

Unifying Corpus Annotations

**Final Report of WG4
TILR**

July 13-15, 2007
Stanford University

Goals of WG

- Come up with a formal specification of the prerequisites and scenarios for mappings
- GOLD: the question as to its role in this mapping.

The Role of Ontologies

1. Consistency and Validation
2. Ontology as part of Annotation
3. Links between annotations
4. Ontology as interlingua:
5. Deeper inference
6. Use promotes use.

Consistency and Validation

- Reference to ontological types provides a platform for checking consistency across multiple annotations of the same corpus.
 - Typed feature structures allow this between markups already.
 - E.g., comparing classic TreeBank annotation with CCG-based CCGBank.

Linking Annotations

- Ontologies help provide integration between annotation levels.
 - PDTB argument typing uses Asher's (1993) Abstract Objects: event, proposition, fact, projected-event
 - Because the party lasted all night, John had a headache.
 - Co-reference annotation can use the same typology
 - It's also why Mary's not here.

Ontology as Interlingua

- Terminology mapping:
 - facilitating linking one tag set to another.
- Taxonomy linking:
 - Node to node

Deeper Inference

- Ontology provides the platform with which or from which to do deeper inference.
 - Text \longrightarrow TimeML \longrightarrow OWL-Time

Use promotes use

- Implementing ontologies in the mapping of annotations actually promotes ontologies and their adoption in the community.

Mapping Scenarios

1. Integrating TimeML and SpatialML. Getting more from integrating two annotations to get a new representation or information structure.
2. Learning more from one annotation to another. Parallel corpora and alignment.
3. Lexicon induction from parallel texts and other examples.
4. Dynamic corpus development and the way ontology can help check consistency

Merging Annotations: Time

The car pulled away from the baggage-claim area and headed south at the first intersection.

- *The car <EVENT id="e1" tense="past" start="t1" end="t2">pulled</EVENT> away from the baggage-claim area and <EVENT id="e2" tense="past" start="t3" end="t4">headed</EVENT> south at the first intersection.*
- *<TLINK id="3" eventID="e1" relatedToeventID="e2" relType="BEFORE"/>*

Merging Annotations: Space

The car pulled away from the baggage-claim area and headed south at the first intersection.

- *The `<PLACE id="pl1" type="VEHICLE">car</PLACE>` pulled away from the baggage-claim `<PLACE id="pl2" form="NOM" type="RGN">area</PLACE>` and headed `<SIGNAL id="s1">south</SIGNAL>` at the first `<PLACE id="pl3" form="NOM" type="ROAD" latLong="40.660465 -73.830276">intersection</PLACE>`.*
- *`<PATH id="p1" source="pl2" />`
`<PATH id="p2" source="pl3" direction="south"/>`*

Merging Annotations: Motion

The car pulled away from the baggage-claim area and headed south at the first intersection.

- The `<PLACE type="VEHICLE" id="pl1">car</PLACE >`
`<EVENT id="e1" tense="past" start="t1" end="t2">pulled</EVENT>`
away from the baggage-claim `<PLACE id="pl2" form="NOM"`
`type="RGN">area</PLACE>` and `<EVENT id="e2" tense="past"`
`start="t3" end="t4">headed</EVENT>`
`<SIGNAL id="s1">south</SIGNAL>` at the first `<PLACE id="pl3"`
`form="NOM" type="ROAD">intersection</PLACE>`.
- `<TLINK id="3" event="e1" relatedToEvent="e2" relType="BEFORE"/>`
`<PATH id="p1" source="pl2" />`
`<PATH id="p2" source="pl3" direction="south"/>`

Merging Annotations: Trajectory

The car pulled away from the baggage-claim area and headed south at the first intersection.

- ```
<Trajectory-Element id="te1">
 <EVENT id="e1" begin="t1" end="t2"/>
 <TrajectoryLink timepoint1="t1" timespoint2="t2"
relType="BEFORE"/>
 <LINK id="l1" source="pl1" linkType="IN" target="pl2"/>
 <LINK id="l2" source="pl1" linkType="IN" target="UNKNOWN"/>
 <TimeStamp linkID="l1" timespoint="t1"/>
 <TimeStamp linkID="l2" timespoint="t2"/>
</Trajectory-Element>
```

# Best Practices

- One uniform data structure for the target corpus.
- Assuming ideally a stand-off annotation strategy
- One annotation is one layer over the data
- Merging two annotation types is:
  - Mapping from one layer to another
  - Define mapping rules or an interlingua for different levels

# WG Action Items

- Complete list of best practices for integrating annotations.
- Map particular annotation schemas to GOLD.