Corpus linguistics and signed languages: no lemmata, no corpus

Trevor Johnston
Department of Linguistics, Macquarie University, Sydney, Australia

3rd Workshop on the Representation and Processing of Sign Languages, June 1
(LREC, Marrakech, Morocco, May 26-June 1)
Outline

1. Corpora and SL linguistics
2. Auslan corpus & Auslan lexical database
3. Notation, transcription, annotation & tagging
4. Lemmatisation & ID-glosses
5. Conventions for glossing different types of signs
6. Using a SL corpus
Corpora and linguistics

➢ The need for SL corpora
  ▪ Endangerment
  ▪ Lack of documentation
  ▪ Problems with introspection & intuitions
  ▪ Issues with native signers
  ▪ Demand for empirical linguistics
Corpora and linguistics

The need for SL corpora

- Endangerment
- Lack of documentation
- Problems with introspection & intuitions
- Issues with native signers
- Need for empirical SL linguistics
Issues with native signers

- most native signers (i.e., deaf of deaf) don’t also have native signing parents (i.e., deaf of deaf of deaf is relatively rare)
  - acquisition environments are rarely optimal
  - so, are they conducive to ‘well-founded’ intuitions, even for native signers?

- native signers in deaf communities are a small minority of all signers
  - usage environments are consistently populated with non-native interlocutors
  - so, is experience conducive to ‘well-founded’ intuitions on what is normal, acceptable or typical?
Need for empirical SL linguistics

- Need for evidence-based generalizations
- Need for testing of descriptions and hypotheses about SLs vocabulary and grammar
- Need for practical and easy access to primary data
  - no widely used and agreed upon ‘IPA’ for SLs
  - idiosyncratic glossing and transcription methods
  - no open archive of naturalistic recordings
  - until relatively recently the GLOSS or transcription was unable to be linked (time aligned) to the source data (recording or media)

Without this, meaningful peer review and/or testing of intuitions against usage data is virtually impossible
What is now meant by corpus?

- Corpus
  - a data set (writings, recordings) on which a particular linguistic analysis is based
    - increasingly ‘old-fashioned’ sense

- Linguistic corpus
  - collection of spoken and written material in a machine-readable form
  - assembled for the purposes of studying the type and frequency of structures/constructions in a language
  - sociolinguistic & sessional data (metadata)
  - uses digitisation, multi-media annotation software

- Signed language corpora?
  - Sociolinguistic variation, e.g., ASL, Auslan? Other?
  - Acquisition, e.g., ASL, HKSL? Other?
  - General, e.g., Auslan, NGT, ISL, BSL, DGS, LSF, and others?
Outline

1. Corpora and SL linguistics
2. Auslan corpus & Auslan lexical database
3. Notation, transcription, annotation & tagging
4. Lemmatisation & ID-glosses
5. Conventions for glossing different types of signs
6. Using a SL corpus
The Auslan corpus

- Source data
  - native deaf signers or near native early learners (before 6 years old)
  - 20 individuals x 5 cities x 3 hours (i.e., 100 participants)
  - language production tasks (interview, survey, conversation, personal narrative, elicited narratives and recounts, language elicitation tasks)

- Raw data
  - Original tapes: 300 digital video tape (300 hours)
  - Digitized backup: 300 iMovies (3 terabytes)

- Edited data
  - Individual .mov files: 1100 ‘task clips’ as annotation media files (100 participants x 11 tasks each) (1 terabyte)

- Annotation files
  - Individual .eaf files attached to each clip
  - only sub-set annotated initially

- Metadata files
  - IMDI metadata files for all clips
Auslan lexical database

- c. 7,000 sign entries (nb: signs, not English equivalents!)
  - Data-base constantly monitored and updated (from 1980s)
  - as internet site www.auslan.org.au since 2004
- sequenced according to formational features of signs
  - i.e. phonologically
- fields for
  - line drawing, video
  - identifying gloss (ID-gloss)
  - lexical and variant status
  - definitions, keywords
  - usage/register
  - semantic fields

Cf. more recent databases, e.g., DanishSL, AustrianSL, NGT, VGT, etc.
<table>
<thead>
<tr>
<th>Sign number</th>
<th>StemSN</th>
</tr>
</thead>
<tbody>
<tr>
<td>12390</td>
<td></td>
</tr>
</tbody>
</table>

**Sense**

**ID-Gloss**

**happen1a**

**HamNoSys**

[Diagram of sign]

**Morph-Gloss**

**MorphHamNoSys**

[Sign illustration]

**Sign movie**

[Stem illustration]

**Stem movie**
Annotator’s view for annotation ID-Gloss used in ELAN

<table>
<thead>
<tr>
<th>Happen</th>
<th>Happen1a</th>
<th>corr: Happon1a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotation ID-Gloss</td>
<td>happen</td>
<td></td>
</tr>
<tr>
<td>Database ID-Gloss</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sense</td>
<td>Sign illustration</td>
<td>Sign movie</td>
</tr>
<tr>
<td>Keywords</td>
<td>happening</td>
<td>happen</td>
</tr>
<tr>
<td></td>
<td>event</td>
<td>occur</td>
</tr>
<tr>
<td></td>
<td>occurrence</td>
<td>appearance</td>
</tr>
<tr>
<td></td>
<td>contingency</td>
<td>appear</td>
</tr>
<tr>
<td></td>
<td>opportunity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chance</td>
<td></td>
</tr>
</tbody>
</table>

**Usual one-handed sign =**
- RH-ID-Gloss WHO
- LH-ID-Gloss

**but made with two hands =**
- RH-ID-Gloss WHO-2h
- LH-ID-Gloss WHO-2h

**Usual two-handed sign =**
- RH-ID-Gloss HOUSE
- LH-ID-Gloss HOUSE

**Usual form of sign =**
- RH-ID-Gloss HOUSE
- LH-ID-Gloss HOUSE

Even when it uses a different handshape (in the database these are written as the ‘a’, ‘b’, ‘c’ variant forms). However, if you want to show that the sign uses a variant handshape, write the handshape symbol after the stem ID-Gloss. For example,
- RH-ID-Gloss HOUSE-1h
- LH-ID-Gloss HOUSE-H

Remember: the ID-Gloss is intended as a unique name for each sign, so keep additional information out of the ID-Gloss and include it on tiers for space, aspect, grammatical class, ‘meaning gloss’, translation, facial expression, and so on.
www.auslan.org.au
As Modifier

1. Used at the beginning or end of the first of two phrases to mean that the action in the first happens earlier than the action of the second phrase (which is signed after the first). English = before.

2. Used at the beginning or end of a sign phrase, or immediately next to a verb (action) sign, to mean that the action took place at a time earlier than the time of speaking or time of point of reference. English = before.

Staff

- Edit the gloss 'before1'
- Include in the Web dictionary?  
- Update
Before

Definition:

Used at the beginning or end of the first of two phrases to mean that the action in the first happens earlier than the action of the second phrase (which is signed after the first). English — before.
Outline

1. Corpora and SL linguistics
2. Auslan corpus & Auslan lexical database
3. Notation, transcription, annotation & tagging
4. Lemmatisation & ID-glosses
5. Conventions for glossing different types of signs
6. Using a SL corpus
Notation

- Writing down some linguistic output (e.g., word or sign) using a dedicated graphic symbol system
  - enables the reader of the notation to reconstruct the form of the word or sign, more or less, depending on the degree of detail in the system
    - i.e., broad or narrow, phonetic or phonemic
Notation using HamNoSys

LINGUISTICS

GREEN

\[ 0 \rightarrow \rightarrow \wedge + + \]

\[ \downarrow \Uparrow 5 \wedge 0 \cap (\leftarrow \rightarrow) + \]
Notation

- Writing down some linguistic output (e.g., word or sign) using a dedicated graphic symbol system
  - enables the reader of the notation to reconstruct the form of the word or sign, more or less, depending on the degree of detail in the system
    - i.e., broad or narrow, phonetic or phonemic
- Notation overlaps somewhat with transcription...
Transcription

- = writing down, using some kind of dedicated graphic symbol system, language which has been signed or spoken
  - usually text rather than isolated words/signs
  - enables the reader of the transcription to “reproduce” the original spoken or signed text
  - once again replicability depends on the comprehensiveness of the transcription system

- = script, when part of a bona fide writing system
  - writing systems usually ignore much of the act of articulation
    - rightly or wrongly certain aspects of language-as-articulated are not considered important (‘paralinguistic’)
  - in contrast, transcription consciously tries to capture much more of the act of articulation than any writing system does
SL transcription?

1. Capitalized glosses alone with translation:

PRO.1 FINISH \_1GIVE\_2 TWO-WEEKS-AGO

I gave it back to you two weeks ago.

2. Interlinear text with transcription, glossing, free translation, and literal translation

PRO.1 \\text{-} finish \ give \ week-\text{-}PL.2-fut.\text{-}TEMP.past

I gave it back to you two weeks ago

I gave it from me to you two weeks ago
Annotation

- linguistic ‘commentaries’ appended to identified units in a language
- add phonological, morphological, syntactic, semantic and discourse information about linguistic forms
- invaluable aid in helping linguists discern patterns in language at many different levels, with or without the aid of computers
Tagging

- no clear cut distinction between an annotation and a tag
  - both are linguistically relevant information appended 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
Annotation/tags in a text

Joanna stubbed out her cigarette with unnecessary fierceness.

- Joanna _NP stubbed _VBD out _RP her _PP$ cigarette _NN with _IN unnecessary _JJ fierceness _NN . _.

- examples of tags used…
  - _NP = singular proper noun
  - _VBD = regular past tense form of lexical verb
  - _RP = adverbial particle
  - _PP$ = possessive pronoun
  - _NN = singular common noun
The hare continued to laugh until suddenly the tortoise turned to look at him with distaste.
Tiers & tags

- RH ID gloss = unique identifying glosses
  - sign-type conventions
    - lexical, depicting, buoys, gestures, points, etc.
- RH-gram cls = grammatical class?
  - NP = plain noun
  - VP = plain verbs
  - VIDir = indicating directional verb
  - VILoc = indicating locatable verb
  - ADJ = adjective
- RH mod = spatially modified?
  - m = yes
  - n = no
  - cg = ‘congruent’
  - na = not applicable
The hare continued to laugh until suddenly the tortoise turned to look at him with distaste.
Annotation ‘parses’

Corpus =
digital movies plus related annotation files

Annotation file for digital movie number 1
becomes richer over time

First annotation parse
e.g., glosses

Second annotation parse
e.g., grammatical class

Third annotation parse
e.g., spatial modification

Fourth annotation parse
e.g., semantic roles

Annotation file 1d

Digital movie 1

Annotation file 1a

Annotation file 1b

Annotation file 1c

Annotation file 1a

Annotation file 1b

Annotation file 1c
Outline

1. Corpora and SL linguistics
2. Auslan corpus & Auslan lexical database
3. Notation, transcription, annotation & tagging
4. Lemmatisation & ID-glosses
5. Conventions for glossing different types of signs
6. Using a SL corpus
ID-glossing

➢ Aim
  ▪ create a text which is itself machine readable

➢ Method
  ▪ identifying (‘naming’) lexical signs uniquely
    ▶ use an ‘ID-gloss’
  ▪ consistent labelling of other types of signs
    ▶ gestures, buoys, depicting signs, points
  ▪ disconnecting ‘naming’ from
    ▶ ‘transcription’ (trying to represent the form of the sign)
    ▶ ‘translating’ (specifying meaning-in-context)
    ▶ ‘morphologizing’ (trying to represent the structure or modification of signs)
Lemmatisation

- Lemmatisation
  - ‘book’, ‘books’ are forms of the lemma BOOK
  - ‘walk’, ‘walks’, ‘walked’, ‘walking’ forms of lemma WALK

- Uniquely identifying signs using an ID-gloss is essentially lemmatisation
  - for SLs, the citation form is more or less the lemma

- Other tiers contain formational and grammatical information about the signs
  - grammatical class (noun, verb, adjective/modifier, etc.)
  - modification (e.g., space, direction, cycles, mouthing)

So no information is lost
Lemma / ID-gloss (example)

➢ Single basic sign, with or without modifications
  ▪ HOUSE (HOUSE-citation, HOUSE-big, HOUSE-left)
    ▶ unless a modified form is lexicalized! e.g.,
      HOUSE-big = MANSION ‘a luxurious house’
      ≠ just ‘a big house’
    ▶ modifications annotated on other tiers

➢ Single sign with different functions
  ▪ DRINK (n, “drink”, “beverage”, “drinking”) or (v, “drink”, “have a drink”)
    ▶ unless a modified form is lexicalised! e.g.,
      DRINK-circular = ALCOHOLIC ‘addicted to alcohol’
      ≠ ‘drink a lot of any kind of beverage’
Corpus lemmatisation & tagging

- corpus lemmatization (e.g., ‘waiting’ → WAIT) & tagging (e.g., n, v, adj.)
  - semi-automatic in languages with standardized orthography and well-described grammar (at least, core grammar) (upto >95% accuracy)
  - however, this is not an option for SL linguists/annotators so it must be done / assigned manually

- which lemma / ID-gloss to assign?
  - it must be consistent within and across texts (annotation files)
  - adhere to the assignment of ID-glosses in a lexical database
Lemmatisation

- Non-unique glosses (‘non-lemmas’) cannot be searched, sorted, or counted consistently within or across annotation files
  - ELAN can constrain searches according to values on more than one tier across multiple annotation files (i.e., the corpus as a whole or identified text-types within the corpus)
  - thus all information can be utilized despite the annotation gloss being ‘lemmatized’ (simplified) because the tags on other tiers constrain searches
Contents of the lexical database

Non-native lexicon
- e.g., fingerspelling, foreign SL borrowings

Native lexicon

Non-core lexicon
- non- or partially specified/lexicalised signs
  - e.g., depicting & pointing signs

Core lexicon
- lexicalised signs

1. dictionary
2. sketch grammar

Initial language description:
- fieldwork, introspection, elicitations, intuitions

Corpus

Subsequent language description with enriched dataset:
- attested, reviewable, quantifiable, attributable usage data
Contents of the lexical database

Non-native lexicon
- e.g., fingerspelling, foreign SL borrowings

Native lexicon

Non-core lexicon
- (non- or partially specified/lexicalised signs)
  - e.g., depicting & pointing signs

Core lexicon
- (lexicalised signs)

ID-glosses

1. dictionary
2. sketch grammar

Initial language description:
- fieldwork, introspection, elicitations, intuitions

Corpus

Subsequent language description with enriched dataset:
- attested, reviewable, quantifiable, attributable usage data
Outline

1. Corpora and SL linguistics
2. Auslan corpus & Auslan lexical database
3. Notation, transcription, annotation & tagging
4. Lemmatisation & ID-glosses
5. Conventions for glossing different types of signs
6. Using a SL corpus
Conventions

- Lexical vs. non-lexical signs*
  - fully specified vs. partially specified
  - frozen vs. productive
  - lexical vs. depicting (‘classifier’) signs
  - standard signs vs. HIS (highly iconic structures) incl. enactment and constructed action

- Signs vs. gestures
  - culturally shared vs. idiosyncratic gestures
  - enactment and constructed action

* Constructions vary from atomic-to-complex & substantive-to-schematic as part of a lexical-to-grammatical construction continuum

NOTE: ‘non-lexical’ ≠ ‘grammatical’ or ‘function’
Depicting (classifier) signs

- PM(handshape):description-of-meaning
  - PM = property marker
    - could use CL or D or anything consistently applied
    - includes handle and trace (possible discrimination in later annotation parses)
    - formationally only handshape currently coded (possible discrimination of orientation in later annotation parses)

- Example
  - PM(1):person-goes-away
  - PM(B):turtle-moves
Other conventions (cont.)

- **Points**
  - PT:
    - PT:PRO, PT:DEM, PT:LOC, PT:POSS
    - PT:PRO1, PT:PRO1sg, PT:PRO1pl
    - PT(B):PRO1sg

- **List buoys**
  - BUOY(handshape):sequence-of-total
    - BUOY(2):second-of-two, BUOY(3):third-of-three

- **List buoys + point**
  - RH tier  BUOY(3):three
  - LH tier  PT(BUOY):second-of-three   [PT(HOLD):second-of-three]

- **Gestures**
  - G:how-stupid-of-me      not      G:hit-forehead-with-palm
Outline

1. Corpora and SL linguistics
2. Auslan corpus & Auslan lexical database
3. Notation, transcription, annotation & tagging
4. Lemmatisation & ID-glosses
5. Conventions for glossing different types of signs
6. Using a SL corpus
Using the corpus & machine-readability

1. Annotate
   - enrich ‘transcription’ with linguistic tags

2. Extract
   - whole corpus / particular text types

3. Identify
   - frequencies, constructions

4. Test
   - intuitions & generalizations

5. Explain
   - linguistic environment and modality

6. Compare
   - other signers, other SLs, SLs & SpLs

7. Propose
   - new generalizations
- All instances
  - concordance view
  - jump to any example
- Automatic extraction of frequency lists
  - exported
  - sorted

- Semi-automatic tagging for frequency
  - find ID-gloss
  - tag on frequency tier
All instances

- concordance view
- understand environment
- jump to any example
- All instances
  - frequency view
  - compare variants
Search for sign with ID-gloss “LOOK” which is a directional indicating verb ("VIDir") which is modified for space ("m")

Repeat search for all signs, using regular expression ("wild card") character $
Repeat both searches for unmodified forms ("n")
Repeat both searches for congruent forms (“cg”)
10 high frequency indicating nouns and 10 high frequency indicating verbs account for over 50% of all tokens of spatially modified lexical signs.
- Point (PT:) before
  - V(erb) m (modified)

- Repeat with
  - Point (PT:) before
    - Verb, not modified
    - Verb, congruent
  - Point (PT:) after
    - m, n, cg
  - PT: before & after
    - m, n, cg
  - c. subtypes of verbs
    - Dir, Loc, Plain
    - High frequency
    - “Iconicity index”
Conclusion

- Demand corpus-based SL research
  - due to the unique sociolinguistic situation of SL-using communities, corpus-based research vitally important

- Prioritize annotation above ‘transcription’
  - preliminary lexical research necessary
  - integrate lexical information into glosses which identify signs uniquely using gloss-based annotations
  - recognize that corpus-data feeds back into lexical data
  - incorporate up-date and revision facility into both corpus annotation files and lexical database

- Remember linguistic corpora should be machine-readable
  - without lemmata / ID-glosses, a SL corpus is not machine-readable in any relevant or practical sense
Acknowledgments

- Hans Rausing Endangered Languages Documentation Program, School of Oriental and African Studies, University of London
  - Grant #MDP0088 awarded to Trevor Johnston

- Australian Research Council
  - grant #LP0346973 awarded to Adam Schembri and Trevor Johnston: *Sociolinguistic Variation in Auslan: Theoretical and applied dimensions*
  - grant #DP0665254 awarded to Louise de Beuzeville and Trevor Johnston: *The linguistic use of space in Auslan: semantic roles and grammatical relations in three dimensions.*
Contact information

A/Prof Trevor Johnston
Department of Linguistics
Macquarie University
Sydney, Australia
(email) trevor.johnston@mq.edu.au