Testing language description through language documentation, archiving and corpus creation: the case of indicating verbs in the Auslan Archive Corpus

Trevor Johnston & Adam Schembri

Proceedings of Conference on

Language Documentation & Linguistic Theory

Edited by Peter K. Austin, Oliver Bond & David Nathan

7-8 December 2007 School of Oriental and African Studies, University of London

School of Oriental and African Studies
Thornhaugh Street, Russell Square
London WC1H 0XG
United Kingdom

Department of Linguistics:
http://www.soas.ac.uk/academics/departments/linguistics

Hans Rausing Endangered Languages Project:
http://www.hrelp.org
elap@soas.ac.uk

ISBN: 978-0-7286-0382-0

This paper can be cited as:

Trevor Johnston & Adam Schembri. 2007. Testing language description through language documentation, archiving and corpus creation: the case of indicating verbs in the Auslan Archive Corpus. In Peter K. Austin, Oliver Bond & David Nathan (eds) *Proceedings of Conference on Language Documentation and Linguistic Theory*. London: SOAS. pp. 145-154.

or:

Trevor Johnston & Adam Schembri 2007. Testing language description through language documentation, archiving and corpus creation: the case of indicating verbs in the Auslan Archive Corpus. In Peter K. Austin, Oliver Bond & David Nathan (eds) *Proceedings of Conference on Language Documentation and Linguistic Theory*. London: SOAS. www.hrelp.org/eprints/ldlt_17.pdf

Testing language description through language documentation, archiving, and corpus creation: the case of indicating verbs in Auslan

TREVOR JOHNSTON[1] & ADAM SCHEMBRI[2]
Macquarie University, Sydney[1] &
University College London[2]

1. INTRODUCTION

For little-researched or unwritten languages, the creation of language documentation and language archives is a vital, if not first, step in linguistic description. The subsequent transformation of documentation into a linguistic corpus adds significant value to an archive (McEnery & Wilson 1996; McEnery, Xiao, & Tono 2006). In this paper, we describe a project[1] to document the signed language of the deaf community in Australia (Australian Sign Language or Auslan), the requirements for transforming an archive into a linguistic corpus, and the results of an initial study based on this corpus.[2] In doing so, we explain why systematic language documentation for signed languages is particularly important, and why it is especially so for Auslan. We also address the special problems and issues that surround building an annotated corpus of a signed language, in particular written representation and lemmatisation.

2. THE NEED TO DOCUMENT AUSLAN

2.1. Scarce primary or secondary sources and endangerment

There are very few primary sources for linguists wishing to study signed languages because the languages have no widely used written form and the technology of recording moving images is little over a century old. Secondary sources (dictionaries, teaching resources, some television broadcasts) only became available for a few signed languages in the last two to three decades. In terms of both primary and secondary resources, documentation in and about Auslan has been minimal.

Auslan is an endangered language and there are probably fewer than 6,500 deaf Auslan users in Australia, in a community that has an inverted age pyramid, with the majority of signers over 40 years of age.

---

[1] The project was funded by the Endangered Language Documentation Program at SOAS, University of London under the Hans Rausing Endangered Languages Project at SOAS, University of London (grant #MDP0088, awarded to Trevor Johnston).

[2] We would like to acknowledge the contribution of the deaf native signers and the Australian deaf community generally to the creation of the Auslan corpus. We would also like to acknowledge the assistance of the following annotators, research assistants and/or fellow researchers: Julia Allen, Karin Banna, Donovan Cresdee, Louise de Beuzeville, Michael Gray, and Della Goswell.

2.2. The problem of the representation of signed languages in a written form
There are two obstacles to adding value to the existing documentation of Auslan. First, linguistically trained annotators who are fluent in these languages are not widely available (and may become less so in the future due to language endangerment). Second, there is no commonly used or widely accepted written form of any signed language and there are only a few signed language notation or transcription systems (Johnston 1991; Miller 2001). These systems are of limited application, however, being used for the representation of individual signs rather than extended texts, having no widely used orthography, nor are they readily compatible with annotation software, such as ELAN (see below).

The standard practice in signed language linguistics, therefore, has been to use spoken language glosses alone with or without additional conventions for representing visual prosody, spatial loci, and some of the more regular internal sign modifications, such as those expressing aspectual meanings on verbs. Texts using these techniques tend be rare and short, being model sentences or phrases, or individual signs. A consequence of these practices, of course, is that the reader usually does not always know with any certainty the form of the sign that is being referred to by the gloss, even if that signed language is their own native language.

Language use and experience in deaf communities is extremely heterogeneous so that native signer intuitions often lack consensus or may actually be in conflict with observation—all the more reason why language documentation and corpus creation is important in signed language research.

2.3. A solution in recent advances in language documentation technology
The recent advent of digitised video technology and improvements in the cost and speed of computer storage and processing power have meant that long-standing inherent problems of analogue video, such as ease of navigation within a video and the difficulty of adding subtitled glosses to video footage, have been overcome in digital multi-media annotation software such as ELAN. It is now possible to collect large amount of signed language data and to transcribe and present this data in a meaningful and machine-readable manner without the necessity of devising dedicated and complex scripts because the primary data is always accessible.

3. THE AUSLAN DOCUMENTATION

The documentation project has recorded samples of Auslan from deaf native and near-native signers across Australia. Footage from 50 three-hour language sessions, each with two participants, has been recorded on 600 hours of digital video. This will ensure that work on a detailed grammatical description of the language remains possible into the future in the light of language endangerment. A subset of the recordings have been transcribed and annotated for initial analysis.

3.1. Archive of digital video from fieldwork sessions

Language recording sessions were conducted involving 100 deaf native and near-native signers of Auslan. The recordings included an interview, the production of narratives, responses to survey questions, free conversation, and other elicited linguistic responses to various stimuli. The sessions were lead by a deaf native signer. The footage has been edited into separate digital movie clips (approximately 17 for each participant) suitable for archiving, individual transcription and annotation.

3.2. From archive to corpus

In order to make the language archive 'corpus-ready' from its inception, it was designed to be representative of native and near-native signers with balanced numbers of males and females, younger signers and older signers, and residents from all five major Australian cities (i.e., Sydney, Melbourne, Brisbane, Perth and Adelaide). The same tasks and texts types were elicited from each for comparability. As a reference corpus of a finite size, value-adding through transcription, annotation and tagging will produce an enriched dataset within the shortest possible time-frame that can be reused for a range of different investigations. Value-adding involves appending to the digital video metadata and annotation files created through multimedia annotation and corpus software. In this way, the lack of an orthography has been partially overcome and a machine readable text created.

3.2.1. The annotation files

The archive data is being transcribed and annotated using the ELAN digital annotation software (Hellwig, van Uytvanck, & Hulsbosch, 2007). The software allows for the precise time-alignment of transcriptions and annotations with corresponding video segments using multiple user-specifiable tiers. It allows one to create, edit and search annotations of the video data, creating frequency lists and counts with collocations. All transcriptions use English-based glosses. The annotations will be added to it over time to make each annotation file and the whole corpus a rich source of data for signed language research.

3.2.2. The transcription and annotation procedure

The transcription is simply intended to create a text which is itself machine-readable and which may then be annotated or tagged with additional information, either manually or automatically. It is not intended to create a text that can be reproduced in full because it carries insufficient information to stand alone from the video recording. By using this technique, signed language researchers are no longer held hostage to the limits of representation imposed by a perceived need to use a dedicated transcription system that attempts to write down the form of signs.

Essentially we use English-based glossing in the transcription system in such a way as to uniquely identify signs and lexemes. This means that when one reads the transcription, one knows that all instances of signs with this gloss are the same sign. Only if a form-meaning pair always has the same ID-gloss can we search, using computers, for how the same sign is used in different ways in the corpus (at

least, until such time as searching the actual digital video file for particular signs becomes possible). The actual form of the sign can be seen from the associated video clip.

The precise ID-gloss to assign to any lexical sign is determined by referring to the Auslan lexical database in which there are currently 7,000 separate illustrated sign entries. Annotators assign new ID-glosses to signs if they feel they have identified a new unrecorded lemma. The database is then subsequently updated.

3.2.3. Glossing, lemmatisation and corpus linguistics

It should be evident from this description of the procedure that without a reference lexical resource the creation of a machine-readable corpus is unlikely to succeed for two reasons. First, it would be extremely difficult, if not impossible, to control the possible multiplication by annotators of glosses referring to the same sign. Non-uniquely identifying glosses are of little or no use as a transcription system in a corpus. Second, even if accepted sign-IPA transcription systems were available and widely used, lexemes still need to be identified. This requires the use of an orthography which needs to ignore variations in individual pronunciations for the lemmatisation of word or sign forms. Otherwise, once again, one would not be able to identify repeated instances of the same sign.

Gloss-based transcription thus combines aspects of transcription and lemmatisation. The identification of lexemes in the transcription, rather than the actual signed production, is essentially lemmatisation of the text. Additional tiers of the annotation file tag for other aspects of sign realisation (e.g., grammatical class, semantic role, sign modification in terms of movement, location, and so on). There is no necessity—indeed, it is counter-productive in terms of machine readability—to include such information in the ID-gloss.

4. DOCUMENTATION LEADING TO NEW CORPUS-BASED RESEARCH

Until the creation of the Auslan archive and then the beginnings of its transformation into a linguistic corpus, it has not been easy to verify empirically claims regarding grammatical patterns and discourse structures of the language. Claims based on elicitation sessions with individual informants or small numbers of signers in signed language research can be problematic (see Johnston, Vermeerbergen, Schembri, & Leeson 2007). This has serious implications for the status of recent claims as to linguistic universals (e.g., Sandler & Lillo-Martin 2006) and our understanding of sign language and cognition derived from behavioural and imaging studies (e.g., Emmorey 2002).

Once the annotation and tagging of the corpus is detailed enough and the data is made publicly accessible, this will enable the description of a grammar of Auslan to be empirically grounded and open to thorough critical peer review. The Auslan corpus also creates the possibility of investigating another area of new research and theoretical interest in sign linguistics which concerns the typical or possible grammaticalisation pathways exploited in sign languages.

4.1. A case study: indicating verbs in the corpus

The ID-gloss transcription technique made it possible to take documents from the Auslan archive and create a small research-specific corpus to investigate the use of indicating verbs. Indicating verbs are signs in Auslan that are able to change their normal citation-form location of articulation in the signing space to show, for example, actor versus undergoer semantic roles. There are two sub-types: (i) those that change their overall location in the signing space ('locatable indicating verbs'); and (ii) those that can also change their initial, final, or initial and final places of articulation, and hence their citation-form direction of movement and not just location ('directional indicating verbs').

The modification of indicating signs in signed languages for semantic roles has long been compared in the literature with grammatical systems of person agreement in spoken languages and, despite the iconicity of the system and the apparent similarities with pointing gestures, this is exactly how it has been analysed by many signed language linguists (Aronoff, Meir, & Sandler 2005; Padden 1988; Sandler & Lillo-Martin 2006). The descriptions of many signed language grammars have thus implied these person agreement systems are obligatory for well-formed and grammatical sentences, as reported by native-signer intuitions and as exemplified in the signing of typical native signers.

If indicating verbs directed at locations in space associated with present or absent referents are indeed a type of person agreement system, then the person, number and/or case features of an associated noun phrase should control the presence of modifications on the target verb (Corbett 2006). Here we investigate the degree to which such modifications of directional signs are grammaticalised by reporting on the frequency with which indicating verbs are actually modified in text, the percentage of tokens which were modified, and co-occurrence of other potentially relevant linguistic features, such as constructed action.[3]

4.1.1. Data

The dataset for this study consisted of 50 narratives. Forty texts were from the Auslan archive: ten spontaneous narratives produced during the periods of free conversation and thirty from elicited recounts of two Aesop's fables ('The Boy who Cried Wolf' and 'The Hare and the Tortoise'). Ten were spontaneous narratives produced during free conversations that were collected as part of a sociolinguistic variation in Auslan project (see above).

Nouns and verbs were identified and tagged according to their spatial modification potential into three categories: plain (unable to be spatially modified), locatable, or directional. Nouns are either only plain or locatable, as there appears to be no directional indicating nouns. Nouns and verbs were then tagged according to their realisation in context: not applicable (i.e. plain),

[3] This is a summary of data presented at the 10th International Cognitive Linguistics Conference from a study by Johnston, de Beuzeville, Schembri, & Goswell (in preparation) based on corpus data from the Auslan archive augmented with data from recordings originally made during an ARC linkage grant awarded to Adam Schembri and Trevor Johnston (#LP0346973 'Sociolinguistic Variation in Auslan').

modified, unmodified, or congruent. The congruent category refers to those spatially modifiable signs whose form in the text is actually identical to their citation form. They thus appear unmodified. However, these forms could be considered to be modified because locating or directing the sign to relevant established locations in the signing space in the given text would actually map directly on to the sign's citation form.

Verbs were also tagged for their frequency in the text in two categories: high frequency and low frequency. In addition, periods of time during the narrative in which there was constructed action—i.e., times when the narrator took on the role of one of the protagonists in the narrative, signalled by actions imitating the referent—were identified so that signs could be tagged for the co-occurrence of modified forms with periods of constructed action. Finally, signs were tagged for the text-type they occurred in: spontaneous narrative, 'The Hare and the Tortoise' story, or 'The Boy who cried Wolf' story.

4.1.2. Results

The annotations and tags were exported from ELAN into Excel, a database program, for quantification of frequency and distribution, and then into GoldVarb, a version of the Varbrul program, for the multivariate analysis of the likely significance of the variables we had coded.

4.1.2.1 Distribution & frequency

The following graphs summarize the distribution of modified indicating verbs and nouns in the dataset.

Figure 1
Verbs modified versus unmodified

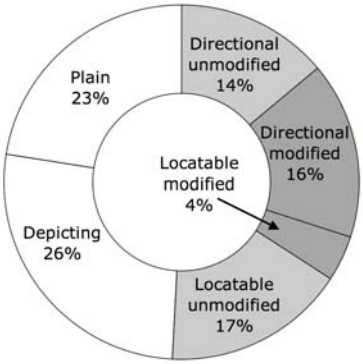


Figure 2
Verbs with congruent as modified

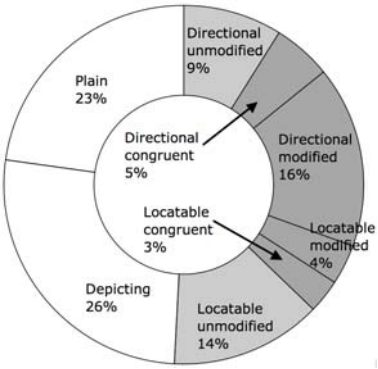


Figure 3
Nouns

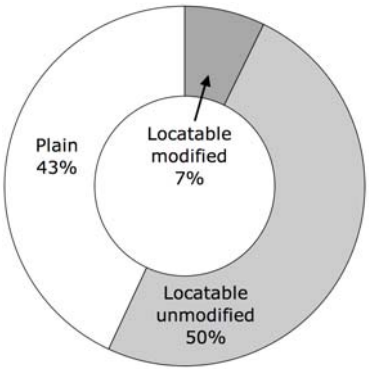
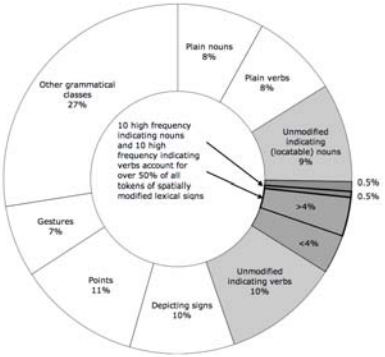


Figure 4
Modified signs in context



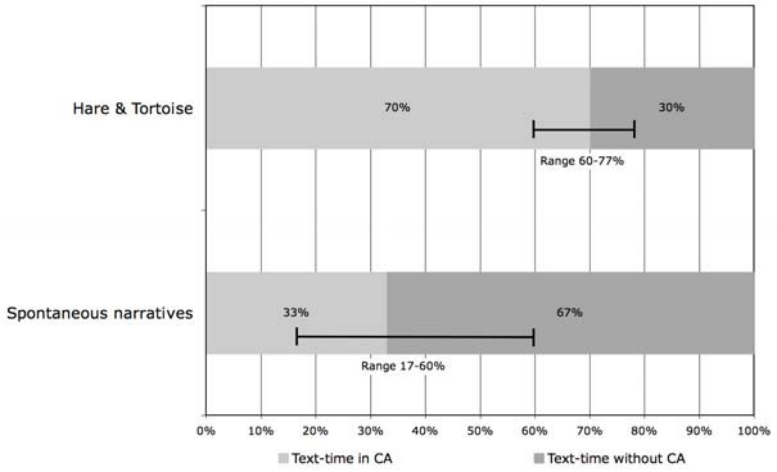
The possible significance of these figures becomes a little clearer if we consider (i) the total number of signs in the texts and (ii) the number of sign types and token frequency (Figure 4).

The final point of interest was the amount of time during a text that signers were engaged in constructed action (Figure 5). It is clear that the mimetic properties of constructed action appear to play a significant role in narrative discourse.

4.1.2.2. Interaction of variables

The results of a Varbrul run are expressed numerically as 'weights' which indicate the degree to which a given factor 'favours' or 'disfavours' the occurrence of the variable. The output also ranks the various factor groups in order of their likely possible significance relative to one another. The output will vary according to the total number of factor groups processed together in any particular run.

Figure 5
Per cent of text time in constructed action



The general results of the application of Varbrul to this dataset were as follows, with the findings ranked in order of importance. The results are reported on in detail in Johnston, de Beuzeville, Schembri, & Goswell (in preparation).

1. Indicating signs that are directional significantly favour modification compared to locatable signs.
2. Locatable verbs significantly favour spatial modification compared to locatable nouns.
3. The most frequent indicating verbs significantly favour modification compared to other indicating verbs.
4. Constructed action significantly favours co-occurrence with spatial modification, especially modified verbs.
5. The Hare & Tortoise recount significantly favours use of constructed action compared to the spontaneous narratives.

When the data were recoded allowing for the category of modified to include 'congruent' signs, outcomes 1, 2, & 4 were still true, but outcome 3 was no longer true, somewhat surprisingly. Recoding for congruent signs has no bearing on observation 5.

4.1.3. Discussion

Cross-linguistically, canonical forms of person agreement in spoken languages tend to be obligatory, although there are many exceptions, and marking tends to be found with most members of the relevant particular grammatical class and not just on a privileged sub-set, and the markers tend to be semantically bleached, if not completely opaque (Corbett 2006). The frequencies and distributions of spatially modified verbs in this data set are not easily compatible with the notion that spatial modification to indicate referent roles is highly grammaticalised. In this Auslan corpus, the use of spatial modification to indicate the involvement of absent referents in actions represented by indicating verbs appears optional, although generally favoured on a set of frequent and iconic signs (e.g., signs of displacement or transference, real or metaphorical).

The use of spatial modifications on indicating signs is also strongly linked to periods of constructed action. One explanation may lie in the fact that modified indicating signs are themselves part of a form of enactment. That is, indicating verbs are directed towards locations associated with absent referents as if the referents were physically present (Liddell, 2000, 2003), just as imitating the facial and bodily actions of a referent in constructed action makes it appear as if an absent referent were present. Hence, it may be no surprise, after all, that they frequently occur during periods of constructed action. In other words, it is possible that the use of spatially modified indicating verbs draws on non-linguistic mental representations of referents more or less directly and what we are seeing in the data is a reflection of this fact.

This first corpus-based analysis of signed language texts is still somewhat limited in size and text-types. The spatial modification of indicating verbs is but one strategy available to signed language users to express various types of

grammatical meanings which must be weighed up against other strategies which are also available to all language users. With the aid of the comprehensive coding of the immediate linguistic environment of indicating verbs, it will be possible to tease apart and weight up the relative importance of the many factors that enable language users to disambiguate or constrain the interpretation of utterances.

5. CONCLUSION

The documentation of face-to-face languages without a written form, and especially the signed languages of deaf communities, is clearly an absolute prerequisite for advancing linguistic description, analysis and theory. It is also evident that adding value to signed language documentation through the creation of corpora will be a labour and time intensive activity, even more so than in spoken language documentation. The creation of digital archives of signed languages is challenging, time-consuming and expensive, as our experience with the Auslan archive reported here shows, but, by exploiting the currently available and rapidly improving digital multimedia annotation software in conjunction with archives and corpora, sign language linguistics is destined to become much more rigorous in the near future.

REFERENCES

Aronoff, M., Meir, I., & Sandler, W. 2005. The Paradox of Sign Language Morphology. *Language, 81*(2), 301-344.

Corbett, G. 2006. *Agreement*. Cambridge: Cambridge University Press.

Emmorey, K. D. 2002. *Language, Cognition, and the Brain: Insights from sign language research*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

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., de Beuzeville, L., Schembri, A., & Goswell, D. (in preparation). On not missing the point: Indicating verbs in Auslan. *Journal of Deaf Studies and Deaf Education*.

Johnston, T., Vermeerbergen, M., Schembri, A., & Leeson, L. 2007. 'Real data are messy': Considering cross-linguistic analysis of constituent ordering in Auslan, VGT, and ISL. In P. Perniss, R. Pfau & M. Steinbach (eds.), *Proceedings of the Workshop on Sign Languages: A Cross-linguistic Perspective, Mainz, Germany, March 25-27, 2004,* pp. 163-205. Berlin: Mouton de Gruyter.

Liddell, S. K. 2000. Indicating verbs and pronouns: Pointing away from agreement. In K. D. Emmorey & H. Lane (eds.), *The signs of language*

revisited: An anthology to honor Ursula Bellugi and Edward Klima, pp. 303-320. Mahwah, NJ: Lawrence Erlbaum Associates.

Liddell, S. K. 2007. *Subjects and referents.* Paper presented at the Typology, Sign and Gesture: the Second International Conference of the Association Française de Linguistique Cognitive, Université Lille, 10-12 May, 2007.

McEnery, T., & Wilson, A. 1996. *Corpus linguistics*. Edinburgh: Edinburgh University Press.

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

Miller, C. 2001. Some reflections on the need for a common sign notation. *Sign Language & Linguistics, 4*(1/2), 11-28.

Padden, C. 1988. Grammatical Theory and Signed Languages. In F. Newmeyer (ed.), *Linguistics: The Cambridge Survey* (Vol. Volume 2, pp. 250-266). Cambridge: Cambridge University Press.

Sandler, W., & Lillo-Martin, D. 2006. *Sign language and linguistic universals*. Cambridge: Cambridge University Press.