Creating a Corpus of Auslan within an Australian National Corpus

Trevor Johnston
Macquarie University

1. Introduction

The creation of signed language (SL) corpora presents special challenges to linguists. They are face-to-face visual-gestural languages that have no widely accepted written forms or standard specialist notation system, making even superficial transcription problematic. SL corpora need to be created taking these facts into account. Using the example of Auslan (Australian Sign Language) this paper describes how multimedia annotation software can now be used to transform a language recording into a machine-readable text without it first being transcribed, provided that conventional linguistic units are systematically and consistently identified, thus making possible the creation of a true linguistic corpus of a SL. Before examining SL annotation in detail, we first review the main features of modern linguistic corpora and introduce the Auslan archive which is the source of the future Auslan corpus. The paper concludes with an assessment of the place of an Auslan corpus within an Australian National Corpus and an evaluation of other recent SL corpus projects elsewhere in the world.

A modern linguistic corpus is something more than just a dataset of written or transcribed texts upon which a description or an analysis of a language is based. This sense of corpus has now essentially been superseded in the literature (e.g., McEnery & Wilson, 2001; Sampson & McCarthy, 2004; Hoey, Mahlberg, Stubbs, & Teubert, 2007). A corpus in the modern sense means a collection of written and spoken texts in a machine-readable form that has been assembled for the purposes of studying the type and frequency of constructions in a language. A modern linguistic corpus contains linguistic annotations and appended sociolinguistic and sessional data (metadata) that describe the participants and the circumstances under which the data were collected. With the development of digitized video recording and multimedia annotation software, a corpus of a signed language (henceforth, SL) can now be described as a subtype of ‘spoken’ language corpora, namely face-to-face language. SL corpora promise to vastly improve peer review of descriptions of SLs and make possible, for the first time, a corpus-based approach to SL analysis.

Corpora are important for the testing of language hypotheses in all language research at all levels, from phonology, through lexis, morphology, syntax, and pragmatics to discourse. There are several reasons why testing is particularly relevant in the field of SL studies. First, SLs, which are invariably young languages of minority communities, lack written forms and the well developed community-based standards of correctness that often accompany literacy. Second, they have interrupted generational transmission and few native speakers. Third, the representation of SL examples using written glosses has meant that primary data have remained essentially inaccessible to other researchers and consequently unavailable for meaningful peer review. Thus, although introspection and observation can still be of valuable assistance to linguists developing hypotheses regarding SL use and structure, one must also recognize that intuitions and researcher observations may fail in the absence of clear native signer consensus of phonological or grammatical typicality, markedness, or acceptability. The previous reliance on the intuitions of small numbers of informants has thus been problematic in the field. As with all modern linguistic corpora, SL corpora should be representative, well-documented, and machine-readable (McEnery & Wilson, 2001; Teubert & Cermáková, 2007). This not only requires dedicated technology and standards (e.g., Crasborn et al., 2007), it also requires a principled methodology for transcription or annotation.

The guiding principle behind the linguistic annotations being created in the initial stages of an Auslan corpus is machine-readability, not transcription narrowly understood. The aim is to create an

© 2009 Trevor Johnston. Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian
National Corpus, ed. Michael Haugh et al., 87-95. Somerville, MA: Cascadilla Proceedings Project.

annotated SL corpus and not a body of SL texts which have been transcribed to a greater or lesser degree of detail. The reason is that one can now use multimedia annotation software to transform a video recording of SL into a machine-readable text without needing to first transcribe that text (i.e., represent in some form of writing what is being uttered). This is an important consideration in building SL corpora because there is no standard or widely accepted SL transcription system. Using the methodology described in this paper, in conjunction with new multimedia annotation software, it is now possible to gain instant and unambiguous access to the actual form of the signs being annotated—the video recording—because annotations and media are time aligned. Without this methodology the attempt to create a SL corpus will be compromised.

The Auslan corpus is being built on a digital video archive consisting of a representative sample of recordings of deaf native signers linked to annotation and metadata files (Johnston & Schembri, 2006). The archive consists of approximately 150 hours of edited footage taken from 100 participants from five Australian capital cities. Each participant took part in three hours of language-based activity that involved an interview, the production of narratives, responses to survey questions, free conversation, and other elicited linguistic responses to various stimuli such as a picture-book story, a filmed cartoon, and a filmed story told in Auslan. The footage has been edited into approximately 1,100 separate digital movie texts, of which approximately 130 have been annotated to various levels of detail.

2. Notation and Transcription Compared with Annotation and Tagging

In order to appreciate the different information that may be encoded in a corpus—and importantly, to determine if all must of necessity be present for a corpus in the modern sense to be created—it is very useful to make distinction between *notation*, *transcription*, *annotation*, and *tagging*. In the creation of the Auslan corpus these distinctions have proved to be very relevant in guiding how and why the data is encoded in a machine-readable text.

2.1. Notation and Transcription

Many scholars make no real distinction between notation and transcription, but it is often useful to do so. *Transcription* is defined here as the graphic representation of a text in face-to-face or 'oral' language, that is, a text which has been signed or spoken. It uses some kind of dedicated script. *Notation* is more narrowly defined as either the writing down of individual words or signs (rather than text as such) or the actual system of symbols used for this purpose ('script' if a bona fide writing system). One of the major purposes of transcription and notation systems is to enable the reader of the graphic symbols to reproduce, with greater or lesser accuracy according to the degree of detail in the notation or transcription system, the original spoken or signed text. Figure 1 is an example of an Auslan sign (illustrated) represented underneath in a dedicated SL notation system called HamNoSys (Hamburg Notation System for signed languages). It was developed at the Institute for German Sign Language, Hamburg University (Prillwitz & Zienert, 1990).

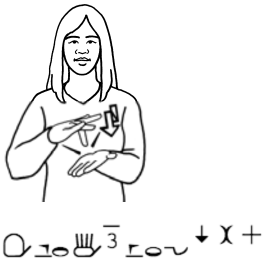


Figure 1. The sign CENTRE in HamNoSys

Generally speaking, transcriptions are usually created as reference points for, or stages in, linguistic analysis, such as in the creation of scripts for writing systems, for phonological analysis, or for grammatical analysis. They also serve as written forms of source texts which are in turn machine-readable and, therefore, able to be processed by computers. Once tokenized, the transcribed and/or written words or signs of a text can then also be annotated for various linguistic features.

Transcription was an absolutely essential step in linguistic analysis before the invention of analogue sound recording in the early 20th century. Without it, the object of study was completely ephemeral. Indeed, the advent of recordings did not reduce the reliance on transcriptions of spoken texts in order to conduct linguistic analysis, as transcriptions could not be time-aligned with recordings using the earlier analogue technology. Recordings did, however, make it possible to 'capture' the ephemeral event so that it could be listened to repeatedly before or in the process of transcription.

The development of digital recording and multimedia annotation software in the late 20th century changed the situation completely, as it enabled annotations to be directly time-aligned with recorded segments. This has become especially relevant in transforming the conduct of SL research, as consideration of annotation and tagging within this environment makes clear.

2.2. Annotation and Tagging

Linguistic annotations are appended to identified units in a language and add phonological, morphological, syntactic, semantic, and discourse information about linguistic forms, depending on the purpose of the analysis. They are an invaluable aid in helping linguists discern patterns in language at many different levels, with or without the aid of computers.

There is no clear cut distinction between an annotation and a tag—both append linguistically relevant information to units of language. However, what is now commonly called 'tagging' refers particularly to the kind of automatic annotations, such as 'part of speech' tags, appended to written texts after they have been digitized and then processed using computers. The process is illustrated in the following sentence taken from the Lancaster-Oslo/Bergen Corpus of English (cited in McEnery & Wilson, 2001, p. 47): *Joanna_NP stubbed_VBD out_RP her_PP$ cigarette_NN with_IN unnecessary_JJ fierceness_NN ._.* It uses underscores and capitalization suffixed to lexical items as its linguistic tags.

2.3. Annotation and Tags in the Proposed Auslan Corpus

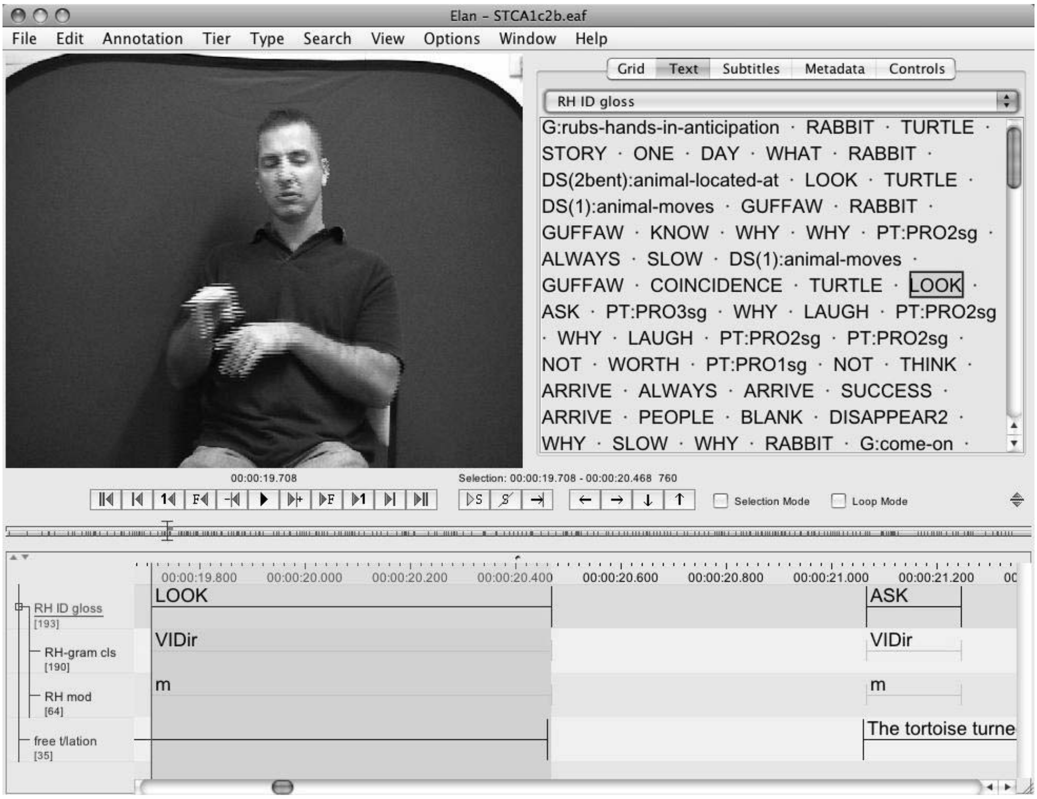
The corpus is annotated using digital video annotation software, such as ELAN (developed at the Max Plank Institute for Psycholinguistics; Hellwig, van Uytvanck, & Hulsbosch, 2007). The software allows for the precise time-alignment of annotations with the corresponding video sources on multiple user-specifiable tiers. Tags are inserted into annotation fields located on various tiers in an ELAN annotation file which are time-aligned to gloss based annotations that identify sign types. These *ID-glosses* are in turn time-aligned with the source media (see Table 1, and Figure 2).

Table 1.
Sample ELAN Tiers in Auslan Corpus

Independent tier ⇒daughter tier	Explanation of abbreviated name
ID-gloss	The ID-gloss for the sign being produced
⇒Gram cls	A tag for grammatical class.
⇒Mod	A tag for sign modification.
Clause	An annotation field that delimits a clause.
Sem.roles	A tag for the semantic role of an argument.
free t/lation	A free translation of an utterance unit.

The Auslan corpus is designed to be added to over time. Each ELAN annotation file is intended to be expanded and enriched by various researchers through repeated annotation 'parses' of individual texts (digital movies). In grammar *to parse* means *to analyze a sentence into its parts and identify their* *syntactic roles*. Here we mean by annotation parse *a pass of the text which identifies sign units and/or attaches a particular type of linguistic annotation to identified units*. This information is placed on dedicated tiers using certain conventions, codes, or controlled vocabularies. Thus, during an annotation parse an annotator will be looking at (and annotating) different aspects of sign structure and grammar on different tiers within the file.

An annotation usually begins with information just on the tiers used to identify and name signs (the gloss tiers). Information can subsequently be added to the identified unit during a second annotation parse that looks at, and tags for, some particular linguistic feature. Over time repeated annotation parses make each annotation file—and the whole Auslan corpus—very detailed and a rich source of data for research. It is expected to take more than five years to comprehensively annotate the 150 hours of video currently in the Auslan archive.



G:rubs-hands-in-anticipation · RABBIT · TURTLE · STORY · ONE · DAY · WHAT · RABBIT · DS(2bent):animal-located-at · LOOK · TURTLE · DS(1):animal-moves · GUFFAW · RABBIT · GUFFAW · KNOW · WHY · WHY · PT:PRO2sg · ALWAYS · SLOW · DS(1):animal-moves · GUFFAW · COINCIDENCE · TURTLE · LOOK · ASK · PT:PRO3sg · WHY · LAUGH · PT:PRO2sg · WHY · LAUGH · PT:PRO2sg · PT:PRO2sg · NOT · WORTH · PT:PRO1sg · NOT · THINK · ARRIVE · ALWAYS · ARRIVE · SUCCESS · ARRIVE · PEOPLE · BLANK · DISAPPEAR2 · WHY · SLOW · WHY · RABBIT · G:come-on ·

00:00:19.708

Selection: 00:00:19.708 - 00:00:20.468 760

00:00:19.800 00:00:20.000 00:00:20.200 00:00:20.400 00:00:20.600 00:00:20.800 00:00:21.000 00:00:21.200 00

RH ID gloss [193]: LOOK ASK

RH-gram cls [190]: VIDir VIDir

RH mod [64]: m m

free t/lation [35]: The tortoise turne

Figure 2. A screen grab from ELAN—the ID-gloss LOOK is tagged with “m” (“modified”) on the RH mod (right hand modification) tier and “VIDir” (for grammatical class “Directional Indicating Verb”) on the RH-gram cls (right hand grammatical class) tier.

3. Creating a Machine-Readable Corpus with Annotation Glosses

In order for a corpus of recordings of face-to-face language in either spoken or signed modalities to be machine-readable, time-aligned annotations need to be appended to the source data. The segmentation and time stamping of the signed units in a recording (i.e., tokenization) is precisely what modern digital multimedia annotation software makes possible.

Prior to the existence of such technology, a transcription of the face-to-face text needed to be made in order to create a medium to which annotations and tags could be appended. Manually or electronically, one read and then processed the transcript. In today’s multimedia digital files, the time-aligned transcriptions and annotations are similarly read by machine, but they are also linked to the source text which is thus always audible or visible. In principle, therefore, one need not have a level of transcription that represents the form of the utterance in order to have a machine-readable corpus that can be researched. This reality and the possibilities it presents in language documentation have even become apparent in recent discussion of the study of endangered spoken languages (Woodbury, 2003; Simons, 2008).

There is little doubt that creating annotations rather than transcriptions will make a larger amount of text, signed or spoken, available for processing within a shorter period of time, especially if the language does not have a writing system or standard orthography. Given that SLs have neither of these, let alone any like the International Phonetic Alphabet (IPA) for spoken languages, it is much more productive to create base level annotations that identify the sign units in the text by using glosses.

A gloss is a kind of annotation. It is a brief one or two word 'translation' in one language for a word or morpheme in another language. The 'translation' must, of course, be relatively crude and simplistic. In the Auslan corpus, the glossing language is English. Importantly, different glosses for the same sign are usually used in different contexts to reflect the meaning of that sign in that context. Consequently, it is often very difficult to know with certainty which sign form is actually being referred to by a particular gloss because a gloss does not contain any information about sign form.

To address this problem signs are identified uniquely and consistently by using an *ID-gloss* (Johnston, 2001). An ID-gloss is an English word used systematically to label a sign within the corpus, regardless of the meaning of that sign in a particular context or whether it has been systematically modified in some way. So an ID-gloss is not a temporary, context-dependent and approximate translation of a sign into the glossing language. Of course, although it is an unavoidable fact that the ID-gloss actually uses an English word that bears a relationship to the meaning of the sign, it is still not intended as a translation. Translations are made on their own dedicated tiers in the ELAN annotation files.

3.1. ID-Glossing and Lemmatization

In assigning an ID-gloss to a sign form one is identifying a sign as a token of a lexical type, so that it can be further annotated or tagged during later annotation parses (e.g., for grammatical class, semantic roles, presence or absence of modifications or 'inflections', co-occurrence with a period of constructed action, and so on). In other words, the process of assigning an ID-gloss to lexical signs in a corpus is essentially lemmatization—just as lemmatization reduces inflected forms of words to their basic forms (lexemes or lemmas), ID-glossing ignores idiosyncratic variants or systematic modifications in the form of signs, provided they are not lexicalized, in favour of the underlying citation form (the lemma).

The major distinction between lemmatization in electronic corpora of SpL languages and the use of ID-glosses in SL corpora is that in the former several different previously existing written word forms in the raw text are annotated or tagged with the one lemma, whereas in the SL recordings there need be no prior written representation or transcription which is then in turn lemmatized. Indeed, I advocate here that one go directly to the lemmatization stage in creating a SL corpus. Not only is this much quicker than attempting a transcription of a three-dimensional visual-gestural language, but a lemmatized SL text can be much more readily searched. Other tiers within the annotation file contain phonological, lexical, or grammatical information about the lemmatized sign that makes it possible to constrain searches according to these values. No information need be lost by assigning ID-glosses, and everything is to be gained in machine-readability.

The use of ID-glosses and standardized glossing procedures in multimedia corpus annotation also ensures the consistency and commensurability of annotations created by different researchers, or even the same researcher on different occasions. The number of sign types in the dataset would proliferate without constraint if distinctive 'meaning-based' glosses are assigned to essentially the same sign form in different contexts. The unique identification of sign types, which is one of the prime motivations for the creation of a linguistic corpus in the modern sense, would thus not be achieved without this approach.

3.2. Fully-Lexical Signs and Partly-Lexical Signs

The signs uttered when communicating in a SL are not all of the same type. From one point of view—just as in spoken languages—the conventionalized units of a SL can be divided into the two broad classes: an open class of content (or lexical) signs/words and a closed class of function (or grammatical) signs/words. Both these types of signs are roughly equivalent to the commonsense notion of *word* generally used to refer to the conventionalized free units of any language. Assigning unique identifying glosses to these types of signs is relatively straightforward, provided the lexicon of the language has been well documented.

From another point of view, however, there is a further word-level distinction that needs to be made for SLs which is particularly relevant for annotation and corpus creation—a distinction between *fully-lexical* and *partly-lexical* signs. The need to make this second distinction stems from the fact that, unlike the phonemes of spoken languages, the five basic formational components of signs in all SLs—handshapes, orientations, locations, movements, and non-manual facial expressions—can be individually meaningful, through iconicity and/or through language-specific form-meaning conventionalization. These components can directly and componentially contribute to the meaning of a given sign form in predictable ways.

The need for the category of *partly-lexical* signs stems from the observation that there are signs that, although conventionalized at the level of the meaningfulness of their components, do not have associated with them a meaning which is additional to or unpredictable from the value of those components when the sign is produced and used in various contexts. These types of signs have also been called *non-lexicalized* signs (Johnston & Schembri, 1999) because they contrast with *fully-lexical* (*lexicalized*) signs whose meaning cannot simply be derived from that sign's form and/or its use in context. However, to avoid confusion of the term *non-lexicalized* or *non-lexical* sign with *grammatical* sign (or word)—in opposition to *lexical* (content) sign—they are referred to here as *partly-lexical* signs in contradistinction to *fully-lexical* signs. In other words, a *fully-lexical sign* may be either a content word/sign or a function word/sign. *Fully-lexical signs* constitute the listable lexicon of a signed language.

3.2.1. Fully-Lexical Signs and ID-Glosses

Fully-lexical signs are identified using an ID-gloss written in upper case (e.g., the British Sign Language sign SISTER is identical to Auslan sign SISTER but completely different to the American Sign Language sign SISTER.) The use of uppercase for all glosses commonly found in SL linguistics is partly due to the fact that doing so helps to distinguish the SL gloss from the surrounding majority language text with which it could easily be confused because it is usually in the same written language. The use of uppercase is also partly due to the fact that simple SL glosses tend to identify citation forms and are thus essentially lemmas. Lemmas are traditionally written in upper case also in linguistic annotation.

The standard ID-gloss for a sign is found by consulting a database of the fully-lexical signs of the language—the lexicon. The Auslan lexical database contains individual sign entries in which short digital movie clips are headwords (i.e., head signs). There are multiple fields coding information on the form, meaning, and lexicalization status of each head sign. The database lists a citation form of a fully-lexical sign as a major stem entry, with common variant forms listed separately. A lexical database of this type is a necessary tool for ID-glossing. It is the result of linguistic research and organized according to linguistic principles (i.e., phonological formational features of signs). Without a lexical database the creation of a corpus using the annotation procedures described here is unlikely to succeed. Linguists need to be able to identify each sign form uniquely and this must be done by sorting sign forms phonologically. Otherwise, one could not locate and compare sign forms in order to determine if a new unique gloss is required for a particular sign form rather than just the association of an additional sense to an existing one.

Conventions for the writing of ID-glosses have been developed to ensure consistency. The conventions deal with lexical and morphological phenomena such as negative incorporation, formational variants, number signs, sign names, the marked used of one or two hands in normally two-handed or one-handed signs respectively, and borrowings from Signed English and other SLs. By way of example, the existence of negative incorporation in Auslan signs needs consistent treatment when glossed using English words in order to avoid potential suppletive or opaque forms in English obscuring the relationship between certain signs that share some important feature (e.g., LIKE-NOT rather than DON'T-LIKE, or WANT-NOT rather than NOT-WANT because each ends in an affix-like negative upturned open handshape). If ID-glosses follow a regular pattern for related types of signs this makes the extraction of statistics for the distribution of these related forms from the corpus much easier.

3.2.2. Annotation Conventions for Partly-Lexical Signs

Unlike content *and* function signs which are *fully-lexical* signs, as explained above, the assignment of ID-glosses to *partly-lexical* signs is not at all straightforward. One cannot simply refer to the database of the lexicon and extract the ID-gloss, because there is no citation form or lemma. However, a relatively small set of annotation and glossing conventions can ensure that tokens of partly-lexical signs are glossed in similar ways. Without such conventions, these categories of signs cannot be easily extracted from the corpus for analysis and comparison because each token is, in a very real sense, unique.

Tokens of partly-lexical signs are glossed using a combination of general and idiosyncratic elements. This makes it is possible to search for all instances of a subtype of partly-lexical signs in the corpus, despite the fact that overall gloss annotations for the same sign form may need to differ from context to context. Searches for frequency and collocation can be conducted using substring matches, based on the component of the gloss which is the general identifier.

The most common type of partly-lexical signs are depicting signs (also known as 'classifier signs' in the literature). Other types include pointing (or index) signs and buoys. In addition, conventions need to be followed for glossing other meaningful manual acts such as gestures and fingerspelling. Annotation glosses for all of these types of signs begin with a fixed string that identifies the subtype: DS for depicting signs, PT for points, B for buoys, G for gestures, and FS for fingerspelling. These types of signs are then further specified by a description of the form and meaning of the sign in that context.

Take pointing or indexical signs as an example. Lexicalized pointing signs are assigned an ID-gloss, for example, pointing to one's ear is lexicalized in Auslan as *hear* and is thus not glossed as PT:EAR, but is assigned the ID-gloss HEAR. However, the overwhelming majority of pointing signs in Auslan are not lexicalized in this way—they essentially remain pointing gestures and are only *partly-lexical* signs, the function or interpretation of which varies according to the context. It is thus usually difficult to establish a context-independent form/meaning pairing for the majority of pointing signs and it would be misleading to assign an ID-gloss—appropriate for fully-lexical signs—to them.

Gestures can be culturally shared or idiosyncratic. Gestures of both types occur commonly in speech and during signed discourse. Even if culturally shared, however, gestures that have not become lexical Auslan signs will not be found in the lexical database and will thus not have an assignable ID-gloss. The gloss for a gesture is prefixed with G for 'gesture' followed by a brief description of the meaning of the gesture, for example, the annotations G:HOW-STUPID-OF-ME or G:STUPID-ME may be used for the gesture of hitting the base of one's palm on one's forehead. As one can see from the example, meaning is initially prioritized over form in the description of the annotation because one can see a sign's form from the time-aligned primary data in the movie clip. By annotating the types of meanings encoded in gestures, it is possible to see both the types of meanings commonly expressed through gesture and the degree of conventionalization a gesture/meaning pairing may be undergoing by comparing annotations of similar meanings.

4. Auslan as a Component of an Australian National Corpus

A subcorpus of the signed language of the Australian deaf community (Auslan) would be an important and valid component of a larger Australian National Corpus, for several reasons. First, Auslan is related to BritishSL as Australian English is to British English. Both languages have undergone changes over the past 200 years and developed Australian forms. Auslan is no less Australian for being in the visual-gestural modality. Second, signed languages exist in a complex sociolinguistic environment in which there are varieties that are very different from the majority of written and spoken language, at one extreme, and others that are virtually signed forms of it, at the other. Third, there is the phenomenon of ‘deaf English’ in the written, and sometimes, spoken, form. This distinctive variety of English produced by deaf people, which would itself be part of the Australian National Corpus, needs to be able to be compared with a corpus of Auslan to distinguish the impact of language interference from the impact of deafness itself.

5. Conclusion

The Auslan documentation project was the first attempt to compile a large machine-readable corpus of a SL. It was begun in 2004. Since that time a number of other SL corpus projects have begun (e.g., the NetherlandsSL corpus[1] and the BritishSL corpus[2]), are about to begin (e.g., the GermanSL corpus and the SwedishSL corpus), or are planned (e.g., the AmericanSL corpus). The NetherlandsSL corpus has been completed in the sense that the archived video recordings have been edited and catalogued and are now openly accessible through a digital video archive on the internet.

However, this paper has tried to show that the creation of SL corpora as corpora in the modern sense involves more than recording, digitising, editing, cataloguing, and archiving video texts. This is not to deny the importance of the creation of reference corpora for SL researchers. After all, there have, to date, been very little publicly available reference texts of any SL. Nonetheless, corpus creation must also involve the transformation of archived material into something that is machine-readable by the principled application of annotation procedures that make optimal use of new digital technologies. Business-as-usual with these new digital archives—so-called enrichment through the addition of transcriptions or ad-hoc glosses—does not add value to the archive in ways that other corpus linguists would assume and expect. The annotation and tagging of ID-glosses, as described in this paper, is not only less time consuming than detailed phonetic or phonological transcription, it is actually much more productive.

6. Acknowledgements

The Auslan archive and corpus was created through support from (i) a Hans Rausing Endangered Languages Documentation Project grant (SOAS, University of London); (ii) an Australian Research Council (ARC) grant for a sociolinguistic variation study of Auslan (with Adam Schembri); and (iii) an ARC grant for a corpus-based study of the linguistic use of space in Auslan (with Louise de Beuzeville). I also wish to acknowledge other research assistants and annotators who contributed to the current body of annotations—Julia Allen, Donovan Cresdee, Karin Banna, Michael Gray, Dani Fired, Della Goswell, and Gerry Shearim—and, especially, the hundreds of deaf signers who contributed material to the Auslan archive.

References

Crasborn, Onno, Johanna Mesch, Dafydd Waters, Annika Nonhebel, Els van der Kooji, & Bencie Woll. (2007). Sharing sign language data online: Experiences from the ECHO project. *International Journal of Corpus Linguistics, 12*(4), 535-562.

Hellwig, Birgit, Dieter van Uytvanck, & Micha Hulsbosch. (2007). *EUDICO linguistic annotator (ELAN)*. Available at: http://www.lat-mpi.eu/tools/elan.

Hoey, Michael, Michaela Mahlberg, Michael Stubbs, & Wolfgang Teubert. (2007). *Text, discourse and dorpora: Theory and analysis*. London: Continuum.

Johnston, Trevor. (2001). The lexical database of Auslan (Australian Sign Language). *Sign Language & Linguistics, 4* (1/2), 145-169.

Johnston, Trevor, & Adam Schembri. (1999). On defining lexeme in a sign language. *Sign Language & Linguistics, 2* (1), 115-185.

[1] http://www.ru.nl/corpusngtuk/

[2] http://www.bslcorpusproject.org/

Johnston, Trevor, & Adam Schembri. (2006). Issues in the creation of a digital archive of a signed language. In Linda Barwick & Nicholas Thieberger (Eds.), *Sustainable data from digital fieldwork* (pp. 7-16). Sydney: Sydney University Press.

McEnery, Tony, & Andrew Wilson. (2001). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Prillwitz, Siegmund, & Heiko Zienert. (1990). Hamburg Notation System for Sign Language: Development of a sign writing with computer application. In Siegmund Prillwitz & Tomas Vollhaber (Eds.), *Current trends in European sign language research: Proceedings of the 3rd European Congress on Sign Language Research* (pp. 355-379). Hamburg: Signum Verlag.

Sampson, Geoffery, & Diana McCarthy. (Eds.). (2004). *Corpus linguistics: Readings in a widening discipline*. London: Continuum.

Simons, Gary. (2008). *The rise of documentary linguistics and a new kind of corpus.* Paper presented at the 5th National Natural Language Research Symposium, 25 November, De La Salle University, Manila.

Teubert, Wolfgang, & Anna Cermáková. (2007). *Corpus linguistics: A short introduction*. London: Continuum.

Woodbury, Tony. (2003). Defining documentary linguistics. In Peter Austin (Ed.), *Language documentation and description 1*. London: Hans Rausing Endangered Languages Documentation Project, SOAS.

Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages

edited by Michael Haugh, Kate Burridge, Jean Mulder, and Pam Peters

Cascadilla Proceedings Project   Somerville, MA   2009