# Annotation Standards

Marc Verhagen

Natural Language Annotation for ML

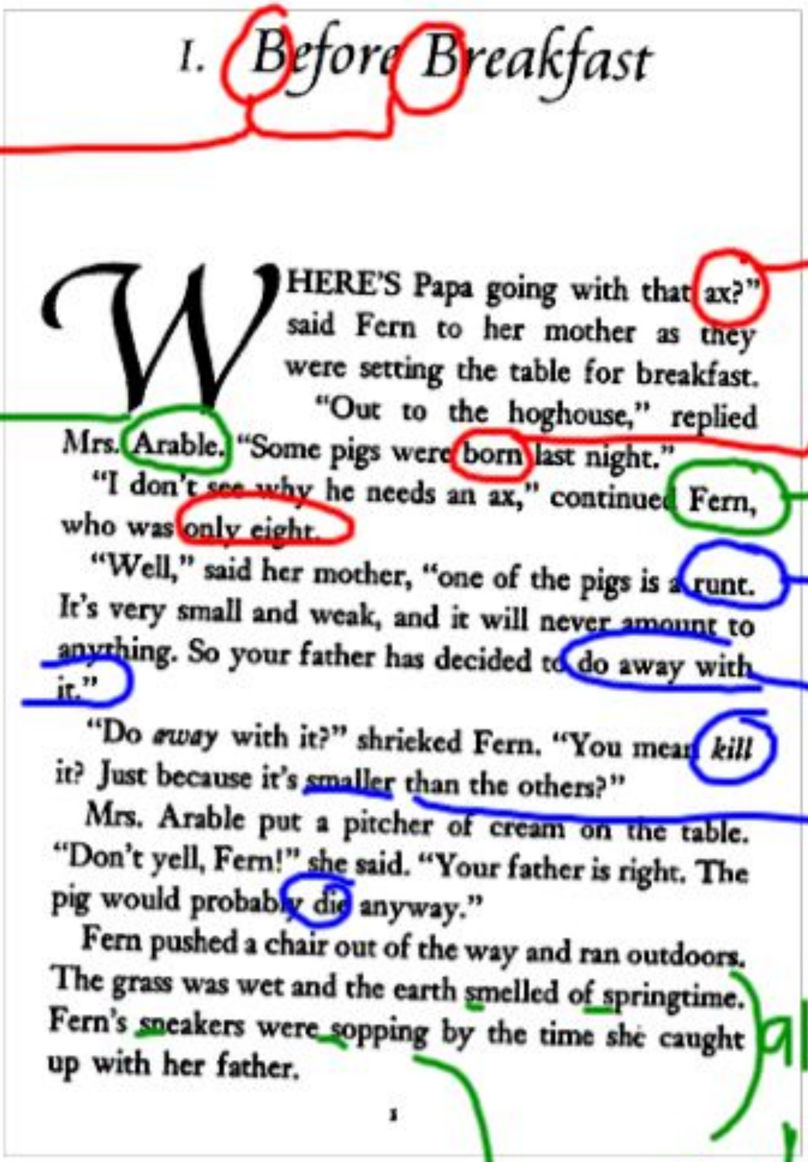
CS 216,  Spring 2016

# Annotation Standards

- Annotation and Annotation Tools
- Standards (why?)
- Linguistic Annotation Framework
- More
  - MASC corpus
  - LAPPS Grid: WSEV and LIF

alliteration

plowable
fertile



## 1. Before Breakfast

"WHERE'S Papa going with that ax?" said Fern to her mother as they were setting the table for breakfast.

"Out to the hoghouse," replied Mrs. Arable. "Some pigs were born last night."

"I don't see why he needs an ax," continued Fern, who was only eight.

"Well," said her mother, "one of the pigs is a runt. It's very small and weak, and it will never amount to anything. So your father has decided to do away with it."

"Do *away* with it?" shrieked Fern. "You mean *kill* it? Just because it's smaller than the others?"

Mrs. Arable put a pitcher of cream on the table. "Don't yell, Fern!" she said. "Your father is right. The pig would probably die anyway."

Fern pushed a chair out of the way and ran outdoors. The grass was wet and the earth smelled of springtime. Fern's sneakers were sopping by the time she caught up with her father.

1

L/D

plant

last born

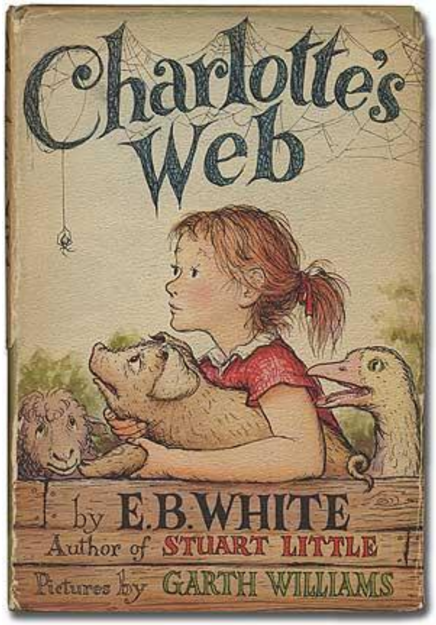
kill it L/D

small but
mighty

alliteration

5 S's

circle of
Life



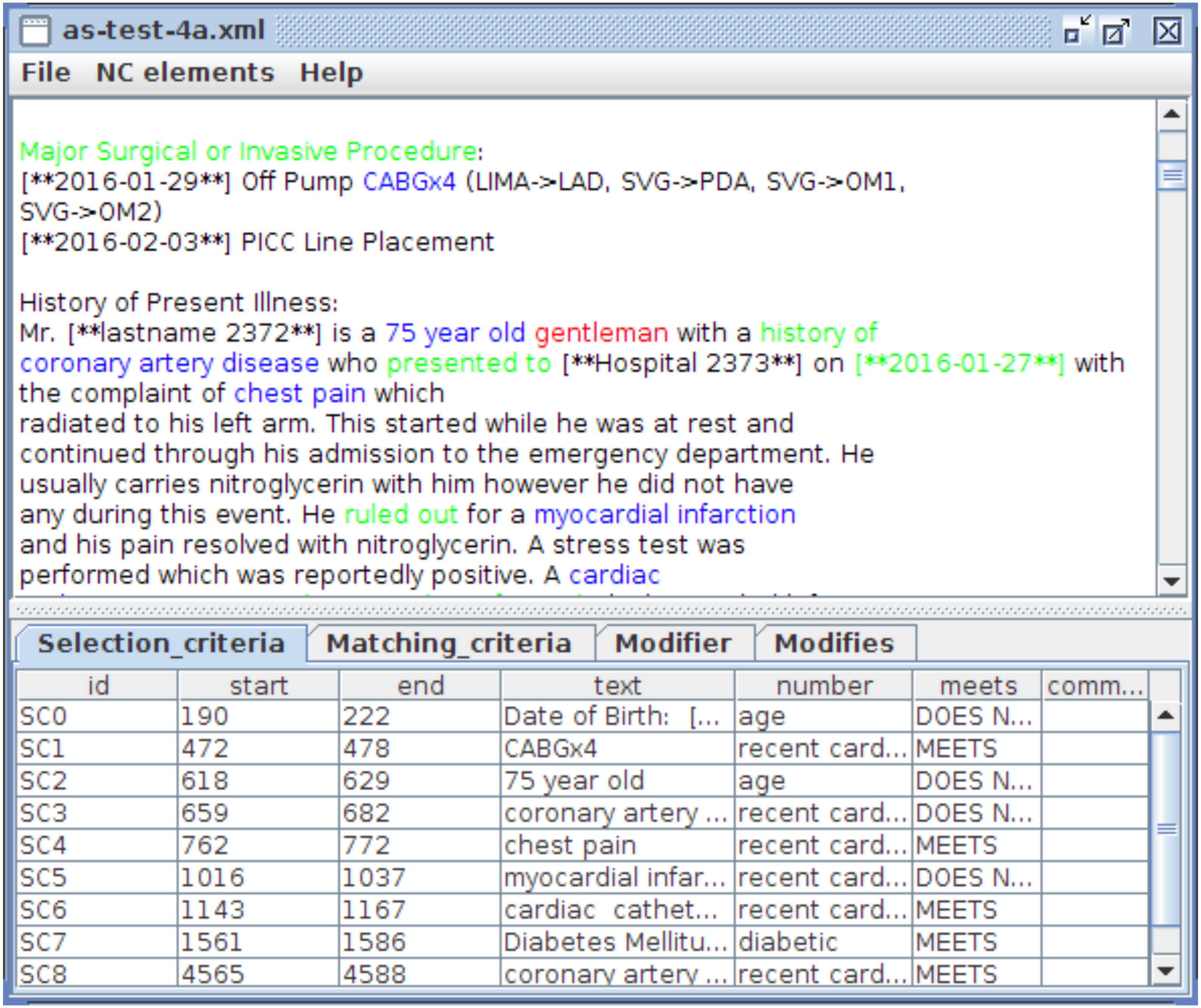
In Washington **<TIMEX3 tid="t1" TYPE="DATE" VAL="PRESENT_REF" temporalFunction="true" valueFromFunction="tf1" anchorTimeID="t0">**today**</TIMEX3>**, the Federal Aviation Administration **<EVENT eid="e1" class="OCCURRENCE">**released**</EVENT>** air traffic control tapes from the night the TWA Flight eight hundred **<EVENT eid="e2" class="OCCURRENCE">**went**</EVENT>** down. There's nothing new on why the plane **<EVENT eid="e3" class="OCCURRENCE">**exploded**</EVENT>**, but you **<EVENT eid="e4" class="OCCURRENCE">**cannot**</EVENT>** **<EVENT eid="e5" class="OCCURRENCE">**miss**</EVENT>** the moment.
ABC's Lisa Stark **<EVENT eid="e6" class="OCCURRENCE">**has**</EVENT>** more.

**MARKABLES**

**<MAKEINSTANCE eventID="e1" pos="VERB" eiid="ei1" tense="PAST" aspect="NONE"/>**
**<MAKEINSTANCE eventID="e2" pos="VERB" eiid="ei2" tense="PAST" aspect="NONE"/>**
**<MAKEINSTANCE eventID="e3" pos="VERB" eiid="ei3" tense="PAST" aspect="NONE"/>**
**<MAKEINSTANCE eventID="e4" pos="VERB" eiid="ei4" tense="PRESENT" aspect="NONE"/>**
**<MAKEINSTANCE eventID="e5" pos="VERB" eiid="ei5" tense="INFINITIVE" aspect="NONE"/>**
**<MAKEINSTANCE eventID="e6" pos="NONE" eiid="ei6" tense="PRESENT" aspect="NONE"/>**

**<TLINK eventInstanceID="ei1" relatedToTime="t1" relType="IS_INCLUDED" rule="2-1" />**
**<TLINK eventInstanceID="ei2" relatedToTime="t1" relType="IS_INCLUDED" rule="2-1" />**
**<TLINK eventInstanceID="ei1" relatedToEventInstance="ei3" relType="BEFORE" rule="3-19" />**
**<TLINK eventInstanceID="ei3" relatedToEventInstance="ei4" relType="BEFORE" rule="6-1" />**
**<TLINK eventInstanceID="ei3" relatedToEventInstance="ei6" relType="BEFORE" rule="3-23" />**

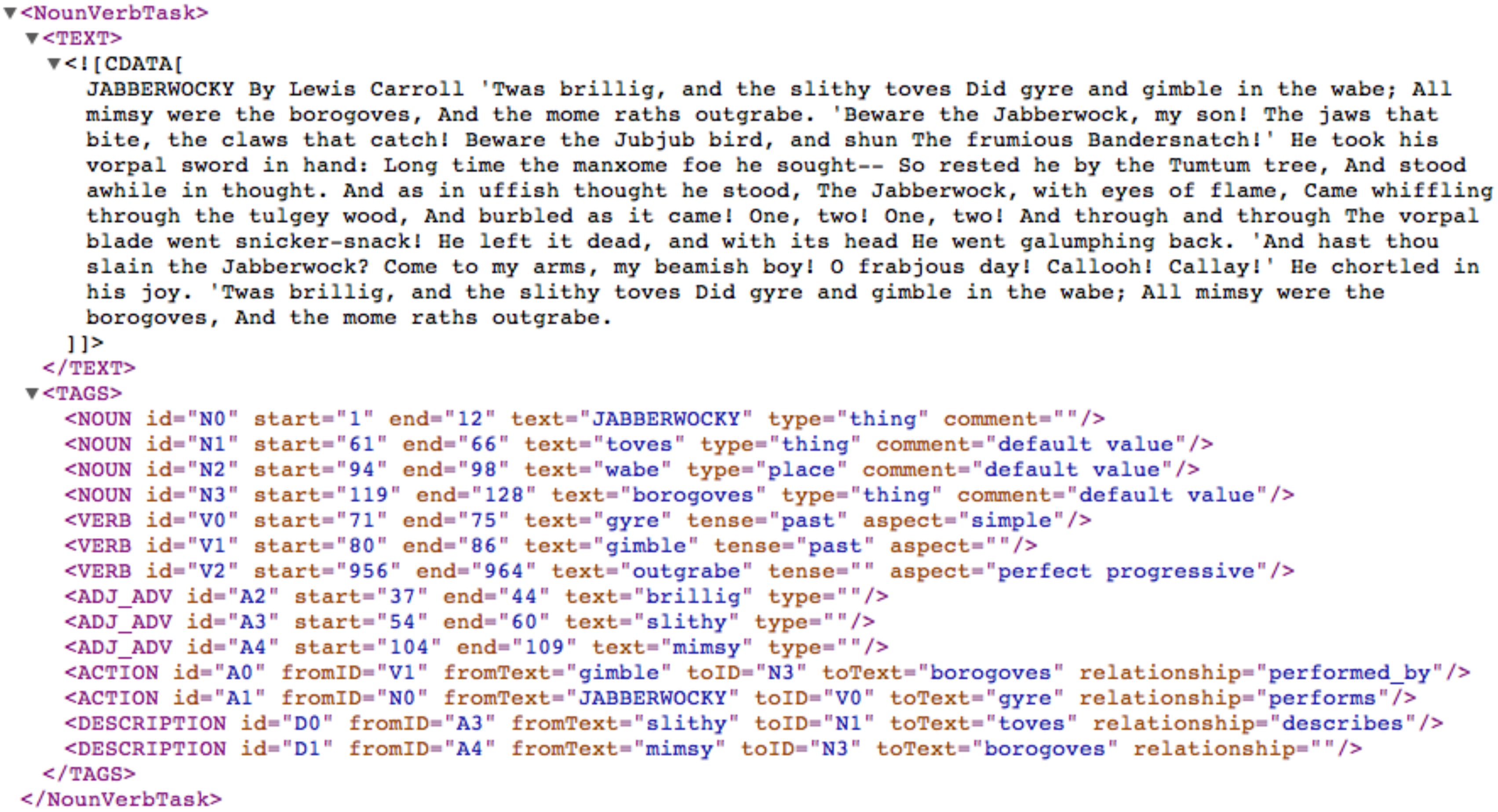
**NON-MARKABLES (EMPTY TAGS)**



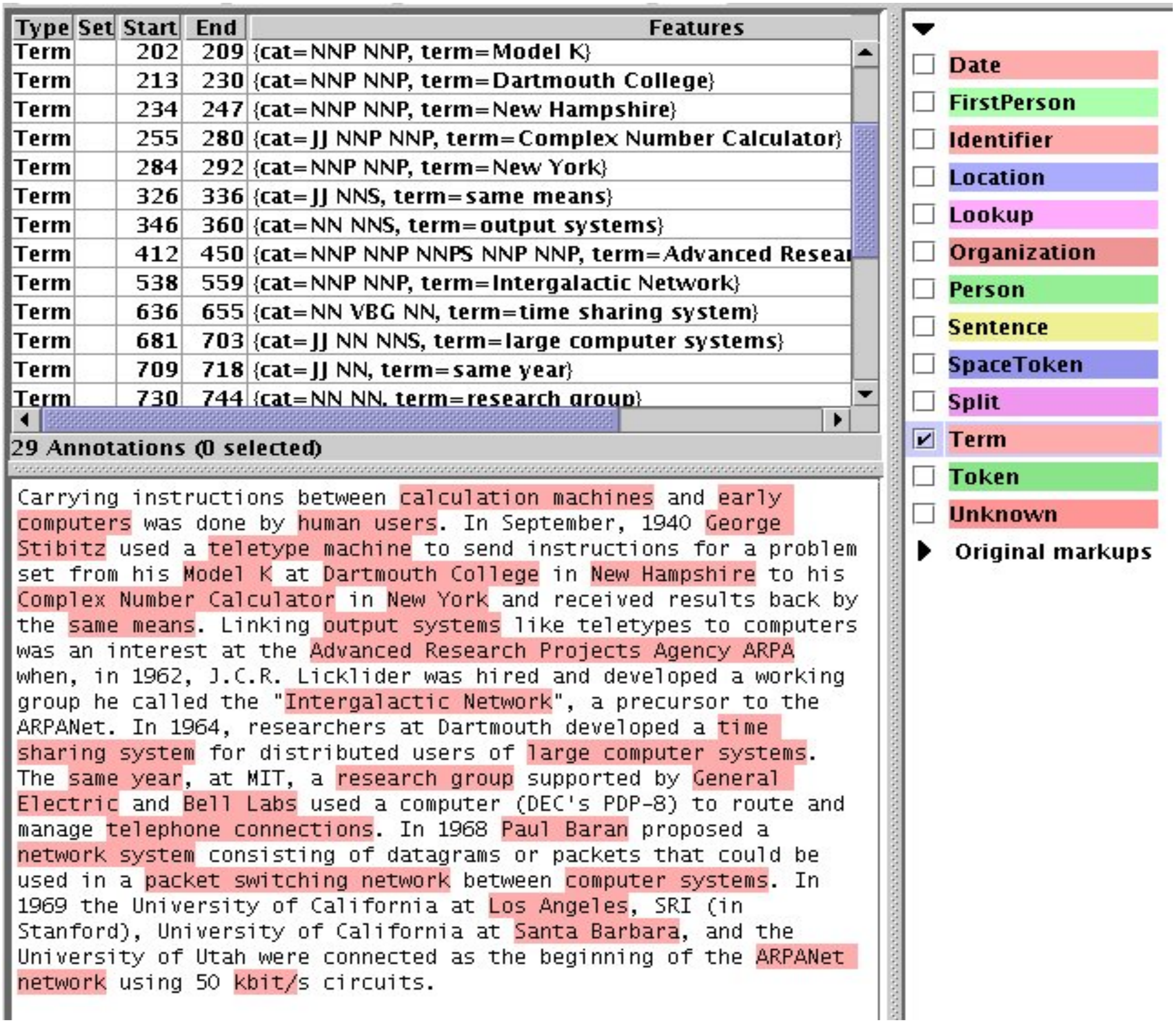
Major Surgical or Invasive Procedure:
[**2016-01-29**] Off Pump CABGx4 (LIMA->LAD, SVG->PDA, SVG->OM1,
SVG->OM2)
[**2016-02-03**] PICC Line Placement

History of Present Illness:
Mr. [**lastname 2372**] is a 75 year old gentleman with a history of coronary artery disease who presented to [**Hospital 2373**] on [**2016-01-27**] with the complaint of chest pain which
radiated to his left arm. This started while he was at rest and continued through his admission to the emergency department. He usually carries nitroglycerin with him however he did not have any during this event. He ruled out for a myocardial infarction and his pain resolved with nitroglycerin. A stress test was performed which was reportedly positive. A cardiac

| id | start | end | text | number | meets | comm... |
|---|---|---|---|---|---|---|
| SC0 | 190 | 222 | Date of Birth:  [... | age | DOES N... | |
| SC1 | 472 | 478 | CABGx4 | recent card... | MEETS | |
| SC2 | 618 | 629 | 75 year old | age | DOES N... | |
| SC3 | 659 | 682 | coronary artery ... | recent card... | DOES N... | |
| SC4 | 762 | 772 | chest pain | recent card... | MEETS | |
| SC5 | 1016 | 1037 | myocardial infar... | recent card... | DOES N... | |
| SC6 | 1143 | 1167 | cardiac  cathet... | recent card... | MEETS | |
| SC7 | 1561 | 1586 | Diabetes Mellitu... | diabetic | MEETS | |
| SC8 | 4565 | 4588 | coronary artery ... | recent card... | MEETS | |



```
▼<NounVerbTask>
  ▼<TEXT>
    ▼<![CDATA[
       JABBERWOCKY By Lewis Carroll 'Twas brillig, and the slithy toves Did gyre and gimble in the wabe; All
       mimsy were the borogoves, And the mome raths outgrabe. 'Beware the Jabberwock, my son! The jaws that
       bite, the claws that catch! Beware the Jubjub bird, and shun The frumious Bandersnatch!' He took his
       vorpal sword in hand: Long time the manxome foe he sought-- So rested he by the Tumtum tree, And stood
       awhile in thought. And as in uffish thought he stood, The Jabberwock, with eyes of flame, Came whiffling
       through the tulgey wood, And burbled as it came! One, two! One, two! And through and through The vorpal
       blade went snicker-snack! He left it dead, and with its head He went galumphing back. 'And hast thou
       slain the Jabberwock? Come to my arms, my beamish boy! O frabjous day! Callooh! Callay!' He chortled in
       his joy. 'Twas brillig, and the slithy toves Did gyre and gimble in the wabe; All mimsy were the
       borogoves, And the mome raths outgrabe.
     ]]>
   </TEXT>
  ▼<TAGS>
     <NOUN id="N0" start="1" end="12" text="JABBERWOCKY" type="thing" comment=""/>
     <NOUN id="N1" start="61" end="66" text="toves" type="thing" comment="default value"/>
     <NOUN id="N2" start="94" end="98" text="wabe" type="place" comment="default value"/>
     <NOUN id="N3" start="119" end="128" text="borogoves" type="thing" comment="default value"/>
     <VERB id="V0" start="71" end="75" text="gyre" tense="past" aspect="simple"/>
     <VERB id="V1" start="80" end="86" text="gimble" tense="past" aspect=""/>
     <VERB id="V2" start="956" end="964" text="outgrabe" tense="" aspect="perfect progressive"/>
     <ADJ_ADV id="A2" start="37" end="44" text="brillig" type=""/>
     <ADJ_ADV id="A3" start="54" end="60" text="slithy" type=""/>
     <ADJ_ADV id="A4" start="104" end="109" text="mimsy" type=""/>
     <ACTION id="A0" fromID="V1" fromText="gimble" toID="N3" toText="borogoves" relationship="performed_by"/>
     <ACTION id="A1" fromID="N0" fromText="JABBERWOCKY" toID="V0" toText="gyre" relationship="performs"/>
     <DESCRIPTION id="D0" fromID="A3" fromText="slithy" toID="N1" toText="toves" relationship="describes"/>
     <DESCRIPTION id="D1" fromID="A4" fromText="mimsy" toID="N3" toText="borogoves" relationship=""/>
   </TAGS>
 </NounVerbTask>
```

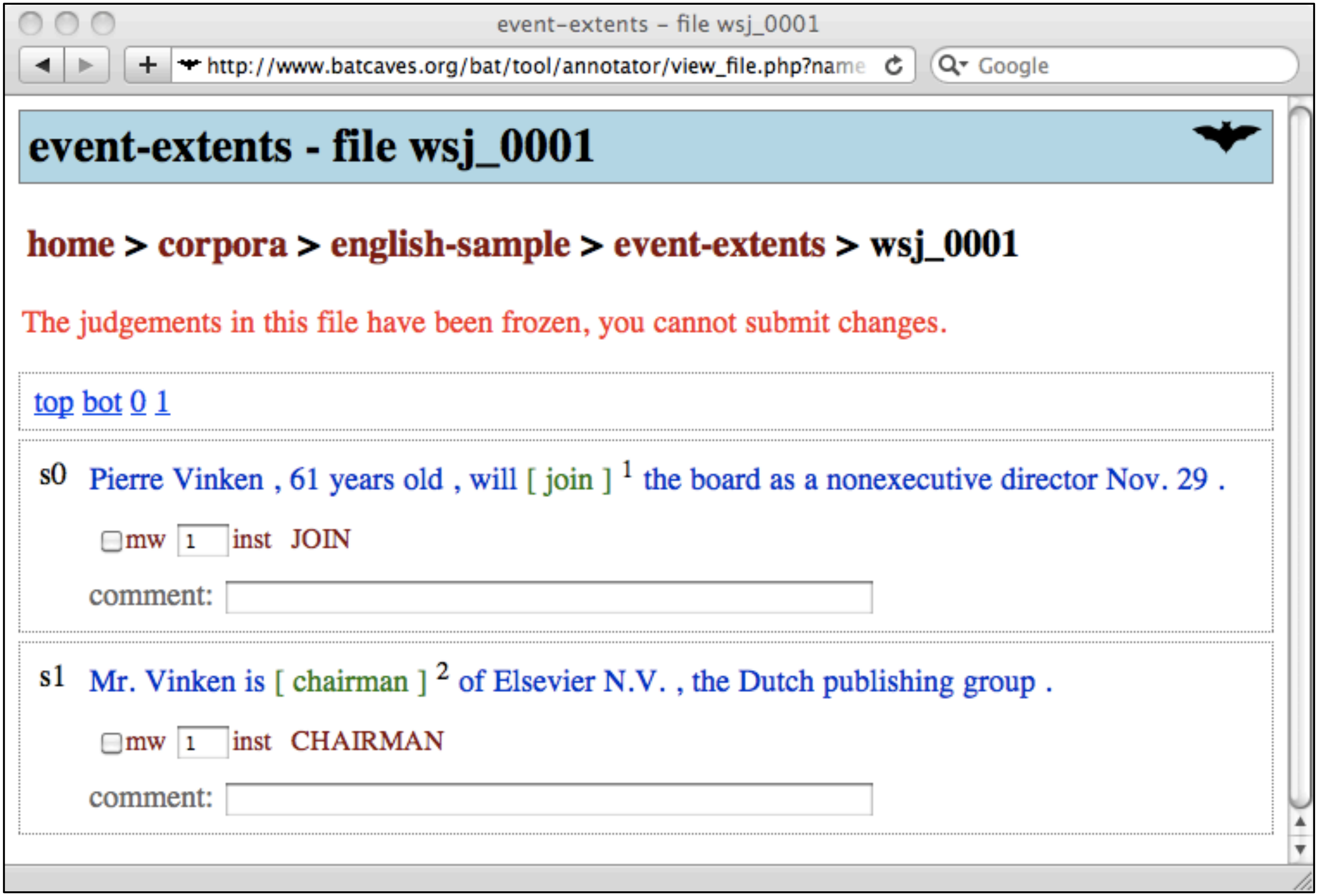


| Type | Set | Start | End | Features |
|---|---|---|---|---|
| Term | | 202 | 209 | {cat=NNP NNP, term=Model K} |
| Term | | 213 | 230 | {cat=NNP NNP, term=Dartmouth College} |
| Term | | 234 | 247 | {cat=NNP NNP, term=New Hampshire} |
| Term | | 255 | 280 | {cat=JJ NNP NNP, term=Complex Number Calculator} |
| Term | | 284 | 292 | {cat=NNP NNP, term=New York} |
| Term | | 326 | 336 | {cat=JJ NNS, term=same means} |
| Term | | 346 | 360 | {cat=NN NNS, term=output systems} |
| Term | | 412 | 450 | {cat=NNP NNP NNPS NNP NNP, term=Advanced Resear |
| Term | | 538 | 559 | {cat=NNP NNP, term=Intergalactic Network} |
| Term | | 636 | 655 | {cat=NN VBG NN, term=time sharing system} |
| Term | | 681 | 703 | {cat=JJ NN NNS, term=large computer systems} |
| Term | | 709 | 718 | {cat=JJ NN, term=same year} |
| Term | | 730 | 744 | {cat=NN NN, term=research group} |

29 Annotations (0 selected)

Carrying instructions between calculation machines and early computers was done by human users. In September, 1940 George Stibitz used a teletype machine to send instructions for a problem set from his Model K at Dartmouth College in New Hampshire to his Complex Number Calculator in New York and received results back by the same means. Linking output systems like teletypes to computers was an interest at the Advanced Research Projects Agency ARPA when, in 1962, J.C.R. Licklider was hired and developed a working group he called the "Intergalactic Network", a precursor to the ARPANet. In 1964, researchers at Dartmouth developed a time sharing system for distributed users of large computer systems. The same year, at MIT, a research group supported by General Electric and Bell Labs used a computer (DEC's PDP-8) to route and manage telephone connections. In 1968 Paul Baran proposed a network system consisting of datagrams or packets that could be used in a packet switching network between computer systems. In 1969 the University of California at Los Angeles, SRI (in Stanford), University of California at Santa Barbara, and the University of Utah were connected as the beginning of the ARPANet network using 50 kbit/s circuits.

☐ Date
☐ FirstPerson
☐ Identifier
☐ Location
☐ Lookup
☐ Organization
☐ Person
☐ Sentence
☐ SpaceToken
☐ Split
☑ Term
☐ Token
☐ Unknown
▶ Original markups


event-extents – file wsj_0001
http://www.batcaves.org/bat/tool/annotator/view_file.php?name
Google

# event-extents - file wsj_0001

home > corpora > english-sample > event-extents > wsj_0001

The judgements in this file have been frozen, you cannot submit changes.

top bot 0 1

s0 Pierre Vinken , 61 years old , will [ join ] 1 the board as a nonexecutive director Nov. 29 .

□ mw 1 inst JOIN

comment:

s1 Mr. Vinken is [ chairman ] 2 of Elsevier N.V. , the Dutch publishing group .

□ mw 1 inst CHAIRMAN

comment:

Pierre Vinken, 61 years old, will join the board

| | | | | |
|---|---|---|---|---|
| 1806815 | wsj_0001 | 0 | 0 | Pierre |
| 1806816 | wsj_0001 | 0 | 1 | Vinken |
| 1806817 | wsj_0001 | 0 | 2 | , |
| 1806818 | wsj_0001 | 0 | 3 | 61 |
| 1806819 | wsj_0001 | 0 | 4 | years |
| 1806820 | wsj_0001 | 0 | 5 | old |
| 1806821 | wsj_0001 | 0 | 6 | , |
| 1806822 | wsj_0001 | 0 | 7 | will |
| 1806823 | wsj_0001 | 0 | 8 | join |
| 1806824 | wsj_0001 | 0 | 9 | the |
| 1806825 | wsj_0001 | 0 | 10 | board |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1806823 | wsj_0001 | 0 | 8 | jane | n | event | 1 | 1 | n |
| 1806823 | wsj_0001 | 0 | 8 | joe | n | event | 1 | 1 | n |
| 1806823 | wsj_0001 | 0 | 8 | judge | y | event | 1 | 1 | n |

# Annotation and Annotation Tools

- Annotation and Annotation Tools
- Standards (why?)
- Linguistic Annotation Framework
- More
  - MASC corpus
  - LAPPS Grid: WSEV and LIF

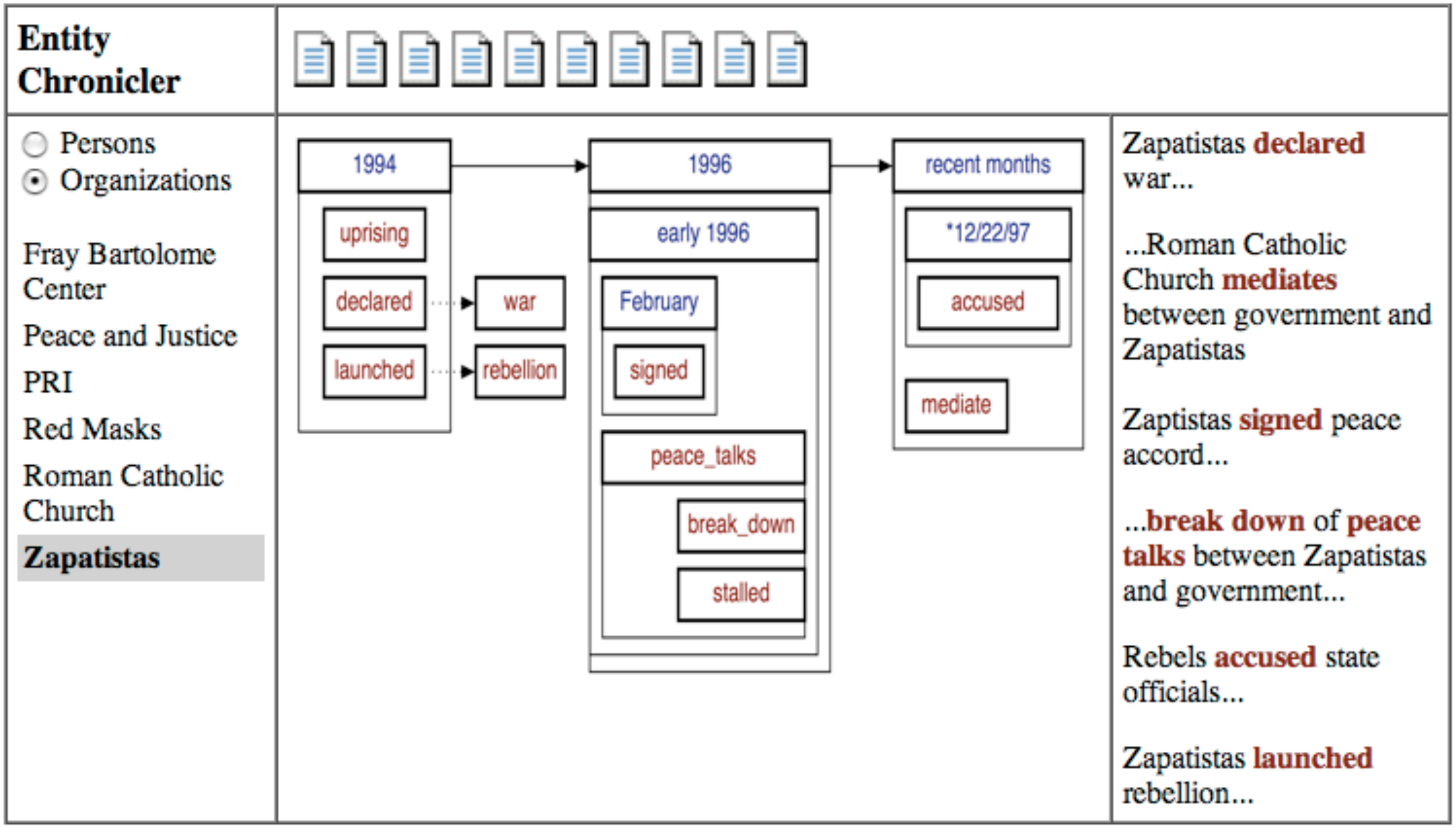
# Merging Annotations

```
( (S
  (NP-SBJ
   (NP (NNP Pierre) (NNP Vinken) )
   (, ,)
   (ADJP (NP (CD 61) (NNS years) ) (JJ old) )
   (, ,) )
  (VP (MD will)
   (VP (VB join)
    (NP (DT the) (NN board) )
    (PP-CLR (IN as)
     (NP (DT a) (JJ nonexecutive) (NN director) ))
    (NP-TMP (NNP Nov.) (CD 29) )))
  (. .) ))
```

```
<ENTITY type=“person”>Pierre
Vinken</ENTITY>, 61 years old,
will join the board as a
 nonexecutive director  Nov. 29.
```

```
Pierre Vinken, 61 years old, will
<EVENT id=“e1”>join</EVENT>
the board as a nonexecutive
director  Nov. 29.
```

```
wsj/00/wsj_0001.mrg 0 8 gold join.01 vf--a 0:2-ARG0 7:0-ARGM-MOD 8:0-rel
9:1-ARG1 11:1-ARGM-PRD 15:1-ARGM-TMP
```



Zapatistas **declared** war...

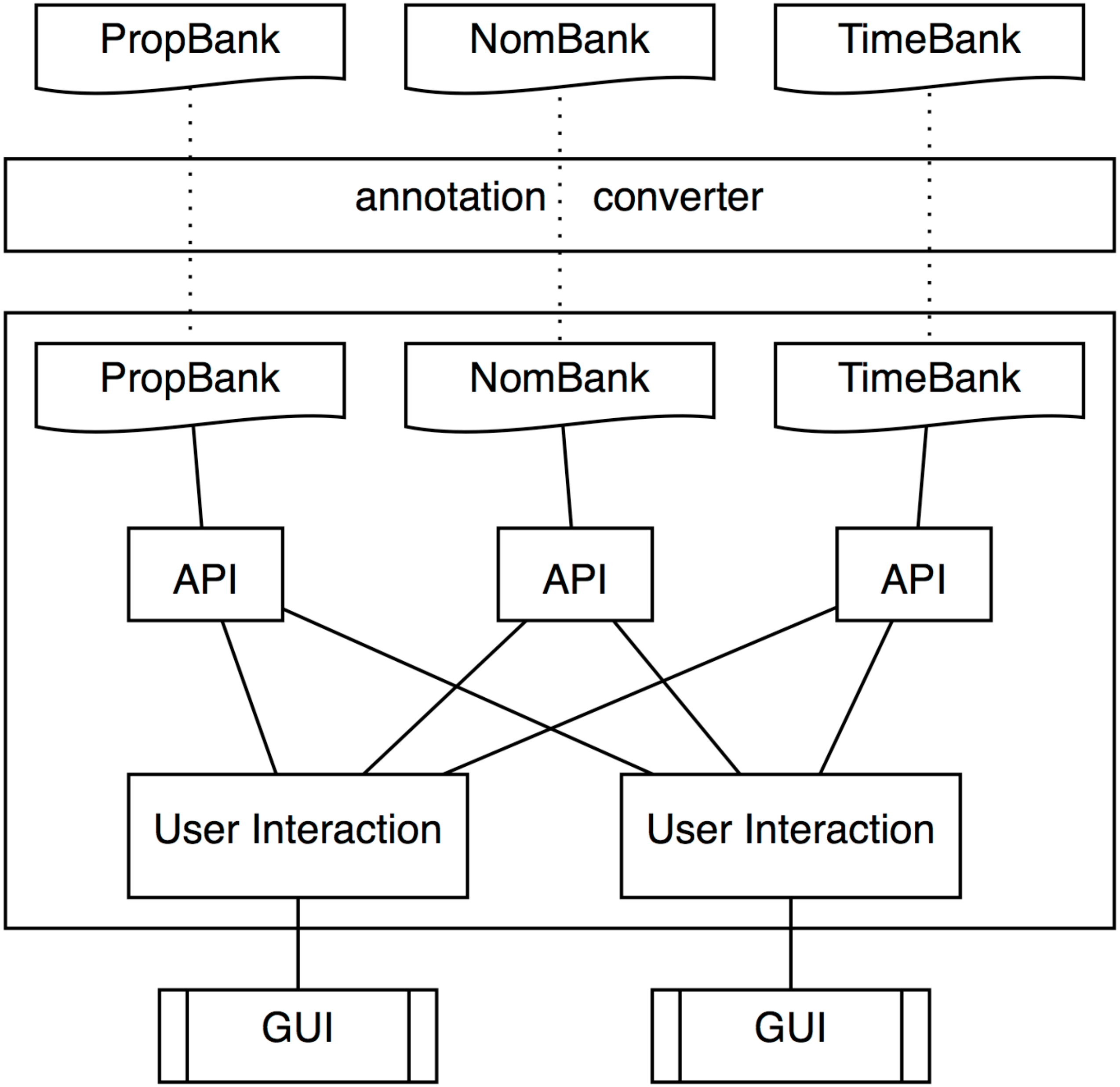
...Roman Catholic Church **mediates** between government and Zapatistas

Zaptistas **signed** peace accord...

...**break down** of **peace talks** between Zapatistas and government...

Rebels **accused** state officials...

Zapatistas **launched** rebellion...


PropBank
NomBank
TimeBank
annotation converter
PropBank
NomBank
TimeBank
API
API
API
User Interaction
User Interaction
GUI
GUI

# Harmonization and Standardization

- Language applications require the integration of varieties of linguistic information which can come from diverse sources

- Interoperability

# Annotation and Annotation Tools

- Annotation and Annotation Tools
- Standards (why?)
- Linguistic Annotation Framework
- More
  - MASC corpus
  - LAPPS Grid: WSEV and LIF

# Linguistic Annotation Framework

- International Standards Organization
- Basis for harmonizing existing language resources, as well as for developing new ones
  - a true standard is impractical
    - Large variety of theoretical and descriptive frameworks
  - Existing resources rendered obsolete if new standards emerge

# What Came Before

- Some recent fundamental representation principles:
  - Stand-off annotation
  - XML
- Generalized annotation mechanisms and formats:
  - XCES, Text Encoding Initiative (TEI)
  - Annotation Graphs

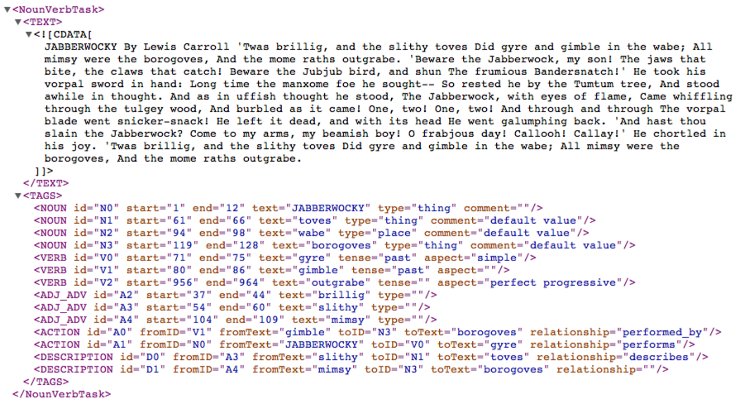
# LAF Design Requirements

- Must allow users to represent their data and annotation in a variety of formats
- Must accommodate all varieties of annotation

```
In Washington <TIMEX3 tid="t1" TYPE="DATE" VAL="PRESENT_REF" temporalFunction="true"
valueFromFunction="tf1" anchorTimeID="t0">today</TIMEX3>, the Federal Aviation
Administration <EVENT eid="e1" class="OCCURRENCE">released</EVENT> air traffic control
tapes from the night the TWA Flight eight hundred
<EVENT eid="e2" class="OCCURRENCE">went</EVENT> down. There's nothing new on why
the plane <EVENT eid="e3" class="OCCURRENCE">exploded</EVENT>, but you
<EVENT eid="e4" class="OCCURRENCE">cannot</EVENT>
<EVENT eid="e5" class="OCCURRENCE">miss</EVENT> the moment.
ABC's Lisa Stark <EVENT eid="e6" class="OCCURRENCE">has</EVENT> more.

<MAKEINSTANCE eventID="e1" pos="VERB" eiid="ei1" tense="PAST" aspect="NONE"/>
<MAKEINSTANCE eventID="e2" pos="VERB" eiid="ei2" tense="PAST" aspect="NONE"/>
<MAKEINSTANCE eventID="e3" pos="VERB" eiid="ei3" tense="PAST" aspect="NONE"/>
<MAKEINSTANCE eventID="e4" pos="VERB" eiid="ei4" tense="PRESENT" aspect="NONE"/>
<MAKEINSTANCE eventID="e5" pos="VERB" eiid="ei5" tense="INFINITIVE" aspect="NONE"/>
<MAKEINSTANCE eventID="e6" pos="NONE" eiid="ei6" tense="PRESENT" aspect="NONE"/>

<TLINK eventInstanceID="ei1" relatedToTime="t1" relType="IS_INCLUDED" rule="2-1" />
<TLINK eventInstanceID="ei2" relatedToTime="t1" relType="IS_INCLUDED" rule="2-1" />
<TLINK eventInstanceID="ei1" relatedToEventInstance="ei3" relType="BEFORE" rule="3-19" />
<TLINK eventInstanceID="ei3" relatedToEventInstance="ei4" relType="BEFORE" rule="6-1" />
<TLINK eventInstanceID="ei3" relatedToEventInstance="ei6" relType="BEFORE" rule="3-23" />
```



```
▼<NounVerbTask>
  ▼<TEXT>
    ▼<![CDATA[
      JABBERWOCKY By Lewis Carroll 'Twas brillig, and the slithy toves Did gyre and gimble in the wabe; All
      mimsy were the borogoves, And the mome raths outgrabe. 'Beware the Jabberwock, my son! The jaws that
      bite, the claws that catch! Beware the Jubjub bird, and shun The frumious Bandersnatch!' He took his
      vorpal sword in hand: Long time the manxome foe he sought-- So rested he by the Tumtum tree, And stood
      awhile in thought. And as in uffish thought he stood, The Jabberwock, with eyes of flame, Came whiffling
      through the tulgey wood, And burbled as it came! One, two! One, two! And through and through The vorpal
      blade went snicker-snack! He left it dead, and with its head He went galumphing back. 'And hast thou
      slain the Jabberwock? Come to my arms, my beamish boy! O frabjous day! Callooh! Callay!' He chortled in
      his joy. 'Twas brillig, and the slithy toves Did gyre and gimble in the wabe; All mimsy were the
      borogoves, And the mome raths outgrabe.
    ]]>
  </TEXT>
  ▼<TAGS>
    <NOUN id="N0" start="1" end="12" text="JABBERWOCKY" type="thing" comment=""/>
    <NOUN id="N1" start="61" end="66" text="toves" type="thing" comment="default value"/>
    <NOUN id="N2" start="94" end="98" text="wabe" type="place" comment="default value"/>
    <NOUN id="N3" start="119" end="128" text="borogoves" type="thing" comment="default value"/>
    <VERB id="V0" start="71" end="75" text="gyre" tense="past" aspect="simple"/>
    <VERB id="V1" start="80" end="86" text="gimble" tense="past" aspect=""/>
    <VERB id="V2" start="956" end="964" text="outgrabe" tense="" aspect="perfect progressive"/>
    <ADJ_ADV id="A2" start="37" end="44" text="brillig" type=""/>
    <ADJ_ADV id="A3" start="54" end="60" text="slithy" type=""/>
    <ADJ_ADV id="A4" start="104" end="109" text="mimsy" type=""/>
    <ACTION id="A0" fromID="V1" fromText="gimble" toID="N3" toText="borogoves" relationship="performed_by"/>
    <ACTION id="A1" fromID="N0" fromText="JABBERWOCKY" toID="V0" toText="gyre" relationship="performs"/>
    <DESCRIPTION id="D0" fromID="A3" fromText="slithy" toID="N1" toText="toves" relationship="describes"/>
    <DESCRIPTION id="D1" fromID="A4" fromText="mimsy" toID="N3" toText="borogoves" relationship=""/>
  </TAGS>
</NounVerbTask>
```

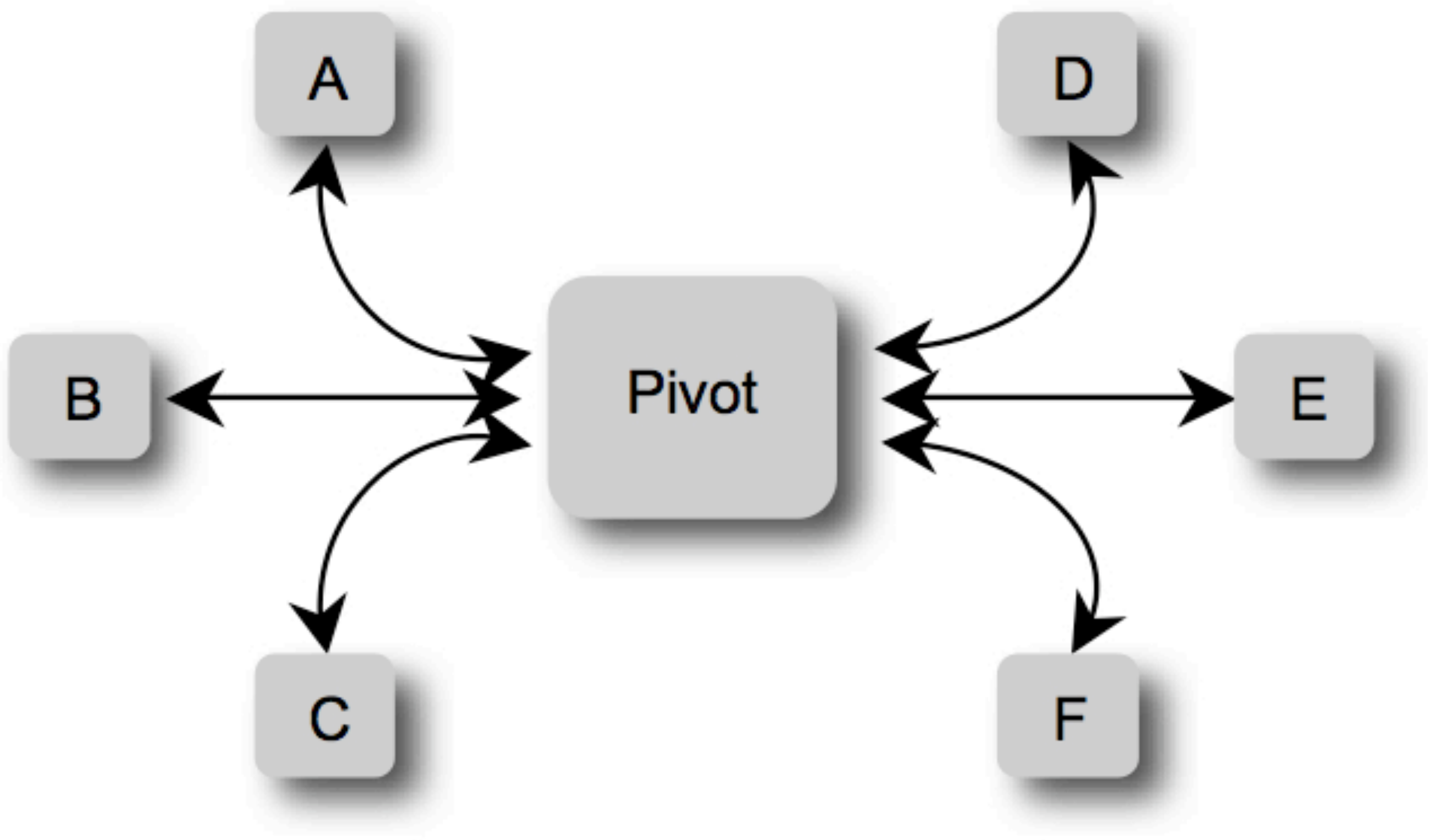
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1806823 wsj_0001 | 0 | 8 | jane | n | event | 1 | 1 | n |
| 1806823 wsj_0001 | 0 | 8 | joe | n | event | 1 | 1 | n |
| 1806823 wsj_0001 | 0 | 8 | judge | y | event | 1 | 1 | n |

# LAF Principles

- Separation of syntax and semantics
  - That is, separation of structure and content
- Separation of data and annotations
  - Read-only primary data & stand-off annotation
- Allow layered annotation
  - an item from one annotation can refer to an item in another layer

# LAF Principles

- An annotation is a graph
- Separation of user annotation formats and exchange format
  - User annotations must be mappable to feature structure based data model instantiated in dump
- Pivot format as an interface to other formats



- Flexible document annotation under user control
- Rigid dump format (pivot)

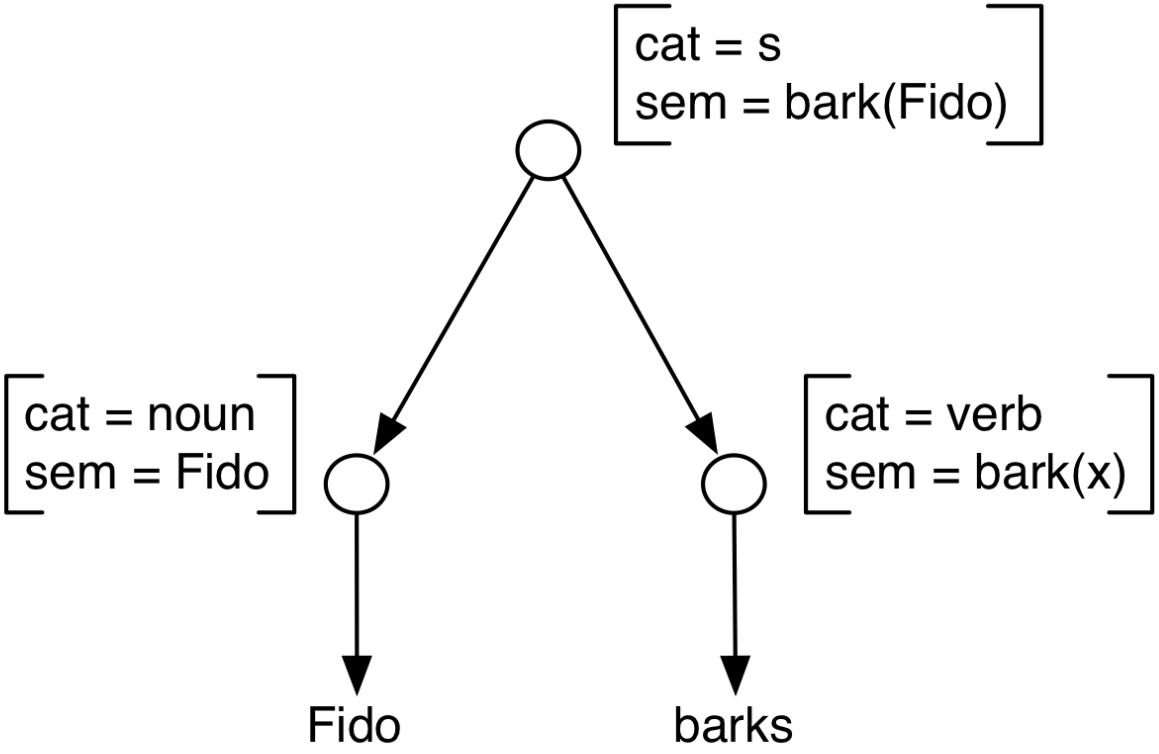
# LAF Core

- An abstract model for annotations …
- instantiated by a pivot format …
- into which annotations are mapped for the purposes of exchange.

# Dump Format

- To map to the pivot, an annotation scheme must be expressible in the abstract model
- Abstract model:
  - a structure that associates stand-off annotations with primary data, instantiated as a directed graph;
  - a feature structure representation for annotation content.

# Abstract Model



cat = s
sem = bark(Fido)

cat = noun
sem = Fido

cat = verb
sem = bark(x)

Fido

barks

# Abstract Model

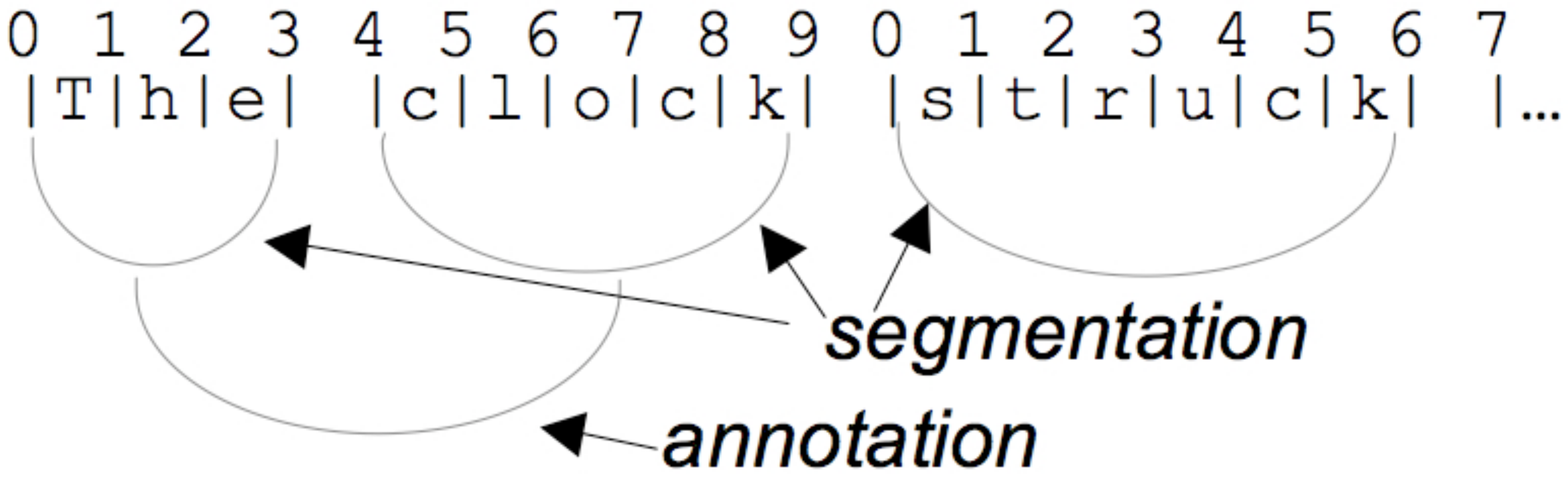
- Graph theory provides a well-understood model for representing objects that can be viewed as a connected set of more elementary sub-objects, together with a wealth of graph-analytic algorithms for information extraction and analysis.
- Feature structures attached to graph nodes
- Nodes versus edges

# Formally…

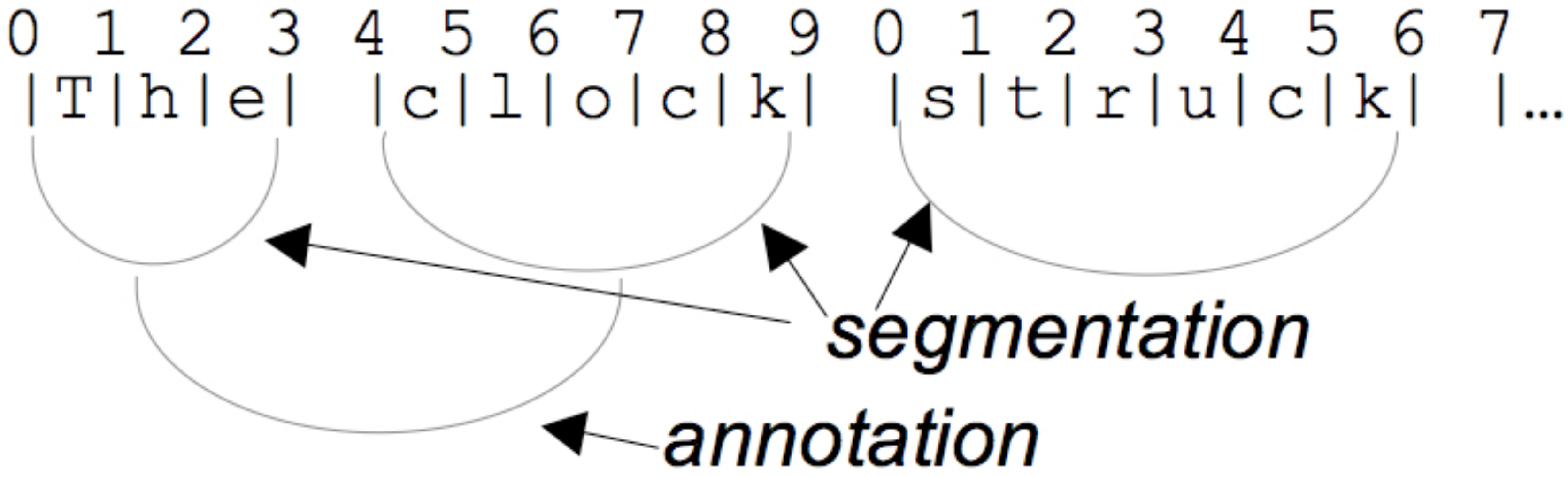
- LAF consists of a data model for annotations based on directed graphs defined as follows:
  - A graph of annotations G is a set of vertices V(G) and a set of edges E(G).
  - Vertices and edges may be labeled with one or more features.
  - A feature consists of a quadruple (G’, VE, K, V) where, G’ is a graph, VE is a vertex or edge in G’, K is the name of the feature and V is the feature value.

# Base Segmentation of Primary Data

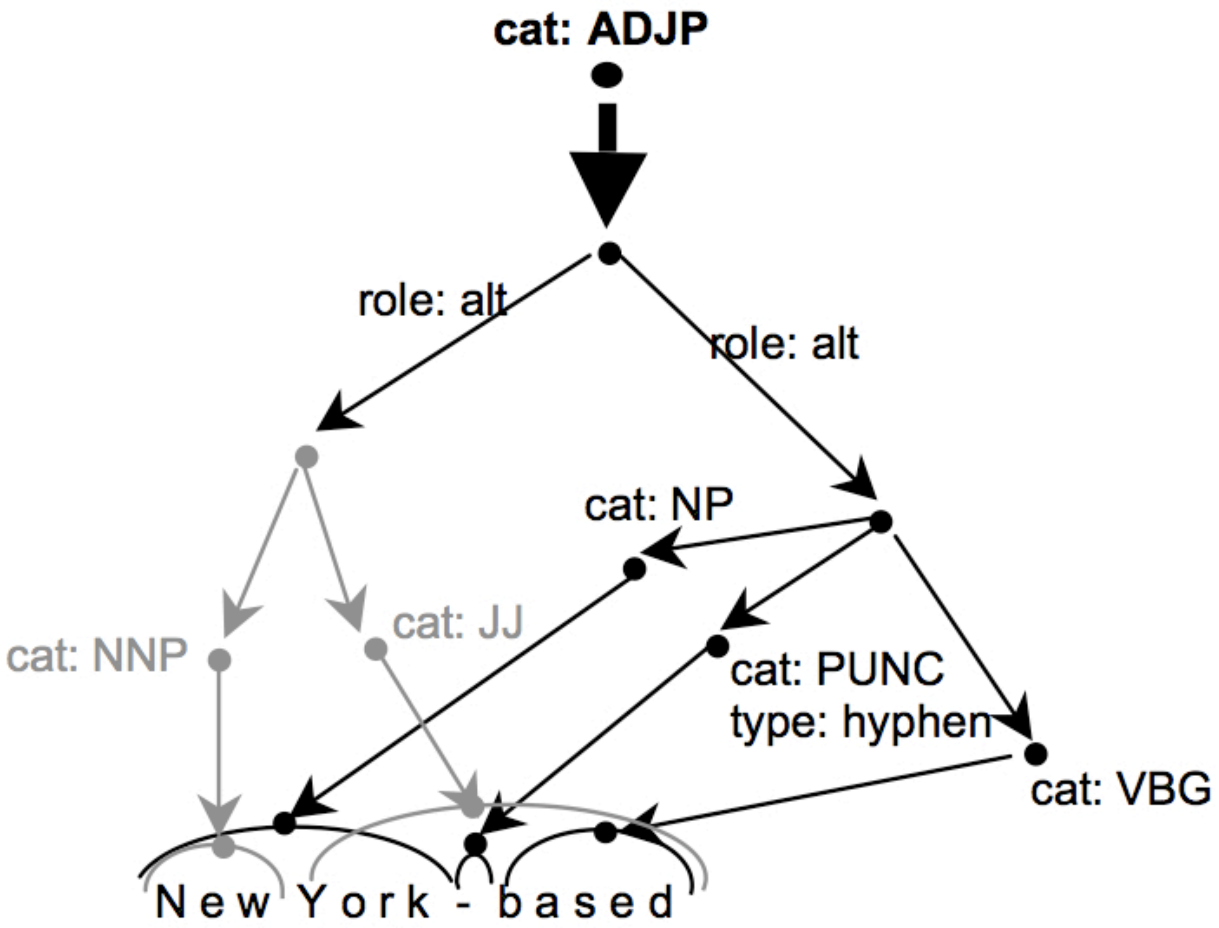
- Defines edges between virtual nodes located between each “character” in the primary data. The resulting graph G is treated as an edge graph G’ whose nodes are the edges of G, and which serve as the leaf (“sink”) nodes.
- These nodes provide the base for an annotation or several layers of annotation. Multiple segmentations can be defined over the primary data, and multiple annotations may refer to the same segmentation.


0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7
|T|h|e| |c|l|o|c|k| |s|t|r|u|c|k| |...
segmentation
annotation

```
<!-- edges over primary data -->
<edge id="e1" from="0" to="3"/>
<edge id="e2" from="4" to="9"/>
<edge id="e2" from="10" to="16"/>
```


0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7
|T|h|e| |c|l|o|c|k| |s|t|r|u|c|k| |...
segmentation
annotation

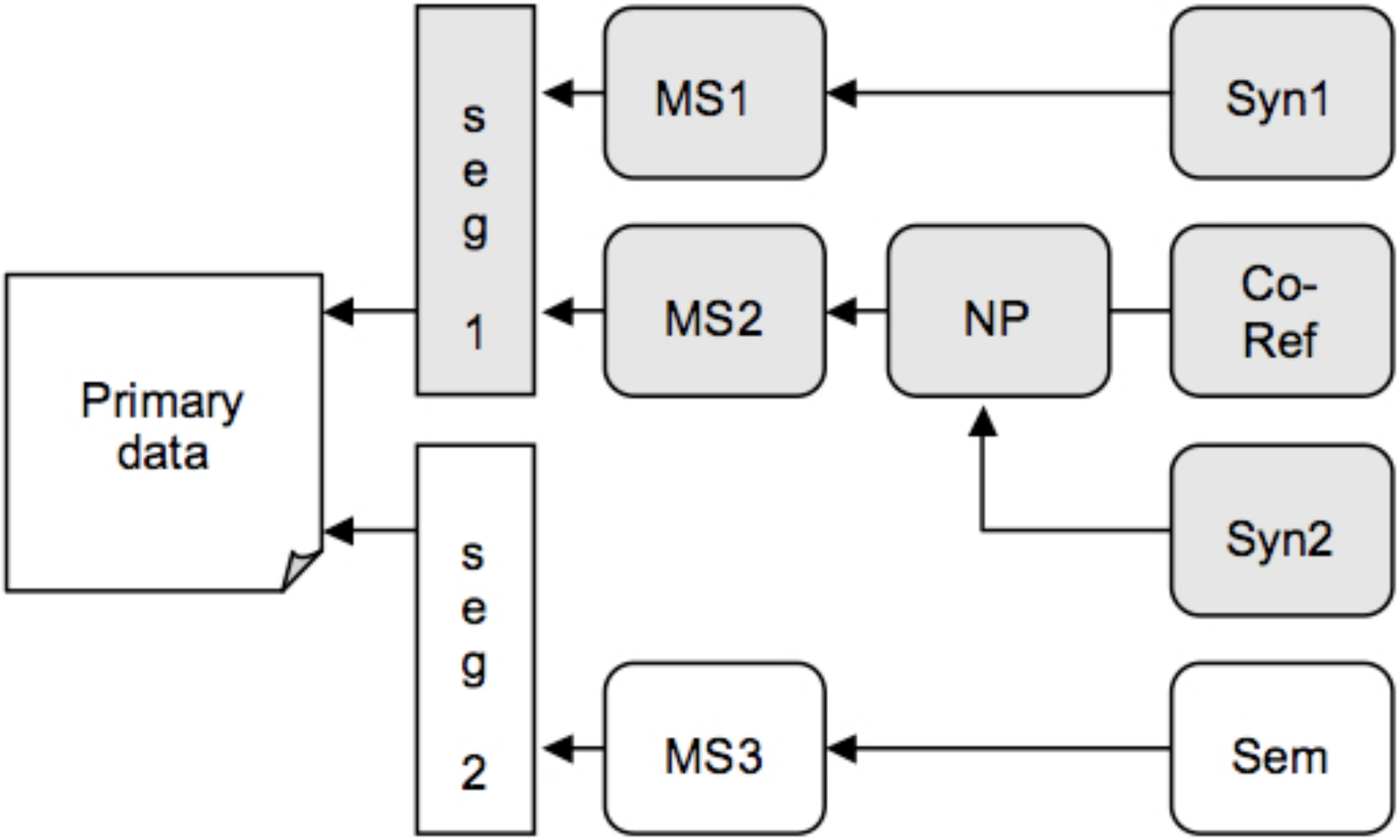
```
<edge id="t2" ref="e2">
  <fs type="token">
      <f name="lemma" sVal="clock"/>
     <f name="pos" sVal="NN"/>
   </fs>
 </edge>
```


cat: ADJP
role: alt
role: alt
cat: NP
cat: JJ
cat: NNP
cat: PUNC
type: hyphen
cat: VBG
New York - based

# Annotation Content

- LAF does not provide specifications for annotation content (the labels describing the associated linguistic phenomena)
  - standardization here is rather tricky
- Data Category Registry (DCR)
  - contains pre-defined data elements and schemas that may be used directly in annotations, together with means to specify new categories and modify existing ones.

# Layered Annotation



# Layered Annotation

*Base segmentation:*

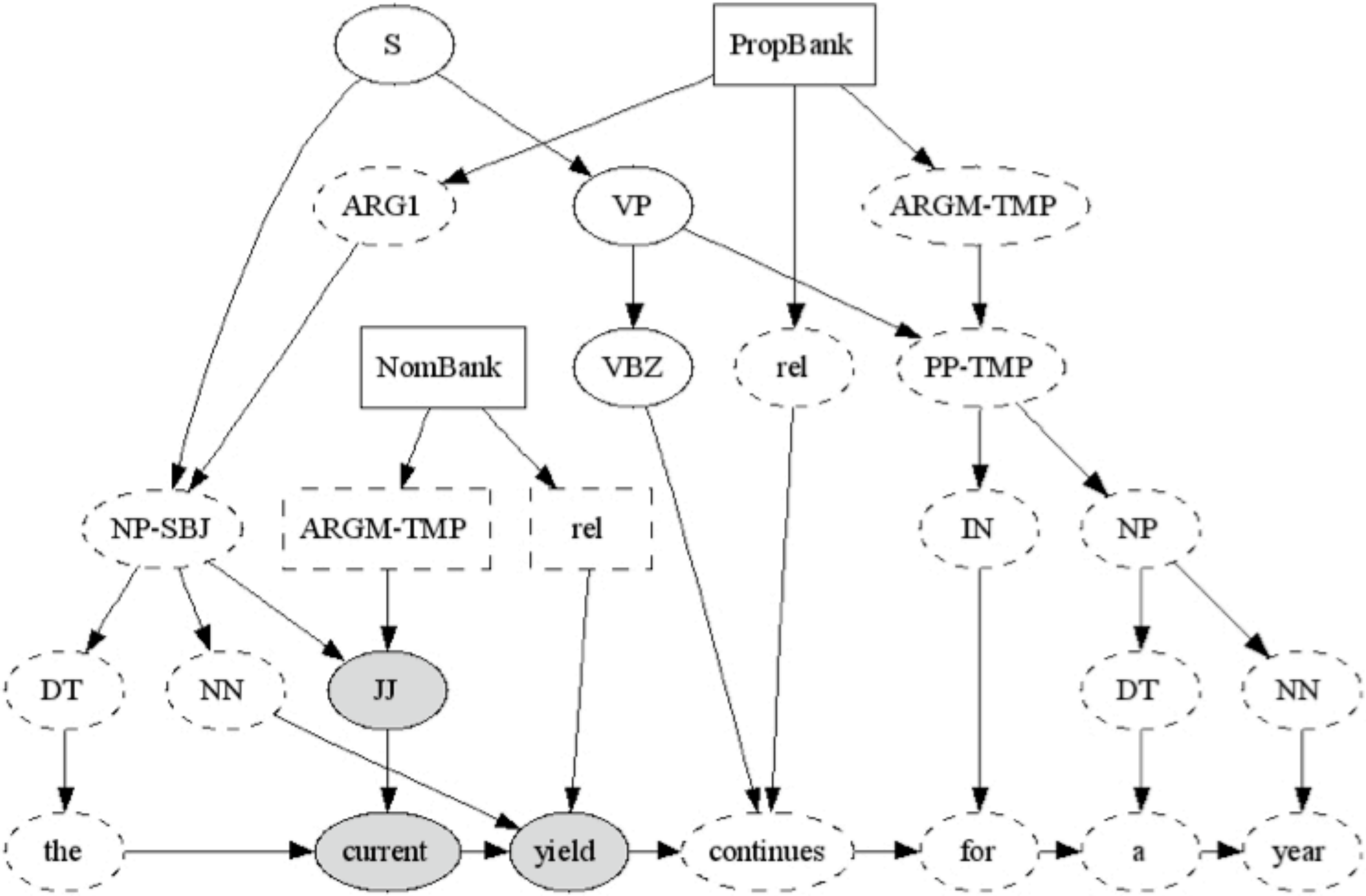
```
  <seg:sink seg:id="42" seg:start="24” seg:end="35"/>
```

*Annotation over the base segmentation:*

```
  <msd:node msd:id=”16”>
    <msd:f name=”cat” value=”NN”/>
  </msd:node>
  <msd:edge from="msd:16" to="seg:42"/>
```

*Annotation over another annotation:*

```
  <ptb:node ptb:id="23">
    <ptb:f name="type" value="NP"/>
    <ptb:f name="role" value="-SBJ"/>
  </ptb:node>
  <ptb:edge from="ptb:23"to="msd:16"/>
```


S
PropBank
ARG1
VP
ARGM-TMP
NomBank
VBZ
rel
PP-TMP
NP-SBJ
ARGM-TMP
rel
IN
NP
DT
NN
JJ
DT
NN
the
current
yield
continues
for
a
year

# Annotation and Annotation Tools

- Annotation and Annotation Tools
- Standards (why?)
- Linguistic Annotation Framework
- More
  - MASC corpus
  - LAPPS Grid: WSEV and LIF

# MASC Files

```
9 SILT/MASC-1.0.3/data/written> ls -al
total 138792
drwxr-xr-x   711 marc  marc    24174 Mar  1  2011 ./
drwxr-xr-x     5 marc  marc      170 Mar  1  2011 ../
-rw-r--r--     1 marc  marc     5287 Sep 19  2010 110CYL067-logical.xml
-rw-r--r--     1 marc  marc    89725 Sep 19  2010 110CYL067-mpqa.xml
-rw-r--r--     1 marc  marc    44185 Sep 19  2010 110CYL067-nc.xml
-rw-r--r--     1 marc  marc     4614 Sep 19  2010 110CYL067-ne.xml
-rw-r--r--     1 marc  marc   147740 Sep 19  2010 110CYL067-penn.xml
-rw-r--r--     1 marc  marc   192338 Sep 19  2010 110CYL067-ptb.xml
-rw-r--r--     1 marc  marc   116178 Sep 19  2010 110CYL067-ptbtok.xml
-rw-r--r--     1 marc  marc     9853 Sep 19  2010 110CYL067-s.xml
-rw-r--r--     1 marc  marc    30898 Sep 19  2010 110CYL067-seg.xml
-rw-r--r--     1 marc  marc    36766 Sep 19  2010 110CYL067-vc.xml
-rw-r--r--     1 marc  marc     3082 Oct 21  2010 110CYL067.anc
-rw-r--r--     1 marc  marc     3094 Sep 19  2010 110CYL067.txt
```

# MASC Header File

```xml
<?xml version="1.0" encoding="UTF-8"?>

<annotations>
  <annotation ann.loc="110CYL067-logical.xml" type="logical">
     Document structure</annotation>
  <annotation ann.loc="110CYL067-ne.xml" type="ne">
     Named Entities</annotation>
  <annotation ann.loc="110CYL067-penn.xml" type="penn">
     Penn part of speech tags</annotation>
  <annotation ann.loc="110CYL067-ptb.xml" type="ptb">
     Penn Tree Bank</annotation>
  <annotation ann.loc="110CYL067-ptbtok.xml" type="ptbtok">
     Penn Tree Bank tokens and part of speech tags</annotation>
  <annotation ann.loc="110CYL067-s.xml" type="s">
     Sentence boundaries</annotation>
  <annotation ann.loc="110CYL067-seg.xml" type="seg">
     Base segmentation (quarks)</annotation>
</annotations>
```

# MASC Base Segmentation File

```
<?xml version="1.0" encoding="UTF-8"?>
<graph xmlns="http://www.xces.org/ns/GrAF/1.0/">
   <region xml:id="seg-r0" anchors="18 26"/>
   <region xml:id="seg-r2" anchors="27 31"/>
   <region xml:id="seg-r4" anchors="38 42"/>
   <region xml:id="seg-r6" anchors="43 55"/>
   <region xml:id="seg-r8" anchors="56 58"/>
   <region xml:id="seg-r10" anchors="59 67"/>
   <region xml:id="seg-r12" anchors="68 72"/>
   <region xml:id="seg-r14" anchors="73 77"/>
   <region xml:id="seg-r16" anchors="78 82"/>
   <region xml:id="seg-r18" anchors="83 87"/>
   <region xml:id="seg-r20" anchors="88 91"/>
   <region xml:id="seg-r22" anchors="92 95"/>
   <region xml:id="seg-r24" anchors="96 100"/>
```

# MASC Penn Tokenization File

```
<?xml version="1.0" encoding="UTF-8"?>
<graph xmlns="http://www.xces.org/ns/GrAF/1.0/">
  <header>
    <dependencies>
      <dependsOn type="seg"/>
    </dependencies>
    <annotationSets>
      <annotationSet name="PTB"
                     type="http://www.cis.upenn.edu/~treebank/"/>
    </annotationSets>
  </header>
  <node xml:id="ptb-n00002">
    <link targets="seg-r0"/>
  </node>
  <a label="tok" ref="ptb-n00002" as="PTB">
     <fs>
       <f name="msd" value="NNP"/>
     </fs>
   </a>
```

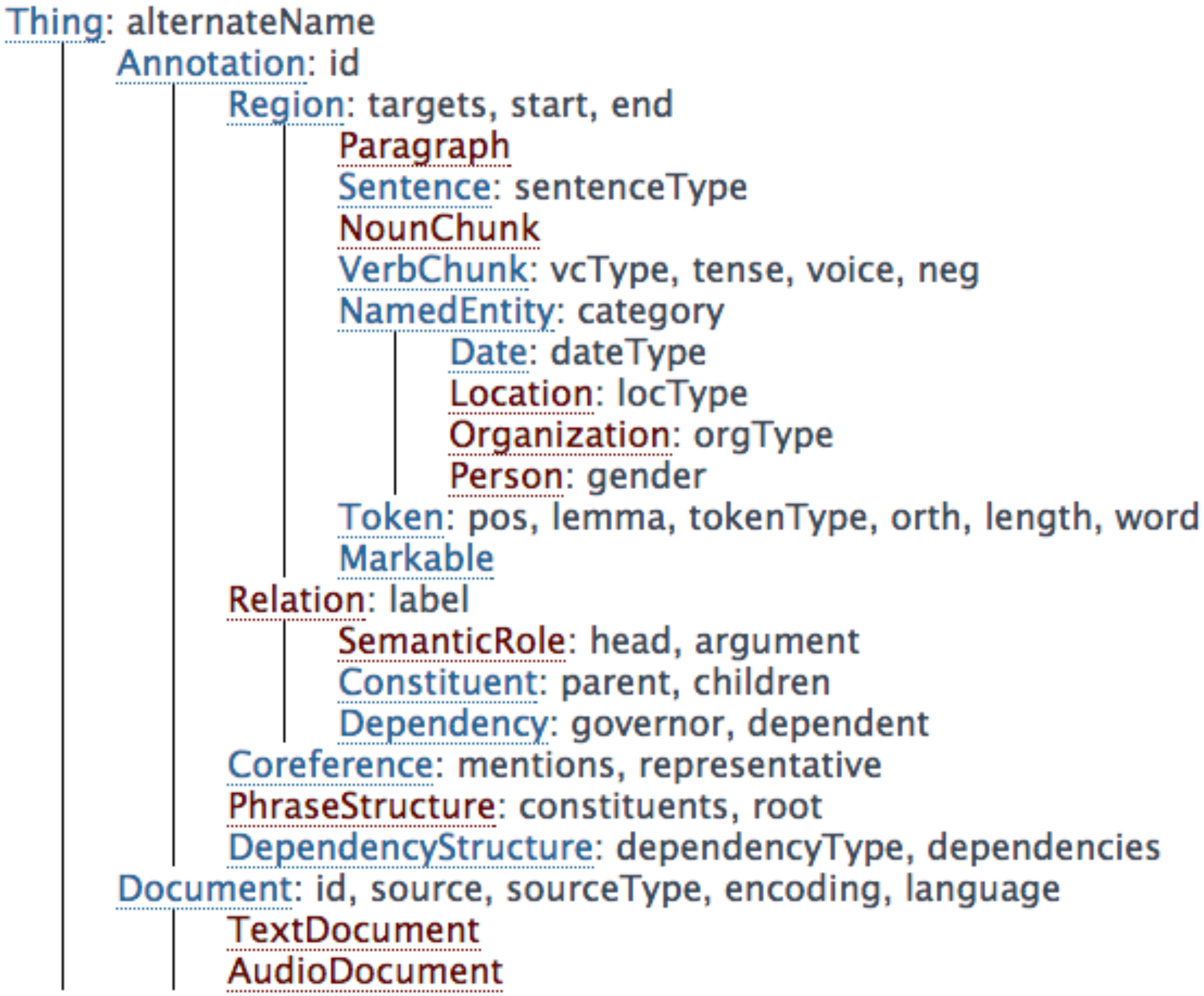
# MASC Penn Syntax File

```
<?xml version="1.0" encoding="UTF-8"?>
<graph xmlns="http://www.xces.org/ns/GrAF/1.0/">
  <header>
    <dependencies>
      <dependsOn type="ptbtok"/>
    </dependencies>
    <annotationSets>
      <annotationSet name="PTB"
                     type="http://www.cis.upenn.edu/~treebank/"/>
    </annotationSets>
  </header>
  <node xml:id="ptb-n00000"/>
  <node xml:id="ptb-n00254"/>
  <a label="S" ref="ptb-n00254" as="PTB">
    <fs>
      <f name="cat" value="S"/>
    </fs>
  </a>
```

# Annotation and Annotation Tools

- Annotation and Annotation Tools
- Standards (why?)
- Linguistic Annotation Framework
- More
  - MASC corpus
  - LAPPS Grid: WSEV and LIF

# LAPPS Exchange Vocabulary Type Hierarchy



- Thing: alternateName
    - Annotation: id
        - Region: targets, start, end
            - Paragraph
            - Sentence: sentenceType
            - NounChunk
            - VerbChunk: vcType, tense, voice, neg
            - NamedEntity: category
                - Date: dateType
                - Location: locType
                - Organization: orgType
                - Person: gender
            - Token: pos, lemma, tokenType, orth, length, word
            - Markable
        - Relation: label
            - SemanticRole: head, argument
            - Constituent: parent, children
            - Dependency: governor, dependent
        - Coreference: mentions, representative
        - PhraseStructure: constituents, root
        - DependencyStructure: dependencyType, dependencies
    - Document: id, source, sourceType, encoding, language
        - TextDocument
        - AudioDocument

# Thing > Annotation > Region > Token

| | |
|---|---|
| **Definition** | A string of one or more characters that serves as an indivisible unit for the purposes of morpho-syntactic labeling (part of speech tagging). |
| **Similar to** | http://www.isocat.org/datcat/DC-1403 |
| **URI** | http://vocab.lappsgrid.org/Token |

# Metadata

| Properties | Type | Description |
|---|---|---|
| posTagSet | String or URI | The definition of the tag set used by the part-of-speech tagger. |

## Metadata from Annotation

| Properties | Type | Description |
|---|---|---|
| producer | List of URI | The software that produced the annotations. |
| rules | List of URI | The documentation (if any) for the rules that were used to identify the annotations. |

# Properties

| Properties | Type | Description |
|---|---|---|
| pos | String or URI | Part-of-speech tag associated with the token. |
| lemma | String or URI | The root (base) form associated with the token. URI may point to a lexicon entry. |
| tokenType | String or URI | Sub-type such as word, punctuation, abbreviation, number, symbol, etc. Ideally a URI referencing a pre-defined descriptor. |
| orth | String or URI | Orthographic properties of the token such as LowerCase, UpperCase, UpperInitial, etc. Ideally a URI referencing a pre-defined descriptor. |
| length | Integer | The length of the token |
| word | String | The surface string in the primary data covered by this Token. |

# LAPPS Interchange Format

```
"views": [
  {
    "@context": {},
    "id": "v0",
    "metadata": {
      "contains": {
        "Token": {
          "producer": "lappsgrid.brandeis.opennlp.Tokenizer:0.0.4",
          "rules": "tokenization:opennlp_basic" }}},
    "annotations": [
      { "@type": "Token",
        "id": "t0",
        "start": 0,
        "end": 5,
        "features": {} } ]
  }
]
```