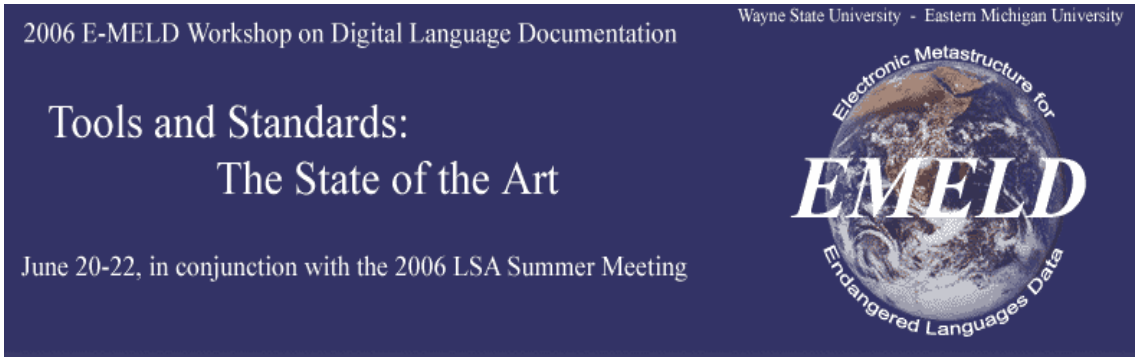


THE USE OF ELAN ANNOTATION SOFTWARE IN THE CREATION OF SIGNED LANGUAGE CORPORA

By

Trevor Johnston & Onno Crasborn

Paper presented at

2006 E-MELD Workshop on Digital Language Documentation
Lansing, MI.
June 20-22, 2006

Please cite this paper as:

Johnston, T. & Crasborn, O. (2006), The use of ELAN annotation software in the creation of signed language corpora, *in* 'Proceedings of the EMELD'06 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art'. Lansing, MI. June 20-22, 2006.

The use of ELAN annotation software in the creation of signed language corpora

Trevor Johnston[1] & Onno Crasborn[2]

[1]Department of Linguistics, Macquarie University, Sydney, Australia
[2]Department of Linguistics, Radboud University, Nijmegen, The Netherlands

Presentation at the 2006 E-MELD Workshop on Digital Language Documentation
Michigan State University in East Lansing, Michigan, June 20-22, 2006

Outline

1. Background
   - Endangerment
   - Documentation of Auslan (Australian Sign Language) & NGT (Nederlandse Gebarentaal or Netherlands Sign Language)
     - documentation of signed languages
   - Linguistics and unwritten languages
2. Language artifacts and archives
3. Corpora
   - Sample Auslan & NGT texts in ELAN

Endangerment

- Johnston, T. (2004). W(h)ither the Deaf Community? Population, Genetics, and the Future of Australian Sign Language. *American Annals of Deaf, 148*(5), 358-375.
- Signing deaf community is historically recent and small
  - 19th century to mid-20th century residential schools
  - “one in one thousand” when really c. one-third this rate (Johnston, 2004)
  - reducing incidence/prevalence
  - technology of assisted hearing
- Today < 6,500 (double to add hearing native signers)
  - < 3.5% (~ 5-10%) are deaf of deaf so that is an even smaller group
  - rubella cohorts dominate and distort
  - rapidly ageing (young community with mostly older members)
  - survival beyond the life-times of the 1970s cohorts unlikely
- Situation in other developed countries vs. developing world

Documentation of Auslan & NGT

- Auslan
  - Primary sources
    - prior to film and television, almost nothing
    - TV news 1980s (now discontinued)
    - film, video recordings private and/or not well described
    - little, if any, digitised to computer files
  - Secondary sources
    - Dictionaries (print, CD, DVD, internet)
    - Sketch grammar
    - L2 teaching materials (video, course books, DVD)
- NGT
  - Primary sources: Visibase archive
    - heterogeneous collection of most video recordings made for research purposes since the late 1980s
    - movies partly digitised to computer files
    - described to variable degrees by IMDI metadata files
  - Secondary sources
    - Video tapes made for deaf children
    - Various dictionary projects (CD, DVD)
    - L2 teaching materials (video)

Documentation of signed languages

- Primary sources
  - there are few archives or collections
    - prior to film and television, almost nothing
    - TV (Britain & other European countries 1980+)
    - film, video recordings private and/or not well described
- Secondary sources
  - effectively made without representative archives or corpora
    - dictionaries
    - sketch grammars
    - L2 teaching materials
- Sign language documentation to address the absence of
  - representative, well-described archival material
  - accessible, machine-readable corpora

Linguistics and unwritten languages

- Study of grammar essentially based on writing*
  - hundreds (even thousands) of years experience/history
  - few people (‘experts’) could write (until 20th century)
  - learned in school, not automatic
  - learned conscious rules for ‘good writing and grammar’
  - careful and planned
  - permanent and public (libraries, literatures)
- Without writing, analysis problematic (impossible?)
  - writing is a form of linguistic analysis
  - no surprise that documentation and/or development of a writing system is a major first step in linguistic description and analysis

Writing and signed languages

- No written forms of signed languages
  (one putative writing system, Sutton SignWriting)
  - no folk linguistics associated with writing
  - no standardization associated with the spread and teaching of writing
  - no written literature (i.e., no reference or sacred texts)
  - no culture of writing (i.e., no elite enforcing standards)
  - no possibility of ‘text mining’
- Unlike spoken languages, there is not even a widely used transcription system like IPA
  (but see Stokoe Notation, Hamburg Notation System)

Automatic annotation

- Automated annotation and tagging of written data not available for signed languages (and not available for foreseeable future)
- Example of text mining and tagging in English *("Joanna stubbed out her cigarette with unnecessary fierceness")*

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

_NP	= singular proper noun
_NN	= singular common noun
_VBD	= past tense form of lexical verb
_IN	= preposition
_RP	= adverbial particle
_JJ	= adjective
_PP$	= possessive pronoun
_.	= full stop

2. Language artifacts and archives

- Reasons for archive
  - Preserved for future generations in circumstances of language endangerment resulting in...
    - ...dramatic linguistic change, or
    - ...disappearance of community entailing language death
- Features of archive
  - recordings of language use by film, video etc.
    - also books, pamphlets, posters, papers (minutes) etc.
  - collected and stored somewhere
  - sorted and categorized (catalogued)
  - historical record
  - able to be consulted/viewed

Limits of simple archive

- Accessibility
  - form, number of sites, permission to view by subject?
- Representativeness
  - sampling, native signers, and natural language (elicitations, recitations)?
- Comparability
  - identifiable text types; numbers of participants doing similar language-based tasks?
- Searchability
  - finding and identifying linguistic entities easily (e.g., words vs. signs, types of signs, non-manual features, etc.)?

3. Corpora

- Corpus: a representative collection of naturalistic written, spoken or signed texts in a machine-readable form
- Accessible and “machine-readable”
  - in an open repository
  - in a digital form and able to be searched by a computer
    - British National Corpus of Spoken English;
    - CGN: Spoken Dutch Corpus
    - OLAC: Open Language Archives Community;
    - ELDP Endangered Languages Documentation Project (SOAS);
    - Linguistic Data Consortium (University of Pennsylvania);
    - PARADISEC (University of Melbourne, University of Sydney, etc.).

Archives vs. corpora in SL linguistics

- impossibility of testing language descriptions of and hypotheses about most SLs (no real access to primary data)
  - (a) idiosyncratic glossing and transcription
  - (b) no open archive of naturalistic recordings
  - and (a) not linked to (b)
- empirical evidence-based signed language linguistics
  - native user intuitions alone problematic
    - in all linguistics but especially in signed language linguistics
  - support/contest intuitions with the corpus examples
- track changes in SLs (widely reported to be rapid)
  - over the past 100 years
  - from now into the future

Do we have SL corpora?

- Despite claims or assumptions to the contrary, most SL researchers appear to have little or no real corpora and little which is easily accessible by other researchers.
- We have incidental or accidental archives, as described above.
- True, many SL linguists having extensive collections of hundreds of hours of video, but in terms of what we mean today by a ‘linguistic corpus’ (see above) most of these recordings or archives would be of limited use or dubious value
- Limitations of collections (where they exist)
  - perishable format (e.g., analogue tape)
  - no explicit releases from participants (i.e., fail to meet contemporary ethical standards and research protocols)
  - lack accurate metadata (i.e., there is little or no information about participants language background and the filming situation)
  - they are not digitised, and without linguistic tags or annotations (i.e., the are not in a machine readable format)
  - they are not in open archives (i.e., able to be consulted for peer review and validation)

The case of Auslan & NGT

- There is no real corpus of Auslan.
- 'Test Battery of Auslan Morphology and Syntax Project' (1999-2000)
  - native signers, similar activities
  - 25 participants x 2 hours
  - coded (not annotated)
- 'Sociolinguistic Variation Project' (2003-07)
  - native signers, similar activities, sociolinguistic metadata
  - 206 participants x 2 hours
  - partially annotate (not ELAN)
- The Visibase project (1996-2002) aimed to digitise all research materials ever collected for NGT.
  - result is an incomplete collection of video recordings; partly digitised to MPEG-1/MPEG-2, partly described by IMDI metadata categories
  - only some recent recordings have been transcribed in an electronic form
- The ECHO project (2003-2004) created a pilot corpus for NGT, British SL and Swedish SL (fable stories, Swadesh list, some poetry)
  - phonetic transcription in ELAN is still limited; basic linguistic annotation is present

The technology of SL corpora

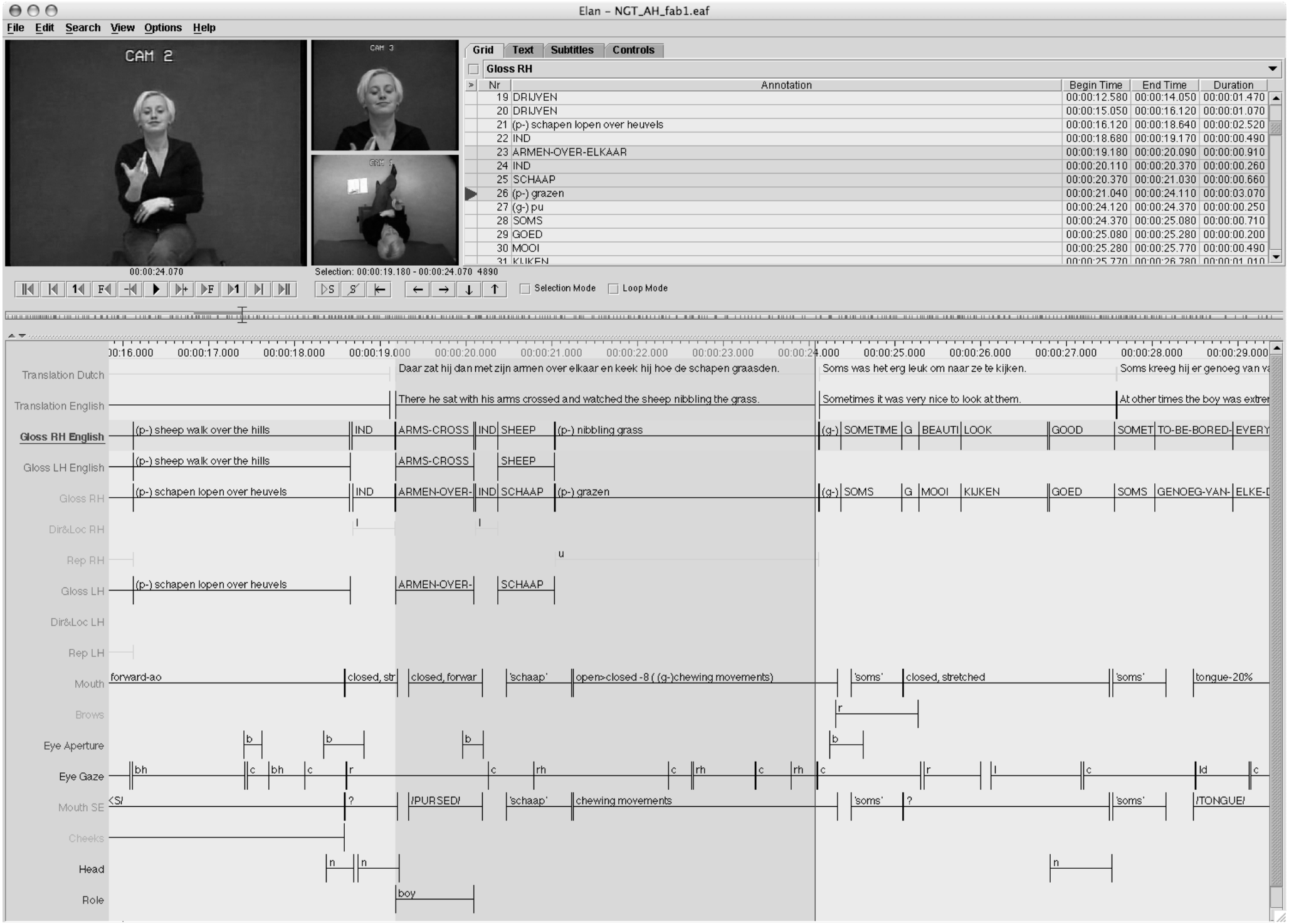
- Digital video technology (quality) & PC processing power limits (memory)
- Software inadequacies (MediaTagger, SyncWRITER, SignStream)
- Now possible, ELAN:
  - annotation and tagging of video clips
    - a kind of substitute for ‘writing’
    - annotation rather than transcription (still lacking a ‘SL IPA’)
    - annotation using linguistic ‘tags’
    - machine-readability based on tags
- Need for strategic use of technology
  - cumulative annotations of a given corpus

The linguistics of SL corpora

- No conventionally agreed, exhaustive set of linguistic descriptors
  - e.g., ‘parts of speech’ (i.e., grammatical word class)
- Annotators (or computer programs) cannot simply apply well established tags to data (assuming it to have already been easily represented through a writing system or a transcription system like IPA).
- Researchers need to have access to the primary data (video clip) for meaningful peer review of claims being made about that data (based on the tagging) because the tagging itself (if not the identification of sign-units themselves) could be disputed, let alone the analysis.

About ELAN

- Linguistic annotation software developed at MPI
  - originally for language and gesture studies (MediaTagger)
  - precise time-alignment of annotations to video/audio sources
  - unlimited no. of user-definable tiers; templates of tier setups
  - linking of annotation to other annotations
    - searching across multiple annotation files (i.e., a corpus)
  - support for different character sets (allows for transcription and glossing in different writing systems)
- Export/import of various forms
  - including tab-delimited text files, CHAT, Transcriber, and Shoebox
  - allows for processing with database programs
  - allows for connectivity with other widely-used linguistic software


Elan – NGT_AH_fab1.eaf
File Edit Search View Options Help
CAM 2
CAM 3
CAM 1
Grid Text Subtitles Controls
Gloss RH
Nr Annotation Begin Time End Time Duration
19 DRIJVEN 00:00:12.580 00:00:14.050 00:00:01.470
20 DRIJVEN 00:00:15.050 00:00:16.120 00:00:01.070
21 (p-) schapen lopen over heuvels 00:00:16.120 00:00:18.640 00:00:02.520
22 IND 00:00:18.680 00:00:19.170 00:00:00.490
23 ARMEN-OVER-ELKAAR 00:00:19.180 00:00:20.090 00:00:00.910
24 IND 00:00:20.110 00:00:20.370 00:00:00.260
25 SCHAAP 00:00:20.370 00:00:21.030 00:00:00.660
26 (p-) grazen 00:00:21.040 00:00:24.110 00:00:03.070
27 (g-) pu 00:00:24.120 00:00:24.370 00:00:00.250
28 SOMS 00:00:24.370 00:00:25.080 00:00:00.710
29 GOED 00:00:25.080 00:00:25.280 00:00:00.200
30 MOOI 00:00:25.280 00:00:25.770 00:00:00.490
00:00:24.070
Selection: 00:00:19.180 - 00:00:24.070 4890
Selection Mode
Loop Mode
00:00:16.000 00:00:17.000 00:00:18.000 00:00:19.000 00:00:20.000 00:00:21.000 00:00:22.000 00:00:23.000 00:00:24.000 00:00:25.000 00:00:26.000 00:00:27.000 00:00:28.000 00:00:29.000
Translation Dutch
Daar zat hij dan met zijn armen over elkaar en keek hij hoe de schapen graasden.
Soms was het erg leuk om naar ze te kijken.
Translation English
There he sat with his arms crossed and watched the sheep nibbling the grass.
Sometimes it was very nice to look at them.
Gloss RH English
(p-) sheep walk over the hills
IND
ARMS-CROSS
IND
SHEEP
(p-) nibbling grass
(g-)
SOMETIME
G
BEAUTI
LOOK
GOOD
SOMET
TO-BE-BORED-
Gloss LH English
(p-) sheep walk over the hills
ARMS-CROSS
SHEEP
Gloss RH
(p-) schapen lopen over heuvels
IND
ARMEN-OVER-
IND
SCHAAP
(p-) grazen
(g-)
SOMS
G
MOOI
KIJKEN
GOED
SOMS
GENOEG-VAN-
Dir&Loc RH
Rep RH
u
Gloss LH
(p-) schapen lopen over heuvels
ARMEN-OVER-
SCHAAP
Dir&Loc LH
Rep LH
Mouth
forward-ao
closed, str
closed, forwar
'schaap'
open>closed -8 ((g-)chewing movements)
'soms'
closed, stretched
'soms'
tongue-20%
Brows
Eye Aperture
Eye Gaze
Mouth SE
?
/PURSED/
'schaap'
chewing movements
'soms'
?
'soms'
/TONGUE/
Cheeks
Head
Role
boy

ELAN annotation files & metadata

- appended/linked to all annotation files
- compatible with IMDI (ISLE MetaData Initiative)
  - aim to use of IMDI standard and IMDI browser (MPI also)
  - sign language specific metadata (the 'Sign language profile' a subset[1] of those proposed by Crasborn & Hanke, 2003[2]) has been embedded in the IMDI editor, e.g.,
    - sign competence (e.g., acquisition age, use, region)
    - bilingualism (e.g, speaking, reading, etc.)
    - education (e.g., deaf school, oral school, hearing school)
    - hearing status? (i.e., not hearing in Auslan Archive/Corpus)
- tiers transcription>glossing>annotation>metadata

1. Crasborn, O. & Hanke, T. 2003. Additions to the IMDI metadata set for sign language corpora. Agreements at an ECHO workshop, May 8 + 9, 2003, Radboud University Nijmegen. http://www.let.ru.nl/sign-lang/echo/docs/SignMetadata_May2003.doc

2. Crasborn, O. & Hanke, T. 2003. "Metadata for sign language corpora: Background document for an ECHO workshop, Nijmegen University" http://www.let.ru.nl/sign-lang/echo/index.html

Current extension of ELAN

- Part of CNGT corpus project (2006-2008)
- Interface
  - enhance viewing of large amounts of tiers
  - increase no. of keyboard shortcuts
- Annotations
  - further facilitate creation of new annotations and manipulation of existing annotations (e.g. easily copying selected annotation to another tier)
- Searching
  - enable complex searches across multiple annotation files
- Collaboration between researchers
  - enable import of selected tiers from another document (created by other researchers)
  - enable adding and visualising of a ‘annotator’ property for each tier

Some key features of a corpus

a) Sampling and representativeness

b) Finite size

c) A standard reference

d) Machine-readable form

a) Sampling and representativeness

- 20 individuals x 5 cities x 3 hours (i.e., 100 participants yielding 200-300 hours of footage)
- native signers only (deaf of deaf or acquired <6 years) talking to each other (two per session), not the entire signing community
- standard tasks
  - narrative (text stimulus Aesop's Fables)
  - recount memorable event
  - attitudes survey (text/sign stimulus)
  - conversation, narrative (non-linguistic video stimulus 'Tweety & Sylvester')
  - narrative recount ("Frog Where Are You" picture book or Auslan stimulus)
  - depicting ("classifier") signs (video stimulus)
  - sign noun-verb derivation (video stimulus)
  - question formation ('spot the difference' picture stimulus)

b) Finite size

- Not necessarily limited to 300 hours, but no plans to extend native signer corpus
  - priority is for empirical grammar based on richly transcribed and tagged corpus (decades?)
- Future extension desirable of complementary groups
  - deaf non-natives, deaf late-learners, hearing natives

c) Standard reference

- Not primary purpose
  - however, language endangerment may mean that relatively soon it will become a ‘standard reference’ for native-like Auslan
- Not meant as a standard reference upon which to prescribe ‘good usage’
  - but if endangerment is real and attrition or linguistic stress become manifest, it could become so

d) Machine-readability

- ELAN interlinear text transcriptions and annotations exportable to database and statistical programs for
  - searching
  - retrieving
  - sorting
  - calculating

Possible tags for SL corpora

- Sign type (lexical status): lexicalised, productive, gesture
- Sign class (‘part of speech’): noun, verb, adjective, etc.
- Verb type: plain, indicating (‘directional’ ‘agreeing’ ‘spatial’), depicting (‘classifier’)
- Sign/stem modification: repetition, slow, fast, etc.
- Clause boundaries
- Semantic or grammatical roles:
  actor, undergoer, locative, instrument, etc.
- Perspective shift/role shift
- ‘Prosody’ (eyebrows, head)
- Expression (head, eyes, mouth gestures)
- Mouthing (of spoken word)
- Spatial placement and/or direction modification:
  lf, rt, up, dn, far, near

The Auslan Corpus

- Rich and detailed annotation of 300 hours could take one person 5 days a week, 48 weeks a year, *more than 50 years* to complete!
- Compromise (for current project)
  - 100% with descriptive meta-data and voice over
  - 10% with above + and annotation of specific signs and tagging for purposes of particular study (e.g., mouthing, ‘pro-drop’)
  - 5% with above + identification of all signs and general annotation
  - 1% with above + exhaustive detailed tagging on the use of space
- The annotation process is meant to be cumulative
  - Subsequent projects will add layers of tagging to corpus
- The usefulness of the corpus is meant to be on-going

The NGT Corpus

- Project started in May 2006; 24 months
- Data recording
  - 12 signers x 2 regions x 2 hours = 48 hours of sign data
  - possible extension to 5 regions (120 hours)
  - use of 3-6 simultaneous video cameras, leading to between 144 and 720 hours of video data
- Data annotation
  - original plan: full glossing (left hand, right hand) and translation in Dutch; partly translate Dutch annotations to English
  - possible diversification of annotations (and/or voice-over translation) similar to Auslan corpus, adding phonetic and/or morphosyntactic annotations

Sample ELAN files: Auslan

- A general purpose richly annotated text
  - Carpentry Story
- A text annotated for ‘classifier’ use
  - Verbs of Motion Production tasks
- Texts annotated for constituent order analysis
  - Picture elicitation task for word order
- A text annotated for use of space with verbs
  - Narrative (“Hare and Tortoise”)

Sample ELAN files: NGT

- ECHO fable story: 3 synchronised camera views
- Annotation of poetry tape by Wim Emmerik