf native or near-native BSL signers who have lived in their local community (i.e., in the sites selected above) for at least ten years, and have thus had sufficient length of exposure to the local dialect of BSL (preference will be given to life long residents where possible). Thus, the participants will be both deaf individuals from deaf families who learned to sign natively in the home as well as deaf individuals who were exposed to sign language before age 7 by mixing with deaf peers in school. Similar numbers of participants will be recruited in four different age groups: (1) 18-35 years of age, (2) 35-50 years, (3) 51-70 years, and (4) 71 years or over. The division of participants into these age groups is partly motivated by changes in language policy in deaf education during the twentieth century (c.f., Lucas, Bayley & Valli, 2001).

The selection of our participants will also take gender, social class and possibly ethnicity into account. Gender and class differences in language use have been a major focus of research on sociolinguistic variation in spoken languages (e.g., Coates, 1986, Labov, 1990) and signed languages (Lucas, Bayley & Valli, 2001; Schembri, Johnston & Goswell, 2006). We will recruit equal numbers of male and female participants from both working and middle class backgrounds in order to determine whether gender and social class are important sociolinguistic factors in the British deaf community. We define 'working class' individuals as those who were employed in unskilled, semi-skilled or skilled manual jobs (e.g., labourer, factory worker, or plumber) or as semi-skilled non-manual workers (e.g., clerk). 'Middle class' participants are those, possibly with a university education, who worked in skilled non-manual jobs (e.g., BSL tutor) or in professional and/or managerial positions (e.g., manager of an interpreting service). Due to the fact that university education has only become widely accessible to deaf people in the UK following disability discrimination legislation enacted since the 1980s, we would not always rely on university qualifications as a defining part of our social class classification (this was a key criterion used in the study of sociolinguistic variation in ASL, see Lucas, Bayley & Valli, 2001). Strength of social network ties is another factor at work in systematic sociolinguistic variation. Research reported in Milroy (1987) showed the degree and nature of social contact between individuals influenced their language use. This is bound to be of great relevance to deaf communities where social networks are comparatively dense and multiplex. We will also investigate the possibility that a signer's ethnic background (e.g., White British, South Asian or Afro-Caribbean background) may be a relevant factor.

2.3 Recruitment

The participants will be recruited in each of the sites by deaf community fieldworkers who will all be deaf individuals living in the 8 target sites with knowledge of the local deaf community. They will be responsible for identifying and recruiting native or near-native BSL users who have lived in the community for at least ten years. Initially, participants already signed up to the DCAL participant database who match our selection criteria will be approached, but additional recruitment will need to be carried out in order to reach our target number of 30 participants at each site. A technique of 'network sampling' will be used, in which the fieldworker begins by recruiting people he or she knows, and then asks these individuals to recommend other individuals matching the project criteria (Milroy & Gordon, 2003). In this way, we hope to ensure that participants will be filmed in pairs consisting of two individuals who will already know each other. All participants will be paid for their time.

Based on our experience collecting data for the Auslan corpus project and on the success of a similar American project (Lucas, Bayley & Valli, 2001), we do not anticipate any problems with finding sufficient number of participants. Exact figures for the number of people in the British deaf community are not available, but it is likely to be somewhere between 20-60,000 people. If only 10% of this population are native signers or early childhood learners of BSL, then even the lowest estimate gives a total population of 2000 deaf signers who meet our criteria. As a result, finding 240 individuals to participate in our project should not be a problem.

It is likely, however, that we may not be able to attract sufficiently balanced numbers of participants from working class and middle class backgrounds from the four age groups listed above, or from different ethnic groups. For social class and ethnicity, we would either run analyses on smaller samples of the participant pool in which the numbers were balanced, or simply not include these factor groups in our analysis and concentrate our analysis on the other social factors. For age, we could simplify the age bands into two groups, younger (e.g., 18-45 years of age) versus older (e.g., 46-90 years of age), rather than analyse the data using the smaller subdivisions into four age groups (i.e., 18-35, 35-50, 51-70, 71 and over).

2.4 Data collection

The data to be collected will be of four types: (1) a set of elicited lexical data, (2) a set of elicited narrative and grammatical data, (3) thirty to forty-five minutes of free conversation, and (4) an interview.

The data in type (3) will be collected first. This will consist of at least thirty minutes of free conversation between the members of a particular dyad in order to collect data that is as naturalistic as possible. To ensure that the deaf participants do not adjust their signing to match the preferences of the researchers (who will not, in most cases, be members of their local deaf community) and to minimise influences from contact with spoken English (Lucas& Valli, 1992), none of the researchers (hearing or deaf) will be present.

After this, data of type (4) will be collected. The field worker will interview the participants about their background, patterns of language use, degree of bilingualism in BSL and English, and attitudes to sign language and deafness.

Next, the data collection will focus on type (1) data. In this case, the field worker will show the participants a set of 50 flashcards to elicit their signs for selected concepts. The set of signs to be elicited will be in part based on earlier work on lexical variation in BSL (Woll, 1991; Brien, 1992) In previous projects, it has been shown that lexical variation was significant in particular semantic fields, such as signs for colour terms, days of the week, and numerals (Deuchar, 1981; Sutton-Spence & Woll, 1999). Our current research design will build on this work by exploring in more detail the correlation between this lexical variation and social factors such as age and region (as has been demonstrated for the use of fingerspelling in BSL, see Sutton-Spence, Woll & Allsop, 1990). In most cases, we will use pictures to elicit the target lexical items (e.g., a flashcard that is coloured green to elicit the sign for this concept), but where this is not possible, the signs will be elicited by the use of flashcards showing a picture together with the sign's closest English equivalent (e.g., a card with the written English word 'people' on it in order to elicit the sign for this concept).

In separate data collection sessions, data of type (2) will be collected. We will use a mixture of tasks to elicit narratives and some of the key grammatical features of the language (e.g., agreement/indicating verbs, constituent order, and non-manual features), but do so indirectly, as the nature of the tasks will involve the participant describing aspects of the stimulus material to another signer, rather than the language features themselves. Some of the tasks will be selected from those used in the Auslan and Sign Language of the Netherlands (NGT) corpus projects (Johnston & Schembri, 2006). Using materials successfully employed on previous projects will ensure that we are able to elicit the kind of linguistic data we need for the purposes of this study, as well as facilitate cross-linguistic comparison.

For all data collection tasks, we will use up to four high definition digital video cameras on tripods so as to provide body length views of the individuals in each group as well as views from above of their use of signing space. Both the chairs and video cameras will be infixed positions so as to ensure that participants are in the centre of the frame at all times. In some cases as well (e.g., where natural light is insufficient), it may be necessary to use studio lighting to ensure the best images of the participant's signed communication are captured.

3. Open-access corpus data and sociolinguistic variation

The BSL Corpus will represent the largest corpus of its kind when it is complete, involving data from 240 participants. As such, the corpus will lend itself naturally to large-scale investigations into sociolinguistic variation

and change in BSL, despite the fact that the data will be limited in terms of situational varieties (i.e., it includes mainly conversational and interview data, together with narratives and some elicitation tasks). Unlike previous large-scale sociolinguistic projects (e.g., Lucas, Bayley & Valli, 2001; Schembri, Johnston & Goswell, 2006; McKee, McKee & Major, 2006), however, the dataset will be archived and made accessible and searchable on-line. As such, we hope that it will become a standard reference and core data source for all researchers investigating BSL structure and use. This means, however, that, unlike previous sociolinguistic projects on ASL, Auslan and NZSL, 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) explained, a specific aim of variationist sociolinguistics is to study the vernacular variety. Labov (1972) defined 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).

Thus, combining sociolinguistic methodology with corpus linguistics objectives creates a unique form of the observer's paradox. Although ideally we want to observe how deaf people use BSL with each other when they are not being observed, participants will in fact be filmed using four cameras, perhaps with lighting equipment, after they have filled in consent forms that make them fully aware that the aim of the project is to create an open-access on-line BSL corpus using the data collected. Thus, participants will not only be aware that their signing will be seen by researchers, but also potentially by anyone with a computer that has access to the internet! It is not yet clear how much this will cause signers to shift away from casual usage.

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, clearly we need 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. As in other projects, we can ensure that participants are comfortable with each other and filmed in the most familiar surroundings possible (e.g., deaf clubs, offices of the British Deaf Association etc). We also will allow the participants to chat with each other about any topic in the free conversation session for a minimum of thirty minutes. This should allow time for the participants to relax in the presence of the cameras. We also need to investigate the possibility of making access to the conversational data, for example, password protected and for use by academic researchers only, while making other parts of the corpus publicly available as part of a dual access archive of BSL. Participants would be informed at the outset that the conversational data will not be part of the open-access archive, and that researchers who wish to see it will have to fill in an online registration form that includes a confidentiality agreement. This may help to make participants feel more relaxed, and thus more likely to produce examples of vernacular BSL.

Conclusion

Corpus-based approaches represent new territory for sociolinguists of both signed and spoken languages. As such, advances in digital video technology open up new possibilities for data sharing and research collaboration, but they also present new challenges. In this paper, I have outlined the methodology to be employed in the BSLCP, and attempted to anticipate some of the problems it may encounter, as well as suggest some possible solutions.

1. Acknowledgements

I would like to thank my co-investigators Kearsy Cormier, Margaret Deuchar, Frances Elton, Donall O Baoill, Rachel Sutton-Spence, Graham Turner and Bencie Woll for assistance in writing the grant proposal from which much of the content of this paper is drawn.

2. References

Brien, D. (Ed.). (1992). *Dictionary of British Sign Language/English*. London: Faber & Faber.

Coates, J. (1986). *Women, men and language*. London: Longman.

Deuchar, M. (1981). Variation in British Sign Language. In B. Woll, J. G. Kyle & M. Deuchar (Eds.), *Perspectives on British Sign Language and Deafness* (pp. 109-119). London: Croom Helm.

Johnston, T. & Schembri, A. (2006). Issues in the creation of a digital archive of a signed language. In: Barwick, L. & Thieberger, N. (Eds.), Sustainable data from digital fieldwork. (pp. 7-16). Sydney: University of Sydney Press.

Labov, W. (1972). *Language in the Inner City*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1990). The intersection of sex and social class in the course of language change. *Language Variation and Change, 2*, 205-254.

Lucas, C., & Valli, C. (1992). *Language contact in the American deaf community*. San Diego: Academic Press.

Lucas, C., Bayley, R., & Valli, C. (2001). *Sociolinguistic Variation in American Sign Language*. Washington, DC: Gallaudet University Press.

McKee, D., & Kennedy, G. (2006). The distribution of signs in New Zealand Sign Language. *Sign Language Studies, 6*(4), 372-390.

McKee, D., McKee, R. & Major, G. 2007. Sociolinguistic variation in New Zealand Sign Language numerals. Paper presented at the *Ninth International Conference on Theoretical Issues in Sign Language Research.* Florianopolis, Brazil, December 6-9.

Milroy, L. (1987). *Language and social networks*. Oxford:

Blackwell.
Milroy, L., & Gordon, M. (2003). *Sociolinguistics: Method and interpretation*. Oxford: Blackwell.
Schembri, A., & Johnston, T. (2007). Sociolinguistic Variation in the Use of Fingerspelling in Australian Sign Language (Auslan): A Pilot Study. *Sign Language Studies, 7*(3).
Schembri, A., Johnston, T., & Goswell, D. (2006). NAME dropping: Location Variation in Australian Sign Language. In C. Lucas (Ed.), *Multilingualism and sign languages: From the great plains to Australia* (Vol. 12). Washington, DC: Gallaudet University Press.
Sutton-Spence, R., & Woll, B. (1999). *The linguistics of British Sign Language: an introduction*. Cambridge, U.K.: Cambridge University Press.
Sutton-Spence, R., Woll, B., & Allsop, L. (1990). Variation and recent change in fingerspelling in British Sign Language. *Language Variation and Change, 2*, 313-330.
Tagliamonte, S. (2006). *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.
Woll, B. (1991). Historical and comparative aspects of British Sign Language. In S. Gregory & G. M. Hartley (Eds.), *Constructing deafness*. London: Printer Publishers.