From archive to corpus

Transcription and annotation in the creation of signed language corpora*

Trevor Johnston
Macquarie University

Annotations are an important resource in corpus-based linguistic research. In fact, the most important feature of a modern signed language corpus should be that it has been annotated rather than simply transcribed. Digital multi-media annotation software can now transform language recordings into machine-readable texts using gloss-based annotations without it first being necessary to transcribe these utterances, provided that sign tokens are identified and discriminated according to type. Further annotations can subsequently be appended to these units. However, unique identifiers of sign types (or 'ID-glosses') can only be used if a comprehensive reference lexical database of the language already exists. In order to create a basic multi-purpose reference signed language corpus, therefore, linguists should prioritize annotation using ID-glosses above transcription. The effort expended in creating a transcription that does not facilitate the unique identification of sign types will not result in a machine-readable corpus in any meaningful sense, contrary to expectations.

Keywords: corpus linguistics, transcription, annotation, sign language, language documentation, Auslan (Australian Sign Language)

1. Introduction

The creation of signed language (henceforth SL) corpora — as modern linguistic 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 which can be used in transcription (i.e. the representation in some form of writing of what is being uttered). In previous research practice, superficial 'transcription' through glossing has proved problematic due to idiosyncratic variable practice and the fact that it gives little or no indication of form.

International Journal of Corpus Linguistics 15:1 (2010), **106–131.** DOI 10.1075/ijcl.15.1.05joh
ISSN 1384–6655 / E-ISSN 1569–9811 © John Benjamins Publishing Company

With few exceptions phonological or phonetic transcriptions have only been made on very small data-sets and then primarily to describe individual signs, rather than being used to transcribe extended utterances. SL corpora, therefore, need to be created taking these facts into account. Using the example of Auslan (Australian SL) 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, it is possible to create modern SL linguistic corpora which conform to the sense of corpus commonly understood in contemporary linguistics.

In this paper, I describe the resources and methodology required to create SL corpora that conform to the goals and practices of corpus linguistics. What is being claimed in this paper is that there are two guiding principles for the creation of modern SL linguistic corpora: first, prioritise annotation above transcription, and, second, identify signs uniquely using gloss-based annotations. Without these principles being implemented the central rationale for corpus-creation will certainly be compromised.

Before examining SL annotation in detail, I first review the main features of modern linguistic corpora with reference to SL corpora. This is followed by a description of the Auslan archive which is the source of the future Auslan corpus. After a discussion of SL corpus annotation, I conclude with a brief evaluation of other recent SL corpus projects elsewhere in the world.

2. Modern linguistic corpora and SL corpora

A modern linguistic corpus is something more than just a data-set 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 et al. 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 multi-media annotation software, corpora of SLs — once they have been created — could be described as sub-types of 'spoken' (or, better, 'face-to-face') language corpora. And, as their spoken language counterparts, they are just as equally in need of 'taming' (Beal et al. 2007a, 2007b).

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 morphology, lexis, syntax and pragmatics to discourse (Baker 2006, Halliday et al. 2004, McEnery et al. 2006, Sampson & McCarthy 2004, Sinclair 1991). There are several reasons why testing is particularly relevant in the field of SL linguistics. 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. 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. Despite the fact that research into SLs has grown dramatically over the past three to four decades, progress in the field has been hindered by these obstacles to data sharing and processing (cf., for example, Johnston & Schembri 2007).

As with all modern linguistic corpora, SL corpora should be representative (e.g. a collected set of texts should accurately reflect the language of an identified entity), well-documented (e.g. with relevant meta-data) and machine-readable (e.g. able to be searched and processed electronically) (McEnery & Wilson 2001, Meyer 2002, Teubert & Cermáková 2007). This requires dedicated technology (e.g. computers and software), standards and protocols (e.g. agreed metadata categories), and shared or at least transparent terminology (e.g. grammatical class labels) (Crasborn et al. 2007).

The guiding principle behind annotations used in the creation of modern SL linguistic corpora should be — and in the case of the Auslan corpus, is — machine-readability, not transcription narrowly understood. In the Auslan corpus, for example, the aim is to create an annotated SL corpus, and not, contrary to the practice of many SL researchers, 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 multi-media annotation software to transform a video recording of SL into a machine-readable text without it first being necessary to transcribe that text.

Using the methodology described in this paper, in conjunction with new multi-media 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. However, the use of multi-media annotation software can only succeed if signed units of the same type are consistently and uniquely identified before detailed linguistic annotations and tags are appended to them. In other words, each token of a type should have the same identifying gloss which is unique to that type. In order to do this one needs a reference lexical database that documents the lexical items (lexical types) of the language.

3. The Auslan corpus

The Auslan corpus is being built on a digital video archive of the language which has been deposited at the Endangered Languages Archive (ELAR) at SOAS (cf. note of acknowledgement). The archive consists of a representative sample of recordings in Auslan (for further details see Johnston & Schembri 2006). The corpus consists of these recordings linked to annotation and metadata files. Access will be initially limited for a period of three years from 2009 to 2011, after which it will be openly accessible, subject to the standard ELAR conditions of use.[1]

The corpus has been augmented with a second data-set which was collected as part of the Sociolinguistic Variation in Auslan research project (SVIAP).[2] Both data-sets are based on language recording sessions conducted with deaf native or early learner/near-native users of Auslan. A native signer is here defined as someone who has acquired Auslan from birth from a signing deaf parent or parents or an older deaf sibling, and an early learner/near-native as someone who has acquired or learned Auslan before the age of seven.

The Auslan corpus consists of approximately 300 hours of unedited footage taken from 100 participants from the five major cities in Australia (Sydney, Melbourne, Brisbane, Adelaide and Perth). 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. This footage has been edited down to around 150 hours of usable language production which, in turn, has been edited into approximately 1,100 separate digital movie texts for annotation. As at March 2009, approximately 201 of these texts have been annotated using ELAN (see Section 5). The supplementary SVIAP corpus consists of films of 211 participants from these same five cities on 140 hours of digital video footage of free conversation, structured interviews, and lexical sign elicitation tasks.[3]

The Auslan corpus annotations that have been created to date are intended primarily for investigations of grammar and discourse, rather than a basic phonological or lexical analysis of the language. The investigation centres on the modification of indicating verbs in terms of frequency of types/tokens, and their environments of occurrence (e.g. the presence or absence of contiguous indexical signs, or the sequential order of nominal arguments of the verb). The focus is on the analysis of the grammatical use of space in Auslan in terms of semantic roles and grammatical relations.[4]

4. Distinguishing between notation, transcription, annotation, and tagging

In order to appreciate the different degree and levels of detail 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 a distinction between notation, transcription, annotation, and tagging (cf. Johnston 1991). In the creation of the Auslan corpus these distinctions have proved to be very relevant in guiding how and why the data are encoded in a machine-readable text.

4.1 Notation and transcription

Although many scholars make no real distinction between notation and transcription, in this context it is useful to do so. 'Notation' tends to be used to refer to the actual system of graphic symbols used to represent or encode some phenomenon. In linguistics, this primarily refers to the symbols for writing down or representing the individual sounds of words, such as the International Phonetic Alphabet (IPA), or if referring to a bona fide writing system, an alphabet or script such as the Roman alphabet (Coulmas 1989, Sampson 1985). 'Transcription' usually refers to the graphic representation of an extended utterance in face-to-face or oral language, i.e. a text which has been uttered. It necessarily uses some kind of dedicated notation system (phonetic or phonological transcription) or script (orthographic transcription) (e.g. Tagliamonte 2007, MacWhinney 2007).

One of the major purposes of notation and transcription systems is to enable the reader of the graphic symbols to know, with greater or lesser accuracy according to the degree of detail in the system being used, the form of what was originally spoken or signed. 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).

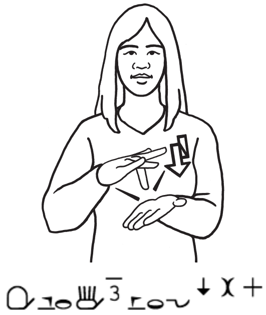


Figure 1. The Auslan sign CENTRE represented in HamNoSys

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 words or signs of a text can then also be further 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. In fact, 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 multi-media annotation software in the late twentieth century changed the situation completely, as it has enabled transcriptions to be directly time-aligned with recorded segments. By so doing *transcriptions* have been "demoted" to a type of *annotation* (see below). In other words, the text can remain the language recording itself, rather than being effectively replaced by its representation in a transcription to which annotations are only subsequently appended. This has implications for the way in which recordings of face-to-face languages can now be best processed in the creation of corpora for the purposes of linguistic analysis (cf. Beal et al. 2007a, 2007b). For example, with respect to SLs, one can now productively use glosses in a digital multi-media environment by exploiting the fact that glosses are "mere" annotations, rather than using them as second-best compromise transcriptions (see Section 7).

4.2 Annotation and tagging

One sense of annotation is any kind of commentary added to an already existing written text. Historically annotations were often found inserted into the margins of classical, learned or religious texts as an aid to understanding. Annotations were often commentaries in the reader's first language on texts which were in a foreign language (e.g. in Latin or Greek). They are still commonly found in publications of literary texts in language which is considered archaic or difficult, even if in the reader's first language (e.g. Shakespeare). Annotations can be added to any type of written text, be it a transcription of a spoken text or a piece of conventional writing (i.e. a text that did not necessarily previously exist as a spoken text or was never intended to become a spoken text by being written down).

For linguists, and especially corpus linguists, annotations have evolved into "mini" linguistic commentaries that are appended to identified units in a language. Annotations add phonological, morphological, syntactic, semantic, pragmatic and discourse information about linguistic forms, depending on the purpose of the analysis. As such, annotations are an invaluable aid in helping linguists discern patterns in language at many different levels, with or without the aid of computers.

In principle, 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 appended to written texts after they have been digitized and then processed using computers. For example, the addition of the word class or part of speech (POS) tags to the written English sentence *Joanna stubbed out her cigarette with unnecessary fierceness* can in a large part be done automatically by utilizing an electronic dictionary in conjunction with information on collocation and distributional patterns. Using the large databases of the most well-described and documented languages, such as English, this process is able to yield accuracy rates of up to 98% (Garside & Smith 1997). The process is illustrated in example (1) which is taken from the Lancaster-Oslo/Bergen Corpus of English (cited in McEnery & Wilson 2001: 47).

(1) Joanna_NP stubbed_VBD out_RP her_PP$ cigarette_NN with_IN
unnecessary_JJ fierceness_NN ._.

The POS tags suffixed to the lexical items use underscores and capitalization. To expand on but a few: *_NP* means singular proper noun, *_VBD* means past tense form of lexical verb, *_RP* means adverbial particle, and *_PP$* means possessive pronoun.

Most tags (or annotations) in the Auslan corpus are not appended to a gloss sequentially as in the above example; rather, they are inserted into annotation fields located on various tiers in the ELAN annotation file which are time-aligned to the ID-gloss annotations. These ID-glosses are in turn themselves time-aligned with the source media. As can be seen from Figure 2 (cf. Section 5.), these are displayed as "vertical" tags (see lower half of the screen shot), rather than "horizontal" tags which are prefixes or suffixes. Of course, annotations of recordings are time-stamped within the electronic database so the notion of spatial alignment is actually irrelevant, except in the sense that following this procedure does avoid creating glosses which are long, complicated, and difficult to read from the human point of view. This is not an unimportant consideration for reasons explained in the detailed discussion of ID-glosses in Section 6. This practice keeps the ID-glosses as simple and short as possible.

4.3 Metadata

'Metadata' refers to any additional and relevant information about a text or dataset which is essentially data about that data as a whole, rather than individual linguistic units within texts. Within linguistics that information is essentially either sociolinguistic in nature or describes the circumstances in which the data were collected. Sociolinguistic metadata appends information about characteristics of the participants such as age, sex, region, class, religion, education, ethnicity, race, dialect, and so on. Metadata on fieldwork sessions or the circumstances of data collection appends information about where and when the data were collected, under what circumstances (e.g. the type of tasks participants engaged in, the number of participants, etc.), and by whom (e.g. another native speaker, a person known or unknown to the participant, a researcher, and so on).

Accurate metadata of both types are essential in good corpus design but they are particularly important in SL corpora not only for reasons outlined in the introduction — SLs are young minority languages without written forms, and experience interrupted generational transmission — but also because of language contact issues and the variable range in, and different types of, hearing loss in the deaf community. Thus, in addition, SL-specific metadata such as the age of first exposure to SL (e.g. from birth, pre-school, at school) and from whom (e.g. deaf parents, deaf siblings, other deaf relatives, deaf peers at school, or teachers) should also be recorded (Crasborn et al. 2004).

4.4 Coding overall

In summary, it should be noted that regardless of the type or degree of detail in the coding or analysis, only behaviours that are (or are assumed to be) linguistically meaningful are identified in transcription and annotation. This means ignoring all articulations and movements that are not (or appear not to be) related to language.

With respect to SLs, for example, a hand scratching a nose or someone leaning forward to pick something up would be ignored, unless these acts are (or are assumed to be) part of a period of constructed action ('role shift'). There are, of course, other behaviours which are not clearly extra-linguistic, especially in SLs which are perforce face-to-face languages. For example, some behaviours may or may not be aspects of the linguistic system (e.g. eye-gaze, facial expressions, movement modifications, etc.) and they will need to be coded *as part of investigations to determine their role within a SL*. Coding for a particular feature of this type is usually based on a reasonable hypothesis about its grammatical function in the language. One must do this type of coding before extracting instances from the corpus to determine if a given hypothesis regarding the form and function of such a feature within the grammar is correct.

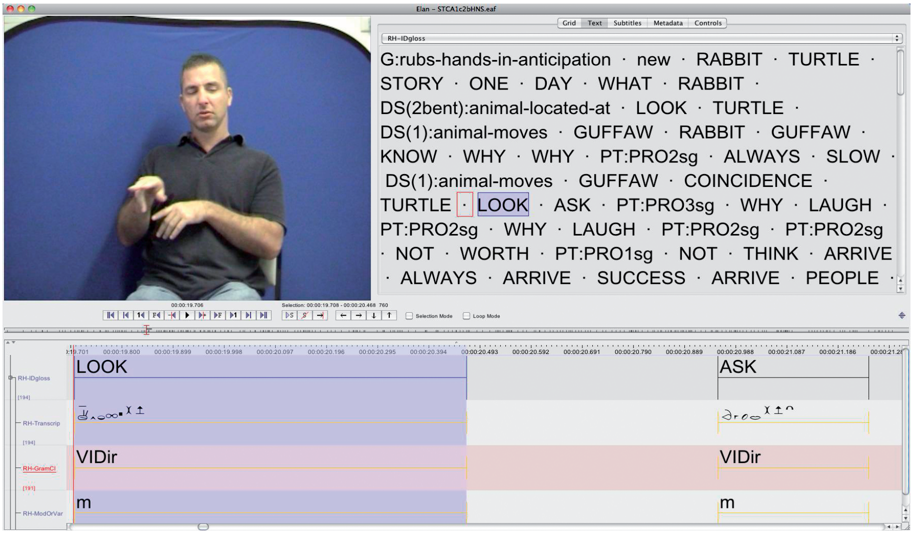
5. ELAN

The Auslan corpus is being annotated using digital video annotation software called ELAN (EUDICO — European Distributed Corpus-linguistic annotator) (MPI/LAT Technical Group 2009). 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; and export and import of annotations into or from several formats. Relevant metadata for the digital recordings are appended to media files which can be visualised within the ELAN software.

Figure 2 is a screen-shot of an opened ELAN annotation file showing the linked media file visible in the top right corner and an ID-gloss tier with several daughter tiers that exemplify the type of "vertical" tags discussed above. The ID-gloss annotation LOOK has itself been tagged with a broad phonetic transcription in HamNoSys of its citation form. LOOK has also been tagged with 'm' (for 'modified') on the *R-ModOrVar* (right hand modification/variant) tier and 'VIDir' (for grammatical class 'Directional Indicating Verb') on the *RH-GramCl* (right hand grammatical class) tier.

5.1 The tiers in ELAN

A small sample of the type of tiers used in the Auslan corpus ELAN annotation files is shown in Table 1.



G:rubs-hands-in-anticipation · new · 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 ·

Figure 2. A screen-shot from ELAN

Table 1. A few of the annotation tiers used the Auslan corpus ELAN template

Independent tier ⇒ daughter tier	Expansion and explanation of abbreviated name
RH-ID-gloss	The ID-gloss for the sign being produced on the right hand.*
⇒ RH-gram cls	A tag for grammatical class.
⇒ RH-mod	A tag for the presence or absence of sign modification.
⇒ RH-loc	A tag for the location a sign has been shifted to, if modified.
RH-mouthing	A gloss of the word being mouthed.
RH-mouth gest.	A tag for a mouth gesture (unrelated to English wording).
Clause	An annotation field that delimits the extent of a clause.
CA/roleshift	An annotation field that delimits the extent of constructed action.
free t/lation	A translation of an utterance unit (based on sense or prosody).

Note. * In SLs it is possible for each hand to articulate a different sign at the same time. For this and other reasons, two sets of tiers need to be specified for some types of sign annotations, one for the right hand (RH) and one for the left hand (LH). For brevity, the sample of tiers in this table only shows the RH tiers even though there is also a LH set.

There are two types of annotation that are absolutely minimally required to begin building a machine readable reference SL corpus: ID-glosses and a written free translation. Together, they give one a reference text which one may then value-add with further annotations. However, the ideal number and type of tiers in a standard ELAN annotation file ('template') for this or other SL corpora is yet to be determined. This is partly due to the fact that a certain amount of trial and error will be needed to determine what should be the most useful number and type of tiers for the majority of files in a SL corpus. Of course, additional, study-specific tiers can always be added at any time. Some of this experience has come from annotations focussing on various aspects of grammar in the Auslan corpus. However, some of this experience will also be derived from international cross-linguistic work and collaboration as new SL corpora are created around the world.

5.2 Annotation passes

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 passes of individual texts (digital movies). In an annotation pass one identifies sign units and/or attaches a particular type of linguistic annotation to already identified units. This information is placed on dedicated tiers using certain conventions, codes, or controlled vocabularies.[5] Thus, during an annotation pass an annotator will be looking at (and annotating) different aspects of sign structure or grammar on different tiers within the file.

Annotating usually begins with inserting 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 pass that looks at, and tags for, some particular linguistic feature. Over time repeated annotation passes make each annotation file — and the whole Auslan corpus — very detailed and a rich source of data for research. The process is represented in Figure 3.

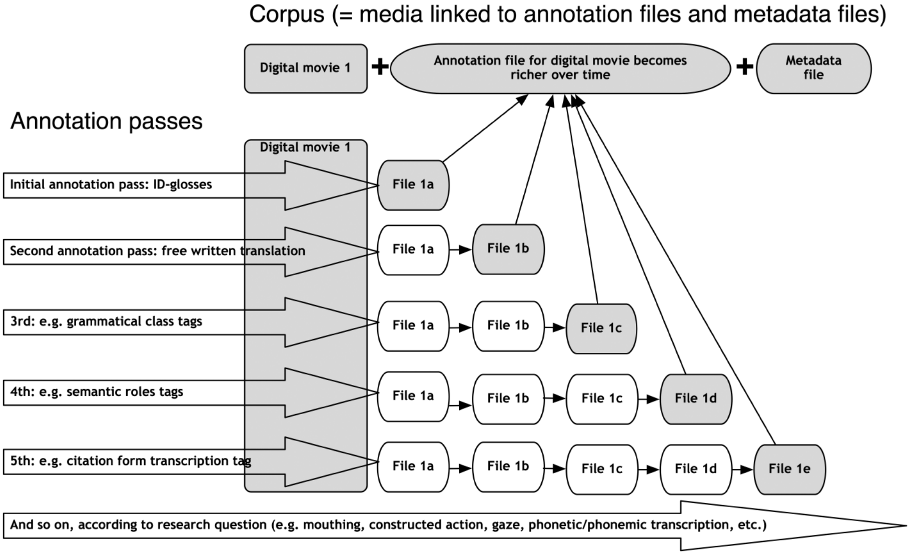

Corpus (= media linked to annotation files and metadata files)
Digital movie 1
+
Annotation file for digital movie becomes richer over time
+
Metadata file
Annotation passes
Digital movie 1
Initial annotation pass: ID-glosses
Second annotation pass: free written translation
3rd: e.g. grammatical class tags
4th: e.g. semantic roles tags
5th: e.g. citation form transcription tag
File 1a
File 1b
File 1c
File 1d
File 1e
And so on, according to research question (e.g. mouthing, constructed action, gaze, phonetic/phonemic transcription, etc.)

Figure 3. Example workflow for repeated annotation passes

5.3 From indeterminate to determinant in subsequent passes

One positive consequence of repeated annotation passes is that they encourage the use of tentative or generalized annotations (or tags) at times when fine-grained linguistic categorization would be premature in the absence of extensive data from the corpus itself. For example, the tags *pred* (meaning "a predicating element which may be a noun, verb, or adjective, but not any other grammatical class") or *norv* (meaning "a noun or a verb, but not any other grammatical class including adjective") are used as interim indeterminate tags. This avoids the need to make a more specific annotation which may force a premature choice between noun, verb or adjective — the final decision on the categorization may not be possible until hundreds of annotation files have been created and thousands of examples are available for comparison. The indeterminate tag at least reduces the set of signs which must be revisited on a subsequent annotation pass for reconsideration and determination.

6. Creating a machine-readable text 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 delimitation of segments within a recording and the time stamping of signed units in a recording (i.e. tokenization) is precisely what modern digital multi-media 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 multi-media 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 which can be researched: the form is ever present in the linked and aligned media. With the new technologies transcription need no longer be seen as a (necessary) substitute for the ephemeral utterance (or a stand-alone recording), but as a type of annotation appended to the digitized media of the linguistic event. Conversely, annotations can also be "transcriptions": they can be used to append form descriptions to identified linguistic segments or behaviours of any size. This reality and the possibilities it presents in Spoken Language (henceforth SpL) corpus-based research have been reported in the literature (e.g. Barbiers et al. 2007, MacWhinney 2007). It has even become apparent in recent discussions of the urgency surrounding the documentation of endangered SpLs (Woodbury 2003; Simons 2008).

These possibilities are especially relevant in their potential to transform the conduct of SL research. Somewhat surprisingly, transcribing a signed recording is still usually regarded as the first and necessary step in creating a multi-media SL corpus — even in the "old-fashioned" sense of simply a data-set — for linguistic research. Although the time alignment of transcriptions and other annotations represents a vast improvement on previous practice where the primary data were virtually inaccessible, this procedure usually entails scores of hours of notation and transcription *per minute of video recording* without producing as output a text which is machine-readable in any meaningful sense as understood by corpus linguistics. By failing to use appropriately the potential of the new technology, this represents both an unproductive use of limited resources and a failure to understand the nature of modern linguistic corpora.

In other words, in multi-media environments phonetic or phonological transcriptions of signed recordings are not necessary in order to create "true" reference texts for research at various levels of linguistic analysis. Actually, time-aligned multi-media gloss-based annotations are also useful in phonetic or phonological research because one can simply append a relevant phonetic or phonological tag for a feature under investigation to a gloss-based annotation, so there is no need for a *transcription* as the reference text, as such. Of course, the use of dedicated transcription (i.e. phonetic or phonological annotation) within ELAN or a similar multi-media annotation program would be necessary in order to carry out detailed phonetic or phonological research on the data. The coding of phonetic or phonological form may be done on a single "transcription" tier or on multiple tiers (each for different aspects of phonetic or phonological form) and it may or may not use a dedicated notation system, such as HamNoSys for SLs or IPA for SpLs. Similar approaches have been adopted in the creation of unconventional SpL corpora (Beal et al. 2007a, Anderwald & Wagner 2007).

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 standardized, widely accepted notation system like IPA, it would be more productive in the first instance to create base level annotations that identify the sign units in the text by using glosses.

Once the source text has been segmented into sign units (i.e. tokenized), the assignment of a unique gloss-based identifier relies on exploiting all known lexical information about the language as well as following standard protocols for glossing partly-lexical signs (described in Section 7) which need to be treated differently.

Indeed, such information could include a transcription of the citation form using a notation system, if it was available (see Figure 2). This is true of entries in Auslan, New Zealand SL and German SL dictionaries (using HamNoSys). However, accessing and entering these data from databases to the annotation program is still easier to do through the mediation of glosses.[6]

6.1 Gloss, ID-gloss and translation

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" is essentially relatively crude and simplistic. In the Auslan corpus, the glossing language is English.

Glosses are used in running text in the sign language linguistics literature (as in *the British Sign Language sign* SISTER *is identical to the Auslan sign* SISTER *but completely different to the American Sign Language sign* SISTER). It is the convention to write glosses in upper case. Importantly, different glosses for the same sign may be 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 usually contain any information about sign form.

In contrast, there needs to be a level in corpus annotation where signs are identified uniquely and consistently. A gloss of this type makes it is possible to search through multiple annotation files and find all instances of a particular sign in order to determine the ways and environments in which it is used. One cannot productively use a corpus in this way with ad-hoc glosses. I call this type of identifying gloss an *ID-gloss* (Johnston 2001).

An ID-gloss is the (English) word that is consistently used 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. For example, if a person signs HOUSE (in Auslan a sign iconically related to the shape of a roof and walls) but actually means *home*; or performs a particularly large and exaggerated form of the sign HOUSE, implying *mansion*, without that modified form itself being a conventionalized lexical item, the same ID-gloss HOUSE is used to identify the sign in both cases. Of course, corpus-based evidence could itself lead to the re-analysis, and hence re-glossing, of a sign (see the discussion of homonyms and pointing signs in Sections 7.1.2 and 7.2 respectively).

With respect to distinguishing between glossing and translation, meaning is assigned to the text through glossing indirectly through the unavoidable fact that the ID-gloss, which is primarily intended to identify a sign, actually uses an English word that bears a relationship to the meaning of the sign. In other words, the

ID-gloss is not chosen arbitrarily or capriciously because the choice of the English word is highly motivated. However, the ID-gloss is still not intended as a translation. Translations are made on their own dedicated tiers in the ELAN annotation files. So, if the signer produces SUCCESS but means "achieve something", it is still annotated with the ID-gloss SUCCESS; and if a person signs IMPORTANT to signify "main", "importance", "importantly", "primary" or "initial" it is still labelled as IMPORTANT. Annotations and tags on other tiers will specify the grammatical class of the sign (noun, verb, adverb, etc.), the presence or absence of mouthing (the simultaneous silent mouthing of related English words such as "main" or "primary"), or its actual meaning in context (e.g. on a translation tier).

6.2 ID-glossing and lemmatisation

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 passes (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). The most common word from the glossing language associated with the simple unmodified citation form of a lexical sign serves as the basis for creating an ID-gloss. 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 lemma or citation form is the form that normally appears as the headword or head sign in a dictionary entry.

There is no principled reason why a broad phonetic or phonemic transcription of each sign's *citation* form could not serve as the unique identifier of sign type. Such a system would also be *de facto* lemmatization (or even pseudo-orthographic transcription) in so far as it would only approximate the actual production found in the text. Indeed, any unique identifying system could be used to identify signs, e.g. arbitrary strings of numbers, symbols, or characters. However, at this time ID-glosses are to be preferred at a practical level as the quickest and most user-friendly way to build a basic reference corpus for most SLs.[7]

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 written or transcribed text are annotated as one lemma, whereas in the latter I suggest there need be no pre-existing written representation or transcription which is then in turn lemmatized. In short, contrary to the assumption that one is obliged to create one, I am advocating here that one go directly to the lemmatization stage in building a basic multipurpose SL corpus. Not only will this prove to be 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 multi-media 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 data-set 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. It would be impossible to use the corpus productively and much of the time spent on annotation would effectively be wasted because the corpus would never become machine-readable in any meaningful sense. The result would not be the type of corpus that linguists aspire to today; rather, it would just be a collection of reference texts — a "corpus" in what is rapidly becoming a superseded sense in the literature.

7. The annotation glosses for 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 SpLs — the conventionalized units of a SL can be divided into 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 SpLs, 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.

Partly-lexical signs do not have associated with them a meaning which is additional to, or unpredictable from, the meaning derived from its combined components when the sign is produced and used in various contexts. There is essentially nothing that could be further specified about the sign's meaning were they to be entered in a dictionary. 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 sign/word or a function sign/word. *Fully-lexical signs* constitute the listable lexicon of a signed language.[8]

7.1 Fully-lexical signs and ID-glosses

Fully-lexical signs are identified using an ID-gloss. In the annotation fields created in ELAN that contain the ID-glosses, the glosses are written in upper case, as is the norm for glossing in SL linguistics. Linguists generally only use capitalised glosses for grammatical morphemes or function words in the interlinear glossing of language examples, as in the following (in the example the source and glossing language are both English — of course, they are usually not the same language):

(2) Source language: He walked home
Glossing language: PRO3.SG.MASC walk-PAST home

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. It 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 in order to distinguish the lemma from the surrounding word forms in the text which is usually in the same language. We continue this practice in the ELAN annotation files. Thus the ID-gloss HOUSE appears on an ID-gloss tier as:

(3) | HOUSE |

(As seen in Figure 2, the boxed annotation field delimits a period of time in the digital media during which a sign is articulated and to which the annotation within the field is time-aligned.)

7.1.1 *Choosing the appropriate ID-gloss*

The standard ID-gloss for a sign is found by consulting the Auslan lexical database. The database contains over 6,600 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 lexical status of each head sign. One field contains a broad phonetic transcription in HamNoSys. Meaning fields include several for definitions, semantic domains, and synonyms and antonyms. Lexical status fields include several for dialect, register, and stem/variant identification. The database lists a citation form of a fully-lexical sign as a major stem entry, with common variant forms listed separately. A public view of the database can be accessed online through *Auslan Signbank* (www.auslan.org.au). Annotators log in to a special researchers' reference view which includes much more information than the public view (including the ID-gloss), as well as many more additional signs (e.g. variant signs and newly identified signs).

Signs can be accessed by searching for any English word which may be commonly associated with a sign form (known as a *keyword* in the database). For example, the sign IMPORTANT could be found by searching under the keywords *important*, *importance*, *main*, or *primary*, all of which are possible meanings or translations of the sign IMPORTANT in various contexts. In addition, entries for signs in the database are ordered formationally, i.e. they are sequenced according to major phonological features of signs, such as handshape and location, so that scrolling through the database records displays formationally similar signs one after the other. This is useful for an annotator who cannot find a particular sign because there is no gloss or keyword match to their initial enquiry (or at least one that is not expected and, hence, not queried by the annotator). In other words, an annotator is able to locate a sign with a similar form whose gloss or keyword is known or matches, and then manually search around that sign to see if the form they have seen in a text is recorded in the database despite there having been no initial gloss or keyword match (i.e. it may be entered under an unexpected gloss or keyword).

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 are 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. The lexical database and its representation in dictionaries in various forms is thus a necessary prerequisite for the creation of a viable SL corpus. Of course, a reference lexical database need not be exhaustive and it is almost certain that a corpus will enable the identification of hitherto unrecorded lexical signs — or even unrecorded senses of already identified signs — which will, in turn, be added to the database.

7.1.2 *ID-glosses and homonyms*

A single sign form can have two entries and two separate ID-glosses if it has been determined that two separate signs exist which are homonyms. The only time an existing sign form will be assigned a different ID-gloss than that which is recorded in the lexical database is when corpus data justify the identification of a completely distinct and unrelated meaning for the sign form in question. In such cases, the sign form receives its own distinctive ID-gloss and the two signs are treated as homonyms. The corpus and database managers then update the lexical database to create a new sign entry.

7.1.3 *Annotation conventions for various other sub-types of fully-lexical signs*

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 use of one or two hands in normally two-handed or one-handed signs respectively, and borrowings from Signed English and other SLs (annotation conventions are downloadable from http://www.auslan.org.au/). 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. not WON'T but WILL-NOT, not DON'T-LIKE or NOT-WANT but LIKE-NOT and WANT-NOT because all these signs are part of a set in Auslan that end 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.

7.2 Annotation conventions for partly-lexical signs

Unlike content *and* function signs which are *lexical* signs, the assignment of ID-glosses to *partly-lexical* signs is not at all straightforward (one cannot simply refer to a lexical database and extract the ID-gloss). There is no citation form or lemma. However, by following a relatively small set of annotation and glossing conventions one can ensure that tokens of sub-types 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.

Instead of using standard identifying glosses unique to a type for each token of a partly-lexical sign in the corpus — as with lexical signs — these tokens are glossed using a combination of general and idiosyncratic elements. This simple convention makes it possible to search for all instances of a sub-type 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. To do this one simply uses sub-string match queries to search for the more general elements found in the glosses of partly-lexical signs in the corpus. The general elements are prefixes of some sort, as explained below.

One sub-type, depicting signs, are prototypical partly-lexical signs and behave essentially as these signs are described above. Other sub-types include pointing (or index) signs, and buoys. (A buoy is a sign that helps track referents in discourse. It usually consists of a sign being produced on the subordinate hand that is held in space as the dominant hand continues to produce other signs.[9]) In addition, conventions need to be followed for glossing other sub-types of partly-lexical "signs", such as fingerspelling and gestures (described below).

Annotation glosses for each of these types of signs begins with a fixed string that identifies the sub-type: 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 (e.g. codes for marked or variant handshapes) and meaning of the sign in that context. For example, a depicting sign representing a piece of paper blowing off a table could be annotated broadly (preferred) as *DS(B):FLAT-SURFACE-LIFTS-AND-TURNS-AWAY* or narrowly as *DS(B):SHEET-OF-PAPER-BLOWS-AWAY*, and a stretch of fingerspelling for the word *electrode* would be annotated as *FS:ELECTRODE*. This allows one to incorporate consistent codes in annotations for these types of signs while at the same time coding the uniqueness of each token using sign-specific glossing elsewhere in the annotation. The set of annotations can thus still be easily read, sorted or otherwise processed by computer.

Lexicalized pointing signs are assigned an ID-gloss, e.g. 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 majority of pointing signs in Auslan, or most SLs, are not lexicalised in this way — they essentially remain pointing gestures, 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 to such signs. This is why the PT prefixing convention for these is used.

Each individual PT annotation can include further levels of specification for sub-type of pointing sign, where contextual evidence makes this clear, e.g. PRO for "pronoun", LOC for "locative", POSS for "possessive", DEM for "demonstrative". Thus an index sign that unambiguously points to a location in the signing space and signifies that location would be glossed as PT:LOC. Further specification for number, person and proximity, can also be added, e.g. PT:PRO3PL would annotate an index sign that points at "third person entities" in the discourse situation. However, generally speaking, it is often difficult to make a more precise grammatical annotation, beyond identifying a PT, on first analysis. Additional specification can thus be deferred to a later annotation pass. The ability to search for, identify and sort *all* these types of signs, based on sub-string matches for prefixes within glosses, enables subsequent analysis (or re-analysis), categorization (or re-categorization), and labelling (or re-labelling) as appropriate. Ultimately, it is the corpus evidence itself that will determine the "final" identification of the types and functions of pointing signs.[10]

8. Annotation conventions for gesture

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 which have not become lexical Auslan signs will not be found in the lexical database and will thus not have an assignable ID-gloss.

When annotated the gloss for a gesture is prefixed with G for "gesture" followed by a brief description of the meaning of the gesture, e.g. 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. When hundreds of annotation files have been created and a large number of examples are available for comparison, some of these gestures may be identified as having subtly distinct forms and/or specific functions that may justify re-categorisation and re-glossing. This is one of the great advantages of using a corpus as part of empirical language description, but in order to do so, it requires that annotators are as consistent as possible in glossing — using ID-glosses for lexical signs and prefixing general type labels to the glosses for partly-lexical signs and gestures.

9. 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. Netherlands SL and British SL), are about to begin (e.g. German SL and Swedish SL), or are planned (e.g. American SL). The Netherlands SL 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. The corpus also includes over ten hours of gloss annotations, although these are not ID-glosses as described in this paper.

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 which 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 that do not identify sign types — does not add value to the archive in ways that other corpus linguists have come to 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 as the first step in creating a basic multi-purpose machine-readable SL corpus.

Notes

* The Auslan archive and corpus was created through support from (i) a Hans Rausing Endangered Languages Documentation Project grant (School of Oriental and African Studies, University of London) awarded to Trevor Johnston, Chief Investigator (#MDP0088, University of Newcastle, Australia); (ii) an Australian Research Council (ARC) grant for a sociolinguistic variation study of Auslan awarded to Trevor Johnston, Chief Investigator, and Adam Schembri, Postdoctoral Research Fellow (#LP09346973, University of Newcastle); and (iii) an ARC grant for a corpus-based study of the linguistic use of space in Auslan awarded to Trevor Johnston, Chief Investigator, and Louise de Beuzeville, Postdoctoral Research Fellow (#DP0665254, Macquarie University). I also wish to acknowledge the 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. Finally, I'd like to thank the two anonymous reviewers who provided very helpful comments which helped improve this paper.

[1] Requests for access to the corpus before the end of the limited access period will be considered on a case by case basis and should be directed to ELAR: http://www.hrelp.org/archive/.

[2] Australian Research Council research grant awarded to Trevor Johnston & Adam Schembri — #LP0346973 "Sociolinguistic Variation in Auslan: Theoretical and Applied Dimensions".

[3] Access to the SVIAP data is subject to separate access restrictions than the ELDP data and requests for access should be directed to either Trevor Johnston or Adam Schembri. Contact details for Adam Schembri: Project Director, British Sign Language Corpus Project, Deafness, Cognition and Language (DCAL) Research Centre, University College London, 49 Gordon Square, London WC1H 0PD, United Kingdom.

[4] An Australian Research Council project grant awarded to Louise de Beuzeville and Trevor Johnston — #DP0665254 "The Linguistic Use of Space in Auslan: Semantic Roles and Grammatical Relations in Three Dimensions". For initial data on indicating verbs see Johnston et al. (2007) and de Beuzeville et al. (2009).

[5] For further details of annotations, tags and controlled vocabularies used in the Auslan corpus please consult the annotation guidelines downloadable from http://www.auslan.org.au/. The guidelines do not attempt to set specific annotation protocols for all signed language corpora. Provided each signed language corpus is internally consistent in its annotation conventions, second-order cross-linguistic comparisons can fruitfully be made after language-internal analyses have been conducted.

[6] The HamNoSys strings can be cut and pasted (manually or semi-automatically) into the appropriate annotation fields. However, though it is possible to display the HamNoSys characters in ELAN, it is difficult and time-consuming to directly input transcriptions using a keyboard, and HamNoSys strings cannot be easily searched or sorted. iLex, a dedicated lexical database software for use with HamNoSys, can be used to manage data but to date it is not interoperable with ELAN. For a description visit http://www.sign-lang.uni-hamburg.de/ilex/.

[7] See also note 6. ID-glosses based on the majority language vocabulary are much easier for researchers to use for searching within ELAN or related lexical databases. For both deaf and hearing researchers, certain words of the majority language are often very strongly associated with individual signs. It would appear that the primary usefulness of ID-glosses tagged with broad phonetic citation form notations may actually be as templates for narrower transcriptions of actual production. The researcher would only need to modify the transcription, rather than write an entirely new one.

[8] Another terminology should be developed for describing the conventional signs of signed languages with respect to form/meaning pairings at the level of individual sign parameters (and whether these parameters are each fully specifiable) and at the level of the sign itself and its degree of lexicalization. For example, it would appear that a construction grammar approach and terminology (Croft 2001, Goldberg 2006) would be more appropriate to describe this lexical cline in signed languages (i.e. as constructions that vary continuously along the two dimensions of the atomic-to-complex and the substantive-to-schematic).

[9] For a detailed discussion of depicting signs and buoys see Liddell (2003).

10. The identification of grammatical classes in Auslan (and possibly many other signed languages) has, to date, proved problematic both in regard to the number and type of classes found in the language and in the assignment of particular signs to these classes in actual texts. An important role for corpus-based SL linguistics is to test the applicability, and universality, of these class labels and/or to provide evidence for alternative categories based on language-specific and language-internal data.

References

Anderwald, L. & Wagner, S. 2007. "FRED — The Freiburg English dialect corpus: Applying corpus-linguistic research tools to the analysis of dialect data". In J. C. Beal, K. P. Corrigan & H. L. Moisl (Eds.), *Creating and Digitizing Language Corpora Volume 1: Synchronic Databases*. New York: Palgrave Macmillian, 35–53.

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Barbiers, S., Cornips, L. & Kunst, J.-P. 2007. "The syntactic atlas of the Dutch dialects (SAND): A corpus of elicited speech as an on-line dynamic atlas". In J. C. Beal, K. P. Corrigan & H. L. Moisl (Eds.), *Creating and Digitizing Language Corpora Volume 1: Synchronic Databases*. New York: Palgrave Macmillian, 54–90.

Beal, J. C., Corrigan, K. P. & Moisl, H. L. 2007a. "Taming digital voices and texts: Models and methods for handling unconventional synchronic corpora". In J. C. Beal, K. P. Corrigan & H. L. Moisl (Eds.), *Creating and Digitizing Language Corpora Volume 1: Synchronic Databases*. New York: Palgrave Macmillian, 1–16.

Beal, J. C., Corrigan, K. P. & Moisl, H. L. 2007b. "Taming digital voices and texts: Models and methods for handling unconventional diachronic corpora". In J. C. Beal, K. P. Corrigan & H. L. Moisl (Eds.), *Creating and Digitizing Language Corpora Volume 2: Diachronic Databases*. New York: Palgrave Macmillian, 1–15.

Coulmas, F. 1989. *The Writing Systems of the World*. Oxford: Basil Blackwell

Crasborn, O., van der Kooij, E., Broeder, D. & Brugman, H. 2004. "Sharing sign language corpora online: Proposals for transcription and metadata categories". In O. Streiter & C. Vettori (Eds.), *Proceedings of the LREC (Language Resources and Evaluation) 2004 Satellite Workshop on Representation and Processing of Sign Languages.* Paris: ELDA, 20–23.

Crasborn, O., Mesch, J., Waters, D., Nonhebel, A., van der Kooij, E., Woll, B. & Bergman, B. 2007. "Sharing sign language data online: Experiences from the ECHO project". *International Journal of Corpus Linguistics*, 12 (4), 535–562.

Croft, W. 2001. *Radical Construction Grammar*. Oxford: Oxford University Press.

De Beuzeville, L., Johnston, T. & Schembri, A. 2009. "The use of space with lexical verbs in Auslan: A corpus-based investigation". *Sign Language & Linguistics*, 12 (1), 53–82.

Garside, R. & Smith, N. 1997. "A hybrid grammatical tagger: CLAWS4". In Garside, R., Leech, G. & McEnery, A. (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora.* London: Longman, 102–121.

Goldberg, A. E. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Halliday, M. A. K., Teubert, W., Yallop, C. & Cermakova, A. 2004. *Lexicology and Corpus Linguistics*. London: Continuum.

Hoey, M., Mahlberg, M., Stubbs, M. & Teubert, W. 2007. *Text, Discourse and Corpora: Theory and Analysis*. London: Continuum.

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

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

Johnston, T. & Schembri, A. 1999. "On defining lexeme in a sign language". *Sign Language & Linguistics*, 2 (1), 115–185.

Johnston, T. & Schembri, A. 2006. "Issues in the creation of a digital archive of a signed language". In L. Barwick & N. Thieberger (Eds.), *Sustainable Data from Digital Fieldwork: Proceedings of the Conference Held at the University of Sydney, 4–6 December 2006*. Sydney: Sydney University Press, 7–16.

Johnston, T. & Schembri, A. 2007. *Australian Sign Language (Auslan): An Introduction to Sign Language Linguistics*. Cambridge: Cambridge University Press.

Johnston, T., de Beuzeville, L., Schembri, A. & Goswell, D. 2007. "On not missing the point: Indicating verbs in Auslan". Paper presented at the *10th International Cognitive Linguistics Conference, Kraków, Poland (15–20 July)*.

Liddell, S. K. 2003. *Grammar, Gesture, and Meaning in American Sign Language*. Cambridge: Cambridge University Press.

MacWhinney, B. 2007. "The Talkbank project". In J. C. Beal, K. P. Corrigan & H. L. Moisl (Eds.), *Creating and Digitizing Language Corpora Volume 1: Synchronic Databases*. New York: Palgrave Macmillian, 163–180.

McEnery, T. & Wilson, A. 2001. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R. & Tono, Y. (Eds.) 2006. *Corpus-Based Language Studies*. London / New York: Routledge.

Meyer, C. F. 2002. *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

MPI/LAT Technical Group: (Head) Wittenburg, P., (Team members) Auer, E., Broeder, D., Gardellini, M., Kemps-Snijders, M. et al. 2009. *EUDICO Linguistic Annotator (ELAN)* (Version 3.8). Nijmegen, Netherlands: Max Plank Institute for Psycholinguistics: Technical Group (Language Archiving Technology). Available at: http://www.lat-mpi.eu/tools/elan/

Prillwitz, S. & Zienert, H. 1990. "Hamburg Notation System for sign language: Development of a sign writing with computer application". In S. Prillwitz & T. Vollhaber (Eds.), *Current Trends in European Sign Language Research: Proceedings of the 3rd European Congress on Sign Language Research Hamburg July 26–29, 1989* Hamburg: Signum Verlag, 355–379.

Sampson, G. 1985. *Writing Systems: A Linguistic Introduction*. London: Hutchinson.

Sampson, G. & McCarthy, D. (Eds.) 2004. *Corpus Linguistics: Readings in a Widening Discipline*. London: Continuum.

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

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Teubert, W. & Cermáková, A. 2007. *Corpus Linguistics: A Short Introduction*. London: Continuum.

Tagliamonte, S. A. 2007. "Representing real language: Consistency, trade-offs and thinking ahead!". In J. C. Beal, K. P. Corrigan & H. L. Moisl (Eds.), *Creating and Digitizing Language Corpora Volume 1: Synchronic Databases*. New York: Palgrave Macmillian, 205–240.

Woodbury, A. C. 2003. "Defining documentary linguistics". In P. Austin (Ed.), *Language Documentation and Description, Volume 1*. London: Hans Rausing Endangered Languages Documentation Project, SOAS, 35–51.

Author's address

Trevor Johnston
Department of Linguistics
Macquarie University
Balaclava Road, Nth Ryde,
NSW, Australia, 2109.

trevor.johnston@mq.edu.au