

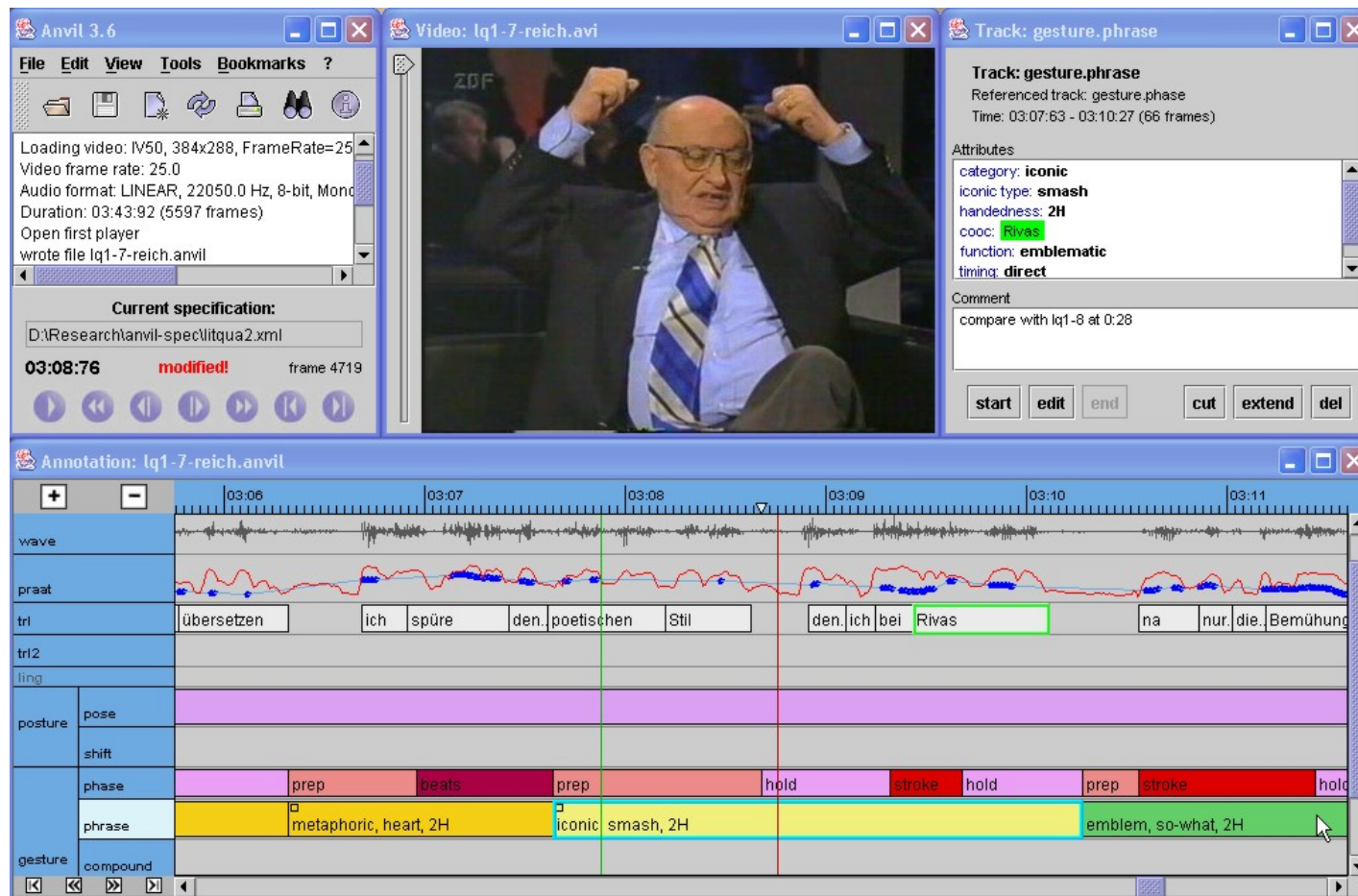
An exchange format for multimodal annotations

Thomas Schmidt, Susan Duncan,
Oliver Ehmer, Jeffrey Hoyt,
Michael Kipp, Dan Loehr,
Magnus Magnusson, Travis
Rose, Han Sloetjes

Background

- International Society for Gesture Studies (ISGS)
 - 2005 Conference in Lyon (‘Interacting Bodies’)
 - User workshop on ‘Multimodal Annotation Tools’
 - 2007 Conference in Chicago (‘Integrating Gestures’)
 - Developer workshop on ‘Annotation Interchange among Multimodal Annotation Tools’
 - Goal: Interoperability between existing tools

Tools (1): Anvil



Developer: Michael Kipp, DFKI Saarbrücken

Tools (2): C-BAS

The screenshot displays the C-BAS software interface. At the top, a video player shows a person sitting in a chair with their arms raised. The video has a timestamp of 00:04:03:02. Below the video player is a control bar with play, stop, and volume icons. To the right of the video player is a 'Video Information' panel with the following details:

Video Information
Title: Baseline, John
Time: 007.74
Current Pos: 243.255

Below the video player is a list of key events:

- Key m: Cue: Right Hand to Face
- Key i: Cue: Left Hand to Face
- Key d: Cue: Both Hands to Face
- Key t: Cue: Hands Together
- Key q: Cue: Right Hand Up
- Key w: Cue: Left Hand Up
- Key e: Cue: Both Hands Up
- Key r: Cue: Right Hand Out
- Key f: Cue: Left Hand Out
- Key g: Cue: Both Hands Out
- Key u: Cue: Right Hand In
- Key v: Cue: Left Hand In
- Key j: Cue: Right Fingers Moving
- Key h: Cue: Left Fingers Moving
- Key o: Cue: Both Fingers Moving
- Key z: Cue: Right Head Tilt
- Key x: Cue: Left Head Tilt

At the bottom of the interface is a table with the following columns: Start Time-End Time, Duration, Event, and Start Frame-End Frame.

Start Time-End Time	Duration	Event	Start Frame-End Frame
(1.429 - 1.677)	0.248	Right Hand to Face	142 - 192
(2.264 - 2.512)	0.248	Left Hand to Face	171 - 199
(3.481 - 4.251)	0.770	Both Hands to Face	1105 - 1421
(5.432 - 13.955)	8.523	Left Hand to Face	1163 - 4159
(14.37 - 14.672)	0.302	Both Hands to Face	431 - 449
(15.701 - 17.922)	2.221	Hands Together	471 - 3638
(16.101 - 16.249)	0.147	Right Hand to Face	483 - 487
(17.13 - 17.275)	0.145	Left Hand to Face	914 - 919
(17.322 - 17)	0.078	Both Hands to Face	828 - 1119
(18.079 - 49)	3.121	Left Hand to Face	1142 - 1449
(49.274 - 59)	9.726	Right Hand to Face	1148 - 1749
(59.431 - 74.435)	15.004	Left Hand to Face	1783 - 2213
(76.919 - 86)	9.181	Left Hand to Face	2385 - 2580
(133.117 - 214.974)	81.857	Both Hands Up	8581 - 7020
(133.117 - 214.974)	81.857	Both Hands Up	1254 - 7219

Developer: Kevin Moffit, University of Arizona

Tools (3): ELAN

The screenshot displays the ELAN software interface for the file 'Elan - r03_v20_s5.eaf'. The interface includes a menu bar (File, Edit, Annotation, Tier, Type, Search, View, Options, Window, Help) and a toolbar with playback controls. A video window on the left shows a scene with several people in a traditional setting. The main area features a 'Gloss2' table with columns for 'Nr', 'Annotation', 'Begin Time', 'End Time', and 'Duration'. Below the table is a timeline with multiple tracks for different tiers: Gloss1, Gloss2, Kp, Pius, K, Ricky, man outside vie, child 1, and loudspeaker. The timeline shows the alignment of these tiers with the audio waveform above it.

Nr	Annotation	Begin Time	End Time	Duration
1	he's speaking into his 1/2 bush knife	00:00:31.708	00:00:32.708	00:00:01.000
2	I am saying that he's talking into a half bushknife	00:00:37.141	00:00:38.501	00:00:01.360
3	everything we are saying is going inside	00:00:39.336	00:00:40.536	00:00:01.200
4	See this fellow: he seems to me like a little frog	00:00:45.426	00:00:47.106	00:00:01.680
5	it's collecting all these things?	00:00:53.400	00:00:54.460	00:00:01.060
6	it's recording	00:00:55.521	00:00:56.056	00:00:00.535
7	You have sung them, sing them again	00:00:59.412	00:01:00.942	00:00:01.530
8	You sing those two parts of the Myää mbwaa	00:01:01.120	00:01:02.462	00:00:01.342
9	You (P) ask him (Kp)	00:01:02.602	00:01:03.982	00:00:01.380
10	that thing, you sung those 2 parts of Myää, sing them again	00:01:05.591	00:01:07.931	00:00:02.340
11	the next one	00:01:32.024	00:01:32.694	00:00:00.670

Developer: Han Sloetjes, MPI Nijmegen

Tools (4): EXMARaLDA Editor

The screenshot displays the EXMARaLDA Partitur-Editor 1.3.4 interface. The main window shows a transcription table with columns for time points (1, 2, 3, 4) and rows for different transcription levels (X [v], X [nv1], X [nv2], X [nv3], Y [v]). The text in the table is as follows:

	1	2	3	4
X [v]	So it starts out with: A	roo ster crows	. ((1,3s)) ((takes breath 0,5s))	And then
X [nv1]	<i>rHA on rKN, iHA on iSH</i>	<i>rHA up and to the right</i>	<i>rHA stays up</i>	<i>rHA back down</i>
X [nv2]			<i>HE nods once</i>	
X [nv3]	<i>emphasizes the crow</i>			
Y [v]				

The interface also includes a menu bar (File, Edit, View, Tier, Event, Timeline, Format, Segmentation, Help), a toolbar with various icons, and a status bar at the bottom that says "Done.". An "Audio/Video panel" is open on the right, showing a video of two people sitting and talking. The panel includes playback controls (Start, Position, Stop) and a timeline with markers at 0.0, 3.2, 5.1, 43.8, and 43.8 sec. The status "Playback halted" is visible at the bottom of the panel.

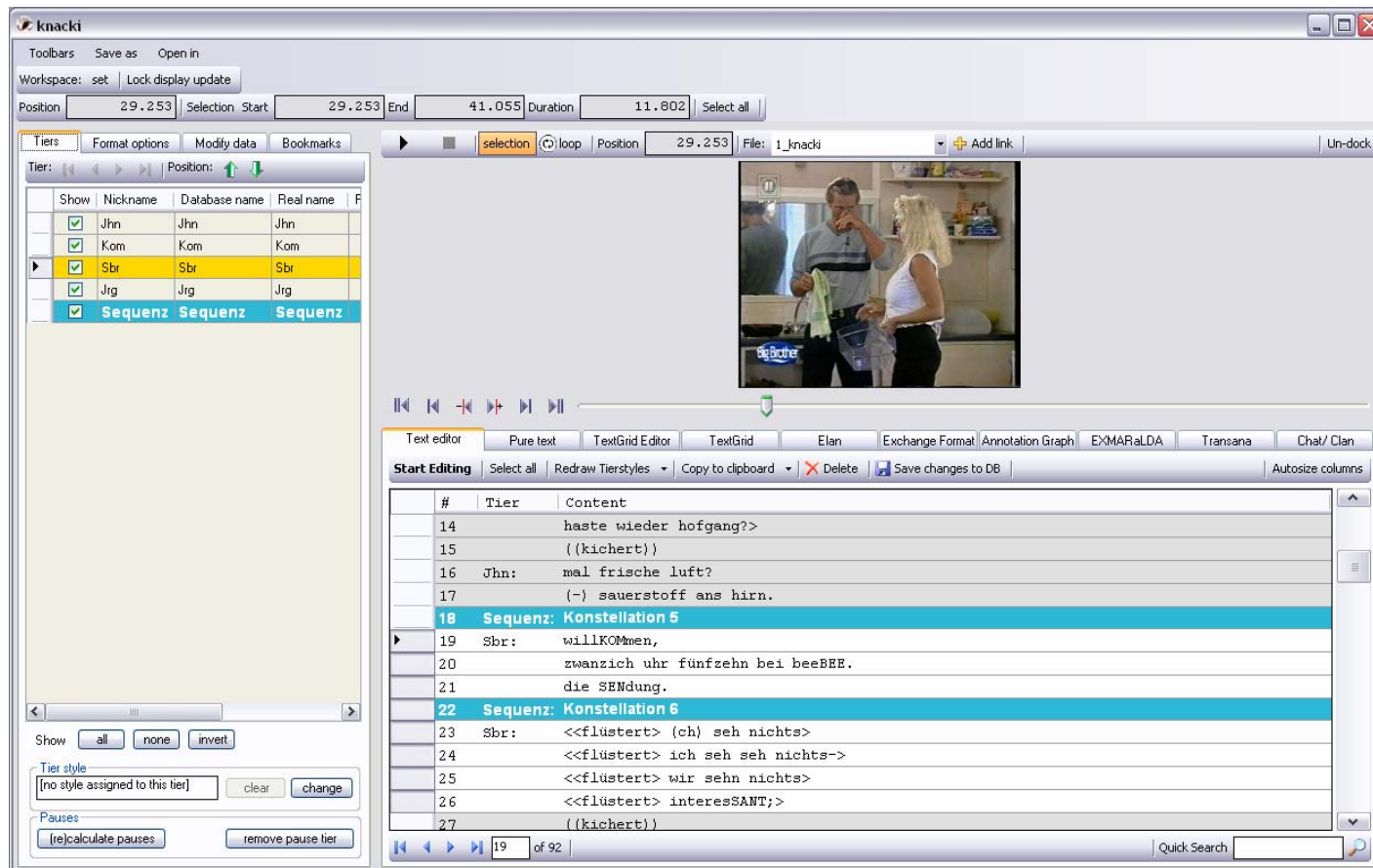
Developer: Thomas Schmidt, University of Hamburg

Tools (5): MacVisSTa



Developer: Travis Rose, Virginia Tech

Tools (6): Transformer

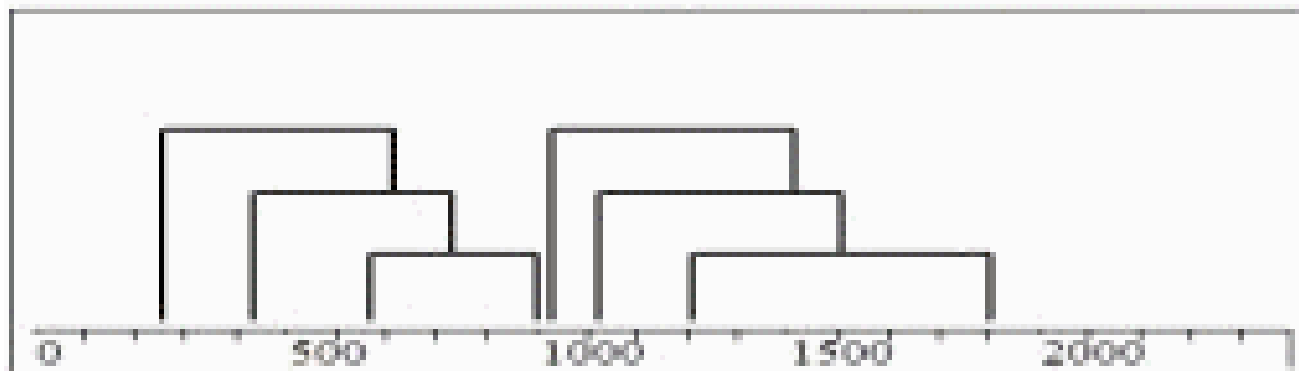
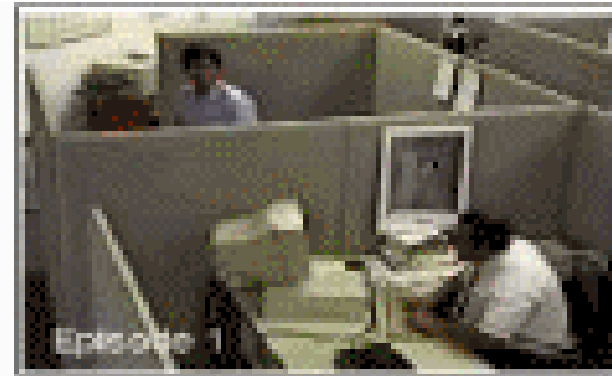
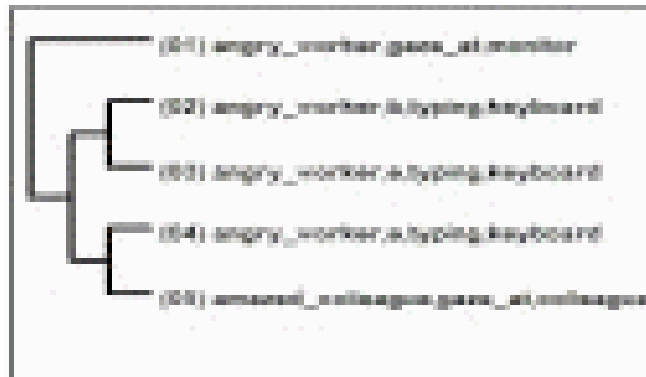


The screenshot displays the 'knacki' software interface. At the top, there are menu options like 'Toolbars', 'Save as', and 'Open in'. Below that, a workspace area shows 'set' and 'Lock display update'. A toolbar contains 'Position', '29.253', 'Selection', 'Start', '29.253', 'End', '41.055', 'Duration', '11.802', and 'Select all'. A 'Tiers' panel on the left lists participants: Jhn, Kom, Sbr, Jrg, and Sequenz. The main area features a video player showing a scene with two people. Below the video is a transcript table with columns for line number, tier, and content.

#	Tier	Content
14		haste wieder hofgang?>
15		{{kichert}}
16	Jhn:	mal frische luft?
17		(-) sauerstoff ans hirn.
18	Sequenz:	Konstellation 5
19	Sbr:	willKOMmen,
20		zwanzich uhr fünfzehn bei beeBEE.
21		die SENDung.
22	Sequenz:	Konstellation 6
23	Sbr:	<<flüstert> (ch) seh nichts>
24		<<flüstert> ich seh seh nichts->
25		<<flüstert> wir sehn nichts>
26		<<flüstert> interesSANT;>
27		{{kichert}}

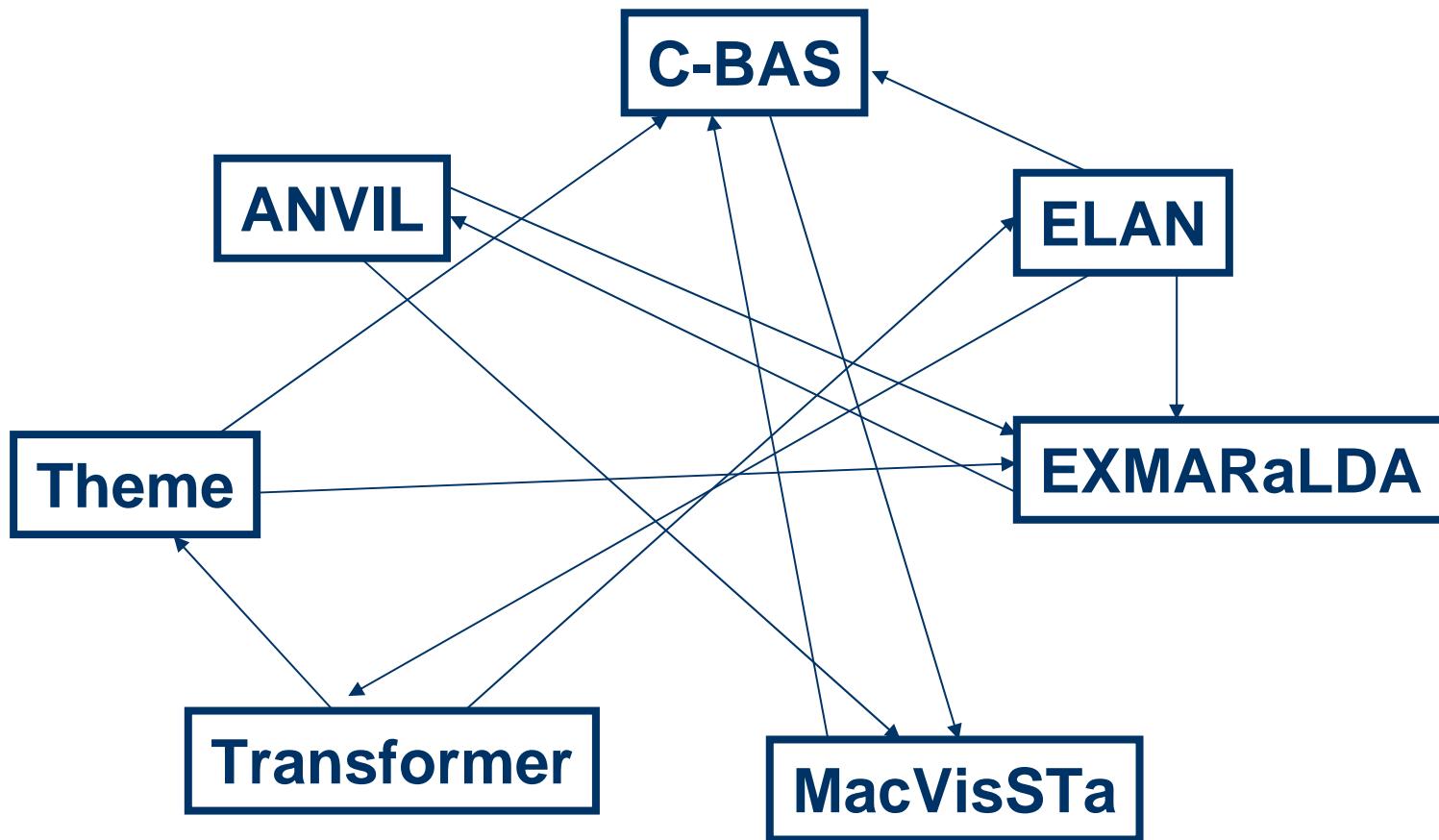
Developer: Oliver Ehmer, University of Freiburg

Tools (7): Theme

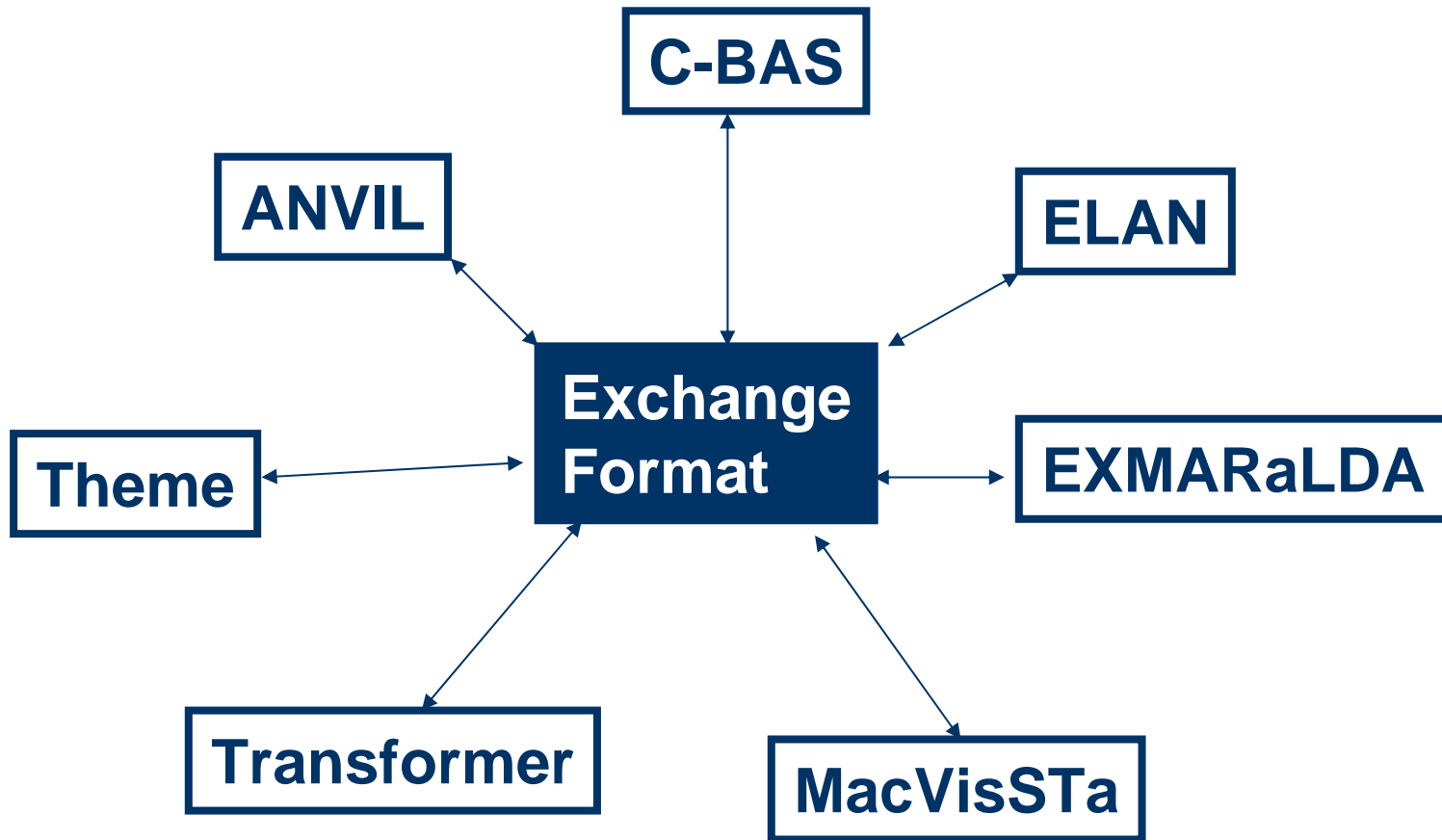


Developer: Magnus Magnusson, NOLDUS

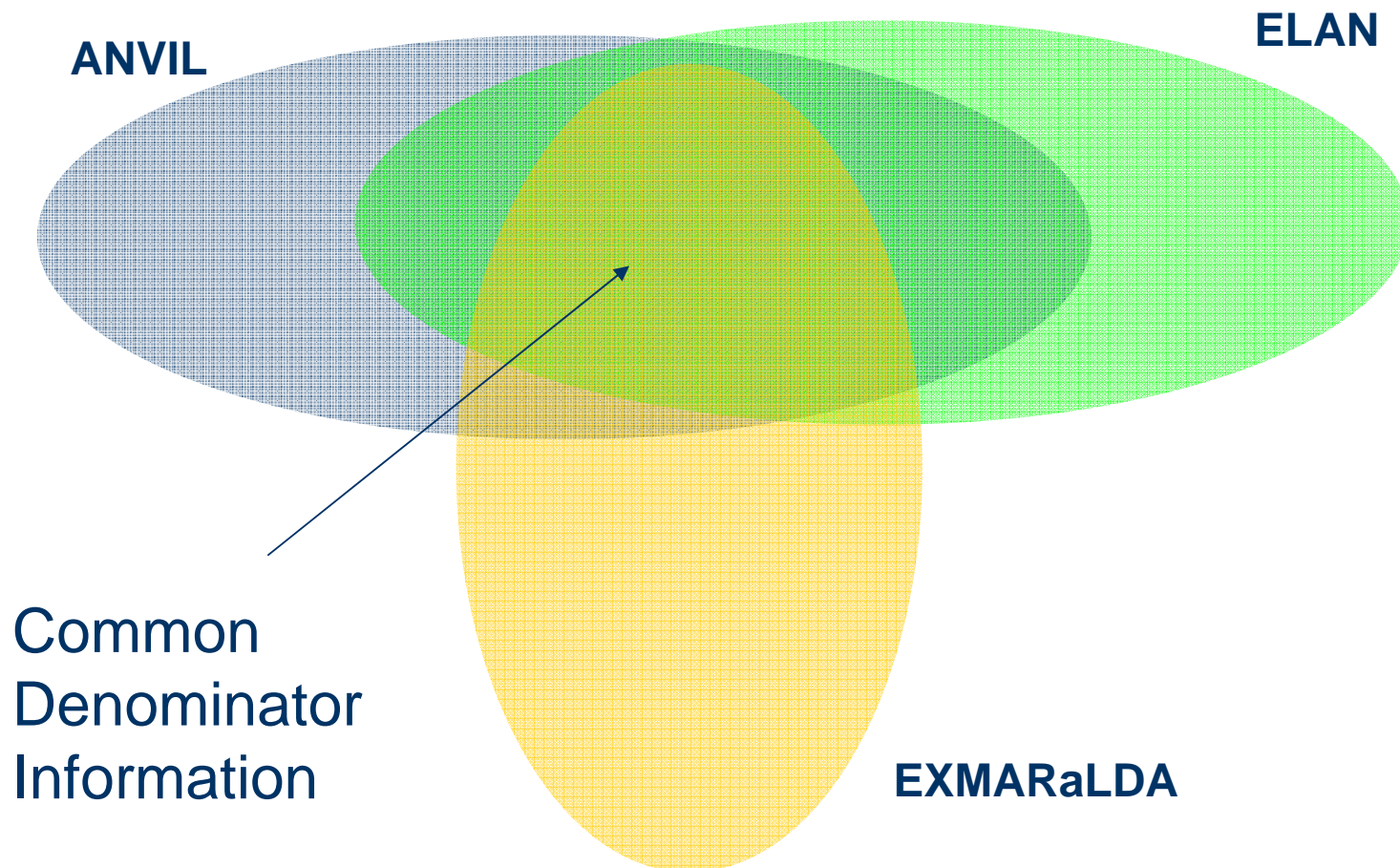
Interoperability



Interoperability



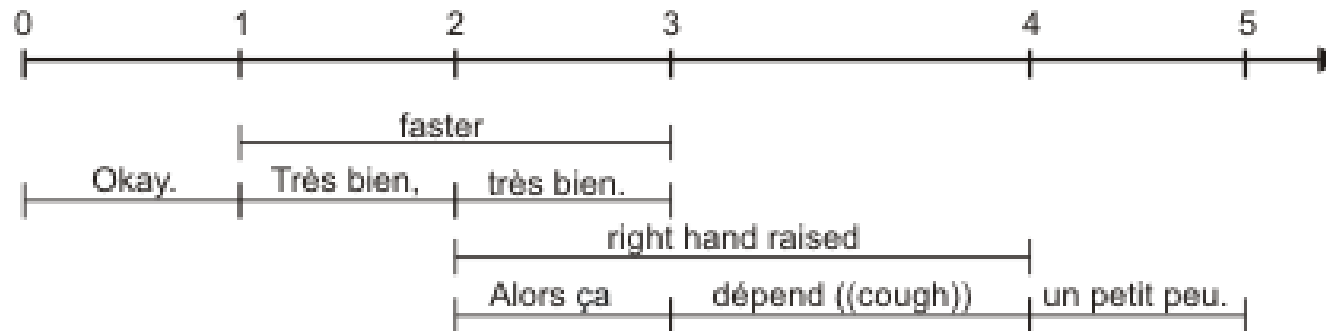
Data model comparison



Data model comparison

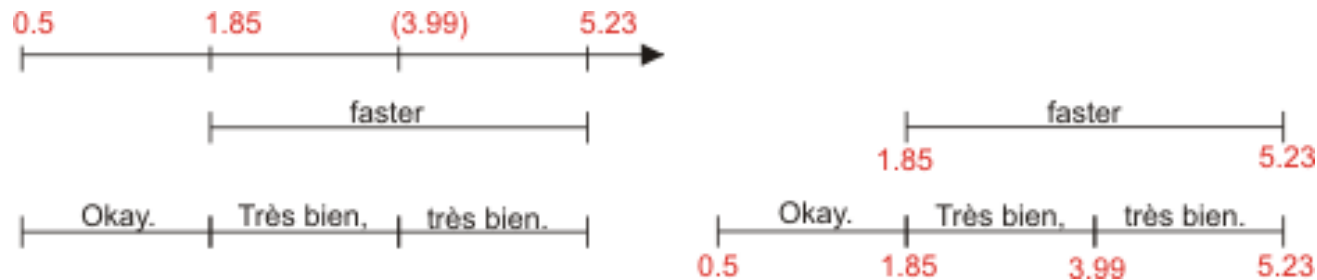
- Basic building blocks: Annotation tuples
 `<start, end, label(s)>`
 - Annotation Graphs as a general framework
 - AG's XML format as the base format
- Differences:
 - General organisation of basic building blocks into larger structural units
 - Semantic specifications and constraints on structural units

Tier-based vs. Non-tier-based



- Tiers = Partition of annotation tuples
 - No temporal overlap within a tier
 - In Anvil, ELAN, EXMARaLDA, Transformer
- Construct partition from other information (e.g. categorisation of labels)

Implicit vs. explicit timeline



- Implicit timeline: annotation tuples refer directly to media times
 - Explicit timeline: annotation tuples refer to points in a timeline which can refer to media times
 - Relative and absolute ordering of timepoints
 - Timepoints without timestamps possible
- Interpolate timepoints without timestamps
- Construct explicit timeline (identical timestamps?)

Tier specifications

- Tier names (all)
- Speaker assignment (ELAN, EXMARaLDA)
- Tier types:
 - ANVIL: primary, singleton, span
 - ELAN/Transformer: time subdivision, included in, symbolic subdivision, symbolic association
 - EXMARaLDA: transcription, description, annotation

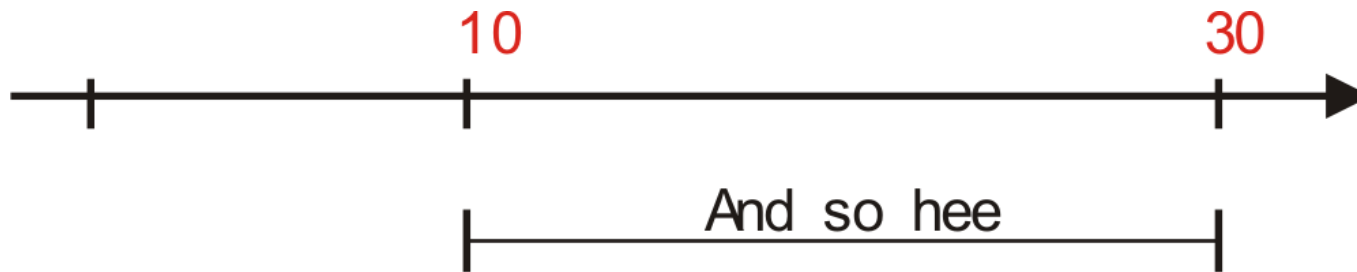
Tier relations and constraints

- Parent/Child relations → tier hierarchy
 - explicit in Anvil and ELAN
 - implicit in EXMARaLDA
- Other constraints arising from tier typing
- Restrictions on label content
 - part of the tools' format?

Exchange Format

- Lossless exchange of common denominator information
 - Uniformly encode all information beyond the common denominator
- no lossless round-tripping, but...
- ... all available information captured and...
- ... lossless exchange in a chain of tools with increasingly complex data formats

Exchange Format



```
<Anchor id="T6" offset="10" unit="milliseconds"/>
<Anchor id="T7" offset="30" unit="milliseconds"/>
[...]
<Annotation type="TIE1" start="T6" end="T7">
  <Feature name="description">
    And so hee
  </Feature>
</Annotation>
```

Exchange Format

```
<MetadataElement name="Tier">
  <MetadataElement name="TierIdentifier">
    TIE1
  </MetadataElement>
</MetadataElement>
[...]
```

```
<Annotation type="TIE1" start=" T6" end=" T7">
```

```
<MetadataElement name="Tier">
  [...]
  <MetadataElement name="TierAttribute">
    <MetadataElement name="Source">
      EXMARaLDA
    </MetadataElement>
    <MetadataElement name="Name">
      speaker
    </MetadataElement>
    <MetadataElement name="Value">
      SPK0
    </MetadataElement>
  </MetadataElement>
</MetadataElement>
[...]
```

Tier definition:
Fixed metadata attribute
,TierIdentifier‘

Tier properties:
Fixed metadata triple
,Source‘
,Name‘
,Value‘

Implementation

- Import / Export routines
 - ANVIL, ELAN: AGLib (Java port)
 - EXMARaLDA: XSLT stylesheets
 - Theme: Perl
 - MacVisSTa: Python
 - Transformer: Visual Basic

Conclusion



Results:

- Commonalities captured
- Differences (better) understood
- Basic interoperability established
- Link to generic framework established



Specific

Abstract

Outlook

- Partial correspondences
 - Simple: e.g. speaker assignments
 - Complex: e.g. parent/child relations
- Modifying/Assimilating tools' formats
- „Process-based“ (as opposed to format-based) interoperability?

