Day 2
Annotation for morphosyntactic research

Trevor Johnston
Centre for Language Sciences, Department of Linguistics
Macquarie University
Sydney, Australia

SLCN Workshop 3: Annotation of sign language corpora
14-16 June, Stockholm.

Outline Day 2

1. Annotation for morpho-syntactic research
2. Auslan Corpus annotation conventions
3. Implementation
   – examples
   – issues & problems
1. Annotation for morpho-syntactic research

- basic annotation vs detailed representation
  - does detailed representation only mean phonetic/phonological detail ("transcription")?
  - "generation" (i.e. reproduction) not the aim
- morpho-syntax & conventional signs/constructions
  - type/token relationship
  - form/meaning relationship
    - annotation as tagging for assessing the systematic co-occurrence of certain potential/assumed morpho-syntactic markers with certain types of meanings

2. Auslan Corpus annotation conventions

- for secondary processing
- for tertiary processing
Secondary processing

- annotations/tags vs ‘glossing’
  - phonological
  - morphological
  - syntactic
  - semantic
  - pragmatic
  - discourse

- constrain searches
  - vertically (tier values)
  - horizontally (environment)
  - domain/metadata
Secondary processing: sign-based

- meaning
- grammatical class (‘part of speech’)
- transcription
  - orientation
- modification and/or variants

Secondary processing: clause-based

- clause
- arguments
- dependency
- macro-roles
- semantic roles
- overt “subject”
Clause

(72) IDgloss
Clause
Clause
ClauseArg

Arguments

(75) IDgloss
Clause
ClauseArg

PT:PRO3SG BUY CAR YESTERDAY
Clause#1

A1 V A2 nonA
Dependency

IF NOT PAY FS: BILL, WILL CUT-WITH-SCISSORS PHONE

Clause | ClauseDep | FreeTransl
-------|-----------|-------------
Clause#1 | beta | If I don't pay the bill, they will disconnect my phone.
Clause#2 | alpha | 

Macro-roles 1

(76) IDgloss | PT:PRO3SG | BUY | CAR | YESTERDAY
---|---|---|---|---
Clause | Clause#1 | 
ClauseArg | A1 | V | A2 | nonA |
MacroRole | ACTR | PROCESS | UNDR |
FreeTransl | He bought a car yesterday.

(77) IDgloss | PT:PRO3SG | GO | SHOP | YESTERDAY
---|---|---|---|---
Clause | Clause#1 | 
ClauseArg | A1 | V | nonA | nonA |
MacroRole | ACTR | PROCESS |
FreeTransl | He went to the shop yesterday.

(78) IDgloss | PT:PRO3SG | SAY | NO |
---|---|---|---
Clause | Clause#1 | 
ClauseArg | A1 | V | A2 |
MacroRole | ACTR | PROCESS | COMPL |
FreeTransl | He said “No” |
Macro-roles 2

(79) IDgloss  WOMAN  DOCTOR
Clause    Clause#1
ClauseArg A1  A2
MacroRole CARRIER  ATTRIBUTE
FreeTransl The woman is a doctor.

Semantic roles 1

<table>
<thead>
<tr>
<th>Tag</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROCESS</td>
<td>The process (verb).</td>
</tr>
<tr>
<td>AGENT</td>
<td>The agent (also 'actor'): instigator of some action, under agents volitional control</td>
</tr>
<tr>
<td>EXPERIENCER</td>
<td>The experiencer: entity experiencing some psychological state</td>
</tr>
<tr>
<td>PATIENT</td>
<td>The patient (also 'theme'): entity undergoing the effect of some action</td>
</tr>
<tr>
<td>UTERANCE</td>
<td>The utterance: the signs or words uttered by someone</td>
</tr>
<tr>
<td>ENACTMENT</td>
<td>The enactment: the non-linguistic actions performed by someone</td>
</tr>
<tr>
<td>SOURCE</td>
<td>The source: entity from which something moves</td>
</tr>
<tr>
<td>GOAL</td>
<td>The goal: entity towards which something moves</td>
</tr>
<tr>
<td>BENEFICIARY</td>
<td>The beneficiary (also 'benefactive'): the entity benefiting from some action</td>
</tr>
<tr>
<td>INSTRUMENT</td>
<td>The instrument: means by which something comes about</td>
</tr>
<tr>
<td>LOCATION</td>
<td>The location (also 'locative') place in which something is situated</td>
</tr>
<tr>
<td>MANNER</td>
<td>The manner: the way in which something is done</td>
</tr>
<tr>
<td>PATH</td>
<td>The path: the route in which something moves</td>
</tr>
<tr>
<td>TIME</td>
<td>The time: the time in which an action takes place</td>
</tr>
</tbody>
</table>

(80) IDgloss  PT:PROSG  GO  SHOP  YESTERDAY
Clause    Clause#1
ClauseArg A1  V  nonA  nonA
MacroRole ACTR  PROCESS
SemRole   AGENT  PROCESS
Semantic roles 2

(81) IDgloss: PT:PRO3SG  
Clause: Clause#1  
ClauseArg: A  
MacroRole: ACTR  
SemRole: AGENT  
CA: CD:girl  
LiteralTrans: She (said) "(it is/I am) fine"

(82) IDgloss: PT:PRO3SG  
Clause: Clause#1  
ClauseArg: A  
MacroRole: ACTR  
SemRole: AGENT  
CA: CA:teacher  
LiteralTrans: He (the teacher) (went) [hit his palm on his forehead in self reproach]

Secondary processing: non-manual

(87) IDGLOSS:  
Mouthing:  
FreeTransl: YES  
BUT  
PRO2SG  
NEG  
DELIBERATE  
deliberate)  
Yes, but he didn't do it deliberately.

(88) IDGLOSS:  
MouthGesture:  
FreeTransl: YES  
BUT  
PRO2SG  
NEG  
DELIBERATE  
pah  
Yes, but he didn't do it deliberately.
Secondary processing: non-manual

(86) IDgloss

<table>
<thead>
<tr>
<th>Body</th>
<th>UNDERSTAND</th>
<th>SCIENCE</th>
<th>UNDERSTAND-NOT</th>
<th>SCIENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>FreeTransl</td>
<td>left:doctor</td>
<td>right:priest</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The doctor understood science, whereas the priest didn’t understand science.

(89) IDgloss

<table>
<thead>
<tr>
<th>Clause</th>
<th>POLICEMAN</th>
<th>SEARCH</th>
<th>FS:CLUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>CA:POLICEMAN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Head</td>
<td>SIDE TO SIDE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eyes</td>
<td>SQUINT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FreeTransl The policeman looked for clues, squinting and moving his head side to side.

Secondary processing: non-manual

(90) IDgloss

<table>
<thead>
<tr>
<th>Clause</th>
<th>BARMAN</th>
<th>G(NNS):WINK</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA tier:</td>
<td>CA:BARMAN</td>
<td></td>
</tr>
<tr>
<td>Eyes</td>
<td>WINK</td>
<td></td>
</tr>
<tr>
<td>LitTrans</td>
<td>The barman (went like this) [wink].</td>
<td></td>
</tr>
<tr>
<td>FreeTransl</td>
<td>The barman winked.</td>
<td></td>
</tr>
</tbody>
</table>

(91) IDgloss

<table>
<thead>
<tr>
<th>Clause</th>
<th>PT:PRO3SG</th>
<th>G:HOLD-ON-TO:SCAFFOLD-AND-SWAY</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>CA:SCHOOLBOY</td>
<td></td>
</tr>
<tr>
<td>Body</td>
<td>SIDE TO SIDE</td>
<td></td>
</tr>
<tr>
<td>FreeTransl</td>
<td>He, the schoolboy, held on the scaffolding and rocked it side to side.</td>
<td></td>
</tr>
</tbody>
</table>
Secondary processing: CA/CD

Previous slide also

<table>
<thead>
<tr>
<th>Clause</th>
<th>F5: EGGS</th>
<th>ROIL</th>
<th>BETTER</th>
<th>DSH(BENT7): TURN-DOWN</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>CA: EGG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Face</td>
<td>STARTLED-AND-WORRIED</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Head</td>
<td>RAPID-LITTLE-SHAKES</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

LitTrans: The eggs are being thrown about everywhere in the boiling water and they are worried [that they'll break]. It would be better to turn the stove down.
FreeTrans: The eggs are boiling too vigorously. It would be better to turn the stove down.

Tertiary processing

- using the corpus to enrich the corpus by
  - implementing primary and secondary processing
  - conducting searches based on primary and secondary processing
  - incorporating these findings into existing annotations
- conducting subsequent searches using the newly incorporated annotations as constraints
Tertiary processing

- frequency tagging
- co-occurrence tagging
  - e.g., constructed action
Outline Day 2

• Annotation for morpho-syntactic research
  – basic annotation vs detailed representation
  – does detailed representation only mean phonetic/phonological detail (“transcription”)?
• Auslan Corpus annotation conventions
  – for secondary processing
  – for tertiary processing
• Implementation
  – examples
  – issues & problems

Examples

• Secondary processing
• Tertiary processing
Examples of secondary searches

• Lexical frequency by grammatical class
• Points (PT)
  – palm orientation and PT-type
  – eye-gaze and PT-type
• Clause types
  – number and order of overt arguments
  – semantic role of argument in position

Palm orientation and PT type

• Is there a relationship with palm orientation and type of point?
• Search for all pointing signs and determine the distribution of their palm orientations
  – IDglosses being with “PT” (^PT)
  – IDglosses overlap an orientation annotation (.+)
  – Export hits into a database, sort and extract statistics
• Answer: strong apparent relationship between palm orientation and point type
Important note: This is not a formal presentation of findings. The data are incomplete. They are from work in progress and are being used here merely to show what can be done using current corpus resources. A formal report is pending.
Eye gaze and PT type

- Is there a relationship with direction of eye gaze and type of point?
- Search for all pointing signs and determine the distribution of the co-temporal eye gaze
  - IDglosses being with “PT” (^PT)
  - IDglosses overlap an eye gaze annotation (+)
  - Get annotation/hit statistics or export hits into a database, sort and extract statistics
- Answer: strong apparent relationship between eye gaze and point type

Important note: This is not a formal presentation of findings. The data are incomplete. They are from work in progress and are being used here merely to show what can be done using current corpus resources. A formal report is pending.
Lexical frequency by grammatical class

- search all lexical glosses on IDgloss tier
  - ^[^QPT\E|^QDS\E|^QFS\E|^QG\E|^QG(\E]
- which overlap an annotation on the grammatical class tier
- view hits in statistic view
- export hits into a database, sort and extract statistics
- all grammatical class annotations incomplete
  – done for “FINISH”

Grammatical class tagging for specific IDglosses across entire corpus (RH)*

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td></td>
</tr>
<tr>
<td>#1</td>
<td></td>
</tr>
<tr>
<td>#1</td>
<td></td>
</tr>
<tr>
<td>#1</td>
<td></td>
</tr>
<tr>
<td>#1</td>
<td></td>
</tr>
<tr>
<td>#1</td>
<td></td>
</tr>
<tr>
<td>#1</td>
<td></td>
</tr>
</tbody>
</table>

* FINISH IDglosses in entire corpus tagged for grammatical class
### Tabular summary of previous hits

<table>
<thead>
<tr>
<th></th>
<th>FINISH-FIVE %</th>
<th>FINISH-GOOD %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Adverb</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>Auxiliary</td>
<td>31</td>
<td>38</td>
</tr>
<tr>
<td>Conjunction</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Discourse marker</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>Noun</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Predicate</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>Verb</td>
<td>19</td>
<td>39</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Important note: This is not a formal presentation of findings. The data are incomplete. They are from work in progress and are being used here merely to show what can be done using current corpus resources. A formal report is pending.

### Argument structure

- What is the range and frequency of argument structures/positions found in Auslan?

- Is there a relationship between position of argument in string and the semantic role of the argument(s)?
Blue = merged Clause boundaries and identified Arguments and Verbs
Orange = merged Arguments and Semantic Roles
Green = merged Argument with Semantic Roles and Clause boundaries
Why merge & concatenate?

We have:

<table>
<thead>
<tr>
<th>ID gloss tier</th>
<th>BOY</th>
<th>EAT</th>
<th>CAKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clause tier</td>
<td>Clause#1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argument tier</td>
<td>A1</td>
<td>V</td>
<td>A2</td>
</tr>
</tbody>
</table>

- Multi-tier co-occurrence (overlap) searches are inadequate. Statistics or exports represent hits as:
  Clause#1, A1
  Clause#1, V
  Clause#1, A2
  etc.

- In order to identify and quantify clause argument sequences we want:
  Clause#1 = A1, V, A2
  Clause#2 = A1, A2, V
  Clause#3 = A2, A1, V
  etc.

Frequency of argument orders in clauses

<table>
<thead>
<tr>
<th>Overt argument order</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>27</td>
</tr>
<tr>
<td>A V</td>
<td>7</td>
</tr>
<tr>
<td>A1 V A2</td>
<td>6</td>
</tr>
<tr>
<td>V A</td>
<td>6</td>
</tr>
<tr>
<td>A1 A2</td>
<td>4</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Overt argument order</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>V (PROCESS)</td>
<td>27</td>
</tr>
<tr>
<td>A (AGENT) V (PROCESS)</td>
<td>6</td>
</tr>
<tr>
<td>A1 (AGENT) V (PROCESS) A2 (PATIENT)</td>
<td>4</td>
</tr>
<tr>
<td>A1 (CARRIER) A2 (ATTRIBUTE)</td>
<td>3</td>
</tr>
<tr>
<td>V (PROCESS) A (PATIENT)</td>
<td>3</td>
</tr>
<tr>
<td>A (ATTRIBUTE)</td>
<td>2</td>
</tr>
<tr>
<td>A (EXPERIENCER) V (PROCESS)</td>
<td>1</td>
</tr>
<tr>
<td>A (UTTERANCE)</td>
<td>1</td>
</tr>
<tr>
<td>A1 (AGENT) V (PROCESS) A2 (GOAL)</td>
<td>1</td>
</tr>
<tr>
<td>A1 (ENTITY) A2 (LOCATION)</td>
<td>1</td>
</tr>
<tr>
<td>A1 (EXPERIENCER) V (PROCESS) A2 (SOURCE)</td>
<td>1</td>
</tr>
<tr>
<td>V (PROCESS) A (ENTITY)</td>
<td>1</td>
</tr>
<tr>
<td>V (PROCESS) A (LOCATION)</td>
<td>1</td>
</tr>
<tr>
<td>V (PROCESS) A (UTTERANCE)</td>
<td>1</td>
</tr>
</tbody>
</table>

Important note: This is not a formal presentation of findings. The data are incomplete. They are from work in progress and are being used here merely to show what can be done using current corpus resources. A formal report is pending.
Examples

• Secondary processing
• Tertiary processing

Incorporating sign frequency

• determine sign frequency
  – as per previous discussion/demonstration
• assign frequency codes to iDGlosses in annotation files
  – tag iDGlosses with frequency code, e.g. “VHF” and “HF” for “very high frequency” & “high frequency”, “LF” for low frequency, etc.
• search for co-occurrence of annotations on
  – iDGloss tier
  – frequency value tier, and
  – any another annotation value under investigation
    • e.g. sign modification, non-manuals, constructed action, etc.
• export hits into a database, sort and extract statistics
Indicating verb modification

  - study based on using this approach
- Further investigations underway for enriched corpus
  - argument sequence/presence & semantic role
  - pointing signs & verb modification
  - overt arguments and constructed action
  - depicting signs and argument structure
  - etc. etc.
  - not the place/time to report here

Outline

- The case for SL corpus linguistics
- Primary processing
  - search examples
- Secondary processing
  - search examples
- Tertiary processing
  - search examples
- Conclusion
  - enabling and streamlining workflows
Enabling and streamlining workflows

Standardization across corpora
• primary processing
  – expectation that minimal processing be implemented before the label ‘corpus’ is applied (in the ‘modern sense’)
• secondary processing
  – expectation that schemas/codes/criteria be publicly available
  – relevance of schemas like Leipzig glossing rules

Standardization within corpora
• mandatory aspect
  – internal consistency (i.e. of labels) is not just “best practice”
  – there is no other way to do empirical language research
• practical aspect
  – internal consistency (i.e. in the ASSIGNMENT of labels)
  – can never be guaranteed, result of review of accessible data

What is the data?

• Too much taken for granted
  – doing linguistics vs assuming linguistic research ‘completed’
  – discovering vs processing
  – rubbish in, rubbish out
Wish list or future developments?

- Multiple files, not single files
- Search on ‘found sets’ (search on search results)
- Search and replace
  - exact match & substring, replace
  - exact match or substring
  - replace/expand annotation values based on a look up table
  - insert a specified value in on a specified tier in a specified annotation search result domain (or associated empty annotation field)
- Copy and/or merge over entire corpus sets, not single files
  - unless improved searching capabilities render this unnecessary
- Overlapping of superordinate annotation (e.g., clause domain) with specified sequence of ‘included’ annotations on another tier (e.g., arguments)
  - (i.e., tier X annotation overlaps sequence of annotation 1 on tier Y with annotation 2 on tier Z, in that specified order)
- and much much more....
  - we need to articulate what we do and why so that processing aids linguistic research and so that sophisticated computational processing is of genuine SL data, not idealised non-representative data-sets