Auslan Corpus annotation conventions compared to BSL & NGT corpora

Trevor Johnston, Adjunct Professor
(Macquarie University & La Trobe University)
Digging into Signs annotation workshop
DCAL, UCL, London 30-31 March 2015
The Auslan Archive & Corpus

As at March 2014

• 256 signers (approx. 200 hours digital video)
  – 100 of which are part of the Endangered Languages Archive project funded by SOAS
  – the remainder (with some overlap) are part of the secondary Sociolinguistic Variation in Auslan project
• all (bar a few errors) are native or near-native (before age 7)
  – 3 erroneously included (quarantined)
• 460 of >1100 clips with some basic annotation (450 ELDP, 10 SVIAP)
  – all have ID glossing, some have free translation
  – c.16 hours
  – c. 110,000 sign tokens
  – c. 250,000 annotations
  – c. 110 spm (cf 119 spm DGS)
• 200 clips (approx. 7 hours) have more detailed annotations, as follows:

Types of annotations

• clause-like units
• grammatical class
• spatial and directional modification
• aspectual modification
• event and situation type (Aktionsart)
• lexical frequency
• constituency
  – argument position and macro- and semantic roles
• clause relationships
  – dependency, embedding
Auslan Archive & Corpus
Initial reaction to proposals
“Towards a gloss annotation standard”

• **Two issues**
  – only gloss?
    • the guidelines also deal with segmentation & hand alignment
  – glossing ‘in isolation’?
    • can one really separate glossing (and segmentation) from the reason a linguistic corpus is created?

• **Auslan annotation practices reflect the goal and desire to**
  – “identify constructional schemas across the lexico-grammar”
    • i.e., not just segment and identify signs in the signing stream

• **Auslan Corpus more than merely ‘glossed signs’**
  – units larger than individual signs
    • clauses (CLUs)
    • argument structure
  – other units beyond lexical (including depicting/classifier) signs
    • gesture
    • constructed action
  – other non-manual activity
    • non-manuals (only recently begun)
    • prosodic aspects (only recently begun)
  – exploitation of space & direction
    • spatial arrangement/placement (only recently begun)
    • reference tracking (only recently begun)
    • directional modification
Similarities & differences to draft proposals

• Auslan practice described in
  – Auslan Annotation Guidelines
    • first version 2005
  – annual (or twice yearly) upgrades/revisions over the past 10 years
    • last revision just prior to this workshop
      – (see workshop dropbox)
    • next revision will undoubtedly be immediately after this workshop
General observation

• Auslan annotation practices very similar to BSL draft proposals
  – partly because the latter were developed out of the former
  – greatest divergence actually occurs on annotations above the level of the individual sign
    • however, this is the very area which is not the focus of this workshop
<table>
<thead>
<tr>
<th>Topic</th>
<th>General practice</th>
<th>Auslan</th>
<th>Example</th>
<th>BSL</th>
<th>NGT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Basic gloss</td>
<td>All lexical signs are annotated using an identifying gloss (ID-gloss), written in upper case. This gloss corresponds to ‘Annotation ID-gloss’ in SignBank</td>
<td>BSL as per Auslan, but also HOUSE HOUSE(W)</td>
<td>SIGN</td>
<td>SIGN02</td>
<td>SIGN-A, SIGN-B, SIGN-C</td>
</tr>
<tr>
<td>4. Lexical variants</td>
<td>Lexical variants (same or similar/related meanings but differ in two parameters or more from each other) are suffixed.</td>
<td>Old BEFORE1 BEFORE2 New BEFORE BEFORE.SUBHAND</td>
<td>SIGN</td>
<td>SIGN02, SIGN03</td>
<td>SIGN-A, SIGN-B, SIGN-C</td>
</tr>
<tr>
<td>12. Number sequences</td>
<td>Number sequences receive one ID-gloss (see also 7. Compounds)</td>
<td>NINETEEN-EIGHTY-SEVEN</td>
<td>NINETEEN-EIGHTY-NINE</td>
<td>1989</td>
<td></td>
</tr>
</tbody>
</table>
Feedback on draft proposal: positive

• Proposals for BSL/NGT overall not radically different than current Auslan Corpus practice
  – some identical ✔
  – other points of divergence can be reconciled with a “search and replace” ✔
    • if combined statistical operations desired of annotated datasets from more than one corpora
  – others are red crosses ✗
Feedback on draft proposal: ‘negative’

- Proposals are very limited in scope (“gloss annotations”)
  - divergence in other more critical areas of corpus processing not addressed and thus not prevented by standardization in this basic area (if achieved)
  - testing of schemas/alternatives not proposed
    - why not see which team (given comparable time and resources) using whichever approach is able to produce the most usable and informative linguistic corpus, i.e., as reflected in analytic output

- Different proposals for the treatment of sign beginning and end times appears to be problematic
  - i.e., the failure to try and discriminate between meaningful and non-meaningful duration—holds, perseveration, buoys etc.—and thus the desired time alignment of strong and weak hands may render some datasets linguistically intractable
  - i.e., it may not be easy or possible to do quantitative and qualitative constructional analysis
Basic glossing conventions for different types of signs

<table>
<thead>
<tr>
<th>Sign type</th>
<th>Example Gloss</th>
<th>Type component</th>
<th>Type-like component</th>
<th>Token-like component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully lexical</td>
<td>WATER</td>
<td>WATER</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Partly lexical</td>
<td>PT:PRO1SG</td>
<td>PT:</td>
<td>PRO1SG</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td>DSM(1-VERT):PERSON-GO-DOWN-HILL</td>
<td>DS</td>
<td>M(1-VERT): PERSON-GO-DOWN-HILL</td>
<td>n/a</td>
</tr>
<tr>
<td>Non-lexical</td>
<td>G(5-UPWARDS):WELL</td>
<td>G</td>
<td>(5-UPWARDS): WELL</td>
<td>FLAILING-ABOUT-IN-WATER</td>
</tr>
<tr>
<td></td>
<td>G:FLAILING-ABOUT-IN-WATER</td>
<td>G:</td>
<td>n/a</td>
<td>FLAILING-ABOUT-IN-WATER</td>
</tr>
</tbody>
</table>

Annotation conventions for different types of signs and for different purposes would benefit by being standardized across SL corpora. However, **the actual symbols/system used is less important than the principle of recognizing different of semiotic sign types in SLs and reflecting this in the differential glossing of signs.**

How does one gloss fully lexical signs? By using the ID gloss.

How does one know what ID gloss to use? By consulting a dictionary of the language.
The annotation tiers

• Core or fixed tiers

• Study specific tiers
  – research questions
    • e.g., aspect
  – temporary & derived data
    • co-occurrence information
    • annotations from overlapping annotations
RH-IDgloss

LH-IDgloss

Free Translation
Primary processing

Basic
• Free translation
  – written (preferable as minimum)
    • spoken is a bonus (potentially quicker), but
    • written is immediately searchable (i.e., spoken has itself to be transcribed before becoming searchable)
• Tokenization of the signing stream
  – identify and gloss
  – discriminate types of signs

Additional possible detailed annotations
• Units larger than the individual sign
  – clause-like units (CLUs)
• Identified (provisionally) through
  – delivery (articulatory units)
  – meaning (coherent proposition or ‘move’ in exchange)
  – form/structure (constructional schema)
• Non-manual features *(relates to Delivery)*
  – body posture, head movements, gaze, facial expressions, mouthing, mouth gestures
Secondary processing

Sign token tagging

• transcription (with or without dedication notation system)
• phonological features (standard parameter model)
• morphological features
  – direction, location etc.

CLU token tagging

• Literal translation
• Situation type
  – state, accomplishment, achievement, activity
• CLU dependency status
  – independent, dependent, embedded
  – manner of marking of dependency/relationship
    • lexis, intonation contour, none (juxtaposition & context)
Secondary processing (cont.)

• Tagging of sign tokens as CLU constituents
  – Grammatical class (PoS)
    • problematic!
  – Argument type and order
  – Macro roles
  – Semantic roles

• Tagging of sign tokens as phrase constituents
  – Not attempted yet in the Auslan Corpus
Tertiary processing

• creating new annotations based on existing annotations

• using the corpus to enrich the corpus by
  – incorporating results of searches, sorting and quantification based on primary and secondary processing into the corpus itself as new annotations

• both enable one to conduct subsequent searches using the newly incorporated annotations as given information or as new constraints
Problems/issues

Segmenting into sign units and glossing at the sign-level only

• Any indication that the sequences identified are CLU-specific
  – i.e., they occur within CLU constructions (and are thus potential schemas), rather than being unrelated juxtapositions of annotations (they span clause boundaries)
  – without multi-sign unit annotation one cannot tell

Separating additional token information on other tiers

• Dispersed information over multiple tiers keeps the gloss annotation clean and simple and “true to type”, but
  • One then needs to exploit this by recombining the data dispersed across several tiers
    – into a single complex annotation (File > Annotations from overlaps)
    – into datasets in spreadsheets in which records (signs or clauses) can be sorted according to various annotation values over multiple tiers (File > Annotation Overlaps Information)
Annotaton overlaps informaon data in Excel

1260 CLUs analyzed for argument structure
- NOTE: sometimes arguments are realised on the weak hand or in CA

Of the 1260, 730 are identiﬁed as single CLUs (not embedded, not in a dependency relationship with another CLU).
Only 24 of the 730 CLUs are A1 V A2, i.e. potential overt S V O order
Exported data can be tagged for additional processing, e.g., in Rbrul.
Conclusion
Use different types of tiers to separate out different types of information that can be combined and recombined in different ways to answer different types of research questions.
Use different types of annotations at different stages of corpus creation

- do not try to do everything at once
- resist the temptation to transcribe for transcription’s sake
- always ask “What am I trying to do?”