Type/token matching in annotated SL corpora: Examples from Auslan and BSL corpus projects

Trevor Johnston¹, Adam Schembri², Kearsy Cormier³, Jordan Fenlon³, Ramas Rentelis³

¹Macquarie University, Sydney, ²La Trobe University, Melbourne, ³University College London

“Building Sign Language Corpora in North America” workshop, Gallaudet University, May 21-22, 2011
Preamble: limited focus

• This presentation builds on a series of presentations, workshops and publications on the topic of SL corpus creation in general and ID-glossing in particular:
  – you can follow up on these if you download the PDF of this PowerPoint from the workshop site
Preamble: some previous work


Preamble: limited focus

• This presentation builds on a series of presentations, workshops and publications on the topic of SL corpus creation in general and ID-glossing in particular
  – you can follow up on these if you download the pdf of this PowerPoint from the workshop site

• There is no time today to go into all related issues or justification in detail
  – the focus here is on clarification of ID-glossing and exemplification of its application
Outline

• Of types and tokens
• Kinds of symbolic units in SLs
• ID-glosses: what they are, what they are not
• ID-glosses at work in corpus-based research
Of types and tokens

• a fundamental characteristic of categorization of phenomena

• a type
  – represents an overall category generally, ideally or abstractly

• a token
  – is any particular or concrete instance of a phenomenon
Token & types in linguistics

• within semiotics and linguistics a symbolic unit is a pairing of a form with a meaning
• symbolic units come in different shapes and sizes (small/large, atomic/complex)
• linguists deal with instances of actual concrete meaningful expressions that are believed to be and analyzed as tokens of linguistic types (phonemes, words, phrases, clauses, etc.)
Kinds of word-level symbolic units in SLs

- lexical signs can be distinguished from other kinds of single sign units in SLs
- researchers have recognized a distinction between core fully-specified lexical signs and other types of signs at the word level in SLs
  - e.g. partly specified signs, signs using classifier handshapes, gestures
  - different labels and differing theoretical perspectives have been used to describe different kinds of signs
Kinds of word-level symbolic units in SLs (2)

• Whatever framework is used, it is essential to distinguish in the glossing and annotation schema between symbolic units which are
  – **tokens** of clearly identifiable and specifiable **types**
    • called “fully-lexical” signs in the following table
  – mixed, with both **type/token** characteristics
    • called “partly-lexical” signs in the following table
  – **tokens only**, they are **not** instances of documented conventionalized **types**
    • called “non-lexical” signs in the following table
One possible overview*

<table>
<thead>
<tr>
<th>Kinds of signed units found in SLs</th>
<th>Some alternative SL terminology</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fully-lexical signs:</strong> numbers, name signs, fingerspelling; all citable signs, incl. spatially unmodified indicating signs (i.e. citation forms).</td>
<td>Frozen signs, lexical signs, standard signs, fingerspelling; citation sign forms of spatial and agreeing signs.</td>
</tr>
<tr>
<td>spatially modified indicating signs, any spatially located/placed sign, (incl. possessives?).</td>
<td>Inflected spatial and agreeing signs, any spatially located/placed sign (‘pragmatic agreement’), possessives.</td>
</tr>
<tr>
<td><strong>Partly-lexical signs:</strong> depicting signs, pointing signs (incl. possessives?).</td>
<td>Indexical signs, pronominal points, pronominal signs, pronouns, possessives, classifier signs, classifier handshapes, classifiers, some highly iconic structures.</td>
</tr>
<tr>
<td><strong>Non-lexical signs:</strong> idiosyncratic gestures, enactments, other vocal or gestural fragments.</td>
<td>Gesture, mime, pantomime, some highly iconic structures.</td>
</tr>
</tbody>
</table>

Corpus gloss-based annotations of different kinds of symbolic units

- fully-lexical signs
  - TYPE-NAME
    - BOY, HOUSE, LOOK, THINK, APPLE, WAIT, BEFORE
      TYPE-NAME = ID-gloss

- partly-lexical
  - TYPE-LIKE-INFORMATION:TOKEN-LIKE-INFORMATION
    - DS:PERSON-PASSES-LEFT-TO-RIGHT
    - CL:PERSON-PASSES-LEFT-TO-RIGHT
    - PM:PERSON-PASSES-LEFT-TO-RIGHT
      DS = depicting sign, CL = classifier, PM = property marker.
      These are all alternative names for similar th

- non-lexical
  - CATEGORY-CODE:TOKEN-INFORMATION
    - G:DISSMISSIVE-DOWNWARD-HANDSTROKE
      G = gesture
ID-glosses

What they are

• Unique identifying gloss-based names for individual sign forms
  – if possible, based on a word in the majority language commonly associated with a fully-lexical sign

• Meant to be used in the annotation of a SL corpus for internal consistency and ease of memory/entry for annotators
  – no need to use any special fonts or transcription systems
  – just to have a primary annotation of the text

What they are not

• They are not the ‘official’ name of a sign, but
  – essential for corpus-based research
  – key in building lexical databases

• They do not replace the normal practice of contextual glossing in linguistic discussion of SL texts/examples.
  – ID-glosses only of use and interest to researchers working with databases and corpora
ID-glosses at work in corpus-based linguistic research

Some recent examples

• Distribution of sign types
  – across corpora
  – according to genre

• Lexical frequency
  – overall lexical frequency
  – lexical frequency by grammatical class

• Morpho-syntactic research
  – indicating verb modification, potential vs actual
  – indicating verb modification, interaction with constructed action

Other & on-going studies

• Gaze behavior
  – direction with types of points
  – indicating verbs

• Palm orientation
  – direction with types of points

• phonological processes (e.g. handshape assimilation)
  – pointing signs

• syntactic distribution
  – points in clauses
  – points with nouns & verbs
  – depicting signs in clauses
  – fingerspelling collocations, e.g. pre- or post-depicting signs
# Distribution of sign types

<table>
<thead>
<tr>
<th>Type</th>
<th>%</th>
<th>Type</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Auslan</strong></td>
<td></td>
<td><strong>BSL</strong></td>
<td></td>
</tr>
<tr>
<td>Fully-lexical (all other)</td>
<td>65.0</td>
<td>Fully-lexical (all other)</td>
<td>60.0</td>
</tr>
<tr>
<td>Fully-lexical (fingerspelling)</td>
<td>5.0</td>
<td>Fully-lexical (fingerspelling)</td>
<td>4.6</td>
</tr>
<tr>
<td>Fully-lexical (name signs)</td>
<td>0.2</td>
<td>Fully-lexical (name signs)</td>
<td>0.3</td>
</tr>
<tr>
<td>Partly-lexical (pointing/indexical, incl. possessives)</td>
<td>12.3</td>
<td>Partly-lexical (pointing/indexical, incl. possessives)</td>
<td>23.0</td>
</tr>
<tr>
<td>Partly-lexical (depicting/classifiers)</td>
<td>11.0</td>
<td>Partly-lexical (depicting/classifiers)</td>
<td>3.0</td>
</tr>
<tr>
<td>Non-lexical (gestures, incl. fragments)</td>
<td>6.5</td>
<td>Non-lexical (gestures, incl. fragments)</td>
<td>9.0</td>
</tr>
</tbody>
</table>
Overall lexical frequency

<table>
<thead>
<tr>
<th>Rank</th>
<th>Auslan</th>
<th>BSL</th>
<th>Percentage</th>
<th>Raw Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SCHOOL</td>
<td>BSL</td>
<td>0.50</td>
<td>0.09</td>
</tr>
<tr>
<td>2</td>
<td>BSL</td>
<td>BSL</td>
<td>0.24</td>
<td>0.04</td>
</tr>
<tr>
<td>3</td>
<td>BSL</td>
<td>BSL</td>
<td>0.18</td>
<td>0.04</td>
</tr>
<tr>
<td>4</td>
<td>BSL</td>
<td>BSL</td>
<td>0.16</td>
<td>0.04</td>
</tr>
<tr>
<td>5</td>
<td>BSL</td>
<td>BSL</td>
<td>0.14</td>
<td>0.04</td>
</tr>
<tr>
<td>6</td>
<td>BSL</td>
<td>BSL</td>
<td>0.09</td>
<td>0.02</td>
</tr>
<tr>
<td>7</td>
<td>BSL</td>
<td>BSL</td>
<td>0.08</td>
<td>0.02</td>
</tr>
<tr>
<td>8</td>
<td>BSL</td>
<td>BSL</td>
<td>0.07</td>
<td>0.02</td>
</tr>
<tr>
<td>9</td>
<td>BSL</td>
<td>BSL</td>
<td>0.07</td>
<td>0.02</td>
</tr>
<tr>
<td>10</td>
<td>BSL</td>
<td>BSL</td>
<td>0.07</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Note: The table above shows the overall lexical frequency for Auslan and BSL, with percentages and raw frequencies provided for each term.
## Overall lexical frequency

### Auslan

<table>
<thead>
<tr>
<th>Rank</th>
<th>ID gloss</th>
<th>% of db</th>
<th>Cumulative %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PT:PRO1</td>
<td>5.08</td>
<td>5.08</td>
</tr>
<tr>
<td>2</td>
<td>G(5-UP):WELL</td>
<td>3.57</td>
<td>8.65</td>
</tr>
<tr>
<td>3</td>
<td>PT:PRO2/PT:PRO3</td>
<td>3.25</td>
<td>11.90</td>
</tr>
<tr>
<td>4</td>
<td>DEAF1/2</td>
<td>1.48</td>
<td>13.38</td>
</tr>
<tr>
<td>5</td>
<td>LOOK</td>
<td>1.44</td>
<td>14.82</td>
</tr>
<tr>
<td>6</td>
<td>BOY</td>
<td>1.27</td>
<td>16.09</td>
</tr>
<tr>
<td>7</td>
<td>PT:LOC</td>
<td>1.25</td>
<td>17.34</td>
</tr>
<tr>
<td>8</td>
<td>DSM/L(BENT2):ANIMATE-MOVES/AT</td>
<td>1.06</td>
<td>18.40</td>
</tr>
<tr>
<td>9</td>
<td>HAVE</td>
<td>1.06</td>
<td>19.46</td>
</tr>
<tr>
<td>10</td>
<td>SAME</td>
<td>1.02</td>
<td>20.48</td>
</tr>
</tbody>
</table>

### BSL

<table>
<thead>
<tr>
<th>Rank</th>
<th>ID gloss</th>
<th>Percentage</th>
<th>Raw figure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PT:PRO1</td>
<td>6.99</td>
<td>869</td>
</tr>
<tr>
<td>2</td>
<td>G:WELL</td>
<td>5.74</td>
<td>714</td>
</tr>
<tr>
<td>3</td>
<td>PT:PRO3</td>
<td>4.20</td>
<td>523</td>
</tr>
<tr>
<td>4</td>
<td>PT</td>
<td>3.10</td>
<td>386</td>
</tr>
<tr>
<td>5</td>
<td>GOOD</td>
<td>2.07</td>
<td>258</td>
</tr>
<tr>
<td>6</td>
<td>PT:LOC</td>
<td>1.74</td>
<td>217</td>
</tr>
<tr>
<td>7</td>
<td>PT:PRO2</td>
<td>1.70</td>
<td>211</td>
</tr>
<tr>
<td>8</td>
<td>PT:DET</td>
<td>1.19</td>
<td>148</td>
</tr>
<tr>
<td>9</td>
<td>SAME</td>
<td>0.98</td>
<td>122</td>
</tr>
<tr>
<td>10</td>
<td>G:HEY</td>
<td>0.88</td>
<td>110</td>
</tr>
<tr>
<td>Rank</td>
<td>ID gloss/lemma</td>
<td>ID gloss/lemma by grammatical class</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>---------------</td>
<td>-------------------------------------</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>DEAF1/2</td>
<td>HAVE</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>LOOK</td>
<td>LOOK</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>BOY</td>
<td>BOY</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>HAVE</td>
<td>PEOPLE</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>SAME</td>
<td>TORTOISE</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>GOOD</td>
<td>SAY</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>WHAT</td>
<td>WOLF</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>THINK</td>
<td>RABBIT</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>NOTHING</td>
<td>THINK</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>NOT</td>
<td>DEAF1/2</td>
<td></td>
</tr>
</tbody>
</table>

Frequency of fully-lexical signs by grammatical class: Auslan data
### Frequency of fully-lexical signs by grammatical class: Auslan data (2)

<table>
<thead>
<tr>
<th>Grammatical class</th>
<th>FINISH.6</th>
<th>FINISH.5</th>
<th>Grammatical class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
<td>38.6%</td>
<td>45.8%</td>
<td>Auxiliary</td>
</tr>
<tr>
<td>Auxiliary</td>
<td>30.8%</td>
<td>13.6%</td>
<td>Adverb</td>
</tr>
<tr>
<td>Interjection</td>
<td>9.2%</td>
<td>13.6%</td>
<td>Verb</td>
</tr>
<tr>
<td>Discourse marker</td>
<td>7.7%</td>
<td>10.1%</td>
<td>Conjunction</td>
</tr>
<tr>
<td>Noun</td>
<td>6.9%</td>
<td>10.1%</td>
<td>Discourse marker</td>
</tr>
<tr>
<td>Conjunction</td>
<td>3.0%</td>
<td>5.1%</td>
<td>Interjection</td>
</tr>
<tr>
<td>Adjective</td>
<td>3.0%</td>
<td>1.7%</td>
<td>Noun</td>
</tr>
<tr>
<td>Adverb</td>
<td>0.8%</td>
<td>0.0%</td>
<td>Adjective</td>
</tr>
</tbody>
</table>

**Ratio of tokens:** 2 : 1
Morpho-syntactic research
(Auslan example)

Frequency of spatial modification of indicating verbs

- **LEFT COLUMN**
  - directional signs spatially un-modified in almost 50% of tokens in corpus

- **RIGHT COLUMN**
  - the ‘locatable’ sub-group is rarely spatially modified
Morpho-syntactic research (Auslan example)

Distribution of spatial modification according to sign frequency group

**LEFT COLUMN**
- very high frequency directional verb types are less than 10% of all directional verb types in corpus

**RIGHT COLUMN**
- yet they are >50% of all modified tokens of all directional verb types
Is ID-glossing mandatory!?

• Are ID-glosses meant to be used in the written representation of signs in all contexts?
• No, contextual glosses or dedicated phonetic or phonological transcription systems are still needed
  – in written representations of SL texts, or
  – in transcriptions embedded in corpora
• The corpus creators do not determine official names for signs
  – ID-glosses are only a tool
Transcription is distinct from tokenization
Normal contextual glossing

• the selective representation of signed text in
  – an academic publication
  – language teaching materials
  – presentations in/to Deaf community
  – informal written communication

• meaning of a sign is highlighted
  – for ease of representation by the author and processing by the audience
Normal contextual glossing is not made redundant by ID-glossing

She never invites me, so I’m not going to invite her.

I never got her invitation

INVITE/INVITATION
– may or may not be formal differences between two tokens
– irrelevant for point being illustrated
– use of ID-gloss would only be confusing for most audiences
– hyperlinks to lexical database (dictionary) can quickly establish gloss as merely convenient or categorical

[Web slide]
Hyperlinking to corpora

- likely to become the norm very soon
  - such that the glossed textual example will merely become a place holder in a written text
  - the “quoted” signed text need not be represented in ID-glosses, it merely needs to be put in a relevant and readable form to illustrate the point at hand.

- The hyperlink would take you directly to ELAN annotation file at the very time point of the glossed example (where the ID-gloss would be visible). This is being developed for Auslan:

RHgloss | BECAUSE PT:DET TEACHER NOT UNDERSTAND SIGN
LHgloss | PT:LOC______________________________
trans | ...because the teacher did not understand sign language.
The assignment of ID-glosses...

- Primarily corpus creators who use the corpus as a research tool
  - linguists, researchers, annotators
- decide on the ID-glosses
  - based on intimate knowledge of the lexical database and existing ID-glosses
- They do this in the light of feedback from
  - the deaf community
  - native-signers
  - the wider Auslan-using community

Conclusion

• ID-glossing is simply a manifestation of the recognition of type/token relationships in linguistic data
  – it helps to identify tokens as instances of types
• ID-glosses are unique identifiers
  – these unique identifiers could be anything (numbers, graphic symbols, even Chinese characters)
  – but standardized glosses based on the words and orthography of the majority language are the best way to achieve a unique identifier AND keep the annotators’ task as simple as possible
  – they have little use outside the corpus and lexical database environment, but are essential in that environment
Acknowledgments

**Auslan archive & corpus**
- Hans Rausing Endangered Languages Documentation Project grant #MDP0088
- Australian Research Council grants #LP09346973 (Schembri, Johnston), #DP0665254 (Johnston, de Beuzeville), #DP1094572 (Johnston)
- Co-researchers, research assistants or annotators who have contributed to the current body of corpus annotations since 2003—Adam Schembri, Julia Allen, Donovan Cresdee, Karin Banna, Michael Gray, Dani Fried, Della Goswell, and Gerry Shearim—and to current post-graduate students who have also contributed to corpus annotations—Lindsay Ferrara, Gabrielle Hodge, Michael Gray and Lori Whynot.
- The 255 deaf signers who participated in the Auslan archive and corpus data collection.

**BSL corpus project**
- UK Economic and Social Research Council grant RES-620-28-6001 (Schembri, Cormier, Deuchar, Elton, O’Baoill, Sutton-Spence, Turner & Woll)
- BSLCP project co-investigators, Deaf community fieldworkers & Deaf community Advisory Group members
- Sally Reynolds for data collection
- The 249 deaf signers across 8 regions of the UK who participated in the BSL Corpus Project data collection