

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.

Sign entry

Sign number

12390

Sense

ID-Gloss

happen1 a

HamNoSys

Q > 0 d ^ o ~ [1 ^ > ~]

Morph-Gloss

MorphHamNoSys

StemSN



Sign illustration

Stem illustration



Sign movie

Stem movie

Annotator's view for annotation ID-Gloss used in ELAN

HAPPEN

Annotation ID-Gloss

happen1a

Database ID-Gloss

Sense



Sign illustration



Sign movie

Stem illustration

When the same word is used as an ID-Gloss for more than one sign then the additional glosses have a number added immediately after (no hyphen!), like this: GLOSS2, GLOSS3 etc.

Corrections & Comments

Keywords

happening	happen
event	occur
occurrence	appearance
contingency	appear
opportunity	
chance	

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

but made with one hand =

RH-ID-Gloss HOUSE-1h

LH-ID-Gloss

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-H

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

Signs used to express the word "before"



[Look up another word](#)

Sign 2 of 5

[◀ prev sign](#)

[next sign ▶](#)



Click on image to replay

Sign Definition

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.

Keywords associated with this sign

- before

Sign Distribution: All States



Provide Feedback

[Provide feedback about this sign](#)

[View feedback for this sign](#)

Staff

- [Edit the gloss 'before1'](#)
- Include in the Web dictionary?
- [Update](#)

Change gloss

History

View on site →

✕ Delete

Save and add another

Save and continue editing

Save

Idgloss:

Annotation idgloss:

Morphemic Analysis:

Sense Number:

Sign Number:

StemSN:

Publication Status (Show)

Lexis & Register: Borrowing (Show)

Lexis & Register: States (Show)

Lexis & Register: Religion (Show)

Lexis & Register: Iconicity (Show)

Lexis & Register: Other (Show)

Phonology (Show)

Morpho-Syntax (Show)

Semantic Domains (Show)

Other (Show)

Definitions

Text	Role	Count
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.	Modifier	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

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



.. 2_0 [→ > ^ #]

GREEN



d_5 ^ 0 n [← →] +

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 FINISH₁GIVE₂ TWO-WEEKS-AGO

I gave it back to you two weeks ago.

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

ᵈˠˣᵣᵝᵞᶜᵀ ᶠᵂᶜᵝᵞᶜᵀᶢᶣᶤ ᶥᶦᶪᶫᶬᶭᶮᶯᶰᶱᶲᶳᶴᶵᶶᶷᶸᶹᶺᶻᶼᶽᶾᶿ
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

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

Annotation using ELAN

The screenshot displays the ELAN software interface for video annotation. At the top, the window title is "Elan - STCA1c2b.eaf". The menu bar includes "File", "Edit", "Annotation", "Tier", "Type", "Search", "View", "Options", "Window", and "Help". Below the menu bar, there are tabs for "Grid", "Text", "Subtitles", and "Controls".

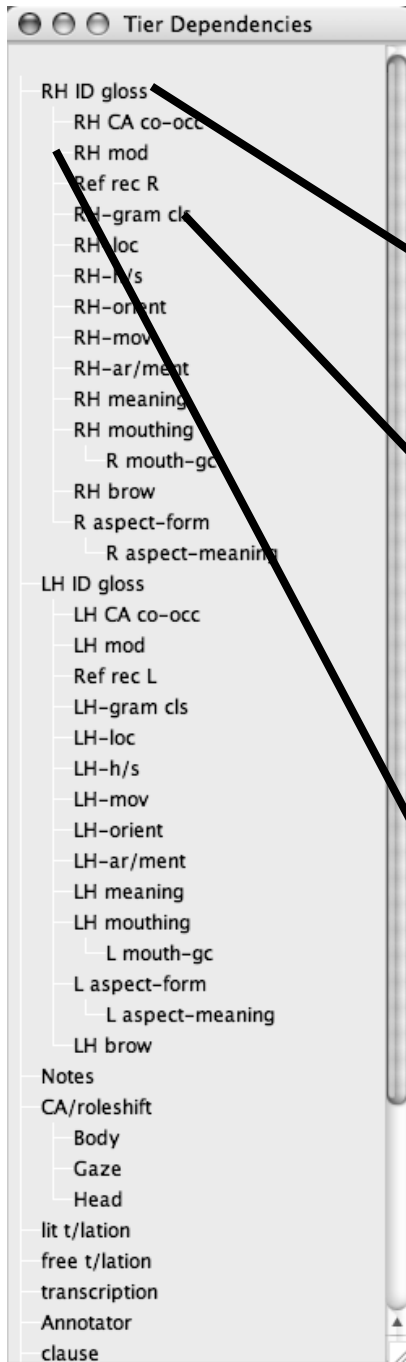
On the left side, a video window shows a man sitting and talking. The video player controls at the bottom of this window show a current time of 00:00:19.708 and a selection range from 00:00:19.708 to 00:00:20.468. Playback controls include buttons for previous, play, next, and stop, along with a seek bar.

On the right side, there is a panel for defining annotation tiers. The "RH ID gloss" tier is currently set to "LOOK". Other tiers include "RH mod" (set to "m"), "RH-gram cls" (set to "VIDir"), and "free t/lation" (set to "free t/lation"). A text box at the bottom of this panel contains the sentence: "The hare continued to laugh until suddenly the tortoise turned to look at him with distaste."

The main area of the interface is a timeline with a scale from 00:00:02.000 to 00:00:26.000. The timeline is divided into several tracks:

- RH ID gloss:** Contains labels like "g:ruhs-hand", "RABBI TU STORY", "pm (2)", "GUFFAW", "GUFFAW", "S", "pm (1)", "GUF", "LOOK", "LA", "L", "WO", "ARRI".
- RH mod:** Contains labels like "n", "cg", "cg", "cg", "m", "n", "n".
- RH-gram cls:** Contains labels like "NP", "Nlo", "Nloc", "VD", "V", "V", "VIDir", "VIDir", "V", "VD", "VIDir", "VIDir", "VP", "V", "VLo", "VIDir".
- LH ID gloss:** Contains labels like "g:ruhs-hand", "RABBI TU STORY", "pm (2)", "GUFFAW", "GUFFAW", "S", "GUF", "WO", "ARRI".
- LH mod:** Contains labels like "cg", "cg", "cg", "n", "n".
- LH-gram cls:** Contains labels like "VD", "VIDir", "VIDir", "VIDir", "VLo", "VIDir".
- CA/roleshift:** Contains labels like "ca:ha", "ca:hare", "ca:tortoi", "ca:h", "ca:tort", "ca:tortoi", "ca:hare".
- lit t/lation:** Contains text segments like "Right, umm...", "The hare and tortois", "One day, a hare sitting relaxing (on the left) look", "The hare laughs, do", "It was because he (po", "(The hare) laughed", "(The tortoise) turned l", "You have no worth, I don".
- free t/lation:** Contains text segments like "Right, umm...", "This is a story about", "One day, the hare was sitting on his haunches r", "He laughed and laugh", "It was because the tor", "The hare continued", "The tortoise turned to t", "The hare replied, "You a".

A black oval highlights the left-hand side of the timeline tracks, specifically the "RH ID gloss", "RH mod", "RH-gram cls", "LH ID gloss", "LH mod", "LH-gram cls", "CA/roleshift", "lit t/lation", and "free t/lation" tracks.



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

Annotation using ELAN

The screenshot displays the ELAN software interface for the file 'STCA1c2b.eaf'. The interface includes a video window on the left showing a man sitting. The main area contains a timeline with various annotation tracks. Two ovals highlight specific parts: one around the 'LOOK' annotation in the 'RH ID gloss' track, and another around the 'LOOK' annotation in the 'RH ID gloss' track and the corresponding 'ca:tort' annotation in the 'CA/roleshift' track.

File Edit Annotation Tier Type Search View Options Window Help

Grid Text Subtitles Controls

RH ID gloss
LOOK
RH mod
m
RH-gram cls
VIDir
free t/lation

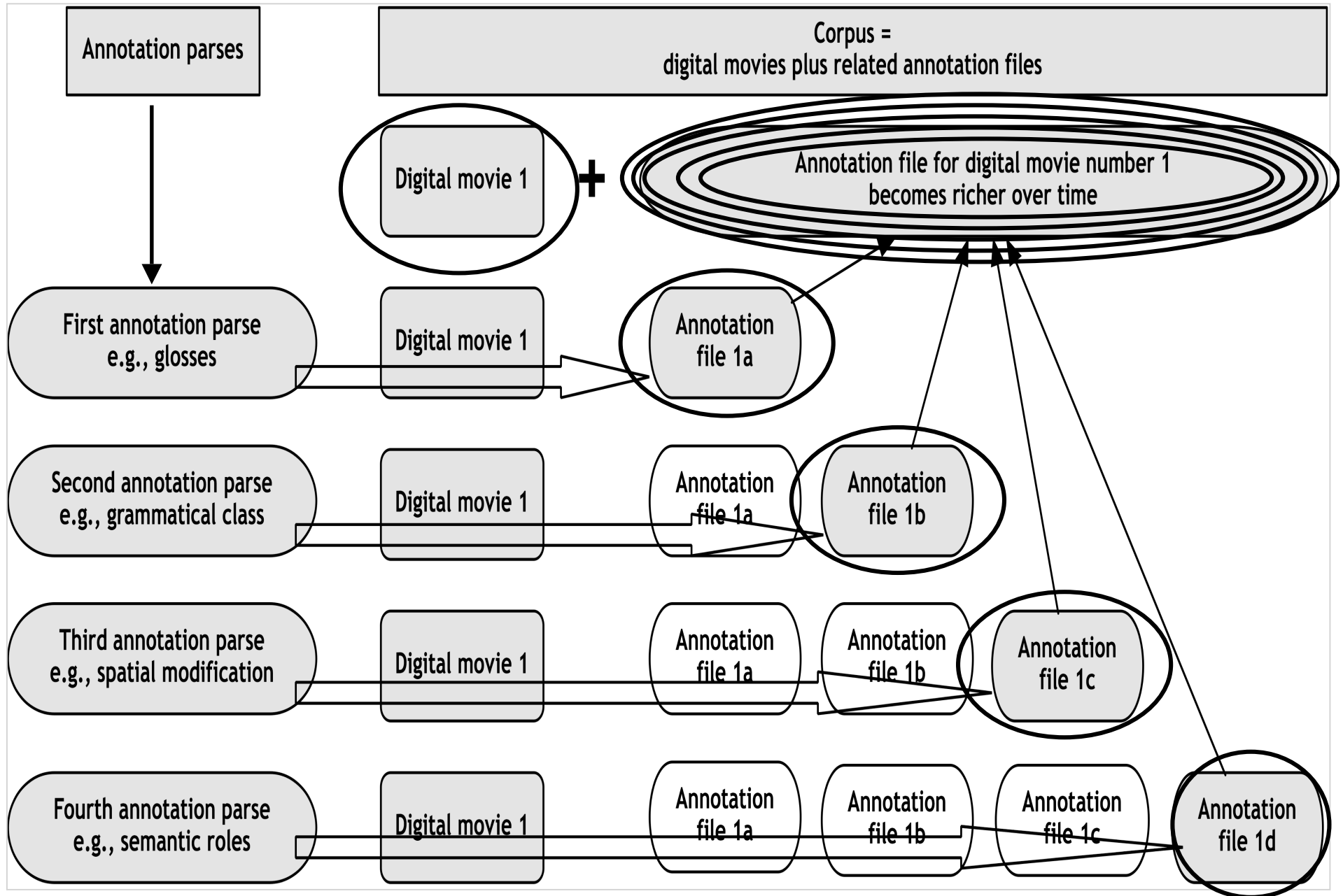
The hare continued to laugh until suddenly the tortoise turned to look at him with distaste.

00:00:19.708 Selection: 00:00:19.708 - 00:00:20.468 760

Selection Mode Loop Mode

Time	RH ID gloss	RH mod	RH-gram cls	LH ID gloss	LH mod	LH-gram cls	CA/roleshift	lit t/lation	free t/lation
00:00:02.000	g:ruhs-hand			g:ruhs-hand				Right, umm...	Right, umm...
00:00:04.000	RABBI TU STORY	n n	NP Nlo Nloc	RABBI TU STORY				The hare and tortois	This is a story about
00:00:08.000								One day, a hare sitting relaxing (on the left) look	One day, the hare was sitting on his haunches r
00:00:10.000	pm (2) L		VD V V	pm (2) L		VD	ca:ha		
00:00:12.000	p GUFFAW	cg	VIDir	GUFFAW	cg	VIDir	ca:hare	The hare laughs, do	He laughed and laugh
00:00:14.000	GUFFAW K	cg	VIDir V	GUFFAW	cg	VIDir		It was because he (po	It was because the tor
00:00:16.000	S		VD	S		VD	ca:tortoi	(The hare) laughed	The hare continued
00:00:18.000	pm (1): GUF		VIDir	GUF		VIDir	ca:tort	The tortoise turned l	The tortoise turned to
00:00:20.000	LOOK	m	VIDir					You have no worth, I don	
00:00:22.000	LA		VP				ca:tortoi		
00:00:24.000	L		V						
00:00:26.000	WO	n	VLo	WO	n	VIDir	ca:hare		The hare replied, "You a
00:00:28.000	ARRI	n	VIDir	ARRI					

Annotation 'parses'



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'

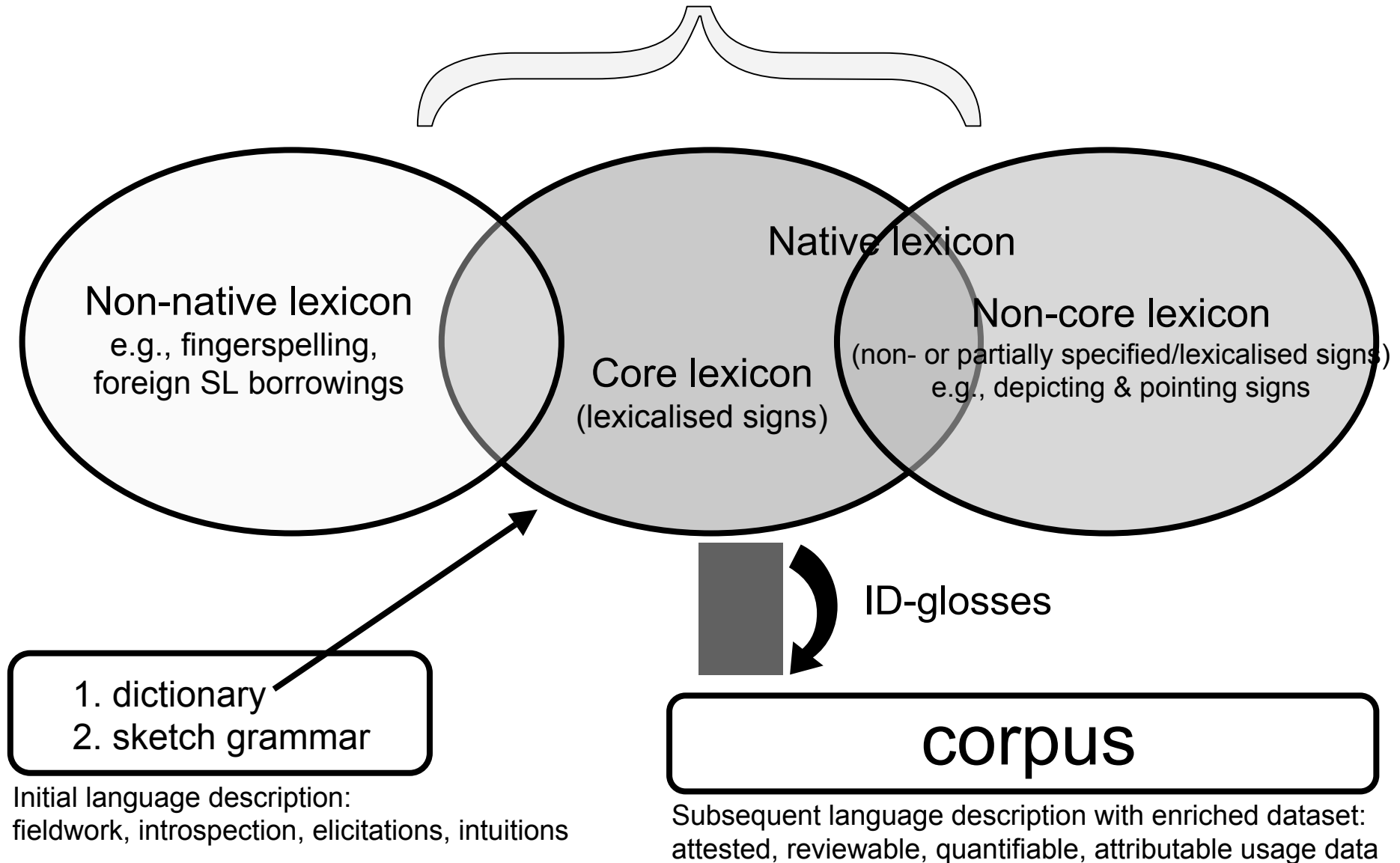
Corpus lemmatisation & tagging

- 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

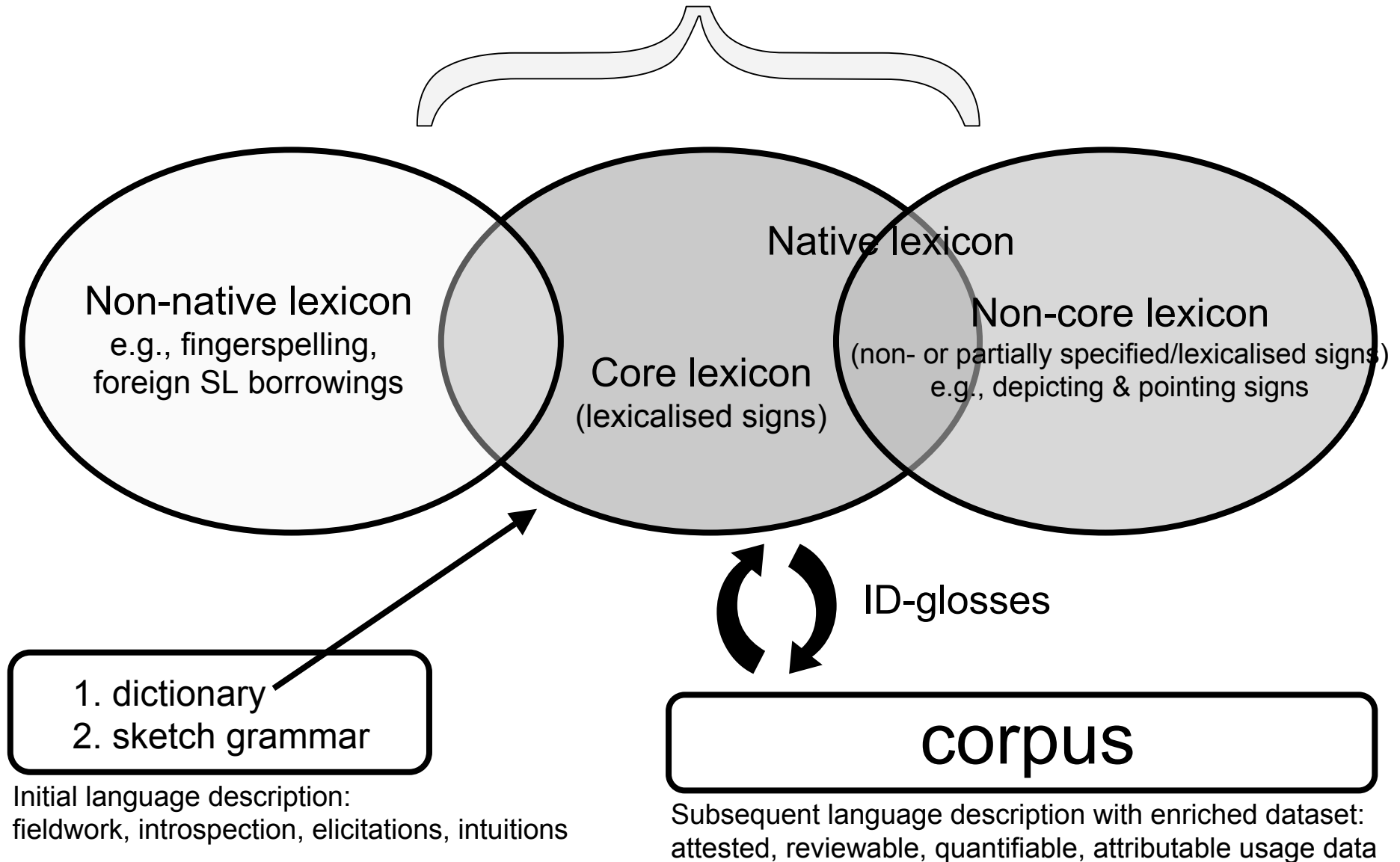
Lemmatisation

- 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



Contents of the lexical database



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

Substring Search Single Layer Search Multiple Layer Search

Domain: 11 eaf files Define New Domain

Query History: < >

Mode: Annotation case sensitive exact match

Find LOOK All Tiers

#hits : 44
 #annotations with a hit : 44
 #annotations investigated : 17972

Ready

hit 1 - 36 of 44

MUSIC INFORM LIPREAD LOOK RESOLUTE LIPREAD MEET
 OTHER FRIEND PM(2"):man-walks-around-and-sits LOOK WANT DANCE WITH
 PT:PRO1sg COINCIDENCE PT:PRO1sg LOOK CATCH PT:PRO3pl CHATTERBOX
 PT:PRO1sg FS:SE(=SO) PM(2"):man-walks-towards-me LOOK HANDSOME PM(2"):man-comes-close-to-me GET-ATTENTION
 NOT THANK-YOU LOOK PT:PRO3sg WHY NOT
 HANDSOME NICE GOOD LOOK LIKE SHOW-OFF PT:PRO1sg
 SHOCK PM(S):head-rotates FATHER LOOK PT:PRO3sg HEARING CRANKY
 PT:PRO3sg PT:PRO1sg NEVER LOOK DEAF FAMILY MANY
 KNOW G:well GOOD LOOK NICE WRITE PT:PRO3sg
 BUT COINCIDENCE LOOK CATCH CHATTERBOX
 WITH LOOK
 FS:STATES PT:LOC PT:PRO3sg(present referent) LOOK PT FS:STA(TES) FIFTY
 PT:PRO3sg JOKE-2 PT:PRO3sg LOOK PUT-UP-WITH NOTHING GIVE
 MOTHER WHERE CHILDREN LOOK AGAIN FAR LOVE
 PT:PRO1sg(B) HAVE PT:DEM LOOK HOLD SLEEP READY
 STOP LAUGH STOP LOOK SN:MISSKENTWELL(HAIR-BUN) FS:MISS FS:KENTWELL MARRY
 PT:LOC PM(bCflat):gap-in-scaffold WRONG-MIND LOOK COME GO-point GET-SE
 HAVE PM(B):long-boards FLUKE LOOK PT:PRO1sg G:arms-out-lying-flat SN:ALFA
 PM(1):boy-far-from-other-boy PT:PRO3sg DEMAND-berate LOOK DEMAND-berate WITH PT:LOC
 LOOK
 GUFFAW LOOK G:gavin-umming
 SPRINT DISAPPEAR2 PM(H):hare-running LOOK TURTLE LAUGH PT
 TURTLE PM(B):turtle-walking RABBIT LOOK LAUGH LOOK G:well
 RABBIT LOOK LAUGH LOOK G:well G:go-away SLOW
 SLOW ALWAYS SLOW LOOK G:go-away LAUGH PM(H):hare-running
 PM(G):hare-running-circuit AGAIN COME LOOK BAD-LUCK PT:PRO2sg ARRIVE
 G:well PT:PRO3sg RABBIT LOOK PT:PRO1pl(2) OPPOSE WANT
 WANT PT:PRO1pl-2 RABBIT LOOK GOOD WHY NOT
 PM(B):hare-running DISAPPEAR2 TURTLE LOOK G:go-away PM(B):turtle-moving FS:GRA(SS)
 TURTLE PM(B):turtle-moving PM(G):hare-diminishing LOOK G:go-away PM(B):turtle-moving RABBIT
 SLEEPY TIRED-claw SLEEPY LOOK TREE AREA GOOD
 TURTLE WHERE GET-UP LOOK G:don't-know HAVE-NOT TURTLE
 G:don't-know HAVE-NOT TURTLE LOOK NOTHING WONDER PM(H):hare-running
 PM(B):hare-running PM(B):location-of-hare PM(V):hare-turning-corner LOOK PT G:well TERRIBLE
 TURTLE G:well DIMINISH LOOK RABBIT SAY
 SAFE PM(4):ground G:out-of-breath LOOK TURTLE DIMINISH

- All instances
 - concordance view
 - jump to any example

Statistics

Annotations Tiers

Tiers

RH ID gloss

Statistics Variables

Annotation	Occurrences	Frequency	Average Duration	Time Ratio	Latency
POSS3pl	1	0.00283567276...	0.13	3.68637459237...	346.36
PREGNANT	1	0.00283567276...	0.54	0.00153126329...	284.23
PRIOR-TO-B	1	0.00283567276...	0.46	0.00130440947...	31.6
PT	3	0.00850701829...	0.33	0.00280731603...	22.44
PT:DEM	3	0.00850701829...	0.29	0.00246703530...	15.81
PT:LOC	3	0.00850701829...	0.21333333333...	0.00181483056...	30.63
PT:PRO3sg	3	0.00850701829...	0.16	0.00136112292...	155.007
PT:PRO1sg	70	0.19849709343...	0.18602857142...	0.03692613072...	5.49
PT:PRO1sg(B)	3	0.00850701829...	0.13666666666...	0.00116262583...	17.54
PT:PRO2sg	7	0.01984970934...	0.17285714285...	0.00343116404...	73.353
PT:PRO3pl	3	0.00850701829...	0.24666666666...	0.00209839784...	63.953
PT:PRO3pl	1	0.00283567276...	0.19	5.38777825038...	59.383
PT:PRO3sg	27	0.07656316461...	0.20814814814...	0.01593648093...	77.773
PT:PRO3sg	1	0.00283567276...	0.14	3.96994186870...	149.547
PUT-UP-WITH	4	0.01134269105...	0.4175	0.00473557351...	138.907
REGULAR	1	0.00283567276...	0.21	5.95491280306...	281.63
RESOLUTE	2	0.00567134552...	0.11	6.23848007939...	42.67
SAME	3	0.00850701829...	0.16666666666...	0.00141783638...	181.746
SAVE-UP	1	0.00283567276...	0.25	7.08918190840...	303.87
SAY	5	0.01417836381...	0.094	0.00133276619...	88.543
SEE	2	0.00567134552...	0.135	7.65631646108...	188.867
SHAKE-HANDS	1	0.00283567276...	1.48	0.00419679568...	60.813
SHOCK	2	0.00567134552...	0.52	0.00294909967...	239.547
SHOW	1	0.00283567276...	0.15	4.25350914504...	230.358
SHOW-OFF	1	0.00283567276...	0.47	0.00133276619...	220.307
SHUT-UP	1	0.00283567276...	0.52	0.00147454983...	126.233
SIGN	4	0.01134269105...	0.49	0.00555791861...	278.9
SISTER	1	0.00283567276...	0.13	3.68637459237...	255.0
SIT	1	0.00283567276...	1.12	0.00317595349...	49.006
SMALL	1	0.00283567276...	0.27	7.65631646108...	302.01
SOCIALISE	1	0.00283567276...	1.09	0.00309088331...	11.09
SOFT	1	0.00283567276...	0.14	3.96994186870...	138.467
SOME	1	0.00283567276...	0.31	8.79058556642...	160.207
SOUND	1	0.00283567276...	0.18	5.10421097405...	106.603
STILL	3	0.00850701829...	0.23666666666...	0.00201332766...	100.213
SURPRISE-claw	1	0.00283567276...	0.35	9.92485467177...	24.24
TALK	2	0.00567134552...	0.366	0.00207571246...	105.583
TALL	1	0.00283567276...	0.17	4.82064369771...	218.247
TEACH	4	0.01134269105...	0.55	0.00623848007...	277.48
THANK-YOU	5	0.01417836381...	0.522	0.00740110591...	75.37
THINK	8	0.02268538210...	0.21125	0.00479228697...	5.76
TIME2	1	0.00283567276...	0.11	3.11924003969...	15.99
TRAVEL	2	0.00567134552...	0.285	0.00161633347...	16.48
TRUE	4	0.01134269105...	0.19275	0.00218630370...	131.056
TWELVE-O'CLOCK	2	0.00567134552...	1.22	0.00691904154...	143.757
TWO	1	0.00283567276...	0.15	4.25350914504...	310.5
WALK	1	0.00283567276...	0.46	0.00130440947...	154.437
WANT	12	0.03402807316...	0.22083333333...	0.00751453282...	53.553
WANT-NOT	2	0.00567134552...	0.21	0.00119098256...	104.383
WE-TWO	2	0.00567134552...	0.785	0.00445200623...	289.42
WHAT	2	0.00567134552...	0.14	7.93988373741...	200.217
WHISKEY	1	0.00283567276...	1.15	0.00226103267...	110.472

- Automatic extraction of frequency lists
 - exported
 - sorted
- Semi-automatic tagging for frequency
 - find ID-gloss
 - tag on frequency tier

Substring Search Single Layer Search Multiple Layer Search

Domain: 11 eaf files Define New Domain

Query History: < >

Mode: Annotation case sensitive substring match

Find PT: Tier Name: RH ID gloss

#hits : 351
 #annotations with a hit : 351
 #annotations investigated : 17972

Ready

hit 1 - 36 of 351

POSS1sg LIFE YOUNG PT:PRO1sg THINK PT:PRO1sg GOOD
 YOUNG PT:PRO1sg THINK PT:PRO1sg GOOD LIFE YES
 GOOD LIFE YES PT:PRO1sg HAVE GOOD WORK
 HAVE GOOD WORK PT:PRO1sg SOCIALISE WITH DEAF
 MEET LOTS PEOPLE PT:DEM TIME2 PT:PRO1sg TRAVEL
 PEOPLE PT:DEM TIME2 PT:PRO1sg TRAVEL AROUND AUSTRALIA
 AROUND AUSTRALIA NOT-YET PT:PRO1sg(B) OVERSEAS PT:PRO1sg TRAVEL
 NOT-YET PT:PRO1sg(B) OVERSEAS PT:PRO1sg TRAVEL PM(1):trace-route-of-holiday AUSTRALIA
 PM(1):trace-route-of-holiday AUSTRALIA NEXT-1 PT:PRO1sg FALL-IN-LOVE G:hold-to-heart POSS1sg
 FALL-IN-LOVE G:hold-to-heart POSS1sg PT: MAN PT: MEET
 POSS1sg PT: MAN PT: MEET SURPRISE-claw PT:DEM
 PT: MEET SURPRISE-claw PT:DEM NIGHT PT:PRO1sg GO-point-2h
 SURPRISE-claw PT:DEM NIGHT PT:PRO1sg GO-point-2h OLD FS:LION
 FS:LION LION FS:HOTEL PT:LOC NORTH ADELAIDE PRIOR-TO-B
 FS:DISCO DANCE PM(bO-5):many-people-move PT:PRO1sg MUST-should MEET POSS1sg
 HEARING FRIEND PT:PRO1sg PM(1):person-move LOUD MUSIC
 RESOLUTE LIPREAD MEET PT:PRO1sg NOT-YET NOT-HAPPEN HAVE-NOT
 NOT-HAPPEN HAVE-NOT WHY PT:PRO1sg ANNOYED G:bit-disappointed HANDSOME
 WANT DANCE WITH PT:PRO1sg PT:PRO1sg COINCIDENCE PT:PRO1sg
 DANCE WITH PT:PRO1sg PT:PRO1sg COINCIDENCE PT:PRO1sg LOOK
 PT:PRO1sg PT:PRO1sg COINCIDENCE PT:PRO1sg LOOK CATCH PT:PRO3pl
 PT:PRO1sg LOOK CATCH PT:PRO3pl CHATTERBOX WANT SHAKE-HANDS
 WANT SHAKE-HANDS FS:BE PT:PRO1sg HAVE MONEY PT:PRO3pl
 PT:PRO1sg HAVE MONEY PT:PRO3pl PT:PRO1sg FS:SE(=SO) PM(2"):man-walks-towards-1
 HAVE MONEY PT:PRO3pl PT:PRO1sg FS:SE(=SO) PM(2"):man-walks-towards-me LOOK
 LIKE DANCE WITH PT:PRO2sg PT:PRO1sg NOT
 DANCE WITH PT:PRO2sg PT:PRO1sg NOT THANK-YOU
 NOT THANK-YOU LOOK PT:PRO3sg WHY NOT THANK-YOU
 LIKE DANCE WITH PT:PRO2sg NICE PT:PRO1sg MUST
 WITH PT:PRO2sg NICE PT:PRO1sg MUST SAY FAIR
 WANT DANCE WITH PT:PRO2sg NOT THANK-YOU PT:PRO1sg
 PT:PRO2sg NOT THANK-YOU PT:PRO1sg STILL ARGUE ALL
 NIGHT WHY ARGUE PT:PRO1sg NOT THANK-YOU PT:PRO1sg
 PT:PRO1sg NOT THANK-YOU PT:PRO1sg WANT-NOT DANCE FIRST-OF-ALL
 TALK EXAGGERATE ARGUE PT:PRO1sg WITH PEOPLE CROWDED
 CAN-NOT MOVE ARGUE PT:PRO1sg WANT DRINK PT:PRO1sg

- All instances
- concordance view
 - understand environment
 - jump to any example

Substring Search Single Layer Search Multiple Layer Search

Domain: 11 eaf files Define New Domain

Query History: < >

Mode: Annotation case sensitive substring match

Find PT: Tier Name: RH ID gloss

#hits : 351
 #annotations with a hit : 351
 #annotations investigated : 17972 Ready

frequency 1 - 23 of 23

Annotation	Percentage	Count
PT:PRO1sg	49.00%	172
PT:PRO3sg	12.25%	43
PT:PRO1sg(B)	7.12%	25
PT:I OC	6.84%	24
PT:	5.98%	21
PT:PRO2sg	5.13%	18
PT:DEM	4.84%	17
PT:PRO3pl	1.99%	7
PT:POSS1sg	1.71%	6
PT:PRO1pl(2)	0.85%	3
PT:PRO1sg(A)	0.57%	2
PT:PRO1pl	0.57%	2
PT:b	0.28%	1
PT:PRO3sg	0.28%	1
PT:PRO3pl	0.28%	1
PT:PRO1sg(B")	0.28%	1
PT:PRO1sg(8)	0.28%	1
PT:PRO1sg(5)	0.28%	1
PT:PRO1sg(2)	0.28%	1
PT:PRO1pl-2	0.28%	1
PT:PRO1pl(B)	0.28%	1
PT:POSS1sg(B)	0.28%	1
PT:(5)	0.28%	1

➤ All instances

- frequency view
- compare variants

Substring Search Single Layer Search Multiple Layer Search

Domain: 11 eaf files Define New Domain

Query History: < >

Mode: Clear

Tier Name: RH ID gloss

Tier Name: RH-gram cls

Tier Name: RH mod

#hits : 33
 #annotations with a hit : 33
 #annotations investigated : 17972

Ready

>

hit 1 - 21 of 33

```
#1 ILOOKI || #2 IVIDir || #3 lml ||
#1 ILOOKI || #2 IVIDir || #3 lml ||
#1 ILOOKI || #2 IVIDir || #3 lml ||
#1 ILOOKI || #2 IVIDir || #3 lml ||
#1 ILOOKI || #2 IVIDir || #3 lml ||
#1 ILOOKI || #2 IVIDir || #3 lml ||
#1 ILOOKI || #2 IVIDir || #3 lml ||
#1 ILOOKI || #2 IVIDir || #3 lml ||
#1 ILOOKI || #2 IVIDir || #3 lml ||
#1 ILOOKI || #2 IVIDir || #3 lml ||
#1 ILOOKI || #2 IVIDir || #3 lml ||
#1 ILOOKI || #2 IVIDir || #3 lml ||
#1 ILOOKI || #2 IVIDir || #3 lml ||
#1 ILOOKI || #2 IVIDir || #3 lml ||
#1 ILOOKI || #2 IVIDir || #3 lml ||
#1 ILOOKI || #2 IVIDir || #3 lml ||
#1 ILOOKI || #2 IVIDir || #3 lml ||
#1 ILOOKI || #2 IVIDir || #3 lml ||
#1 ILOOKI || #2 IVIDir || #3 lml ||
#1 ILOOKI || #2 IVIDir || #3 lml ||
```

Search for sign with ID-gloss “LOOK” which is a directional indicating verb (“VDir”) which is modified for space (“m”)

Substring Search Single Layer Search Multiple Layer Search

Domain: 11 eaf files Define New Domain

Query History: < >

Mode: Clear

Tier Name: RH ID gloss

Tier Name: RH-gram cls

Tier Name: RH mod

#hits : 140
 #annotations with a hit : 140
 #annotations investigated : 17972

Ready

>

frequency: 140/17972

Annotation	Percentage	Count
#1 ILOOKI #2 IVIDir #3 lml	23.57%	33
#1 IGO-ATTENTIONI #2 IVIDir #3 lml	5.71%	8
#1 ISAYI #2 IVIDir #3 lml	5.00%	7
#1 ICOMEI #2 IVIDir #3 lml	3.57%	5
#1 ISHUT-UPI #2 IVIDir #3 lml	2.86%	4
#1 IPUTI #2 IVIDir #3 lml	2.86%	4
#1 IGO-point #2 IVIDir #3 lml	2.86%	4
#1 ITEACHI #2 IVIDir #3 lml	2.14%	3
#1 ISMSI #2 IVIDir #3 lml	2.14%	3
#1 IGO-point-2I #2 IVIDir #3 lml	2.14%	3
#1 ICATCHI #2 IVIDir #3 lml	2.14%	3
#1 IASKI #2 IVIDir #3 lml	2.14%	3
#1 IARRIVEI #2 IVIDir #3 lml	2.14%	3
#1 IWILL-NOTI #2 IVIDir #3 lml	1.43%	2
#1 ISHOWI #2 IVIDir #3 lml	1.43%	2
#1 ISEND-OUTI #2 IVIDir #3 lml	1.43%	2
#1 IPT:PROIsgl #2 IVIDir #3 lml	1.43%	2
#1 IPHONEI #2 IVIDir #3 lml	1.43%	2
#1 IMOTHER-FATHERI #2 IVIDir #3 lml	1.43%	2

Repeat search for all signs, using regular expression (“wild card”) character \$

Substring Search Single Layer Search Multiple Layer Search

Domain: 11 eaf files Define New Domain

Query History: < >

Mode: case sensitive exact match Clear

Minimal Duration Maximal Duration Begin After End Before

LOOK Tier Name: RH ID gloss

Overlap

VIDir Tier Name: RH-gram cls

Overlap

n Tier Name: RH mod

Find

#hits : 2
 #annotations with a hit : 2
 #annotations investigated : 17972 Ready

hit 1 - 2 of 2

#1 |LOOK| || #2 |VIDir| || #3 |nl| ||
 #1 |LOOK| || #2 |VIDir| || #3 |nl| ||

Substring Search Single Layer Search Multiple Layer Search

Domain: 11 eaf files Define New Domain

Query History: < >

Mode: case sensitive regular expression Clear

Minimal Duration Maximal Duration Begin After End Before

\$ Tier Name: RH ID gloss

Overlap

VIDir Tier Name: RH-gram cls

Overlap

n Tier Name: RH mod

Find

#hits : 71
 #annotations with a hit : 71
 #annotations investigated : 17972 Ready

frequency 1 - 19 of 33

Annotation	Percentage	Count
#1 SAY #2 VIDir #3 nl	23.94%	17
#1 GO #2 VIDir #3 nl	8.45%	6
#1 ARRIVE #2 VIDir #3 nl	7.04%	5
#1 TELEPHONE #2 VIDir #3 nl	5.63%	4
#1 DISAPPEAR2 #2 VIDir #3 nl	4.23%	3
#1 CANE #2 VIDir #3 nl	4.23%	3
#1 VISIT #2 VIDir #3 nl	2.82%	2
#1 SUN-SHINES #2 VIDir #3 nl	2.82%	2
#1 SNOW #2 VIDir #3 nl	2.82%	2
#1 MEET #2 VIDir #3 nl	2.82%	2
#1 LOOK #2 VIDir #3 nl	2.82%	2
#1 GO-OUT #2 VIDir #3 nl	2.82%	2
#1 WARN #2 VIDir #3 nl	1.41%	1
#1 SUSPEND #2 VIDir #3 nl	1.41%	1
#1 SIGN #2 VIDir #3 nl	1.41%	1
#1 SEQUENCE #2 VIDir #3 nl	1.41%	1
#1 SEE #2 VIDir #3 nl	1.41%	1
#1 SCREAM #2 VIDir #3 nl	1.41%	1
#1 SAY-2 #2 VIDir #3 nl	1.41%	1

Repeat both searches for unmodified forms ("n")

Substring Search Single Layer Search Multiple Layer Search

Domain: 11 eaf files Define New Domain

Query History: < >

Mode: case sensitive exact match Clear

Minimal Duration Maximal Duration Begin After End Before

LOOK Tier Name: RH ID gloss

Overlap

VIDir Tier Name: RH-gram cls

Overlap

cg Tier Name: RH mod

Find

#hits : 4
 #annotations with a hit : 4
 #annotations investigated : 17972 Ready

hit 1 - 4 of 4

#1 ILOOKI || #2 IVIDirI || #3 lcgI ||
 #1 ILOOKI || #2 IVIDirI || #3 lcgI ||
 #1 ILOOKI || #2 IVIDirI || #3 lcgI ||
 #1 ILOOKI || #2 IVIDirI || #3 lcgI ||

Repeat both searches for congruent forms ("cg")

Substring Search Single Layer Search Multiple Layer Search

Domain: 11 eaf files Define New Domain

Query History: < >

Mode: case sensitive regular expression Clear

Minimal Duration Maximal Duration Begin After End Before

\$ Tier Name: RH ID gloss

Overlap

VIDir Tier Name: RH-gram cls

Overlap

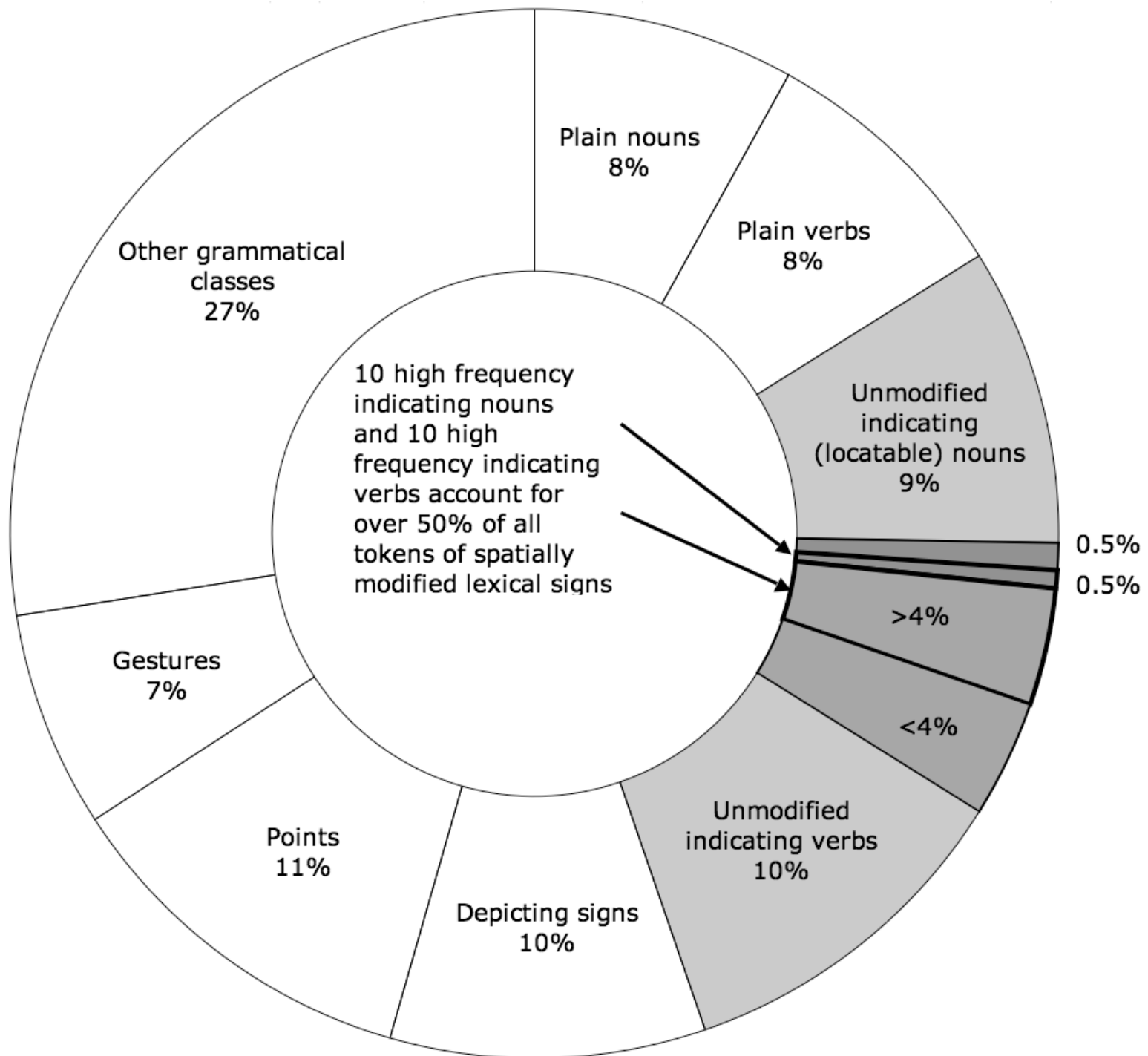
cg Tier Name: RH mod

Find

#hits : 53
 #annotations with a hit : 53
 #annotations investigated : 17972 Ready

frequency 1 - 20 of 20

Annotation	Percentage	Count
#1 IGOI #2 IVIDirI #3 lcgI	13.21%	7
#1 IGUFFAWI #2 IVIDirI #3 lcgI	9.43%	5
#1 IDRIVE-2I #2 IVIDirI #3 lcgI	9.43%	5
#1 ISTOPI #2 IVIDirI #3 lcgI	7.55%	4
#1 ISEEI #2 IVIDirI #3 lcgI	7.55%	4
#1 ILOOKI #2 IVIDirI #3 lcgI	7.55%	4
#1 IKNOW-YOUI #2 IVIDirI #3 lcgI	7.55%	4
#1 IDRIVEI #2 IVIDirI #3 lcgI	7.55%	4
#1 ISPEED-DUSTI #2 IVIDirI #3 lcgI	3.77%	2
#1 IRECEIVEI #2 IVIDirI #3 lcgI	3.77%	2
#1 IFINISH-6I #2 IVIDirI #3 lcgI	3.77%	2
#1 IBETI #2 IVIDirI #3 lcgI	3.77%	2
#1 IWITNESSI #2 IVIDirI #3 lcgI	1.89%	1
#1 ITEACHI #2 IVIDirI #3 lcgI	1.89%	1
#1 ILIPREADI #2 IVIDirI #3 lcgI	1.89%	1
#1 IGO-2I #2 IVIDirI #3 lcgI	1.89%	1
#1 IDAREI #2 IVIDirI #3 lcgI	1.89%	1
#1 ICOMMUNICATEI #2 IVIDirI #3 lcgI...	1.89%	1
#1 IARRIVEI #2 IVIDirI #3 lcgI	1.89%	1
#1 IANNOUNCEI #2 IVIDirI #3 lcgI	1.89%	1



Substring Search Single Layer Search Multiple Layer Search

Domain: 11 eaf files Define New Domain

Query History: < >

Mode: case insensitive regular expression Clear

Minimal Duration Maximal Duration Begin After End Before

PT: = 0 annotations . Tier Name: RH ID gloss

Overlap

V Tier Name: RH-gram cls

Overlap

m Tier Name: RH mod

Find

#hits : 29
 #annotations with a hit : 29
 #annotations investigated : 17972

Ready

frequency 1 - 21 of 27

Annotation	Percentage	Count
#1 IPT: IWILL-NOT #2 VIDir #3 lml	6.90%	2
#1 IPT:PRO3sgl LOOK #2 VIDir #3 lml	6.90%	2
#1 IPT:PRO3sgl JOKE-2 #2 VIDir #3 lml	3.45%	1
#1 IPT:PRO3sgl GET-ATTENTION #2 VIDir #3 l...	3.45%	1
#1 IPT:PRO3sgl ATTRACT #2 VIDir #3 lml	3.45%	1
#1 IPT:PRO3sgl ARGUE #2 VILoc #3 lml	3.45%	1
#1 IPT:PRO3pl CHATTERBOX #2 VILoc #3 ...	3.45%	1
#1 IPT:PRO2sgl ARRIVE #2 VIDir #3 lml	3.45%	1
#1 IPT:PRO1sgl TEACH #2 VIDir #3 lml	3.45%	1
#1 IPT:PRO1sgl STAY #2 VILoc #3 lml	3.45%	1
#1 IPT:PRO1sgl ISAY #2 VIDir #3 lml	3.45%	1
#1 IPT:PRO1sgl IPUT #2 VIDir #3 lml	3.45%	1
#1 IPT:PRO1sgl PM(F):eye-twitches #2 VID #3 l...	3.45%	1
#1 IPT:PRO1sgl LOOK #2 VIDir #3 lml	3.45%	1
#1 IPT:PRO1sgl LOOK-AFTER #2 VIDir #3 l...	3.45%	1
#1 IPT:PRO1sgl HAVE #2 VILoc #3 lml	3.45%	1
#1 IPT:PRO1sgl GO-point-2h #2 VILoc #3 lm...	3.45%	1
#1 IPT:PRO1sgl GO-TO #2 VIDir #3 lml	3.45%	1
#1 IPT:PRO1sgl BOTHER #2 VIDir #3 lml	3.45%	1
#1 IPT:PRO1sgl ARGUE #2 VILoc #3 lml	3.45%	1
#1 IPT:PRO1sg(B) SHAKE-HANDS #2 VILoc #3 ...	3.45%	1

- 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