From archive to corpus: transcription and annotation in the creation of signed language corpora[*]

Trevor Johnston
Department of Linguistics, Macquarie University, Sydney, Australia
trevor.johnston@mq.edu.au

Abstract. The essential characteristic of a signed language corpus is that it has been annotated, and not, contrary to the practice of many signed language researchers, that it has been transcribed. Annotations are necessary for corpus-based investigations of signed or spoken languages. Multi-media annotation software can now be used to transform a recording into a machine-readable text without it first being necessary to transcribe the text, provided that linguistic units are uniquely identified and annotations subsequently appended to these units. These unique identifiers are here referred to as *ID-glosses*. The use of ID-glosses is only possible if a reference lexical database (i.e., dictionary) exists as the result of prior foundation research into the lexicon. In short, the creators of signed language corpora should prioritize annotation above transcription, and ensure that signs are identified using unique gloss-based annotations. Without this the whole rationale for corpus-creation is undermined.

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

1 Introduction

A modern linguistic corpus is something more than just a reference dataset of written or transcribed texts of a particular language on which a description of a language is based. If anything, this is an old fashioned sense of *corpus*. Rather, a corpus in the modern sense is a collection of spoken and spoken texts *in a machine-readable form* that has been assembled for the purposes of studying the type and frequency of lexical items and grammatical structures and constructions in a language (McEnery and Wilson, 2001). 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, signed language corpora can now be described as sub-types of 'spoken' (i.e., 'face-to-face') language corpora.

Signed language corpora promise to vastly improve peer review of descriptions of signed languages and make possible, for the first time, a corpus-based approach to signed language analysis. Corpora are important for the testing of language hypotheses in all language research at all levels, from phonology, through lexis, morphology and syntax to discourse (Baker, 2006; Halliday *et al.*, 2004; Hoey *et al.*, 2007; McEnery *et al.*, 2006; Sampson and McCarthy, 2004; Sinclair 1991). There are several reasons why this is especially true of deaf signing communities. First, signed languages—inevitably young minority language communities—lack written forms

[*] Acknowledgments: research towards this paper was supported by a Hans Rausing Endangered Languages Documentation Project (University of London) grant #MDP0088 to Trevor Johnston. Additional archived corpus material was contributed through Australian Research Council (ARC) grant #LP09346973 to Adam Schembri and Trevor Johnston. Additional corpus annotation were partially supported by Australian Research Council grant #DP0665254 to Trevor Johnston and Louise de Beuzeville. I also wish acknowledge the support of the Australian deaf community native signer participants who have contributed to the Auslan archive and corpus and research assistants and annotators who contributed to the current body of annotations: Julia Allen, Donovan Cresdee, Karin Banna, Michael Gray, and Gerry Shearim.

and well developed community-based standards of correctness. Second, they have interrupted generational transmission and few native speakers. Third, the representation of signed language examples using written gloss-based text has meant that these data have remained essentially inaccessible to other researchers for meaningful peer review. Thus, although introspection and observation can still be of valuable assistance to linguists developing hypotheses regarding signed language 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 signed languages 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.

As with all modern linguistic corpora, signed language corpora should be representative, well-documented (i.e., with relevant metadata) and machine-readable (i.e., able to be annotated and tagged consistently and systematically) (McEnery and Wilson, 1996; Meyer, 2002; Teubert and Cermáková, 2007). This requires dedicated technology (e.g., ELAN), standards and protocols (e.g., IMDI metadata descriptors), and transparent and agreed grammatical tags (e.g., grammatical class labels) (Crasborn *et al*, 2007). One aim of this paper is to describe these resources and to identify the principles that need to be adhered to in the creation of signed language corpora such that the goals and practices of corpus linguistics, as now generally understood, can be achieved and implemented with respect to signed languages.

The guiding principle behind the annotations being created for the Auslan (Australian Sign Language) corpus is machine-readability, not transcription narrowly understood. The aim is to create an annotated signed language corpus, and not, contrary to the practice of many signed language researchers, a body of signed language 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 signed language into a machine-readable text without it first being necessary to transcribe that text. Transcription is defined here as the encoding of face-to-face language (signed or spoken) using a recognized notation system that represents the phonetic or phonological form of the signal, or using a dedicated writing script that represents the conventional units of the language. This is an important consideration in building signed language corpora because there is no standard or widely accepted signed language transcription system. Using this type of multi-media annotation software it is thus now possible to gain instant and unambiguous access to the actual form of the signs being annotated—the video recording—because they are both time aligned. In saying this, it should be noted, however, that this type of multi-media annotation software can only profitably be used to create a machine-readable corpus if signed units are consistently and uniquely identified before more detailed linguistic annotations and tags are appended to them. A secondary aim of this paper is to describe how this can be achieved, using the example of the Auslan corpus.

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., during periods of constructed action, with or without contiguous pointing signs, or with reference to the sequential order of related nominal arguments). The focus is on the analysis of the grammatical use of space in Auslan in terms of semantic roles and grammatical relations.[1]

[1] 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* (submitted). For details of the controlled vocabularies and codes (tags) used in these annotated texts the reader is referred to annotation guidelines produced as part of the above mentioned project (see Johnston and de Beuzeville, 2008).). 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 conven-

What is being claimed in this paper is that there are two principles which all signed language corpora should adhere to in order to facilitate their optimal use: prioritise annotation above transcription, and identify signs uniquely using gloss-based annotations. Without this the whole rationale for corpus-creation is undermined.

2 The Auslan Corpus

The Auslan Corpus is a digital video archive of Australian Sign Language (Auslan). The archive is an Endangered Languages Documentation Project funded through the Hans Rausing Endangered Languages Documentation Programme at the School of Oriental and African Studies (SOAS), University of London (grant #MDP0088 awarded to Trevor Johnston). The corpus was deposited during 2008 at the Endangered Languages Archive (ELAR) at SOAS. 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.[2]

The corpus brings together into one digital archive a representative sample of Auslan in video recordings to which are appended annotation and metadata files. It consists of two sub-corpora: data collected through the Endangered Languages Documentation Project (ELDP), mentioned above, and data collected as part of the Sociolinguistic Variation in Auslan Project (SVIAP).[3] Both datasets 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 (Johnston and Schembri, 2006).

The ELDP corpus consists of approximately 300 hours of unedited footage taken from 100 participants from the same five 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. 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. To date approximately 130 of these texts have been annotated using ELAN (see below for more details).

The SVIAP corpus consists of films of 211 participants from the five major cities in Australia (Sydney, Melbourne, Brisbane, Adelaide and Perth). This yielded over 140 hours of unedited digital video footage of free conversation, structured interviews, and lexical sign elicitation tasks.[4]

3 Distinguishing between notation, transcription, annotation, tagging and metadata

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 distinction between *notation*, *transcription*, *annotation* and *tagging* (cf. Johnston, 1991a). In the creation of the Auslan corpus these distinctions

tions, second-order cross-linguistic comparisons can fruitfully be made after language-internal analyses have been conducted.

[2] 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/.

[3] Australian Research Council research grant awarded to Adam Schembri and Trevor Johnston — #LP0346973 *Sociolinguistic Variation in Auslan: Theoretical and applied dimensions.*

[4] Access to the SVIAP data to subject to separate access restrictions than the ELDP data and requests for access should be directed to either Trevor Johnston or Adam Schembri. Contact 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.

have proved to be very relevant in guiding how and why the data is encoded in a machine-readable text.

3.1 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, i.e., 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 represented in a dedicated signed language notations system (HamNoSys[5]).

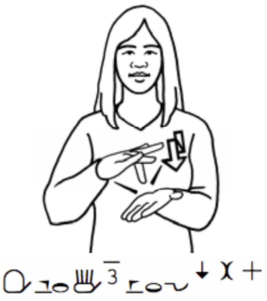


Figure 1: The Auslan sign CENTRE represented in HamNoSys.

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 isolated, 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. The development of digital recording and multi-media annotation software in the late twentieth century changed the situation further, as it now enabled annotations to be directly time-aligned with recorded segments. This has become especially relevant in transforming the conduct of signed language research. Somewhat surprisingly, it is still often assumed by signed language researchers that transcribing a signed text is a first and necessary step in the creation of a signed language corpus. Given that this usually entail scores of hours of notation and transcription *per minute of video recording* without producing as output a useable, machine-readable text, this practice not only represents a significant waste of resources by failing to use appropriately the potential of the new technology, it also represents a fundamental misunderstanding of the nature of modern linguistic corpora.

3.2 Annotation and tagging

Initially, an annotation was any kind of ‘additional commentary’ added to an already existing written text, be it a transcription of a spoken text or a piece of conventional writing (i.e., a text

[5] HamNoSys = Hamburg Notation System for signed languages, developed at the Institute for German Sign Language, Hamburg University, Germany.

that did not necessarily previously exist as a spoken text or was never intended to become a spoken text by being written down). Annotations were often bilingual commentaries on 'difficult' foreign or ancient texts and were intended as aids to understanding.

For linguists, annotations have evolved into 'mini' linguistic commentaries that are appended to identified units in a language. Annotations add phonological, morphological, syntactic, semantic 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 a unit 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 tags to the written English sentence *Joanna stubbed out her cigarette with unnecessary fierceness* can in a large part be done automatically with reference to computerized dictionary of English in conjunction with the application of simple rules of word collocation and distribution.[6] The process is illustrated in example (1), below, taken from the Lancaster-Oslo/Bergen Corpus of English (cited in McEnery and Wilson, 2001 p. 47). It uses underscores and capitalization suffixed to lexical items as its linguistic tags (see Table 1).

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

Table 1: Key to tags used in example (1)

Tag	Meaning	Tag	Meaning
_NP	singular proper noun	_NN	singular common noun
_VBD	past tense form of lexical verb	_IN	preposition
_RP	adverbial particle	_JJ	adjective
_PP$	possessive pronoun	_.	full stop

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 with are time-aligned to the ID-glosses that are located on separate tiers in the ELAN annotation file. As can be seen from Figure 2, they are vertical tags, rather than sequential ones.

3.3 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 that dataset. Within linguistics that information is essentially sociolinguistic and sessional in nature. Sociolinguistically, metadata appends information about characteristics of the participants such as age, sex, region, class, religion, education, ethnicity, race, dialect and so on. Sessional metadata appends information about where and when the data was collected, under what circumstances and by whom.

3.4 Coding overall

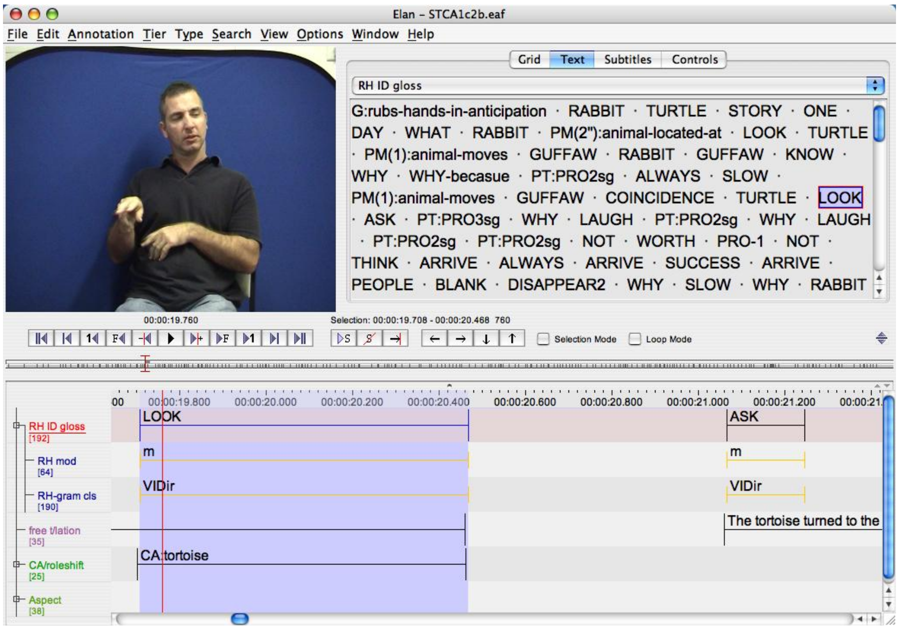
In summary, it should be noted that regardless of the type or degree of detail in the coding or analysis, only behaviours that are (or are assumed to be) linguistically meaningful are identified in transcription and annotation. This means ignoring all articulations and movements that are not (or appear not to be) related to language. With respect to signed languages, 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 role shift or constructed action. There are, of course, other behaviours which are not clearly extra-linguistic, especially in signed

[6] 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 and Smith, 1997).

languages which are perforce face-to-face languages. For example, some behaviours may or may not be aspects of the linguistic system (eye-gaze, facial expressions, movement modifications, etc.) and they will need to be encoded in the first instance *as part of investigations to determine their role within the language*. Coding for a particular feature of this type is usually based on a reasonable hypothesis about its grammatical role in the language. One must do this type of coding before extracting instances from the corpus to determine if a given hypothesis is correct. Only then could the coding for the feature be continued or discontinued on a principled basis.

4 ELAN

The corpus is being annotated digital video annotation software called ELAN (EUDICO – European Distributed Corpus – linguistic annotator) (Hellwig *et al.*, 2007). The software allows for the precise time-alignment of annotations with the corresponding video sources on multiple user-specifiable tiers. It allows one to create, edit, visualise and search annotations for video data. It supports display of video with its annotation; time linking of annotations to media streams; linking of annotation to other annotations; unlimited number of annotation tiers defined by users; different character sets; export of annotations as tab-delimited text files and a complementary ability to import text file annotations. Relevant metadata for the digital recordings is appended to media files. The following screen-shot of an opened ELAN annotation file shows an ID-gloss tier with several daughter tiers that exemplify the type of vertical tags, discussed above.



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

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

4.1 The tiers in ELAN

The tiers currently available in the Auslan corpus ELAN template are shown in Figure 3. Since, the majority of tiers have yet to be used with a very large collection of texts, the number and type of tiers in a standard ELAN annotation file is yet to be fixed. This is partly due to the fact that a certain amount of trial and error will be needed to determine what should be the minimum core number and type of tiers for all files in the corpus. Cumulative experience from repeated annotation parses focussing on different aspects of grammar will be needed before this can be done.

Tier map	Expansion*
QuickTime™ and a TIFF (Uncompressed) decompressor are needed to see this picture.	
	RH ID gloss: retrieved from databases
	RH meaning: a temporary gloss for the meaning of a sign when ID-gloss is unknown
	RH-gram cls: the grammatical class of the sign
	RH mouthing: the word (or parts of a word) being mouthed during the sign
	RH mouth-gc: the grammatical class of the mouthed word (not the sign)
	RH brow: eyebrow behaviour
	RH mod: spatial modification
	RH aspect-form: movement modification which is assumed to be aspectual
	RH aspect-meaning: the meaning of an aspectual modification
	RH-loc: modified (non-citation) location or direction
	RH-h/s: the handshape(s) on the RH, for phonological analysis (cf transcription)
	RH-mov: the movement(s) of the RH, for phonological analysis (cf transcription)]
	RH-ar/ment: codes singling/doubling of hands (cf transcription) [now included in ID-gloss] see “Marked use of one or two hands” in guidelines).
	Ref rec RH: the recoverability of the referent related to a spatial modification [now discontinued]
	RH CA co-occ: codes co-occurrence of CA during the sign performance [now discontinued]
	RH-sem.roles: the semantic role being played by the sign
	LH ID-gloss: retrieved from databases
	LH meaning: a temporary gloss for the meaning of a sign when ID-gloss is unknown
	LH-gram cls: the grammatical class of the sign
	LH mouthing: the word (or parts of a word) being mouthed during the sign
	LH mouth-gc: the grammatical class of the mouthed word (not the sign)
	LH brow: eyebrow behaviour
	LH mod: codes for spatial modification
	LH aspect-form: a movement modification which is assumed to be aspectual
	LH aspect-meaning: the meaning of an aspectual modification
	LH-loc: modified (non-citation) location or direction
	LH-h/s: the orientation on the LH, for phonological analysis (cf transcription)
	LH-mov: the movement on the LH, for phonological analysis (cf transcription)
	LH-orient: the orientation on the LH, for phonological analysis (cf transcription)
	LH-ar/ment: codes singling/doubling of hands (cf transcription) [now included in ID-gloss]
	Ref rec LH: the recoverability of the referent related to a spatial modification [now discontinued]
	LH CA co-occ: codes co-occurrence of CA during the sign performance [now discontinued]
	LH-sem.roles: the semantic role being played by the sign
	CA/roleshift: the start and finish of a period of body shift (BS), role shift (RS), constructed action (CA) or constructed dialog (CD) (with character/role specified), as appropriate
	Body: body shift behaviour including ‘swivelling’
	Head: head movements and facial expressions
	Gaze: direction of gaze and eyebrow movements
	transcription: transcription using dedicated fonts and systems (e.g. HamNoSys)
	clause: identifies a clause
	phrase: identifies a phrase
	part/situ: [to be deleted]
	affect-form: any facial expression features no elsewhere coded, especially spreading over more than one sign
	affect-meaning: the meaning of that facial expression
	aspect: tier replaced by equivalent RH and LH tiers [now discontinued]
	aspect meaning: tier replaced by equivalent RH and LH tiers [now discontinued]
	free t/lation: free translation in prosodic or meaning units
	lit t/lation: literal translation in prosodic or menaing units
	metadata: possible string of metadata codes
	annotator: possible annotator ID code
	notes: any notes or queries about the annotation (time aligned to queried annotation)

Figure 3: Current list of all tiers in the ELAN template used for the Auslan (under revision)
* RH = right hand; LH = left hand.

5 Annotation parses

The Auslan corpus is designed to be added to over time. Each ELAN annotation file (file extension *.eaf* for *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 analyse 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 *ID-gloss* tier). 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 makes each annotation file—and the whole Auslan corpus—very detailed and a rich source of data for research. The process is represented in Figure 4.

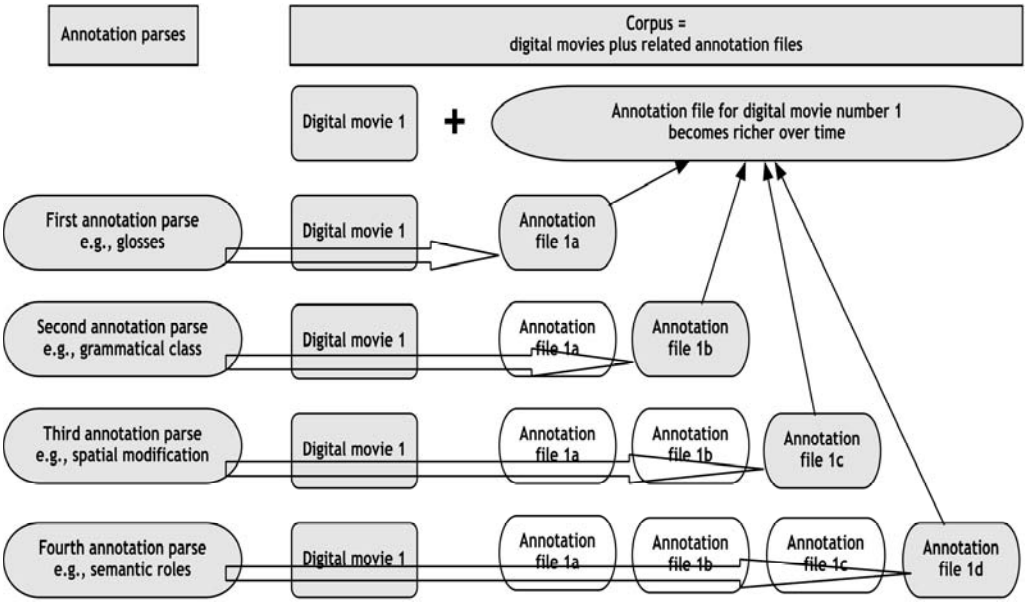

Annotation parses
Corpus = digital movies plus related annotation files
Digital movie 1
+
Annotation file for digital movie number 1 becomes richer over time
First annotation parse e.g., glosses
Digital movie 1
Annotation file 1a
Second annotation parse e.g., grammatical class
Digital movie 1
Annotation file 1a
Annotation file 1b
Third annotation parse e.g., spatial modification
Digital movie 1
Annotation file 1a
Annotation file 1b
Annotation file 1c
Fourth annotation parse e.g., semantic roles
Digital movie 1
Annotation file 1a
Annotation file 1b
Annotation file 1c
Annotation file 1d

Figure 4: Repeated annotation parses in ELAN of a given video text produces an ever richer annotation file.

5.1 From indeterminate to determinant in subsequent parses

One positive consequence of repeated annotation parses is that it encourages the use of tentative or generalized annotations (or tags) at times when fine-grained linguistic categorization may be difficult to make, if not 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 available as 'interim' tags (for details of controlled vocabularies and tags, see Johnston and de Beuzeville, 2008). 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 interim tag at least reduces the set of signs which must be revisited on a subsequent annotation parse for reconsideration.

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 using some form of multi-media annotation software. In the first instance, these annotations are simply glosses that identify the sign units in the text.

The identification of segments within a recording 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. In today's multi-media digital files, the time-aligned annotations appended to segments of the text are read by machine, not the text (i.e., source data) itself, nor, necessarily, a transcription of the source text, whether spoken or signed. Thus, despite what many signed language researchers continue to believe, a full written phonetic or phonological transcription of signed texts is no longer essential in order to conduct corpus-based research at various levels of linguistic analysis, even the phonological. Time aligned multi-media gloss-based annotations are adequate for this task because one need not transcribe all signs in their entirety—one could append a relevant phonetic or phonological tag for a feature under investigation. Of course, the use of a dedicated transcription tier(s) within ELAN would be necessary in order to carry out detailed phonetic or phonological research. Overall, however, there is little doubt that the use of time aligned gloss-based annotations is superior to transcription-based annotation in terms of the time it takes for a sizable amount of text to be make available for processing.

In order to segment the source data into sign units to which gloss-based or linguistic annotations can subsequently be appended, it is essential to integrate all *available* lexical information about the language into the common identifier for each lexical sign, and to follow standard protocols for glossing non-lexical signs.

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' must, of course, be 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 (e.g., the British Sign Language sign SISTER is identical to 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 contain any information about sign form.[7]

By contrast, there needs to be a level in corpus annotation where signs ars identified uniquely and consistently because one cannot productively use ad-hoc glosses which may vary from context to context. In the Auslan Corpus project, this type of identifying gloss is referred to as the *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 (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 recognized and distinctive lexeme of the language), the ID-gloss HOUSE is used in both instances to identify the sign in the gloss annotation. A consistently applied label of this type means it is possible to search through multiple annotation files and find all instances of a par-

[7] The additional use of a dedicated signed language notation/transcription system, such as the Hamburg Notation System, can overcome this.

ticular sign in order to determine the ways and environments in which it is used. We can only do this if all relevant signs have the same ID-gloss in the corpus. Of course, corpus-based evidence could itself lead to the re-analysis (and hence re-glossing) of certain signs (e.g., see the discussion of homonyms and pointing signs below).

With respect to distinguishing between glossing and translation, meaning is assigned to the text through glossing only 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 but means 'main' or 'importance', it is still labelled IMPORTANT.

6.2 Summary of ID-glossing

In assigning an ID-gloss to a sign form one is simply labelling a sign so that it can be uniquely and quickly identified or tagged (e.g., for grammatical class, sign modification potential, presence or absence of constructed action, semantic roles, and so on) during a later annotation parse, or searched for with or without these tags being taken into consideration (i.e., as search constraints). Apart from the obvious motivation of the English word used to gloss a sign, no serious attempt is being made in the assigning of an ID-gloss to translate a sign.

Failure to use ID-glosses and standardized glossing procedures in multi-media corpus annotation would create two problems. First, the consistency and commensurability of data that is annotated (i.e., glossed) by different researchers (or even by the same researcher on different occasions) can not be assured any other way. Second, the dataset would become effectively unbounded if there was no constraint on 'meaning-based glossing' because each sign articulation which may be distinctive in form could potentially have its own distinctive gloss reflecting its meaning in each context. The unique identification of sign types, which is one of the prime motivations for the creation of a linguistic corpus in the modern sense (e.g., for the purposes of searching and quantification of types and token), would thus not be achieved.

Without consistency in using the ID-gloss, it will be impossible to use the corpus productively and much of the time spent on annotation will be effectively wasted because the corpus will cease to be, or never become, machine-readable in any meaningful sense. It will not actually be the type of corpus that linguists aspire to today. Rather, it will 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 lexical signs, non-lexical signs, and pointing signs

Not all of the signs produced when a signer is communicating in a signed language are of the same type. There are two major types of signs which have been described as *lexical* signs and *non-lexical* signs (Johnston and Schembri, 1999; Sandler and Lillo-Martin, 2006).[8]

A *lexical sign* is a signed form whose meaning in context is more than the conventionalised and/or iconic value of its components (handshape, location, etc.) within the inventory of meaning units of a given signed language in a given context; and that meaning is stable or consistent across contexts. A lexical sign is, essentially, equivalent to the commonsense notion of word (Sandler and Lillo-Martin, 2006) and should thus not be confused with *lexical word* (or *content*

[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).

word) and not opposed to *grammatical word* or *grammatical sign.* A *non-lexical sign* in this terminology is thus a signed form that has little or no conventionalised or language-specific meaning value beyond that of its components in a given context (e.g., depicting or 'classifier' signs). The annotation conventions for lexical and non-lexical signs are described below.

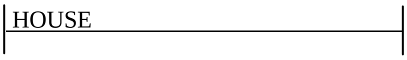
7.1 Lexical signs and ID-glosses

Lexical signs are identified using an ID-gloss. In the annotation fields created in ELAN that contain the ID-glosses, the lexical glosses are written in upper case, as is the norm for glossing in signed language linguistics. In mainstream linguistics, it is usually only glosses for grammatical morphemes or function words which are written in upper case, as in the following example. (In the example, the source and glossing language are both English. Commonly, when interlinear glossing is used, the source and glossing language are different.)

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

The use of uppercase for all glosses commonly found in signed language linguistics is partly due to the fact that doing so helps to distinguish the signed language gloss from the surrounding English text with which it could easily be confused. We maintain this convention in the ELAN annotation files. Thus the ID-gloss HOUSE appears on an ID-gloss tier as:

(3) (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 7,000 individual sign entries in which short digital movie clips are headwords (i.e., headsigns). There are multiple fields coding information on the form, meaning and lexical status of each headsign. 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 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 in 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 (know as a *keyword*). 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, 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 an unavoidable prerequisite for creation of a viable corpus. However, it need not be exhaustive. After all, it is highly likely a corpus will actually reveal unrecorded lexical signs which need to be added to the reference lexical 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 what is recorded in the database is when corpus data justifies 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 lexical signs

In order to maximise the consistency and uniqueness of annotations of lexical signs it has been necessary to develop and implement conventions for the treatment of various lexical or morphological phenomena found in Auslan (and other signed languages). For example, the existence of negative incorporation in Auslan signs (in both suppletive and ‘affixed’ forms) 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. Details can be found in the guidelines for Auslan corpus annotators (Johnston and de Beuzeville, 2008). These comprehensive guidelines deal with phenomena such as negative incorporation, variant forms, the marked used of one or two hands in normally two-handed or one-handed signs respectively, numbers, sign names, and borrowings from Signed English and other signed languages.

7.2 Annotation conventions for non-lexical signs

As with ID-glosses, a relatively small set of annotation and glossing conventions need to be followed in order to ensure that similar types of non-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. Non-lexical signs are primarily represented by depicting signs (also known as ‘classifier signs’ in the literature), but also included fingerspelled signs and buoys of various types: list, theme, fragment and pointer buoys (Liddell, 2003). Annotation conventions for non-lexcial signs include prefixing all depicting sign glosses with the tag PM, and all fingerspelling strings with FS. By following these simple conventions one can incorporate consistent codes in annotations for these types of signs while at the same time using sign specific—and potentially ‘inconsistent’—codes for individual signs, and yet create a set of annotations that can be easily read, sorted or otherwise processed by machine.

7.3 Annotation conventions for pointing signs

All ID-glosses for points begin with PT (for ‘point’) in upper case. By following this convention it is possible to search for all instances of pointing signs in the corpus and on this basis analyse (or re-analyse), categorize (or re-categorize), and label (or re-label) them, as appropriate. In other words, corpus evidence itself will assist in understanding the function of these types of signs. Generally speaking, it is not unusual for an annotator to be unable to make a detailed grammatical annotation, beyond identifying a PT, with any certainty during a first parse. It can be a difficult task. All points are thus coded with PT (followed by a colon) at minimum, with additional specification being made during later annotation parses.

8 Annotation conventions for gesture

Gestures can be culturally shared or idiosyncratic. Even if culturally shared, however, gestures which have not become lexical Auslan signs will not be found in the dictionary database and will thus not have an assignable ID-gloss. Gestures of both types occur commonly in speech and during signed discourse. When annotated the gloss for a gesture is prefixed with G: for ‘gesture’ followed by a brief description of the meaning of the gesture. One can see a sign’s form from the associate movie clip in the annotation file, so it is not essential to have that information separately encoded in an annotation. 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 has having subtly distinct forms and/or specific functions that may justify recategorisation and reglossing. 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 assigning ID-glosses or glossing conventions to various other types of non-lexical signs and gestures.

9 Conclusion

The Auslan corpus project was one of the first to attempt to compile a large machine-readable corpus of a signed language. It was begun in 2004. Since that time a number of other signed language corpus projects have begun (e.g., the NGT or Netherlands Sign Language corpus and the BSL or British Sign Language corpus), are about to begin (e.g., the DGS or German Sign Language corpus), or are planned (e.g., the ASL or American Sign Language corpus). Some, like the NGT corpus, have 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. A small percentage of these texts have also been transcribed using ELAN.

However, this paper has tried to show that the creation of signed language 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 signed language researchers. After all, there have, to date, been very little publicly available reference texts of any signed language. Nonetheless, corpus creation must also involve the transformation of this archived material into a machine-readable corpus by the principled application of annotation procedures that make optimal use of the new digital technologies. Business-as-usual with these new digital archives—so-called enrichment through the addition of transcriptions—does not add value to the archive in ways that corpus linguists would assume and expect. Happily, 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.

References

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

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

Cormier, Kearsy and Jordan Fenlon. To appear. Possession in British Sign Language. In *Expression of possession,* ed. by W. B. McGregor. Berlin: Mouton de Gruyter.

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

de Beuzeville, L., Johnston, T., and Schembri, A. (submitted). The use of space with lexical verbs in Auslan: a corpus-based investigation. *Sign Language & Linguistics.*

Dudis, P. G. 2004). Body partitioning and real-space blends. *Cognitive Linguistics*, 15(2), 223-238

Garside, R., and Smith, N. 1997. A hybrid grammatical tagger: CLAWS4, in Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, pp. 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., and Cermakova, A. 2004. *Lexicology and Corpus Linguistics*. London: Continuum.

Hellwig, B., van Uytvanck, D., and Hulsbosch, M. 2007. EUDICO Linguistic Annotator (ELAN). http://www.lat-mpi.eu/tools/elan/

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

Hoting, N., and Slobin, D. I. 2002. Transcription as a tool for understanding: The Berkeley Transcription System for sign language research (BTS). In G. Morgan and B. Woll (Eds.) Directions in Sign Language Acquisition (pp. 55-75). Amsterdam/Philadelphia: John Benjamins.

Johnston, T. 1991b. Spatial syntax and spatial semantics in the inflection of signs for the marking of person and location in Auslan. *International Journal of Sign Linguistics, 2*(1), 29-62.

Johnston, T. 1991a. 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., and de Beuzeville, L. 2008. *Researching the linguistic use of space in Auslan: guidelines for annotators using the Auslan corpus*. Manuscript, Department of Linguistics, Macquarie University, Sydney, Australia. Downloadable at www.auslan.org/about/corpus/.

Johnston, T., de Beuzeville, L., Schembri, A., and 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).

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

Johnston, T., and Schembri, A. 2006. Issues in the creation of a digital archive of a signed language. In L. Barwick and N. Thieberger (Eds.), *Sustainable data from digital fieldwork: Proceedings of the conference held at the University of Sydney, 4-6 December 2006* (pp. 7-16). Sydney: Sydney University Press.

Johnston, T., and Schembri, A. 2006. *The use of ELAN annotation software in the Auslan Archive/Corpus Project.* Paper presented at the Ethnographic Eresearch Annotation Conference, University of Melbourne, Victoria, Australia (Feburary 15-16).

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. London and New York: Longman.

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

McEnery, T., and Wilson, A. 2001. *Corpus linguistics*. Edinburgh: Edinburgh University Press.

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

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

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

Sinclair, J. 1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

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

van der Hulst, H., Crasborn, O., and van der Kooij, E. 1998. *How SignPhon addresses the database paradox.* Paper presented at the Second Intersign Workshop, Leiden, The Netherlands, December, 1998.