MACQUARIE UNIVERSITY
FACULTY OF HUMAN SCIENCES

The reluctant oracle: annotating a sign language corpus for answers to questions we can't ask any other way

Trevor Johnston
Macquarie University
Sydney, Australia

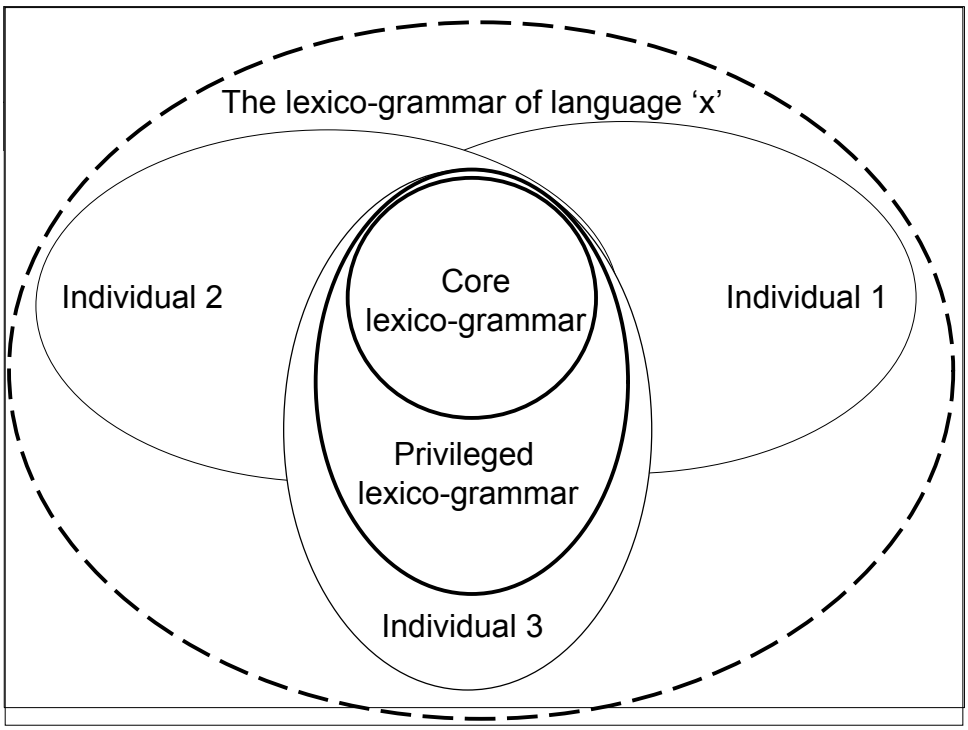
[Sign language corpora: linguistic issues, DCAL, London, 24-25 July 2009]

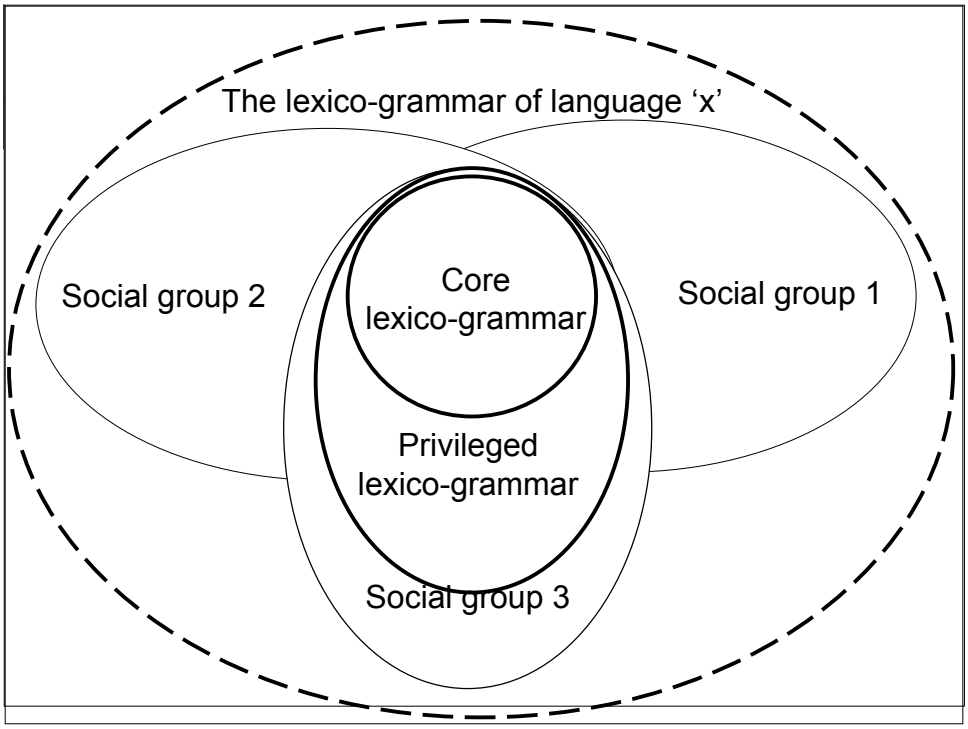
Outline

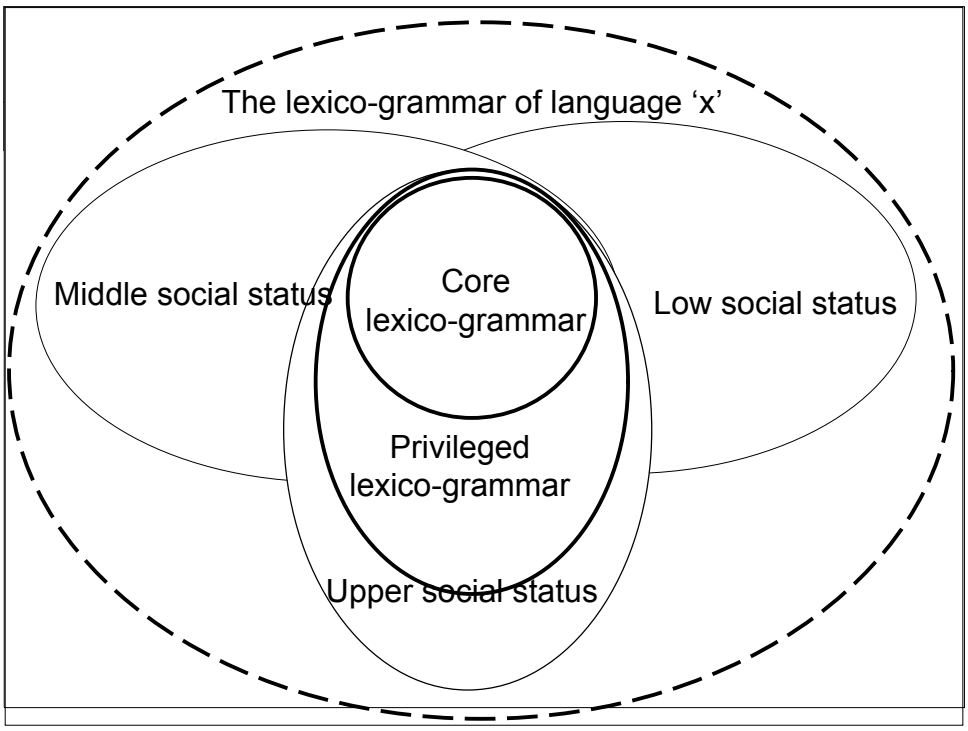
- Introduction
- The case for SL corpus linguistics
- Corpus-based SL research
- Conclusion

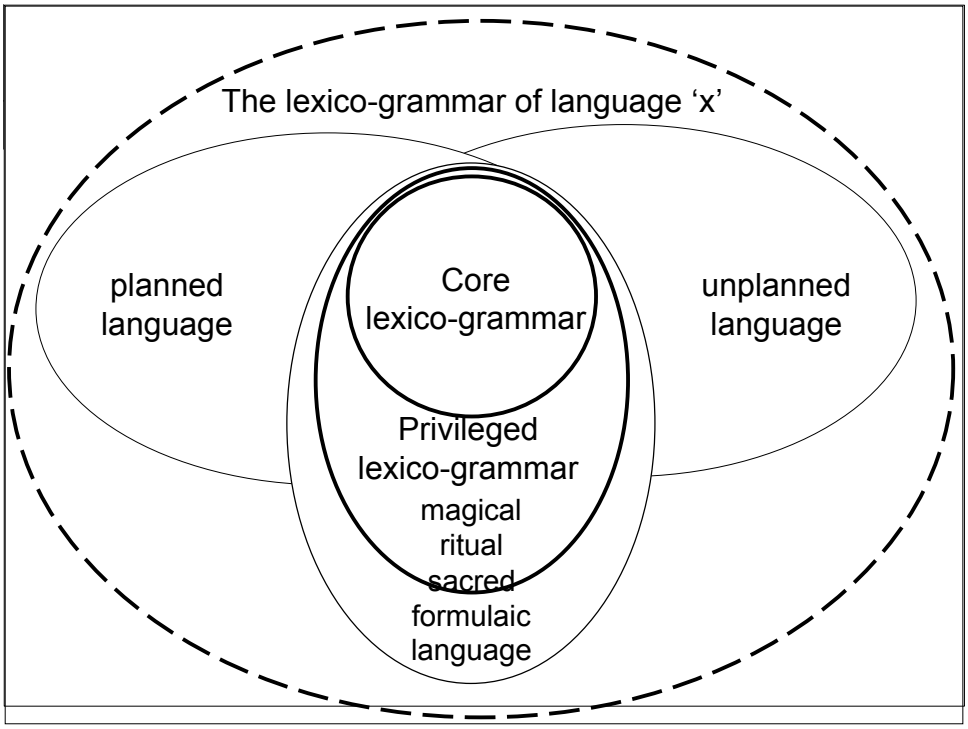
Introduction

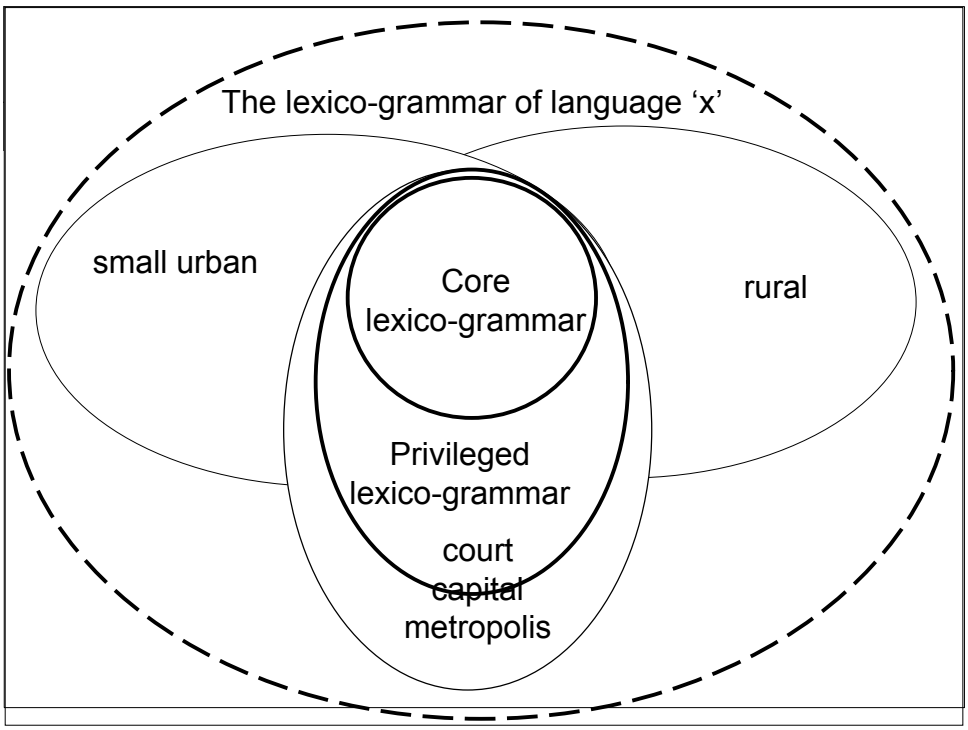
Linguistic atheism

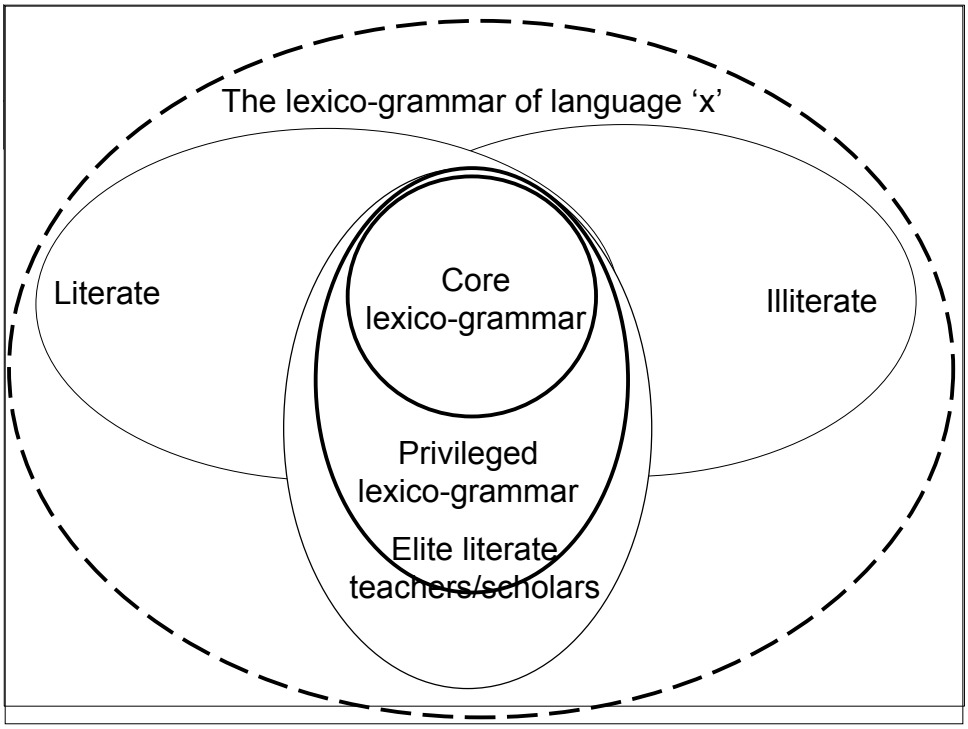


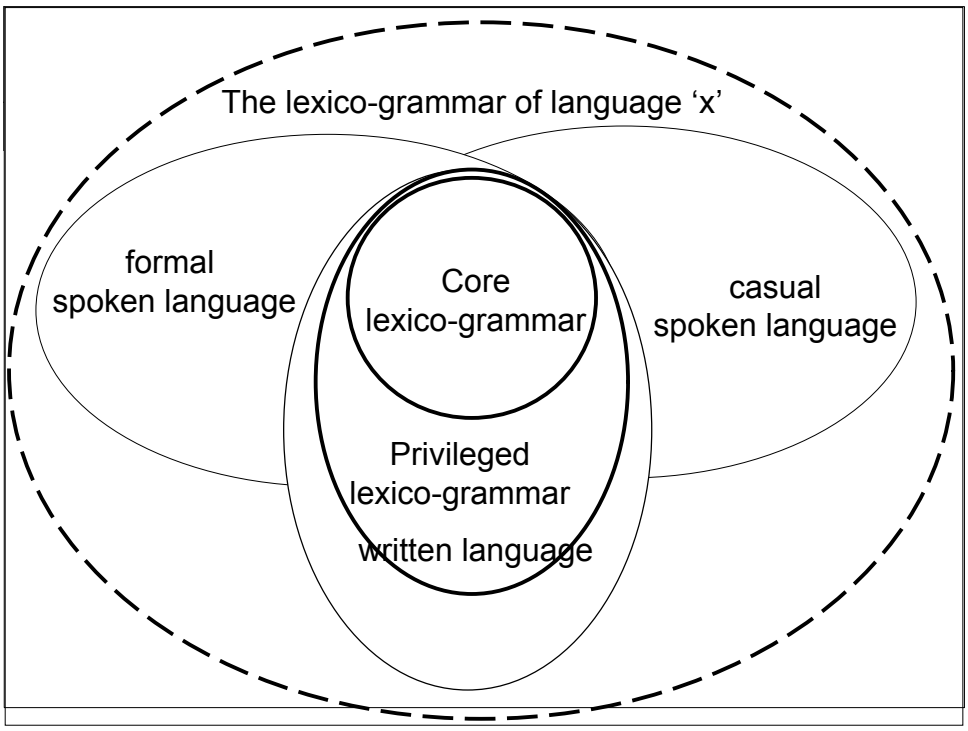

The lexico-grammar of language ‘x’
Social group 2
Core
lexico-grammar
Social group 1
Privileged
lexico-grammar
Social group 3

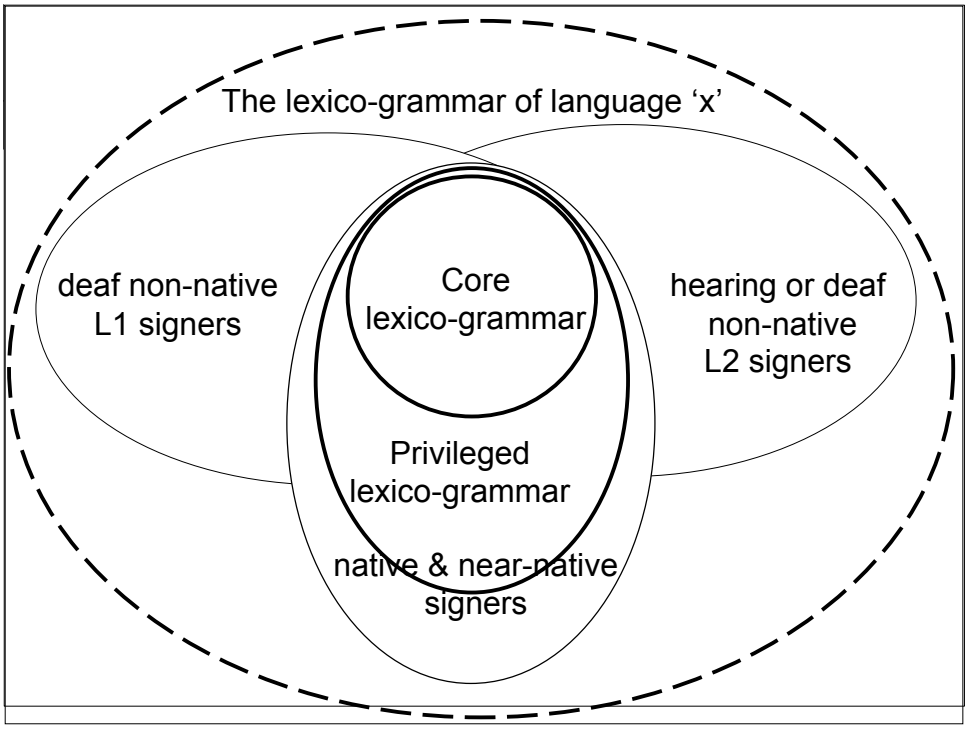

The lexico-grammar of language ‘x’
Middle social status
Core
lexico-grammar
Low social status
Privileged
lexico-grammar
Upper social status


The lexico-grammar of language ‘x’
planned language
Core lexico-grammar
unplanned language
Privileged lexico-grammar
magical
ritual
sacred
formulaic language


The lexico-grammar of language ‘x’
small urban
Core lexico-grammar
rural
Privileged lexico-grammar
court
capital
metropolis


The lexico-grammar of language ‘x’
Literate
Illiterate
Core
lexico-grammar
Privileged
lexico-grammar
Elite literate
teachers/scholars


The lexico-grammar of language ‘x’
formal
spoken language
casual
spoken language
Core
lexico-grammar
Privileged
lexico-grammar
written language



The case for SL corpus linguistics

- What do we want to do?
- Why do we want to do it?
- How do we do it?

What do we want to do?

- empirically ground SL description
- validate previous research
- generate new observations
- document linguistic community
- create teaching/learning resources

Why do we want to do it?

- no easily or commonly used written form
- lack of language documentation
  - cf. preservation
- language endangerment
  - cf. maintenance, revitalization
- limits to intuitions and introspection
- unique usage/acquisition environments
- difficult for learners to gain exposure

How do we do it?

- create language archives
  - i.e., documentary linguistics
- adopt a corpus-based approach
- value-add to language archives using
  - multi-media annotation software
  - annotation, not necessarily transcription
  - systematic linguistic tagging
  - controlled gloss-based annotations (ID-glosses)
- open access for researchers and community
  - learners and teachers

Annotation, not necessarily transcription

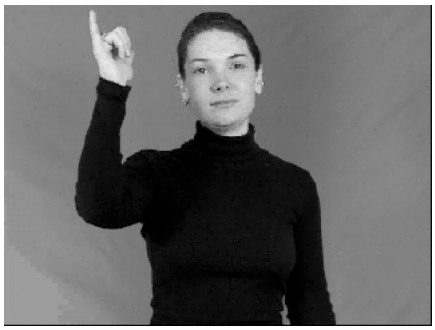
- Notation = Symbol system
- Transcription = Writing system
- Annotation = Appended notes
- Tagging = Appended codes

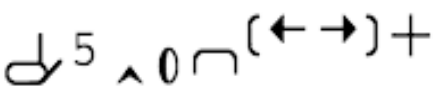
Notation & transcription

- Notation: representation of language units (e.g., phonemes, morphemes, words or signs) using a dedicated graphic symbol system
  - enables the reader reconstruct the uttered unit, depending on the degree of detail in the system

Notation using HamNoSys

GREEN

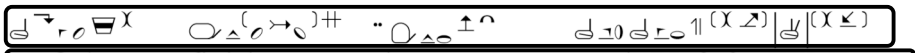




Notation & transcription

- Notation: representation of language units (e.g., phonemes, morphemes, words or signs) using a dedicated graphic symbol system
  - enables the reader reconstruct the uttered unit, depending on the degree of detail in the system
- Transcription overlaps with notation, but
  - usually refers to representation of extended utterances (texts) rather than just isolated words/signs
  - consciously tries to capture much more of the act of articulation than any writing system ever does

SL transcription



PRO.1 finish give week-PL.2-fut.TEMP.past

I gave it back to you two weeks ago

I gave it from me to you two weeks ago

The above is an example of interlinear text with

1. transcription
2. glossing
3. free translation
4. literal translation

Is notation/transcription necessary?

- YES, notation is required
  - for detailed phonological analysis
  - for sorting lexical entries by form (pronunciation)
- NO, transcription is not necessary
  - a (written) text is not essential prerequisite for multi-media corpus linguistics
    - sign form can be seen in time-aligned video
  - one simply needs to identify relevant linguistic units (words/signs) and one can then undertake morphosyntactic, phrase, clause, utterance or discourse level analysis of constructions or structures
    - i.e., the sign or extended utterance does not have to be represented (transcribed) before it can be analysed

Is this transcription or annotation?

PRO1sg FINISH 1-GIVE-2 TWO-WEEKS-AGO

I gave it (back) to you two weeks ago

It is neither:

≠ transcription

because apart from the attempt to specify the beginning and end points of GIVE (as “1” and “2”) nothing indicates the form of the utterance

≠ annotation

because there are no utterance units (no recording or no transcription) to which the annotations are attached or appended

Is this transcription?

NO, IT IS ANNOTATION

Annotation

- linguistic ‘commentaries’ appended to identified units in a language
- add phonological, lexical, morphological, syntactic, semantic, pragmatic 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
  - e.g., part of speech tagging

Tags: horizontal v. 'vertical'

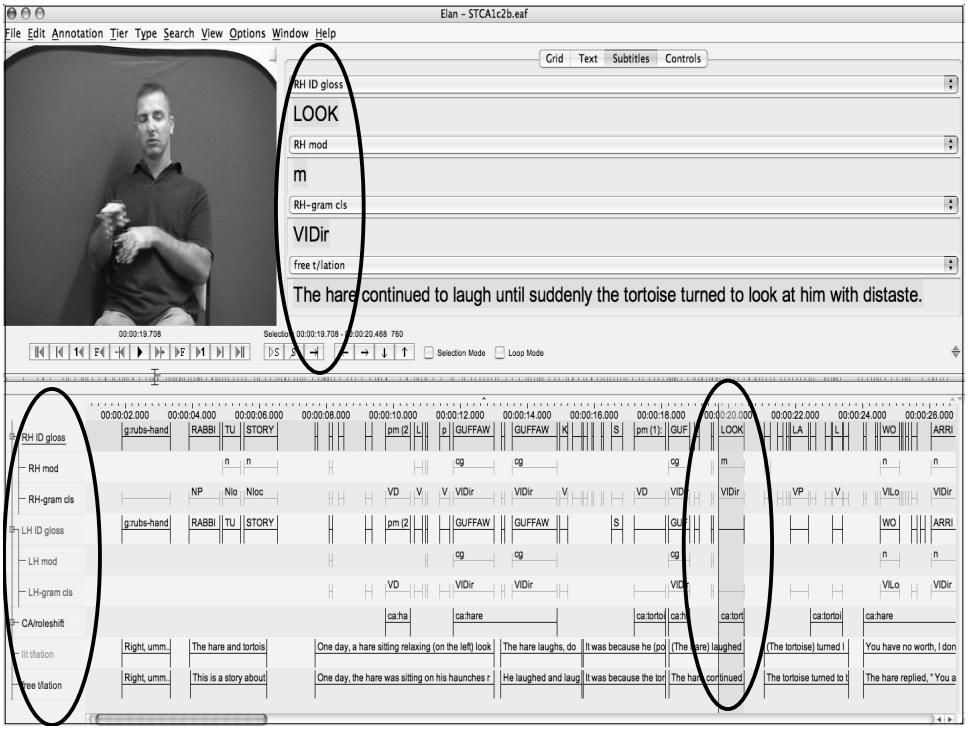
Horizontal, e.g.,

Joanna_NP stubbed_VBD out_RP her_PP$ cigarette_NN with_IN unnecessary_JJ fierceness_NN ._.

- tags, e.g. _NP for singular proper noun appended to the written text

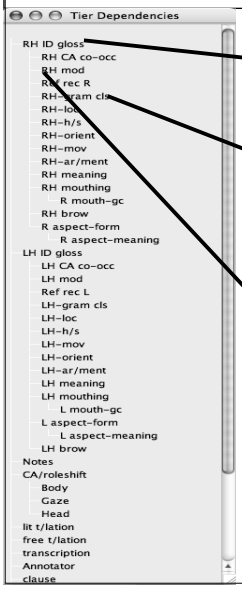
Vertical, e.g.,

ELAN annotations/tags are tiered or 'vertical' rather than sequential.



Trevor Johnston	annotating a sign language corpus
Introduction	
The case for SL corpus linguistics	What do we want to do?
Corpus-based SL research	Why do we want to do it?
Conclusion	How do we do it?

Tiers & annotation/tags



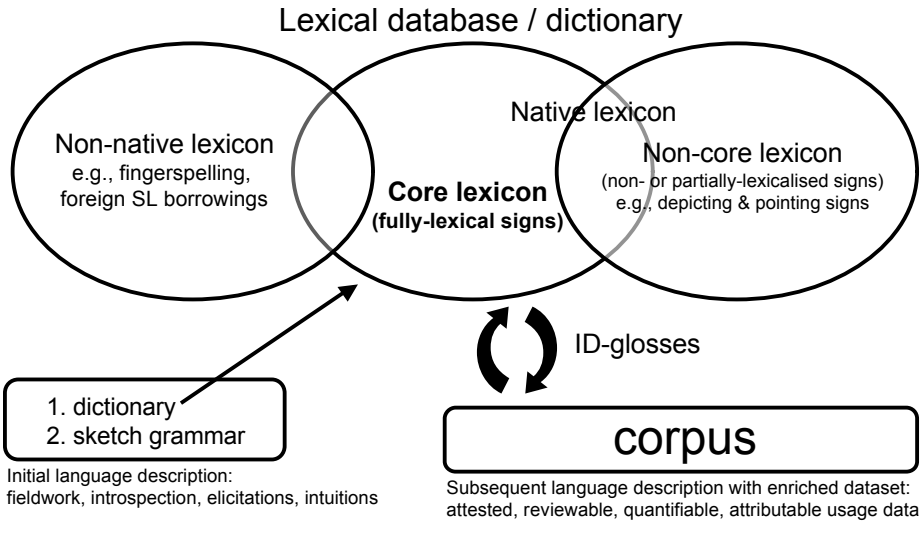
- RH ID gloss = unique identifying gloss
- RH-gram cls = grammatical class
  - NP = plain noun
  - VP = plain verbs
  - VIDir = indicating directional verb
  - VILoc = indicating locatable verb
  - ADJ = adjective
- RH mod = spatial modification
  - m = modified
  - n = not modified
  - na = not applicable

ID-glossing as lemmatization

- Lemmatization
  - 'book', 'books' are forms of the lemma BOOK
  - 'walk', 'walks', 'walked', 'walking' forms of lemma WALK
- ID-glossing ("lexical annotation") is essentially lemmatization
  - for SLs, the citation form is analogous to the lemma
  - note: explicit lexical annotation conventions are needed for use with partly- or non-lexicalized signs (e.g., points, depicting signs, etc.)
- Other tiers contain formational and grammatical information about the signs
  - grammatical class
  - grammatical/semantic/thematic roles
  - modification
  - phonetic/phonological transcriptions (or simply tags)

So no information is lost

ID-glossing (lemmatisation) using a lexical database



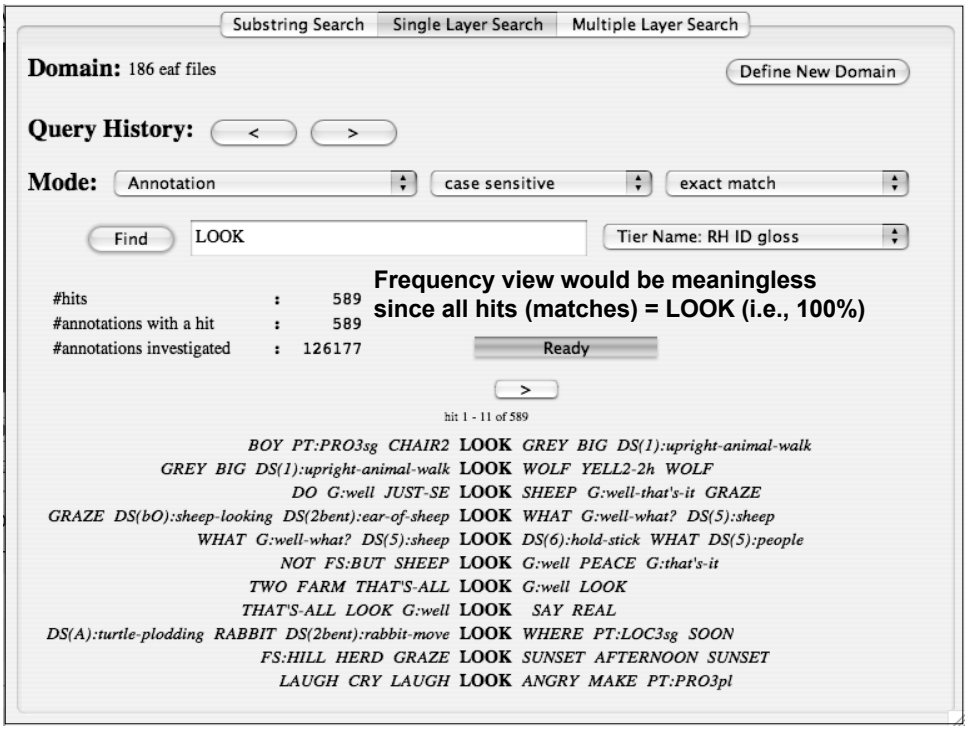
Corpus-based SL research

Some example searches based on annotations

- Single sign searches
  - Types/tokens
  - Frequency statistics
- Multiple sign searches
  - Concordance patterns and/or constructional schemas
  - Contextual constraints

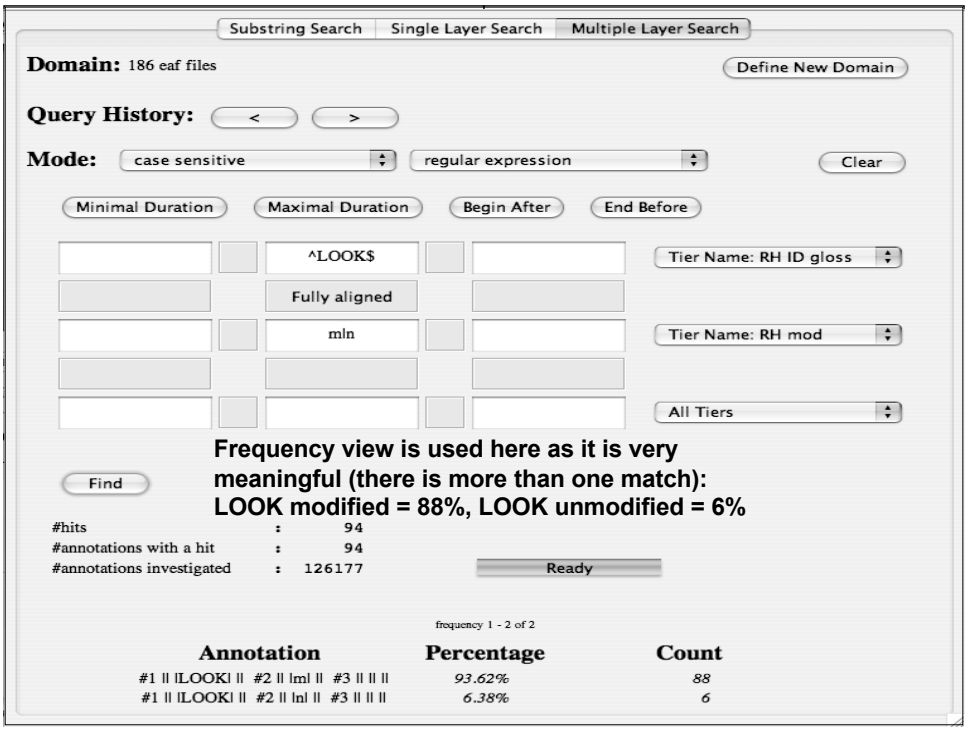
Single sign searches: types/tokens

- A search for any ID-gloss is a search based on a type
  - the hits are the tokens which may be viewed in context (concordance)



Single sign searches: types/tokens

- A search for any ID-gloss is a search based on a type
  - the hits are the tokens which may be viewed in context (concordance)
- Searches may be constrained by features of the token tagged on other tiers
  - e.g., 1. RH ID-gloss = “x”
    2. RH mod = n or m (m|n)



Frequency view is used here as it is very meaningful (there is more than one match): LOOK modified = 88%, LOOK unmodified = 6%

Single signs searches: frequency

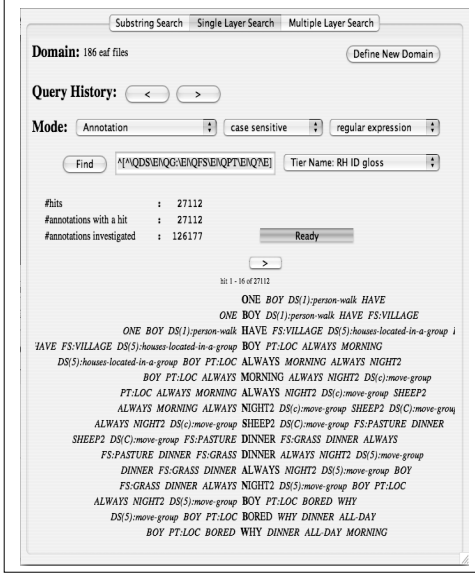
- Hits from a search which finds more than one type or sorts different tokens can be viewed (meaningfully) as a frequency list

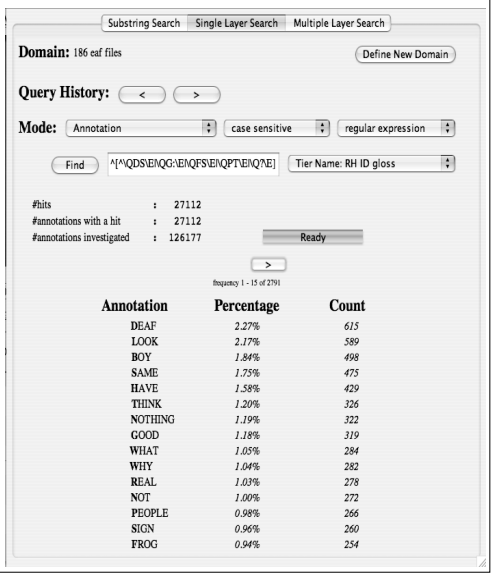
Auslan corpus frequency (50 text subset of 202)

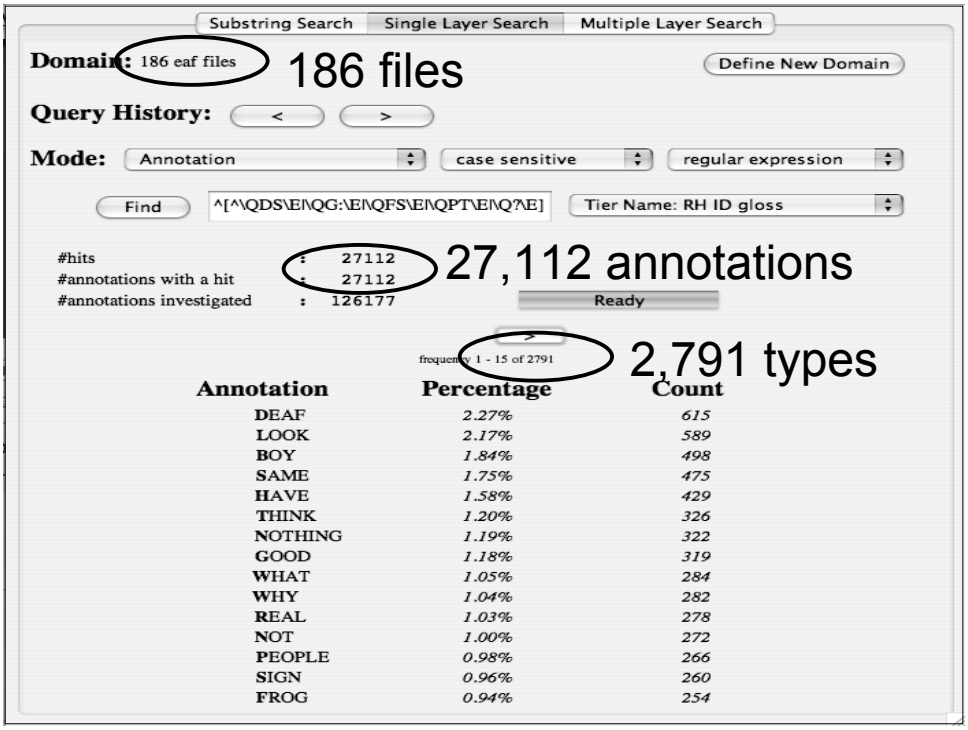
rank	gloss	numberBOTH
1	PT:PRO1	406
2	PT:PRO3	233
3	PT:DEM/LOC	221
4	G:well	218
5	LOOK	121
6	TURTLE	116
7	GOOD	115
8	BOY	107
9	WOLF	100
10	RABBIT	90
11	PEOPLE	86
12	SAY	80
13	REAL	70
14	HAVE	67
15	PT:PRO2	67
16	ONE	64
17	SHEEP	61
18	AGAIN	58
19	SAME	56
20	WHAT	56
21	LAUGH	53
22	DAY	51
23	PT:POSS1	51
24	NOTHING	49
25	THINK	48
26	WHY	48
27	ALL	45
28	CALL	44
29	COME	44
30	NOT	44
31	SLEEP	44
32	WANT	43
33	ARRIVE	42
34	WITH	40
35	SLOW	37
36	KNOW	35
37	SEE	35
38	ALWAYS	33
39	WILL	33
40	GO	32
41	SPRINT	32
42	FS:VILLAGE	31
43	CONTINUE	29
44	NIGHT	28
45	DS(B):turtle-moving	27
46	MORNING	27
47	STORY	25
48	BAD	24
49	CAN-NOT	23
50	GRAZE	23
51	PAST	23
52	STOP	23

- 5,000 type hit limit!
- 50 texts (c. 10,000 tokens)
- 202 texts (c. 43,000 tokens)
- Signs ranked 1, 2, 3, 4, 15, 23, 30, 39, 51 are grammatical (function) signs
- summed they are comparable to most frequent as % of all tokens in SpL corpora where grammatical or function words occupy the top frequencies

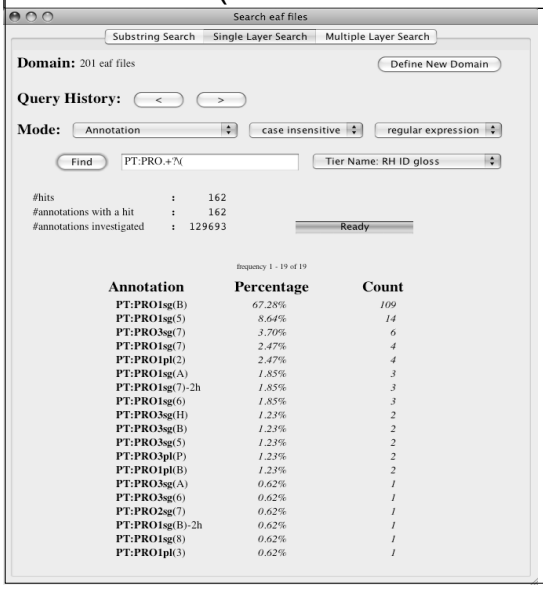
^[^\QDS\E|\QG:\E|\QFS\E|\QPT\E|\Q?\E]



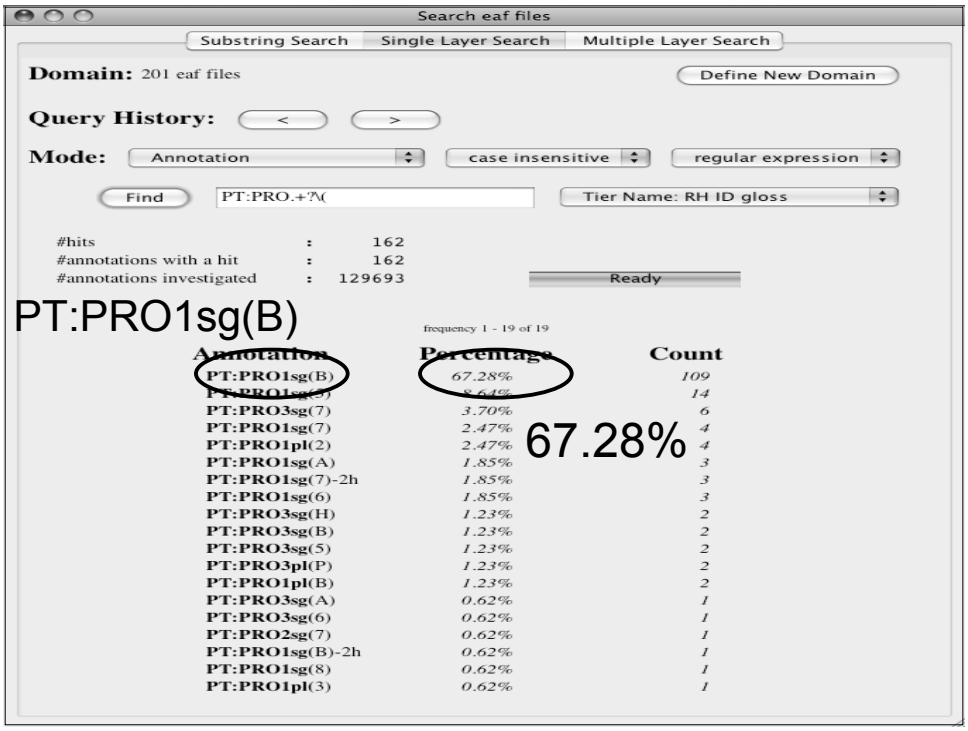




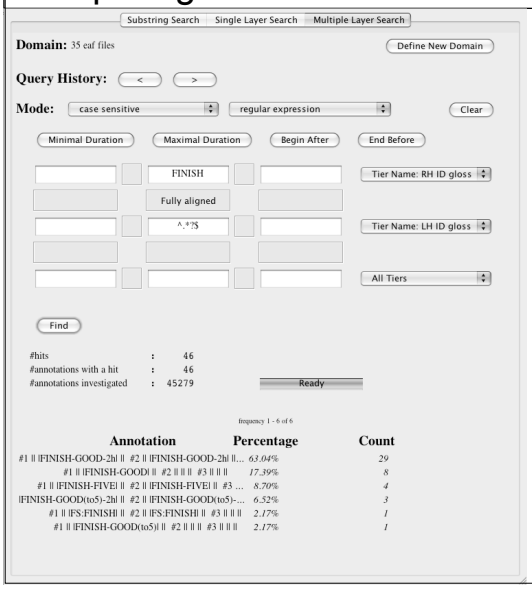
PT:PRO.+?\(



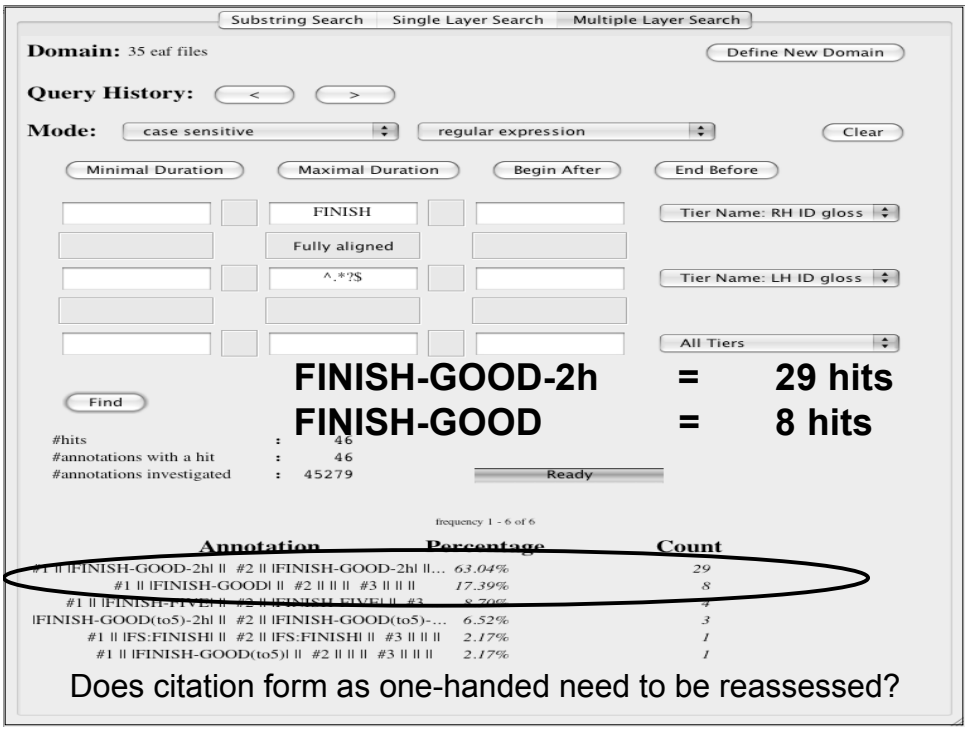
- Finds all pronouns coded with a variant handshape
  - i.e., ~1 handshape.
  - Of course, variants must be coded FIRST!
- Can also be narrowed further:
  - ^PT:PRO1.*?\(
  - ^PT:PRO1sg.*?\(



Comparing subordinate hand annotation



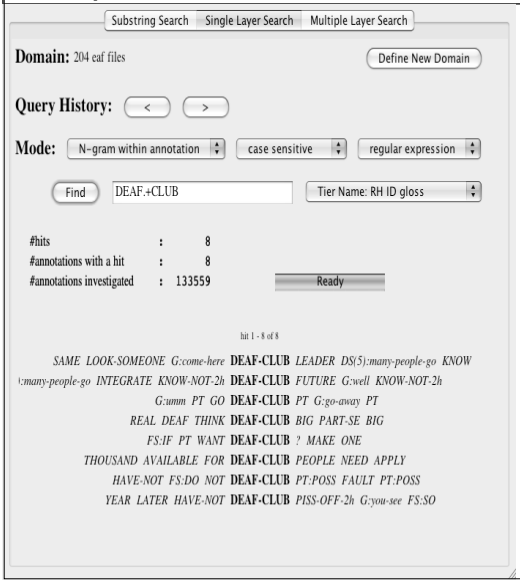
- Reg. exp. ^.*?$ will find an empty or full field
  - it finds a beginning to a field and an end to a field with or without anything in between
- Useful for identifying alternative 1 and 2 handed forms of signs (cf. weak drop, weak prop)



Multiple sign searches: concordance patterns

- Searches may also be constrained for signs occurring before or after a specific sign

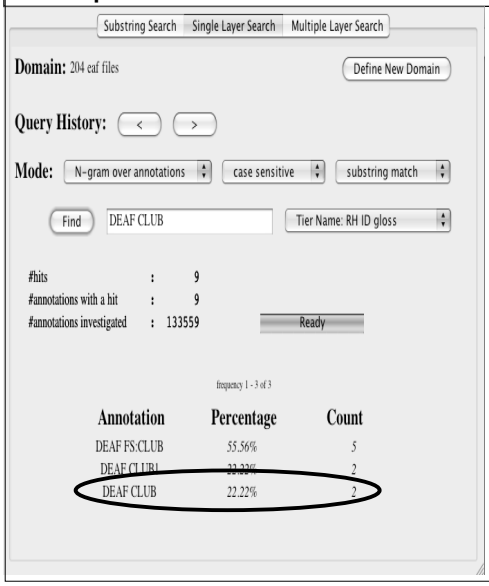
Compound or collocation? 1



- Search for a sequence of DEAF and CLUB
  - search in single layer
  - looking WITHIN annotations
    - "N-gram within annotation"
  - separated by "any kind of matter"
    - regular expression = .+
  - X-Y is one convention for writing a compound
    - the other is to use a unique gloss (such as TOMATO for RED-BALL)
  - one can view hits individually to confirm compound status
    - e.g., is there really phonological reduced in one of the compounded units?

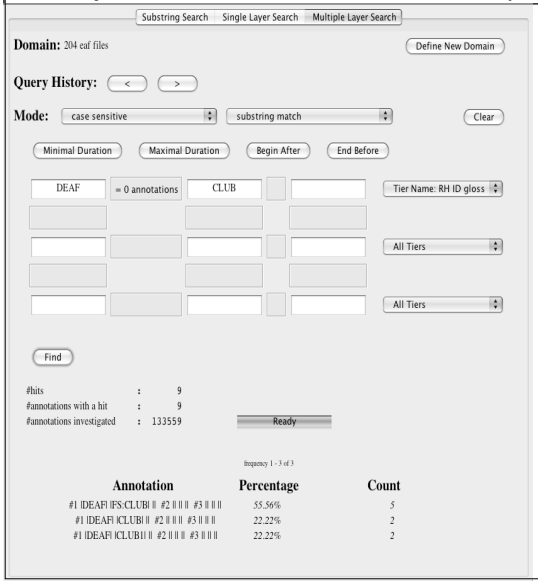
Trevor Johnston annotating a sign language corpus
Introduction
The case for SL corpus linguistics Single sign searches: types/tokens
Corpus-based SL research Single sign searches: frequency statistics
Conclusion Multiple sign searches: concordance patterns
Multiple sign searches: contextually constrained

Compound or collocation? 1



- Search for a sequence of DEAF and CLUB
  - search in single layer
  - looking OVER annotations
    - "N-gram over annotation"
  - i.e. one separate annotation after another
  - a sequence of DEAF CLUB is a potential candidate for a compound
    - there are 2 that fit the criteria
  - one can view hits individually to investigate
    - e.g., if there is phonological reduction in one of the elements in these two annotations (DEAF CLUB) then it may need to be combined as one annotation (DEAF-CLUB)

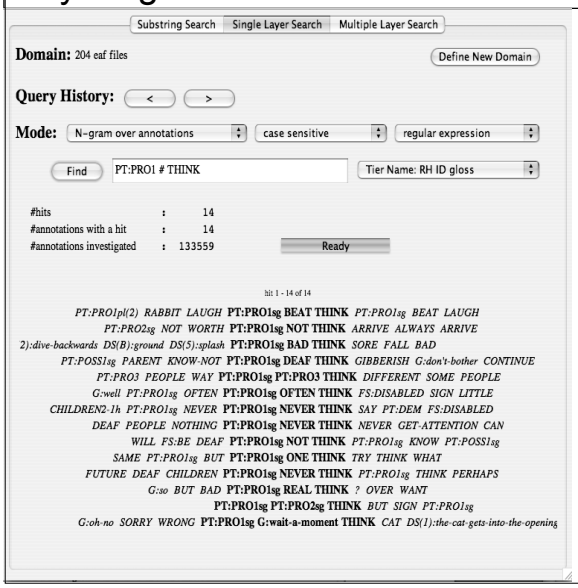
Compound or collocation? 2 (alternative search)



- Search finds potential candidates for compound status
  - e.g., a sequence of DEAF and CLUB over two contiguous annotations.

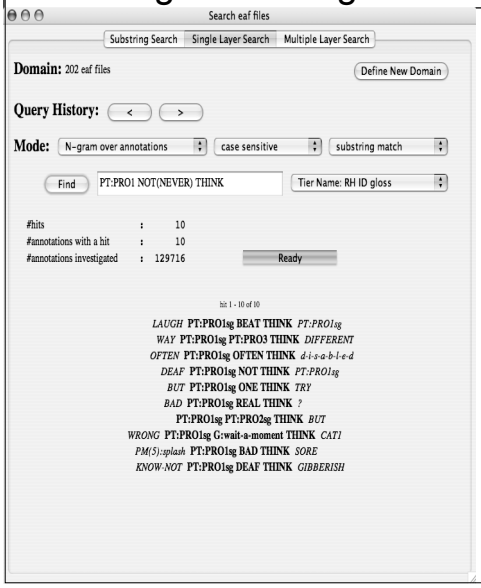
Trevor Johnston annotating a sign language corpus
Introduction
The case for SL corpus linguistics Single sign searches: types/tokens
Corpus-based SL research Single sign searches: frequency statistics
Conclusion Multiple sign searches: concordance patterns
Multiple sign searches: contextually constrained

Anything between two know annotations



- Finds all strings between two specified annotations
  - uses the wildcard symbol (#)
- Can do the same thing in multiple layer search grid
  - often more than one way in ELAN

Excluding something between two known annotations



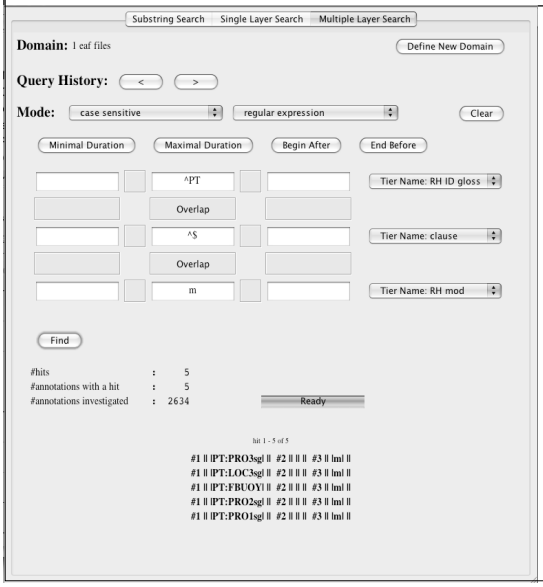
- Finds a string with anything between two stated annotations (substrings) except THE stated annotation
  - instead of # (as in previous search) it uses
  - NOT(the unwanted gloss)
- Thus NOT(NEVER) will find all
  - PRO1 *any gloss* THINK

  but not
  - PRO1 *NEVER* THINK

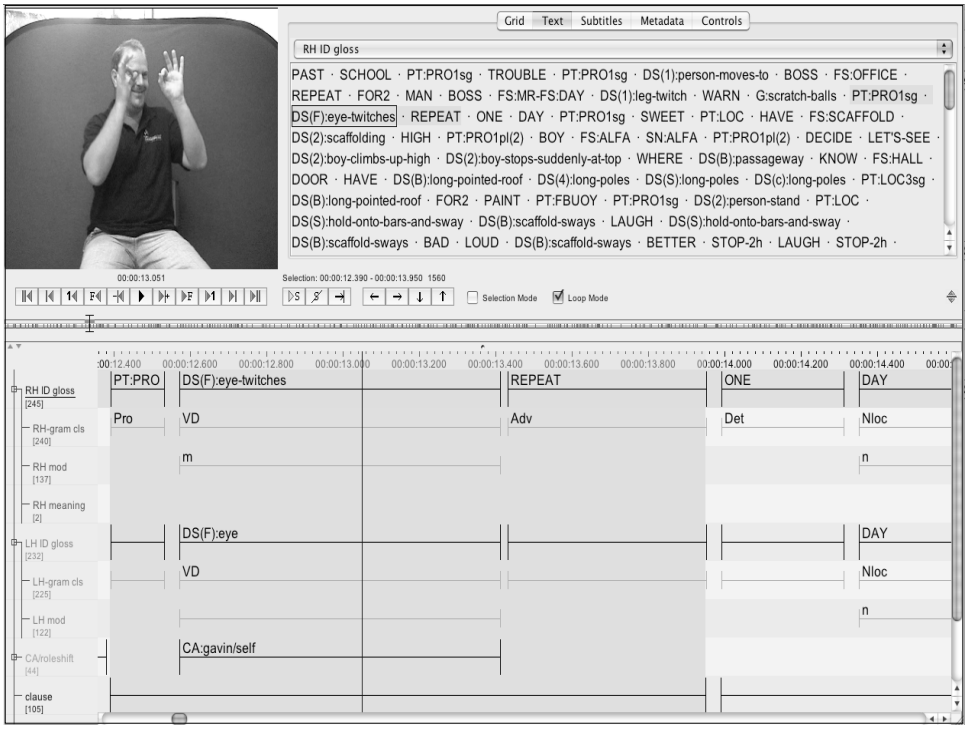
Multiple sign searches: contextual constraints

- Searches may also be constrained for signs occurring before or after a specific sign
- These sequential constraints can be combined with simultaneous constraints

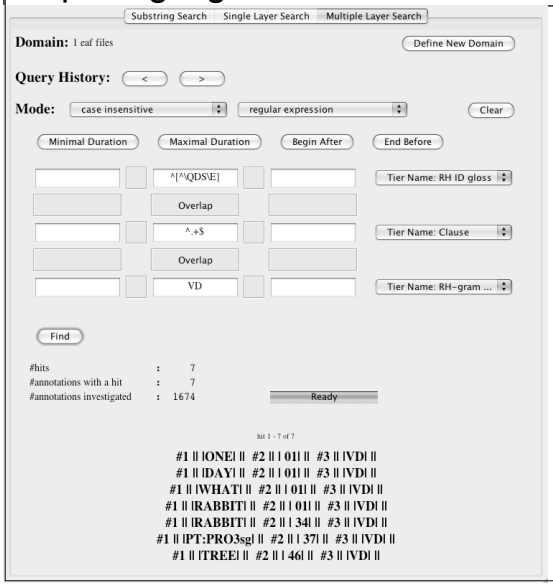
PT in same clause as modified sign



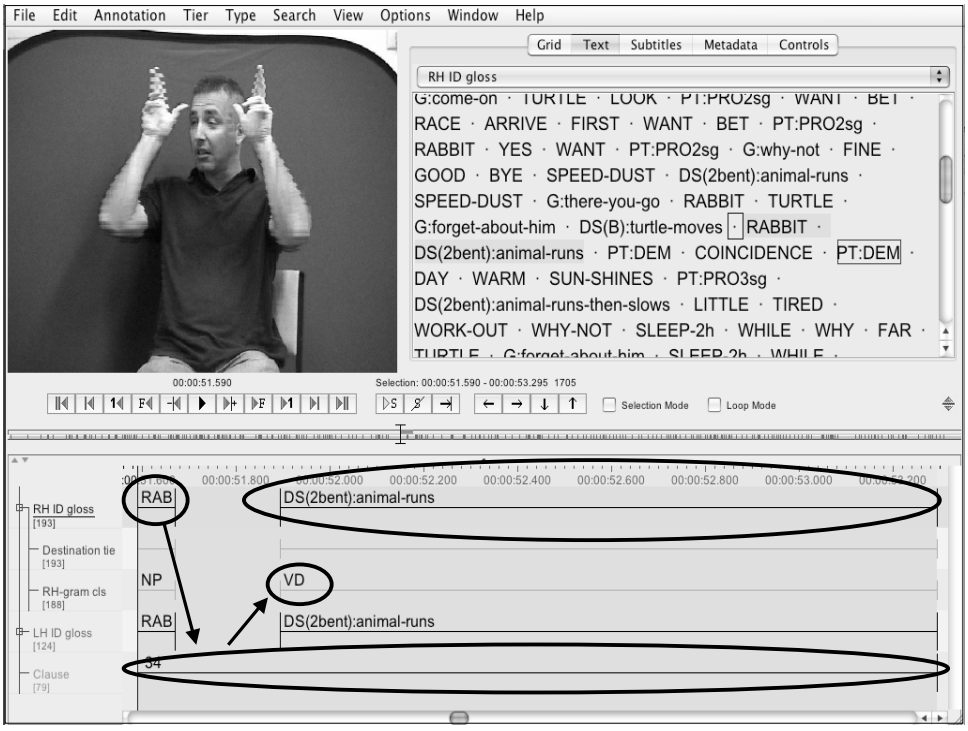
- Finds a pointing sign (^PT) which is part of a clause (^$) which contains a modified sign (m)
  - see next slide for example ELAN hit



Depicting sign in clause with other signs



- Finds a RH ID gloss that DOES NOT begin with “DS” (i.e., any gloss but a depicting sign gloss)
  - Reg. exp. = ^[^\QDS\E]
- which overlaps a clause tier annotation which is empty or has something in it (e.g. # or α or β)
  - Reg. exp. = ^.+$
- which ALSO overlaps a grammatical class tier label which is VD (i.e., which IS a depicting sign, after all).
- Finds all clauses that have a DS and at least one other sign. With numbered clauses, one can export the hits to Excel and sort the hits temporally.
  - Very useful for finding DS-containing clauses that have other signs as well, so one can inspect them. Can make manual coding and analysis much quicker and easier.



Clause arguments tier

A	a single overt argument of a verb
A1	a first overt argument of a verb (when there are more than one)
A2	a second overt argument of a verb
A3	a third overt argument of a verb
A4	a fourth overt argument of a verb
V	a verb
V1	a first verb in a serial verb construction
V2	a second verb in a serial verb construction
V3	a third verb in a serial verb construction
V4	a fourth verb in a serial verb construction, and so on.
nonA	an element of a clause which cannot be construed as an argument. It contributes temporal, location, purposive/reason, verbal auxiliary etc. information to the clause, but is not a 'participant' (argument) or 'process' (verb), as such.

Trevor Johnston annotating a sign language corpus
Introduction
The case for SL corpus linguistics Single sign searches: types/tokens
Corpus-based SL research Single sign searches: frequency statistics
Conclusion Multiple sign searches: concordance patterns
Multiple sign searches: contextually constrained

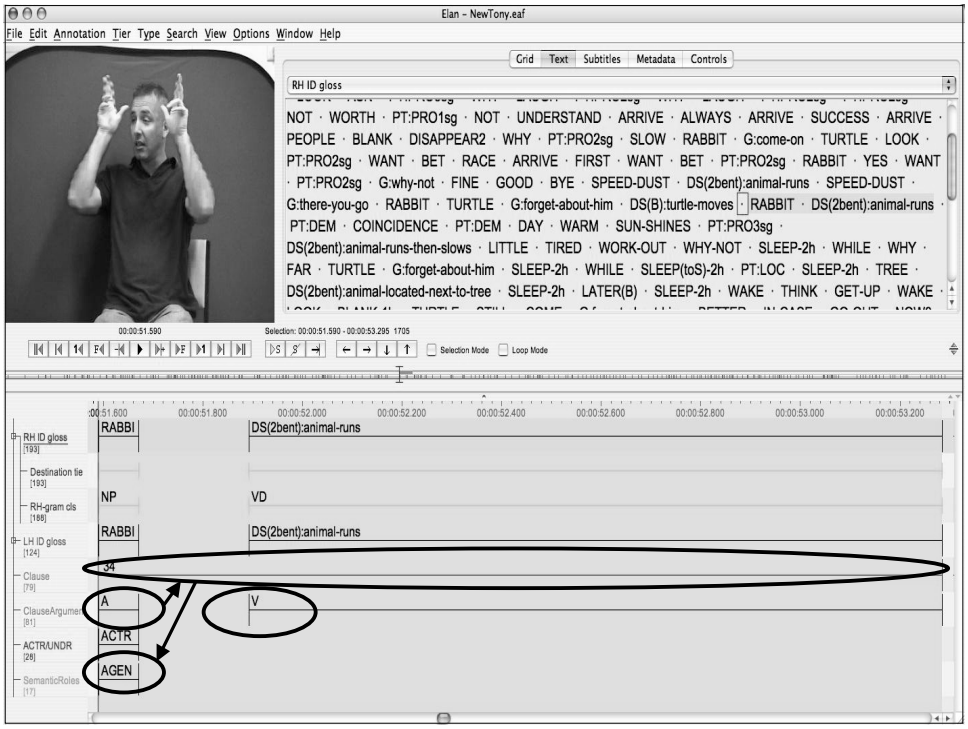
Semantic-macro roles

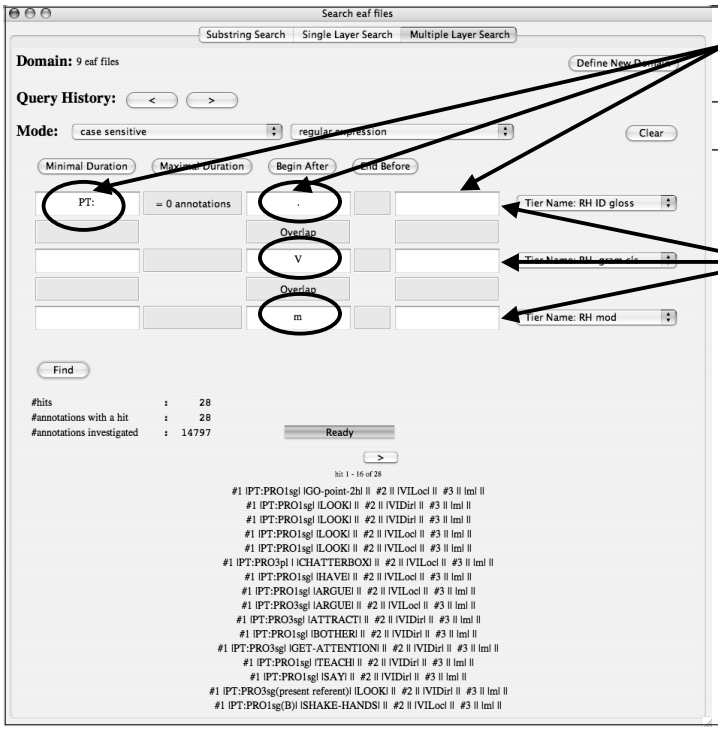
ACTR	an Actor-like argument of a verb ('Subject'*)
UNDR	an Undergoer, i.e., a non-Actor-like argument of a verb ('Object')
UNDR1	a first Undergoer when there is more than one ('Indirect Object')
UNDR2	a second Undergoer ('other Object')
UNDR3	a third Undergoer ('yet another Object').
CARRIER	argument in verbless clause of which the other argument is the attribute
ATTRIB	argument in verbless clause which names an attribute of the other argument

* Note: 'Subject' and 'Object' terminology is meant in only the most general possible way. Essentially, at this level of analysis the terminology is misleading. It does not mean the grammatical relations of subject and object.

Basic semantic roles

AGENT	agent
BEN	benefactive, recipient
EXP	experiencer
GOAL	goal
INST	instrument
LOC	locative
PATIENT	patient
SOURCE	source

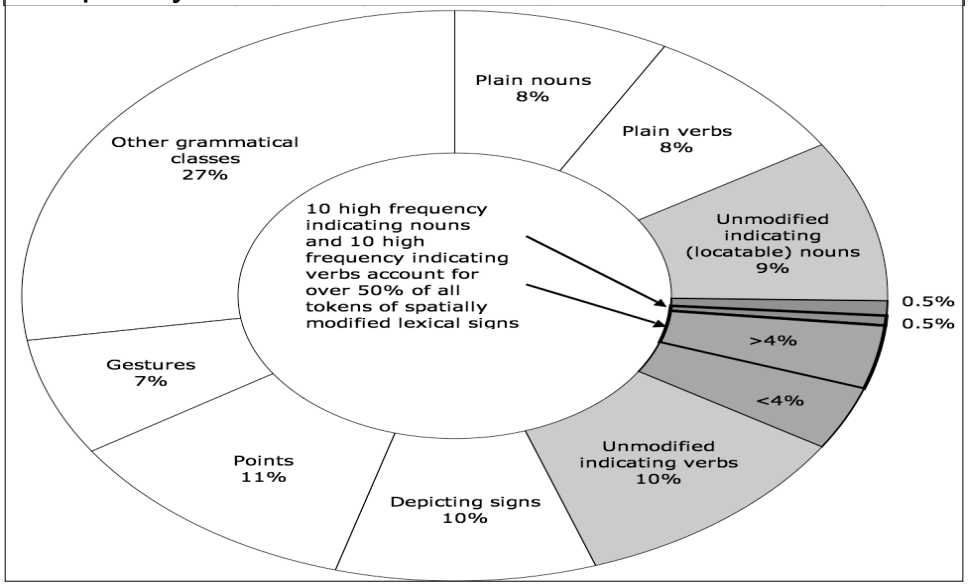




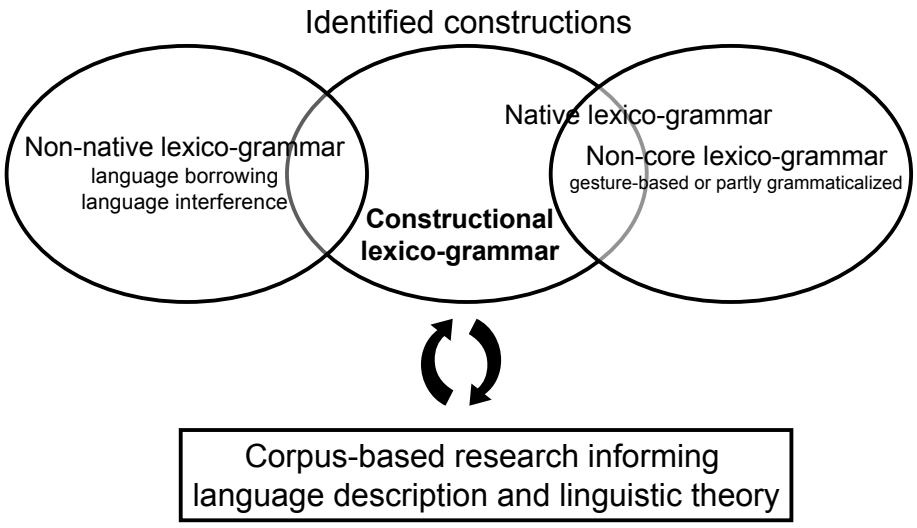
Three columns

Three rows

Frequency and constrained searches combined



Constructions in the lexico-grammatical continuum



Trevor Johnston
Introduction
The case for SL corpus linguistics
Corpus-based SL research
Conclusion
annotating a sign language corpus
Single sign searches: types/tokens
Single sign searches: frequency statistics
Multiple sign searches: concordance patterns
Multiple sign searches: contextually constrained

Discovering, not searching for patterns (constructions)

- Pattern testing (existing capabilities)
  - Enriching the corpus
  - Testing hypotheses
  - Research observations
- Pattern recognition (desirable capabilities)
  - e.g., CREAGEST team (e.g., Antonio Balvet)
  - need for plug in or software improvement to detect patterns/constructions constrained both 'vertically' and 'horizontally' by 2, 3 or more values
    - ➢ linguistic analysis, new hypotheses etc.

Conclusion

- Cross-linguistic & typological SL research
- Towards a SL corpus linguistics

Cross-linguistic & typological research

- Consistency
  - needed at two levels
    - language-internal & cross-linguistic consistency
  - documented practice, guidelines or standards?
    - standards desirable, but well-documented internally consistent local practice must not be neglected in the meantime
- Comparability
  - descriptive adequacy & typological observations
    - cross-linguistic comparisons are only as strong (valid) as the weakest language-specific description
  - validation > comparison > re-evaluation
    - testing and validation of language-specific observations should precede cross-linguistic generalization
    - cross-linguistic comparison nonetheless vital to open new perspectives enabling possible re-evaluation of local descriptions and leading more robust typological generalizations

Towards a SL corpus linguistics

- Insist upon corpus-based SL research
  - due to the unique sociolinguistic situation of SL-using communities, corpus-based research is vitally important
- Create true corpora
  - a linguistic corpus is not simply a data-set
  - it is a collection of language which has accurate metadata and is representative, machine readable, accessible and able to be further enriched
- Prioritize annotation above transcription
  - preliminary lexical research necessary to do this effectively
  - use ID-glosses and restricted set of conventions for partly-lexical and non-lexical signs
  - use other tiers to annotate for linguistically salient information
- Use in-built search routines and SQL query language to extract patterns or test generalizations

MACQUARIE UNIVERSITY
FACULTY OF HUMAN SCIENCES

Acknowledgments

- Hans Rausing Endangered Languages Documentation Program, School of Oriental and African Studies, University of London
  - Project grant #MDP0088 awarded to Trevor Johnston (2004-07)
- Australian Research Council
  - Project grant #LP0346973 awarded to Adam Schembri and Trevor Johnston: *Sociolinguistic Variation in Auslan: Theoretical and applied dimensions* (2003-05)
  - Project 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* (2006-09)
- Macquarie University
  - Grant #MQSNG9200729255 awarded to Trevor Johnston (2008-09): *The emergence of grammar in human languages: evidence from video-recordings of variation in vocabulary and grammar in a signed language (pilot study).*

MACQUARIE UNIVERSITY
FACULTY OF HUMAN SCIENCES

Contact information

A/Prof Trevor Johnston
Centre for Language Sciences
Department of Linguistics
Macquarie University
Balaclava Road, North Ryde
Sydney, NSW, Australia 2109

trevor.johnston@mq.edu.au