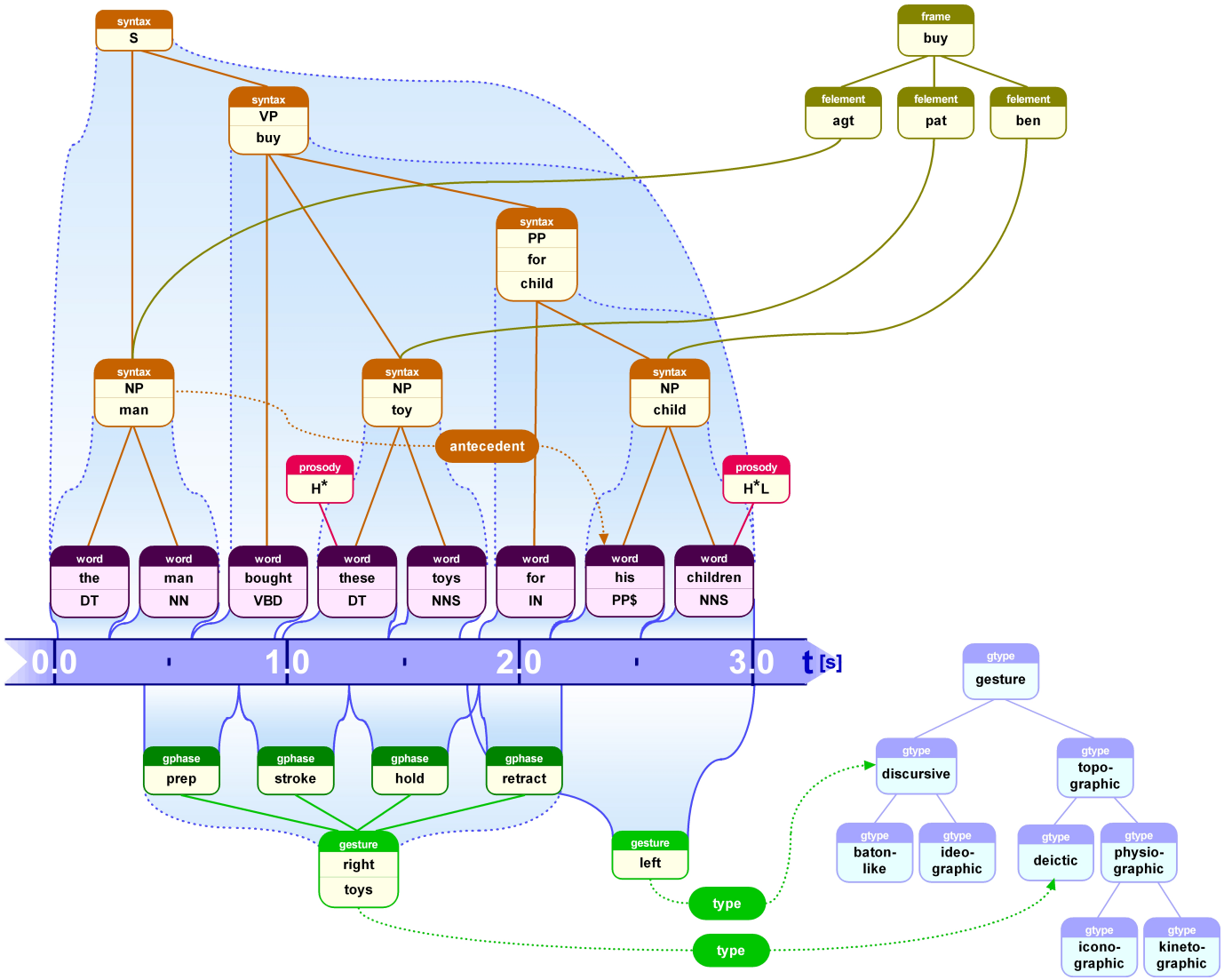
The NITE XML Toolkit

Jonathan Kilgour and Jean Carletta

University of Edinburgh

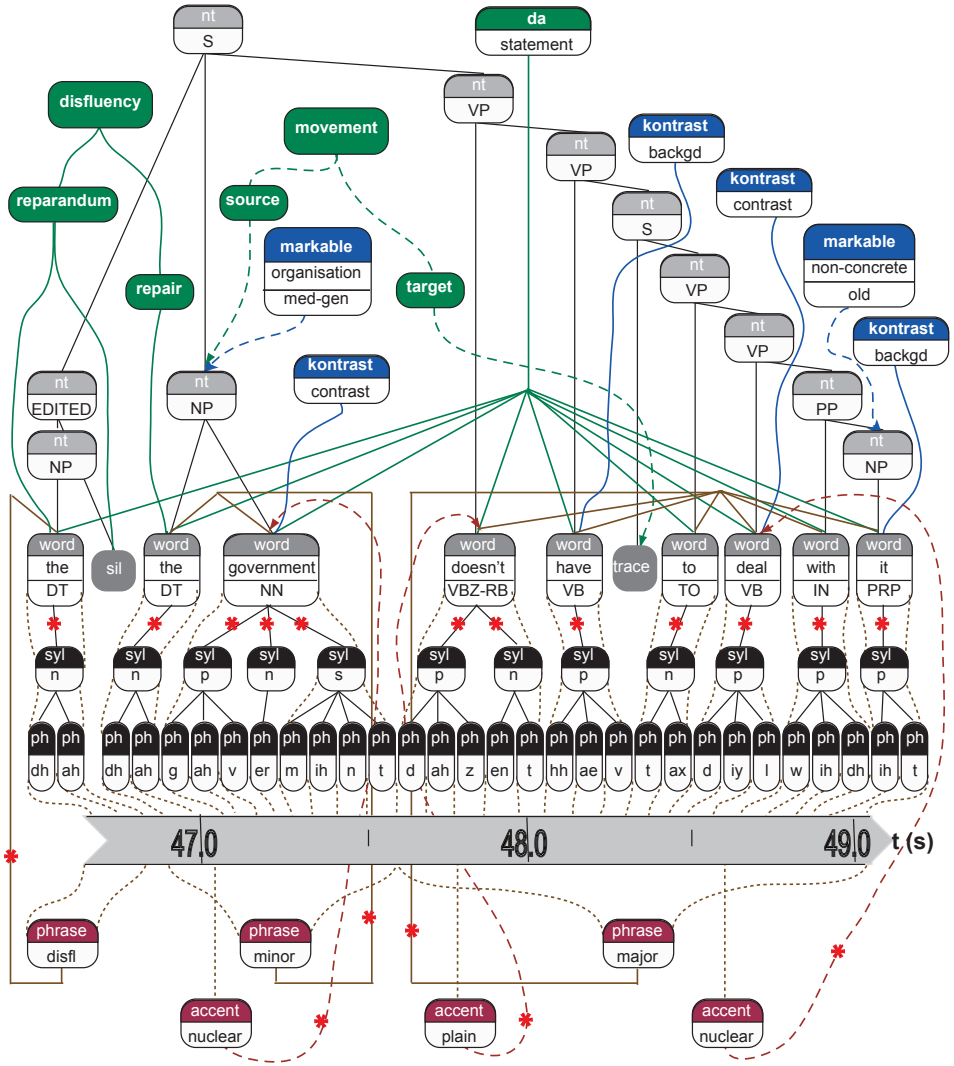
Dialogue Interest Group
Dec 2009

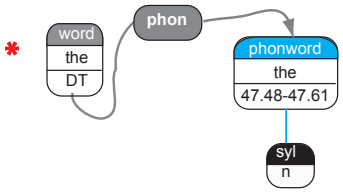
A toy example of linguistic data

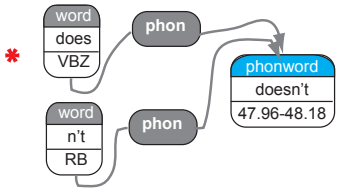


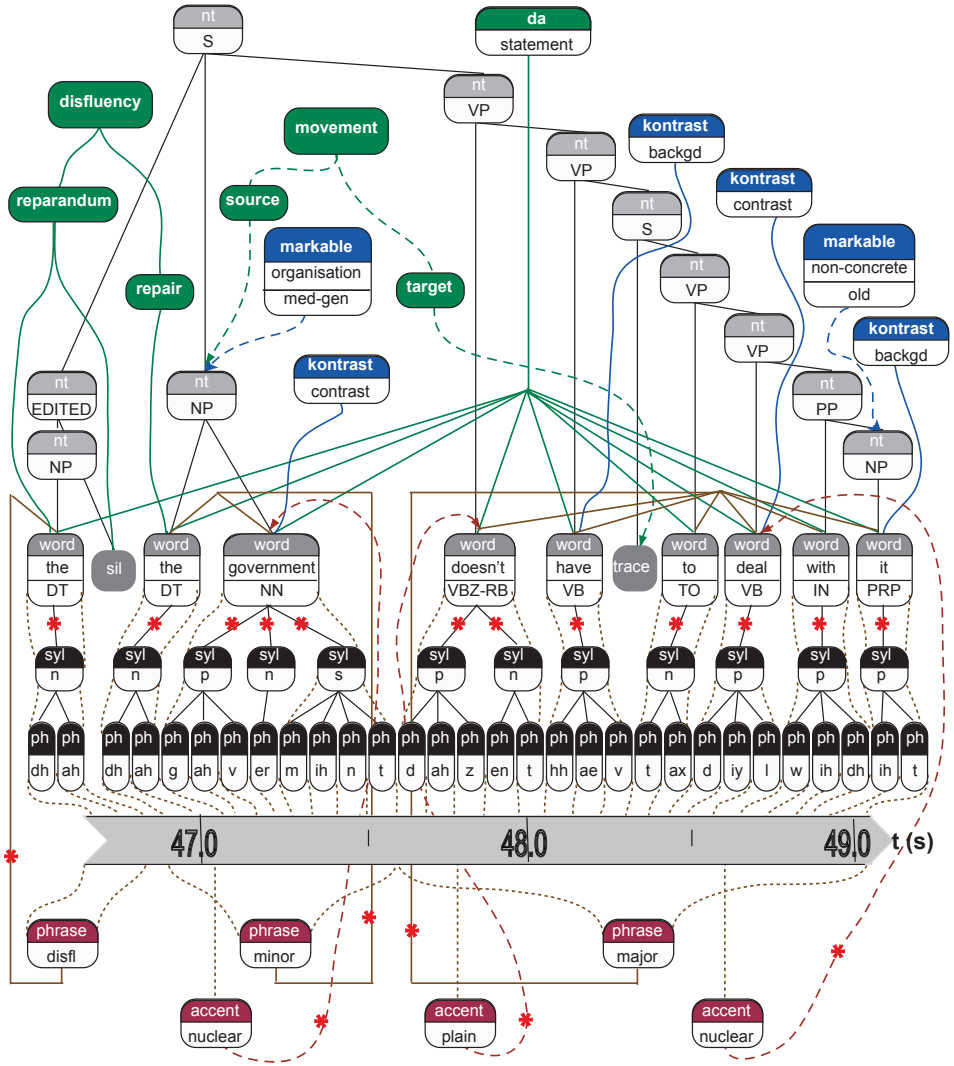
NITE XML Toolkit

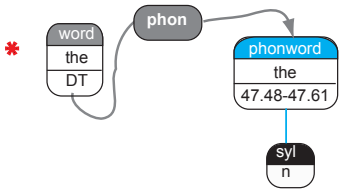
- Open source toolkit for handling annotations with temporal ordering and full structural relations
- Data storage format designed to support distributed corpus development
  - Libraries for data handling, query, and writing graphical user interfaces
  - Configurable end user browsing and annotation tools for common tasks
  - Command line utilities for analysis, feature extraction

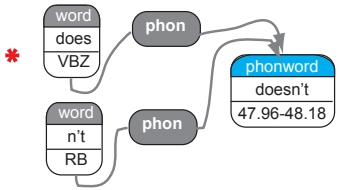

nt S
da statement
nt VP
disfluency
movement
kontrast backgd
nt VP
reparandum
source
kontrast contrast
nt S
markable organisation med-gen
markable non-concrete old
nt VP
repair
target
nt VP
kontrast backgd
nt EDITED
nt NP
kontrast contrast
nt NP
nt PP
nt NP
word the DT
sil
word the DT
word government NN
word doesn't VBZ-RB
word have VB
trace
word to TO
word deal VB
word with IN
word it PRP
syl n
syl n
syl p
syl n
syl s
syl p
syl n
syl p
syl n
syl p
syl p
syl p
ph dh
ph ah
ph dh
ph ah
ph g
ph ah
ph v
ph er
ph m
ph ih
ph n
ph t
ph d
ph ah
ph z
ph en
ph t
ph hh
ph ae
ph v
ph t
ph ax
ph d
ph iy
ph l
ph w
ph ih
ph dh
ph ih
ph t
47.0
48.0
49.0
t (s)
phrase disfl
phrase minor
phrase major
accent nuclear
accent plain
accent nuclear


word the DT
phon
phonword the 47.48-47.61
syl n


word does VBZ
phon
phonword doesn't 47.96-48.18
word n't RB
phon


nt S
da statement
nt VP
disfluency
movement
kontrast backgd
nt VP
reparandum
source
kontrast contrast
nt S
markable organisation med-gen
markable non-concrete old
nt VP
repair
target
nt VP
kontrast backgd
nt EDITED
nt NP
kontrast contrast
nt NP
nt PP
nt NP
word the DT
sil
word the DT
word government NN
word doesn't VBZ-RB
word have VB
trace
word to TO
word deal VB
word with IN
word it PRP
syl n
syl n
syl p
syl n
syl s
syl p
syl n
syl p
syl n
syl p
syl p
syl p
ph dh
ph ah
ph dh
ph ah
ph g
ph ah
ph v
ph er
ph m
ph ih
ph n
ph t
ph d
ph ah
ph z
ph en
ph t
ph hh
ph ae
ph v
ph t
ph ax
ph d
ph iy
ph l
ph w
ph ih
ph dh
ph ih
ph t
47.0
48.0
49.0
t (s)
phrase disfl
phrase minor
phrase major
accent nuclear
accent plain
accent nuclear


word the DT
phon
phonword the 47.48-47.61
syl n


word does VBZ
phon
phonword doesn't 47.96-48.18
word n't RB
phon

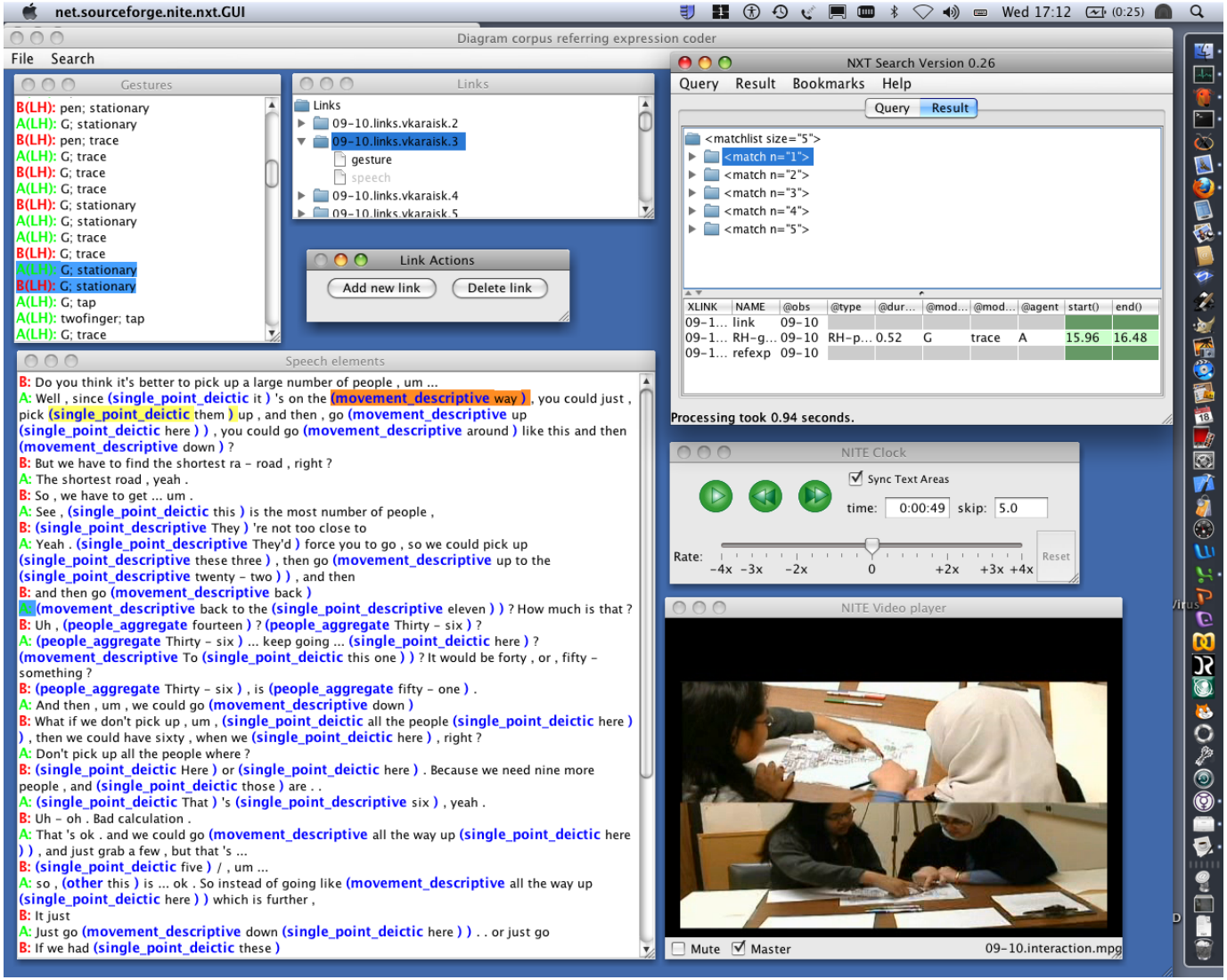
Community support

- stand-off annotation using multiple files under version control
- dependency structure for keeping track of which annotations rely on which versions of which other annotations
- multiple competing annotations for the same thing (different humans for a reliability assessment, different automatic processes for a competition)
- logical query language - because this is the only way to analyse this kind of data

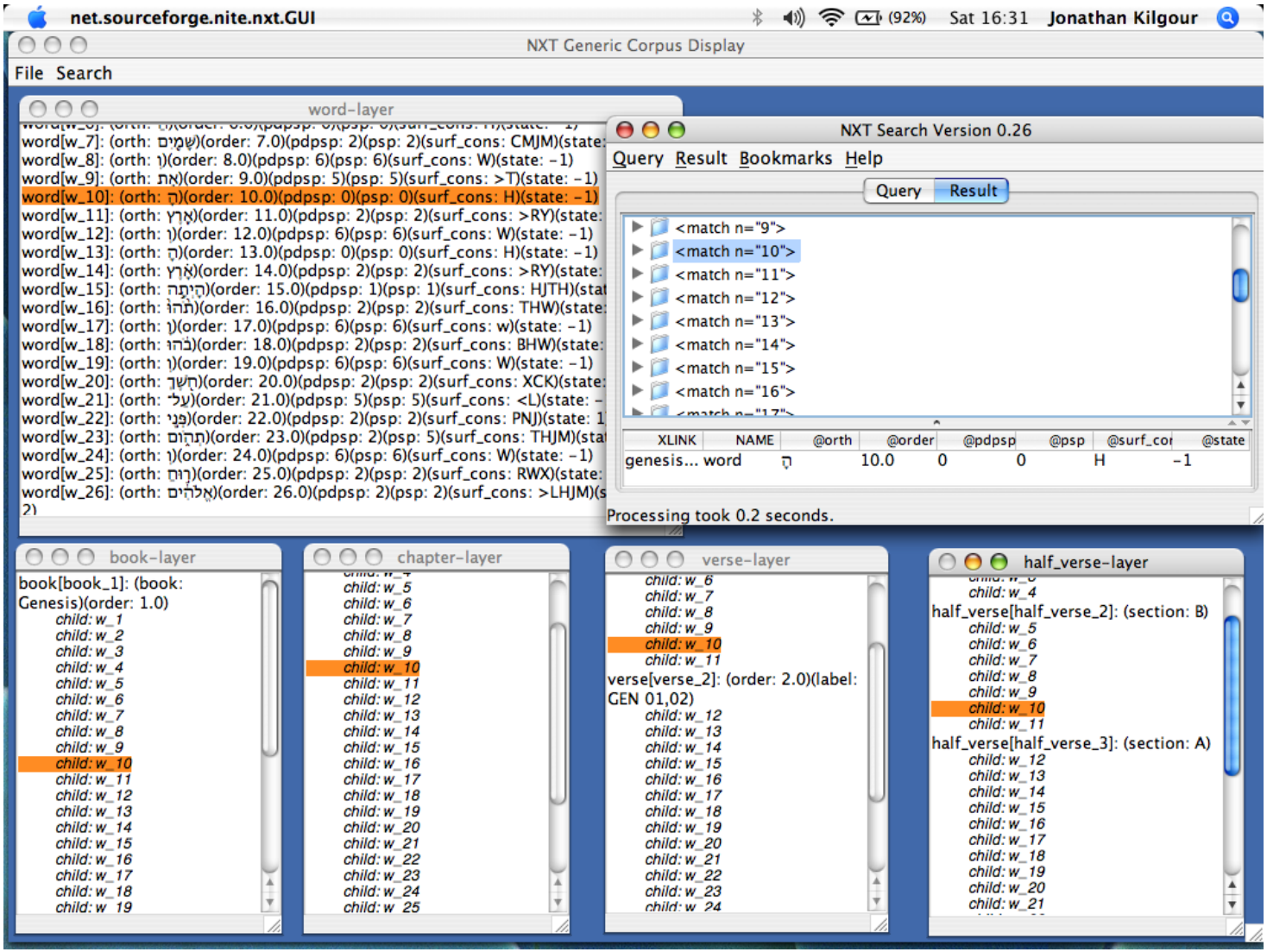
What's wrong with NXT

- Flexibility makes it harder to just start using it
  - need to formally describe corpus structure
  - some users struggle with logic
- no indexing locations within still images or video frames
- Not enough packaging (connection to automatic tools, authoring corpus structure description)
- Not "sold" enough, not known very well in America

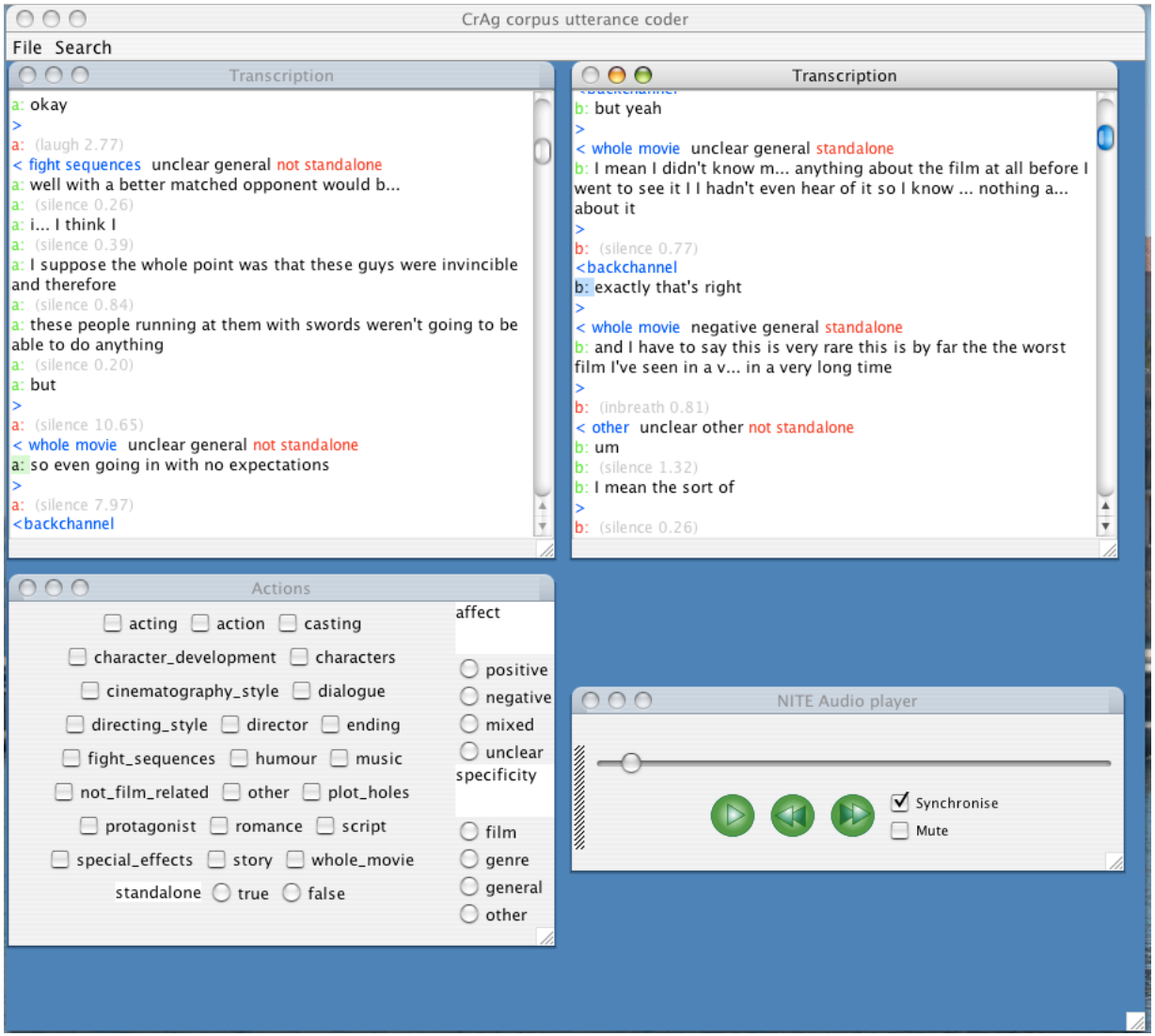
Butterflies: deixis



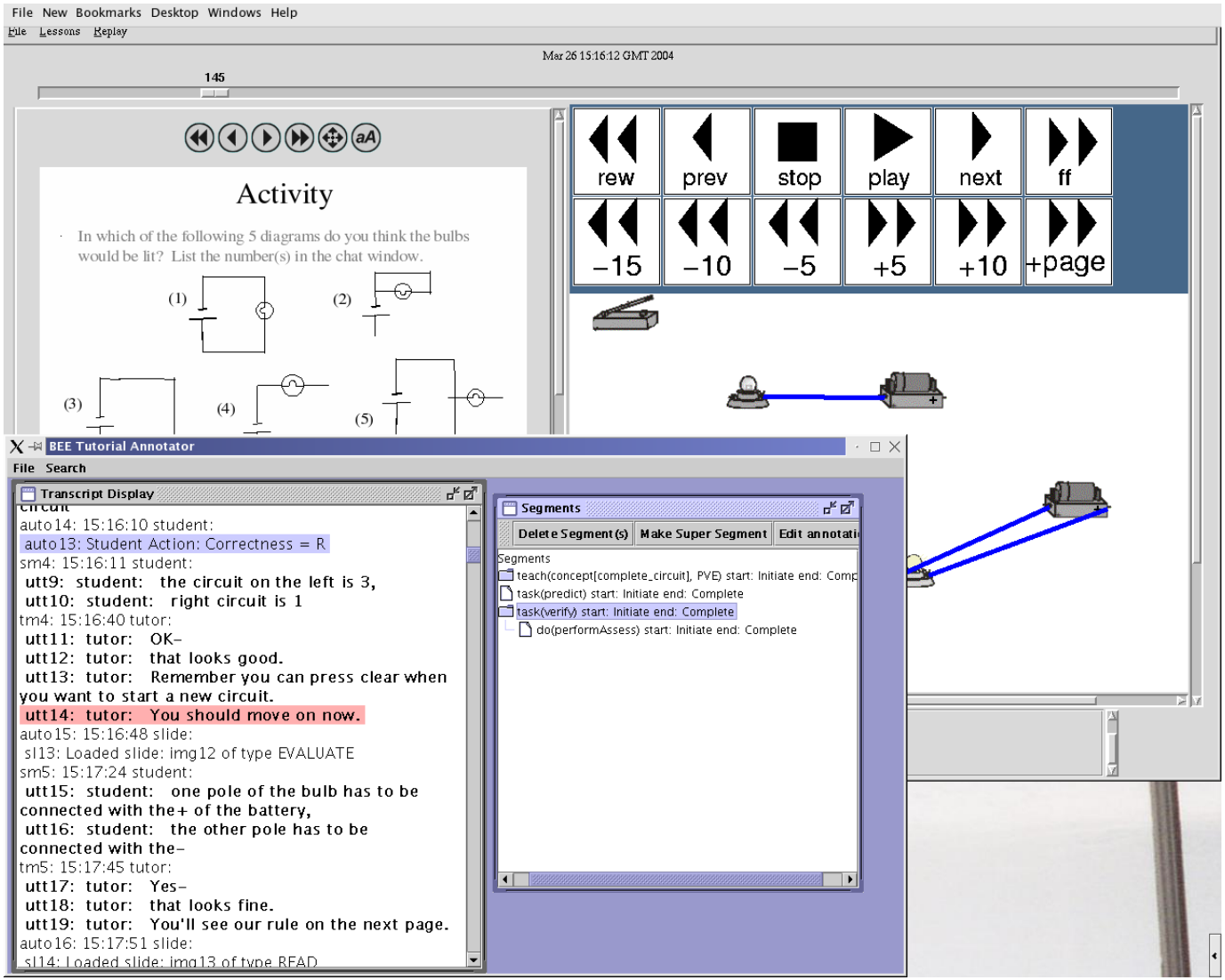
Butterflies: Bible studies



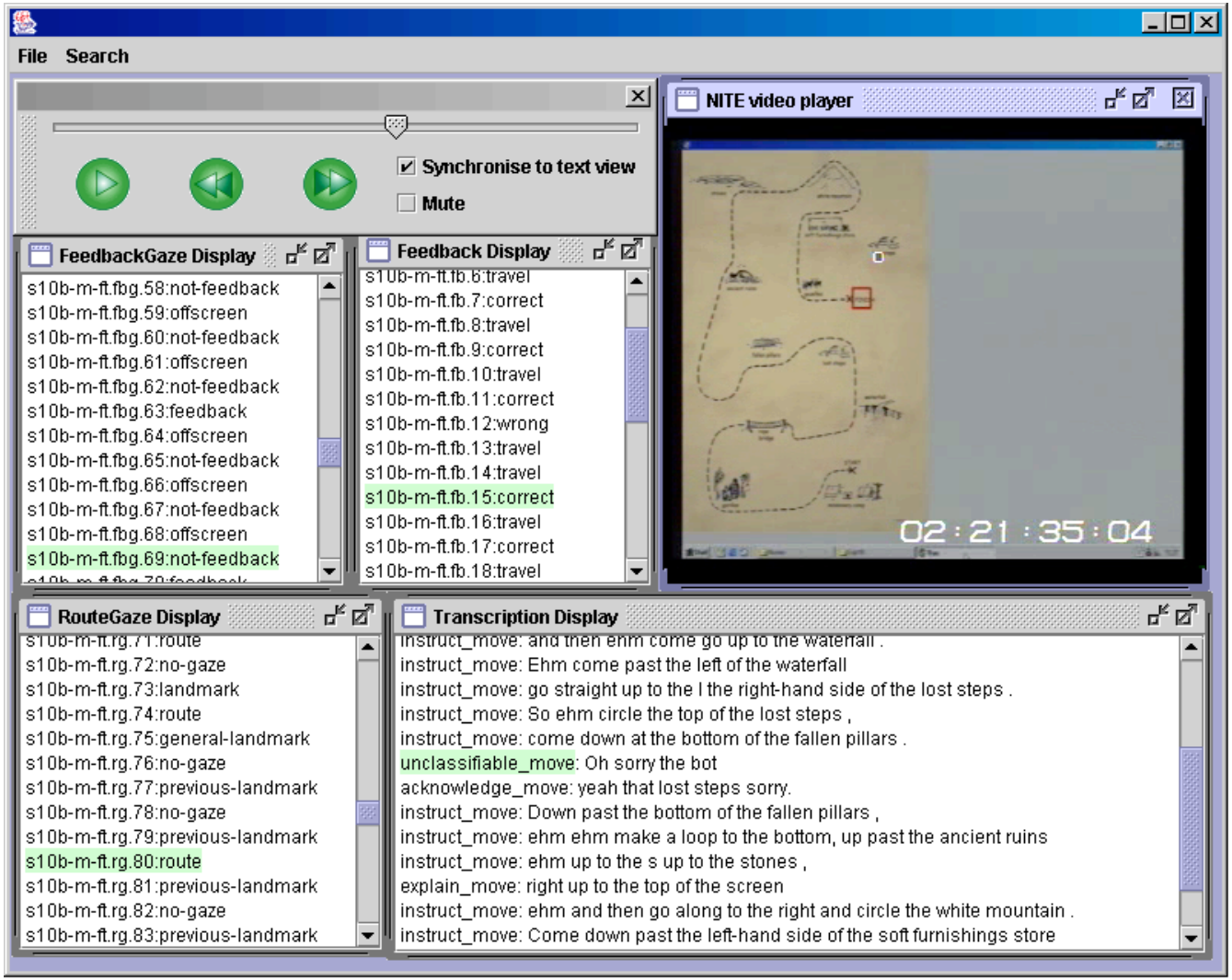
Butterflies: movie review analysis



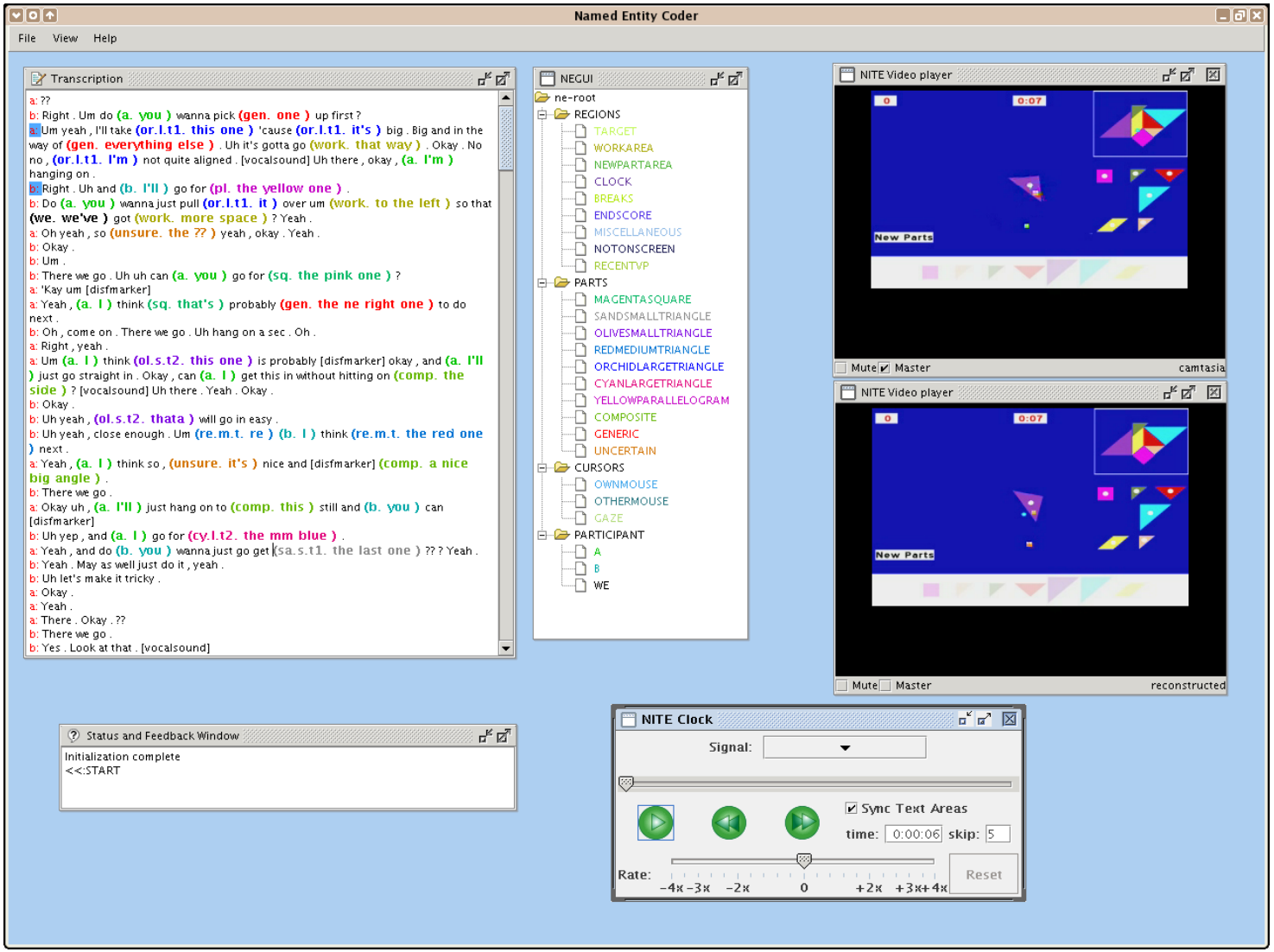
Butterflies: dialogue system strategy



Butterflies: eyetracking



Butterflies: eyetracking



Flock of birds

Google Earth mashup