ISSUES IN THE CREATION OF A DIGITAL ARCHIVE OF A SIGNED LANGUAGE

Trevor Johnston
Department of Linguistics, Macquarie University, Sydney

Adam Schembri
Deafness, Cognition and Language Research Centre, University College London

Introduction

In this paper we summarise the fieldwork involved in creating the Auslan (Australian Sign Language) archive and corpus. We briefly discuss some of the technical and ethical problems in data collection and management associated with visually based linguistic data for an archive intended to be as open as possible. We also discuss the new research questions opened up by the existence of signed language data in this new form. The Auslan archive is the output of a project funded by the Endangered Languages Documentation Program within the School of Oriental and African Studies (SOAS) at the University of London and is to be submitted in mid-2007.

The Archive

The Auslan documentation project has recorded, collated, annotated and selectively transcribed naturalistic, controlled and elicited signed language text from deaf native and near-native users of the language across Australia. Footage from 50 three-hour language-use sessions, each with two participants, has been recorded on 300 hours of digital videotape. This footage has been backed up to disk totalling approximately 3 terabytes of data. Overall, the primary aim of the project is to create a set of digital archive of Auslan as used by deaf native (or near-native) signers to ensure that future corpus-based grammatical description of the language remains possible in the face of rapidly decreasing numbers of signers. The archived will be located at Macquarie University, Sydney, Australia (primarily to guarantee and facilitate access to the results to members of the Australian deaf Auslan-using community) and at the Endangered Languages Archive (ELAR) which is housed at SOAS.

Language endangerment

Auslan is an endangered language (Johnston, 2004). There are probably fewer than 6,500 deaf Auslan users in Australia. No more than 5-10% would have acquired Auslan as a first language in the home from deaf signing parents (that is, no more than 600 signers), though perhaps a sizeable minority learned Auslan as a somewhat delayed first language as young school children (around six years of age) in centralised residential schools for the deaf (many of which ceased to exist several decades ago). The community has an inverted age pyramid, with the vast majority of signers over 40 years of age. Due to improved medical care, fewer deaf children are being born and, of these, a large proportion are receiving cochlear implants and are not being exposed to Auslan at all, or only much later in life. Falling enrolments in schools for deaf children, lack of exposure to deaf community linguistic role models and the scarcity of teachers proficient in Auslan is seeing the use of the language in this area already becoming problematic. With smaller numbers and with decreasing ongoing use in a variety of functional domains, including the socialisation and enculturation of new signers, the language must clearly be regarded as endangered.

Fieldwork

Deaf native signers of Auslan were employed to co-ordinate and organise local deaf community language participants during data collection sessions in Sydney, Melbourne, Brisbane, Adelaide, and Perth. These individuals were themselves long-term residents of each city, with a thorough knowledge of their local deaf community. All of them worked for us on a related, earlier project investigating sociolinguistic variation in Auslan (Schembri, Johnston & Goswell, 2006).

Fieldwork sessions

Language recording sessions were conducted involving 100 deaf native and near-native signers of Auslan (20 participants in each of five sites). Participants were involved in three hours of language-based activity that involved an interview, the production of narratives, a survey, free group conversation, and other elicited linguistic responses to various stimuli. Two people were recorded in each session. The sessions were lead by a deaf native signer, usually a person well known to the participants. The footage of the tasks, listed below, is currently being edited into separate

digital movie clips (approximately 17 for each participant) as indicated (for example, c1a, c1b, and so forth).

1. A brief identification and background interview in Auslan conducted by the session leader (c1a). Each person also gave their sign name (that is, their sign 'nickname' commonly given to individual signers by other members of the signing community) and explained its origin (c1b). (10 minutes)
2. The production of a narrative in Auslan based on a text stimulus, which was one of two Aesop's Fables: 'The Boy Who Cried Wolf' (c2a) or 'The Hare and the Tortoise' (c2b). The fables had been read one week prior to the recording session and practiced individually be each participant, each of whom had one fable. Each participant retold the other participant the fable they had prepared. (15 minutes)
3. The session leader asked each participant to tell the other a memorable event that had happened to them or someone close to them (c3). (10 minutes)
4. A questionnaire/survey was conducted on each participant's attitudes regarding various issues of relevance to the deaf community, such as issues related to the genetics of deafness, the use of cochlear implants, the availability and cost of sign language interpreters, the use of Auslan in schools for deaf children, and the future of the deaf community. The session leader encouraged participants to elaborate and give reasons for their answers (c4). (15 minutes)
5. The two participants were left alone and allowed to converse freely on any topic (c5). (15 minutes)
6. An episode of the Warner Brothers cartoon series 'Tweety and Sylvester' was edited and sections shown to each participant alternatively and separately. Then each retold in Auslan to the other participant what they had seen. Four episodes were described (c6ii, c6iii, c6iv, c6v). The cartoon had no subtitles or sound track so as to minimise influence from English. (25 minutes)
7. Each participant recounted a narrative to the other that they had just seen. The first was based on a textless picture book ('Frog Where Are You' (c7a)). The second was based on a story in Auslan told by an unknown deaf person (c7b). The Auslan story had been pre-recorded on video and was shown to the participant. (20 minutes)

8. One participant watched a series of filmed vignettes designed to elicit depicting signs (also known as 'classifier' constructions, see Emmorey, 2003). They then described what they had seen to their addressee (c8a). Next, the other participant watched a series of video vignettes designed to elicit distinct forms of related nouns and verbs in Auslan and then described what they seen to the first participant (c8b). (20 minutes)
9. One participant described a series of 18 pictures depicting a series of simple events, situations and actions categorised as 'reversible' (either participant in the illustration could be the agent), 'non-reversible' (only one participant in the illustration could be easily construed as the agent), or 'locative' (describing a locative relationship, not an action)(c9). (15 minutes)
10. Each participant was given one of two very similar drawings. They asked each other questions to establish the exact differences between the two pictures ('spot the difference'). This task was designed to elicit interrogative and negated constructions (c10). (15 minutes)

Issues with fieldwork

Cameras. Two digital video cameras were used in each recording session. As it is possible to synchronise and display two camera views in a single ELAN file (ELAN is discussed below), it was originally thought that a third camera, to capture the interaction of both signers, was therefore unnecessary. However, additional cameras do in fact appear to be necessary to capture other aspects of signed interactions that may be difficult to view and code using a single camera that only captures the entire upper torso of a signer. Namely, additional cameras for a detailed view of the face for recording head and eye movements and facial expressions and a single overhead camera (facing downwards) to record the horizontal use of the space around the signer's body. The overhead camera would significantly improve the accuracy and speed with which annotators are able to specify the use of locations in the signing space. (ELAN can support at least 4 simultaneous camera views when annotating data.)

Lighting. Modern commercially available cameras are very good and the normal quality of recordings is excellent. However, when filming in people's homes and/or for long periods of time (such as over a 3 hour period), lighting conditions may change sufficiently to cause the initial set up to be no longer optimal if not constantly monitored. In several filming sessions, the third hour of tape was of poor, though still useable, quality

because it had, for instance, recorded an afternoon session that had extended into the evening.

Tasks. The final task of the language elicitation sessions appears not to have worked well. Unlike users of a spoken language in which the 'spot the difference' task does indeed elicit question forms, when conducted with deaf signers this did not work well. Participants found it difficult not to point to their pictures and/or produce simple one or two sign utterances without question marking (because the nature of the data collection activity made this redundant).

Issues with data management post-fieldwork, pre-archiving

Backup. Large amounts of digital video data must be backed up both in case of accidental damage or loss and as a means of distributing clips of a manageable size to annotators and other researchers involved in analysis of the data. This was not properly anticipated in the timetabling and budgeting of this project. Commercial backup services for 300 hours of video data would have cost the equivalent of almost an entire year's wages for a basic research assistant. To 'save' money, we had to use one quarter of the allocated annotation time of a research assistant just to back up the video data. In addition, the vast storage requirements for the video necessitated the purchase of 5 terabytes of disk storage space that had not also been budgeted for. Neither university-based servers nor local Australian language archives were prepared to provide such large storage—even on a temporary basis (for example, as a short-term measure during the editing process).

Editing. In order to allocate and manage the annotation of or the appending of appropriate metadata to video files it was also necessary to edit the hour long digital tapes/recordings into 'task episodes' which are easier and less time-consuming to manage (to burn to disk, copy to computers, and so forth) for use by annotators. To manage the data for ongoing long-term annotation and analysis, the 'fieldwork' component of digital language archiving, especially if it involves video data, must allow for considerable amounts of time just for editing of the data in this way. When submitted the Auslan archive will now actually consist of approximately 1,700 individual video clips (ranging between 2 mins and 20 mins in length) with linked ELAN annotation or IMDI metadata files (IMDI is explained below). This has yielded approximately another 1-2 terabytes of edited digital video. Of course, the archive will also include the complete set of original backup digital videos, itself 3 terabytes in size.

Annotation at archive submission. The amount of time required for annotation of signed language texts is enormous. Consequently, the initial proposal only aimed at annotating a small proportion of the archive. However, the target amount of annotation has been further reduced during the life of the project due to (a) the demands on research assistant's time for digital backup of the data and (b) the lack of available annotators in Australia (that is, people with linguistic training, knowledge of Auslan, and computer skills). Therefore, only a small proportion of individual video clips have been or will be annotated for (i) the use of space with verb signs, and (ii) the word classes of all signs within a given text. A larger amount of archive footage will have a free English translation in voice over or written text aligned to various natural breaks in the text.

The archive and new corpus-based research

Until the creation of this archive/corpus, there had been no systematic, widespread and exhaustive collection of a representative sample of Auslan as used by deaf native and near-native signers. Consequently, it has not been easy to verify empirically some claims regarding grammatical patterns and discourse structures of the language that result from elicitation sessions with individual informants or small numbers of signers (see Johnston, Vermeerbergen, Schembri & Leeson, in press). (Somewhat surprisingly, the situation is little better for most of the world's signed languages.) This has serious implications for the status of recent claims as to linguistic universals (for example, Sandler & Lillo-Martin, 2006) and our understanding of processing of language in the brain derived from signed language studies (for example, Emmorey, 2002). Neither grammatical descriptions of language in the visual-gestural medium, nor the impact of modality on the processing of language in the brain can be properly understood and evaluated without representative corpora of signed languages.

Our intention is to limit the size of the Auslan archive to the original 300 hours collected as part of this project, with the priority being to create a richly annotated and tagged corpus within the shortest possible time. However, this will still take several years. Once the annotation and tagging of the corpus is detailed enough and the data is made publicly accessible, this will enable the description of a grammar of Auslan to be empirically grounded and open to thorough critical peer review.

The annotation procedure

A subset of the archive data is being annotated using the ELAN (EUDICO – European linguistic annotator) digital video software. The software allows for the precise time-alignment of annotations with the corresponding video sources on multiple user-specifiable tiers. It allows one to create, edit, visualise and search annotations for video data. It supports display of video with its annotation; time linking of annotations to media streams; linking of annotation to other annotations; unlimited number of annotation tiers defined by users; different character sets; export of annotations as tab-delimited text files and a complementary ability to import text file annotations. Relevant metadata for all the digital recordings is being appended to all annotation files. All annotations are be in English and use tags or descriptors developed in a project conducted by Johnston, Woll and their colleagues during 2005.[1] The Auslan archive is using established standards that are derived from spoken language corpora (such as those defined under the EAGLES project) and standards that are fully compatible with them (such as ISLE MetaData Initiative or IMDI). IMDI offers detailed categories to describe corpora and lexical and has been developed and defined at the Max Planck Institute for Psycholinguistics in The Netherlands. IMDI comprises different sets of metadata, such as sessional metadata (descriptions of combinations of files and linguistic annotation files), catalogue metadata (description on a more abstract level of the corpus as a whole) and lexicon metadata (descriptions of lexicons). It also now includes a subset of metadata descriptors specific for signed languages. These standards are compatible with other sign language corpora projects, including the current corpus project on Sign Language of the Netherlands (Nederlandse Gebarentaal or NGT,) and planned projects on German Sign Language (Deutsche Gebardensprache or DGS) and British Sign Language (BSL).

At the time of submission of the archive, a report will also be made on one morpho-syntactic feature of Auslan—the use of spatial modifications of signs (placement, re-orientation and/or direction) to encode semantic roles (Johnston et al, 2006). The report will be based on data extracted from an annotated subset of this digital archive.

Most importantly, The EthnoER project is developing software and protocols (such as Annodex) that will enable digital video data, like that found in the Auslan archive/corpus, to be remotely annotated, using ELAN or other software, whilst still being under the control and management of a central repository. Once this is developed and

operational, the Auslan archive will be able to move into the second and most important phase of its development.

Research questions to be addressed in planned corpus-based research

In the first instance, research using the newly established corpus will concentrate on finding support for observations already made about Auslan lexis and grammar (such as word order, use of non-manual features, the linguistic use of space, and so forth).

However, one current area of research and theoretical interest in sign linguistics concerns the typical or possible grammaticalisation pathways exploited in sign languages (for example, the possible grammaticalisation of auxiliaries from content signs, linguistic uses of space from gesture, and syntactic markers from affective facial expressions), and the Auslan archive will enable investigation of this question. A study is planned (in cooperation with the planned BSL corpus project) in which variant forms of some aspect marking signs (such as two lexical variants of the sign FINISH used in both Auslan and BSL) found throughout the corpus can be identified. The environments in which they are located will be annotated and tagged to determine the extent to which this variation can be attributed to the process of grammaticalisation. There is a well-established relationship between language variation and change (Labov, 1994). We are particularly interested in differential rates of grammaticalisation of a lexical form and its variants in these two related sign language varieties in different parts of the world. It has been previously demonstrated that lexical and grammatical variation in Auslan is often sociolinguistic in nature —reflecting regional, educational, social class, age and gender characteristics of its users (Schembri, Johnston & Goswell, 2006; Schembri & Johnston, in press). However, some variation in the form and use of particular signs such as FINISH may also reflect different degrees of grammaticalisation of this sign as a perfective aspect maker in certain subgroups of Auslan users (for example, in older versus younger signers, or signers in particular regions of in Australia and Britain). Data will be extracted from the corpus on the variable form and use of signs used to signal perfective aspect to address this question. This type of study would simply have been impossible without the Auslan archive.

The archive will also facilitate a second study on lexical frequency in Auslan. It is planned to create an annotated corpus of 100,000 lexical items from the archive to identify out which items are the most frequent in Auslan conversation. The frequency of particular linguistic items in

Auslan is unknown. Lexical frequency is an important factor in phonological variation (Schembri, Johnston & Goswell, 2006) and psycholinguistic studies, for example, but no information about this aspect of Auslan usage is currently available.

Ethical issues in creating visual ethno-linguistic archives

The most important ethical issue in this particular project centred on securing the active consent of all participants to have visual recordings of themselves freely available through an open archive—one that would potentially be accessible through the internet. Signed language data cannot be made anonymous and remain primary data. Because of the importance of the face in signed languages (the use of particular facial expressions and mouthing play a vital role in its grammar and lexis) is impossible to access digital video recordings without seeing the participants. There can be no empirical validation of signed language research if this fact is not acknowledged and accommodated. Accordingly, we asked for, and received, the consent from all participants to allow open access to the video archives in which they appeared. It was explicitly stated in the consent forms that this did not entail permission to researchers to show or use video data or stills from video data in secondary publication or to access any personal or identifying information about participants. Additional, separate and optional consent was also sought from those participants who would permit the project researchers to use segments of digitised video in academic and learned presentations. Not one potential participant objected and all gave their dual consents.

Conclusion

The creation of digital archives of signed languages is challenging, time-consuming and expensive, as our experience with the Auslan archive reported here shows. It is hoped that improvements will be able to be made in similar archive projects that are able to learn from our experience. Overall, however, there is little doubt that the creation of a digital archive is well worth the effort. By exploiting the currently available and rapidly improving video annotation software in conjunction with digital archives and corpora, sign language linguistics destined to become much more rigorous within a relatively short period of time.

Endnotes

[1] A/Prof Trevor Johnston & Prof. Bencie Woll. *Exploring tagging agreement for comparative analyses in Australian (Auslan) and British (BSL) sign language copora.* Project supported by Australian Academy of the Humanities, Academy of the Social Sciences in Australia, British Academy (Australian-British Joint Projects).

References

Emmorey, K. D. (2002). *Language, Cognition, and the Brain: Insights from Sign Language Research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Emmorey, K. D. (Ed.). (2003). *Perspectives on Classifier Constructions in Sign Languages*. Mahwah, NJ: Lawrence Erlbaum Associates.

Johnston, T. (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, T., Vermeerbergen, M., Schembri, A., & Leeson, L. (in press). 'Real Data Are Messy': Considering Cross-Linguistic Analysis of Constituent Ordering in Australian Sign Language (Auslan), Vlaamse Gebarentaal (VGT), and Irish Sign Language (ISL). In P. Perniss, R. Pfau & M. Steinbach (Eds.), *Proceedings of the Workshop on Sign Languages: A Cross-Linguistic Perspective, Mainz, Germany, March 25-27, 2004*. Berlin: Mouton de Gruyter.

Labov, W. (1994). *Principles of Linguistic Change: Internal Factors*. Oxford: Blackwell.

Sandler, W., & Lillo-Martin, D. (2006). *Sign Language and Linguistic Universals*. Cambridge: Cambridge University Press.

Schembri, A., & Johnston, T. (in press). Sociolinguistic Variation in the Use of Fingerspelling in Australian Sign Language: A Pilot Study. *Sign Language Studies*.

Schembri, A., Johnston, T., & Goswell, D. (2006). NAME Dropping: Location Variation in Australian Sign Language. In C. Lucas (Ed.), *Sociolinguistics in Deaf Communities (Vol. 12)*. Washington, DC: Gallaudet University Press.