WP3 Task 3.9

Automatic, Context-of-capture based Categorization, Structure Detection and Segmentation of News Telecasts

DELOS All Tasks Meeting 2006-01-30, Paris

Partners involved

- Center for Computing Technologies (TZI), University of Bremen, Germany
  Arne Jacobs, George Ioannidis
- Laboratory of Distributed Multimedia Information Systems and Applications, Technical University of Crete, Chania, Greece
  Nektarios Moumoutzis , Stavros Christodoulakis
- Fraunhofer Institute for Media Communication (IMK), Sankt Augustin, Germany
  Martha Larson

Overview

- Motivation
- Goals
- System Architecture
  - Stochastic parser
  - Audio-/visual recognizers
  - Semantic recognizer
- Plans for JPA3

Motivation

- Domain: News telecasts
- News telecasts are structured
- All instances of one news format follow the same structure

=> News formats can be modeled

- News is composed of story units
- A story unit can be associated with a topic

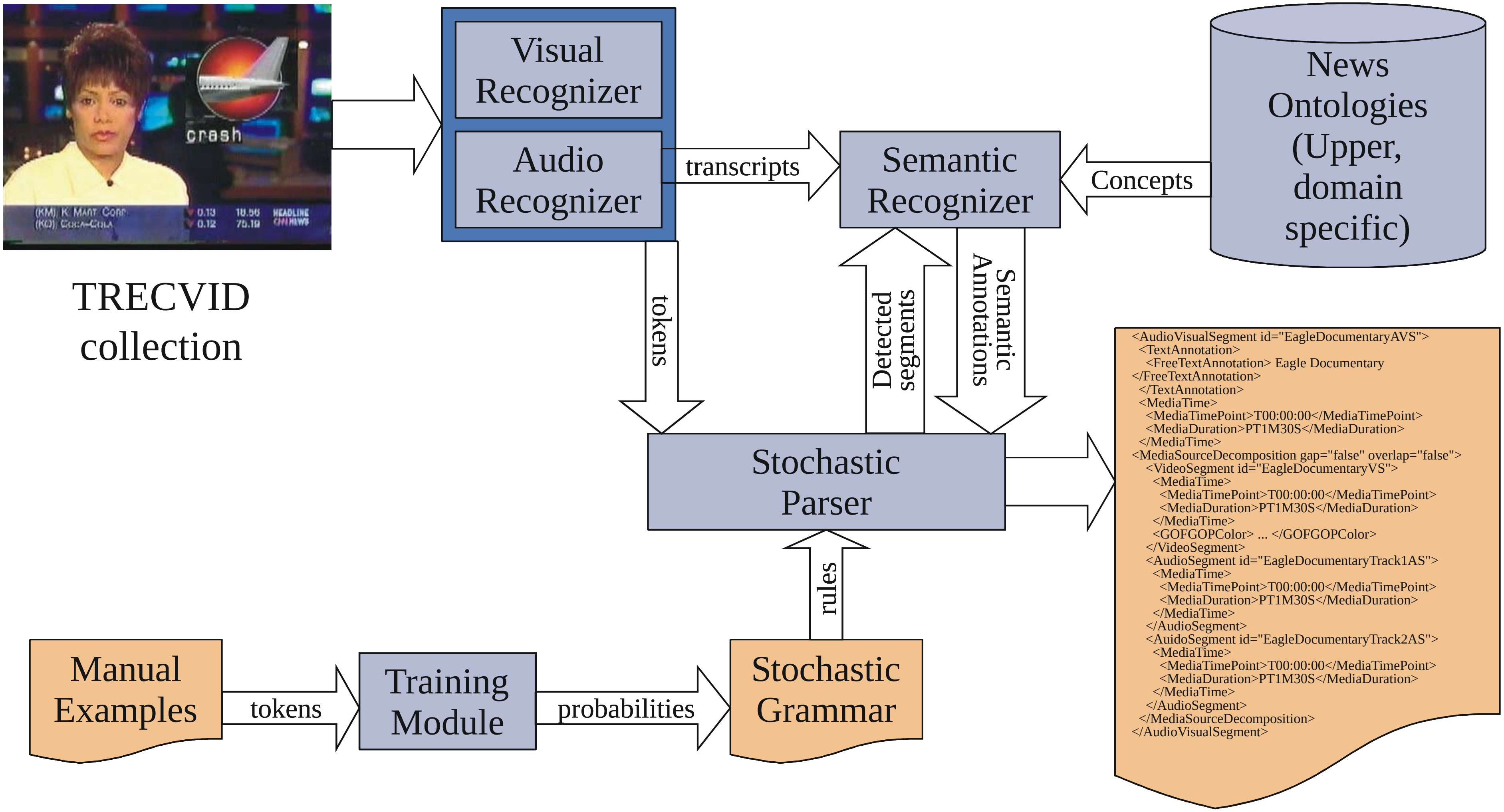
Goals

- Main goal: Make news structure explicit for the user to be usable for search and retrieval
- Derived goals:
  - News format modeling
  - Structure detection
  - Segmentation
  - Context-of-Capture-based categorization

Overview

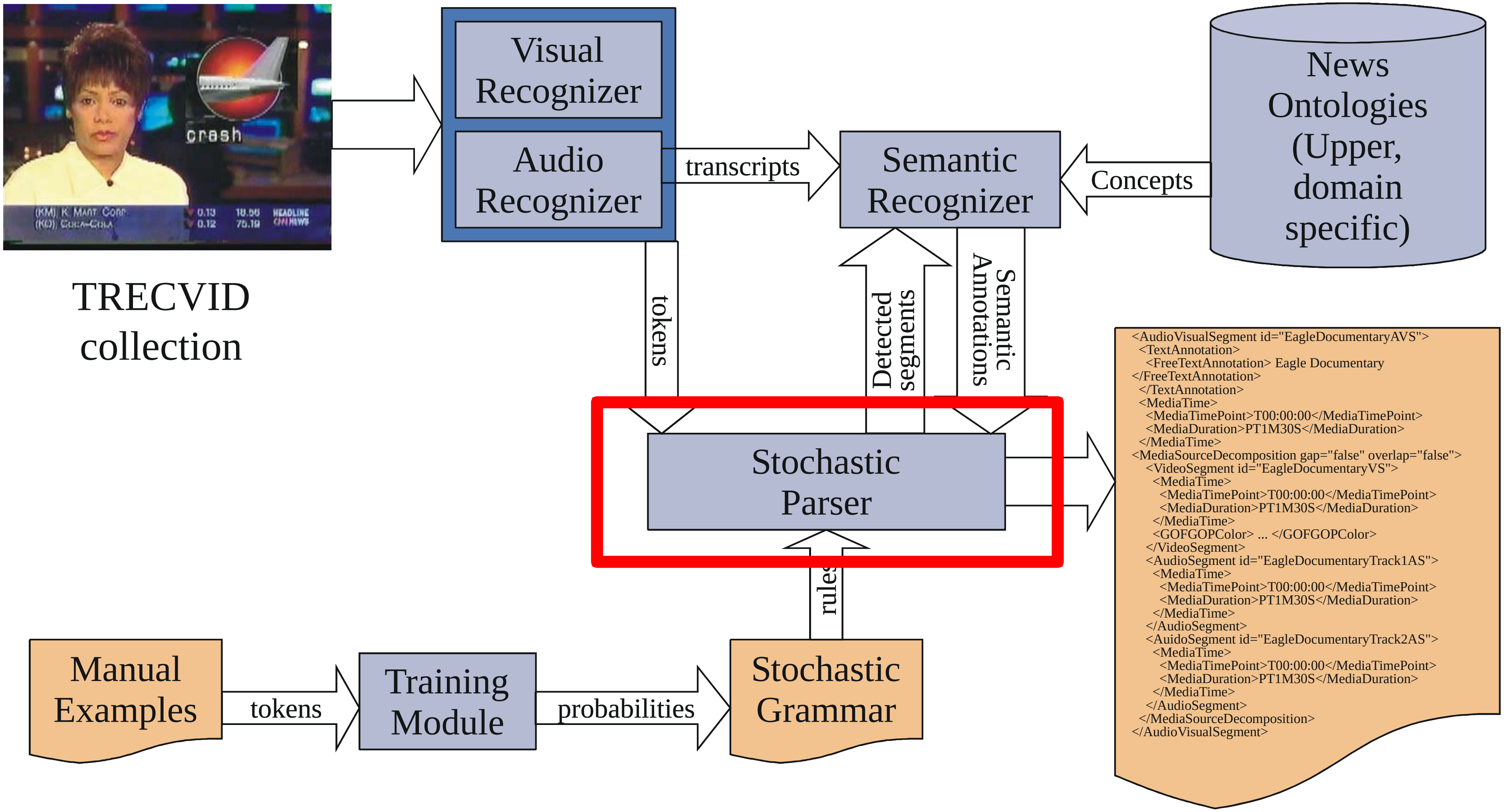
- Motivation
- Goals
- System Architecture
  - Stochastic parser
  - Audio-/visual recognizers
  - Semantic recognizer
- Plans for JPA3

System architecture



```
<AudioVisualSegment id="EagleDocumentaryAVS">
  <TextAnnotation>
    <FreeTextAnnotation> Eagle Documentary
</FreeTextAnnotation>
  </TextAnnotation>
  <MediaTime>
    <MediaTimePoint>T00:00:00</MediaTimePoint>
    <MediaDuration>PT1M30S</MediaDuration>
  </MediaTime>
<MediaSourceDecomposition gap="false" overlap="false">
    <VideoSegment id="EagleDocumentaryVS">
      <MediaTime>
        <MediaTimePoint>T00:00:00</MediaTimePoint>
        <MediaDuration>PT1M30S</MediaDuration>
      </MediaTime>
      <GOFGOPColor> ... </GOFGOPColor>
    </VideoSegment>
    <AudioSegment id="EagleDocumentaryTrack1AS">
      <MediaTime>
        <MediaTimePoint>T00:00:00</MediaTimePoint>
        <MediaDuration>PT1M30S</MediaDuration>
      </MediaTime>
    </AudioSegment>
    <AuidoSegment id="EagleDocumentaryTrack2AS">
      <MediaTime>
        <MediaTimePoint>T00:00:00</MediaTimePoint>
        <MediaDuration>PT1M30S</MediaDuration>
      </MediaTime>
    </AudioSegment>
  </MediaSourceDecomposition>
</AudioVisualSegment>
```

System architecture



```
<AudioVisualSegment id="EagleDocumentaryAVS">
  <TextAnnotation>
    <FreeTextAnnotation> Eagle Documentary
</FreeTextAnnotation>
  </TextAnnotation>
  <MediaTime>
    <MediaTimePoint>T00:00:00</MediaTimePoint>
    <MediaDuration>PT1M30S</MediaDuration>
  </MediaTime>
<MediaSourceDecomposition gap="false" overlap="false">
    <VideoSegment id="EagleDocumentaryVS">
      <MediaTime>
        <MediaTimePoint>T00:00:00</MediaTimePoint>
        <MediaDuration>PT1M30S</MediaDuration>
      </MediaTime>
      <GOFGOPColor> ... </GOFGOPColor>
    </VideoSegment>
    <AudioSegment id="EagleDocumentaryTrack1AS">
      <MediaTime>
        <MediaTimePoint>T00:00:00</MediaTimePoint>
        <MediaDuration>PT1M30S</MediaDuration>
      </MediaTime>
    </AudioSegment>
    <AuidoSegment id="EagleDocumentaryTrack2AS">
      <MediaTime>
        <MediaTimePoint>T00:00:00</MediaTimePoint>
        <MediaDuration>PT1M30S</MediaDuration>
      </MediaTime>
    </AudioSegment>
  </MediaSourceDecomposition>
</AudioVisualSegment>
```

Stochastic parser

- News formats can be modeled using a context-free grammar
  - Terminals correspond to audio-/visual structuring elements (here “tokens”)
- But: Token detection may be erroneous
- Some rules are more probable than others

=> Use of a stochastic context-free grammar

- Robust against misdetections

Stochastic parser (2)

- (Manually) created a grammar for “CNN Headline News”
- including identification of 21 structural tokens

```
Broadcast --> Intro Stories Weather Stories Misc
    Sports Stories PreviewPresentation
Intro --> MainTitle HeadlinesPresentation
Intro --> MainTitle
Intro --> HeadlinesPresentation
Stories --> Story
Stories --> Story Stories
Story -->Presentation
Weather --> USMap IslandMap ExtendedForecast
USMap --> TemperatureMap PressureMap
USMap --> PressureMap TemperatureMap
Misc --> PreviewPresentation ComingUpYourHealth
    Sponsored_Commercials TopStories
    DollarsAndSense Commercials
ComingUpYourHealth --> ComingUp YourHealth
ComingUpYourHealth --> YourHealth ComingUp
ComingUpYourHealth --> ComingUp
YourHealth --> YourHealthScreen YourHealthScreen
Sponsored_Commercials --> TechTrendsSponsor
Sponsored_Commercials --> Commercials
Commercials --> CommercialsIntro BlackFrames
TopStories --> TopStoriesIntro Presentation
DollarsAndSense --> DollarsAndSenseIntro
    DollarPresentation Presentation
DollarsAndSense --> DollasAndSenseIntro
    Presentation Presentation
Sports --> SportsBlock PlayOfTheDayblock
    Commercials
SportsBlock --> SportsIntro SportsOutro
PlayOfTheDayblock --> PlayOfTheDayIntro
    PlayOfTheDayOutro
```

Stochastic parser (3)

- The grammar is trained with manually created token sequences

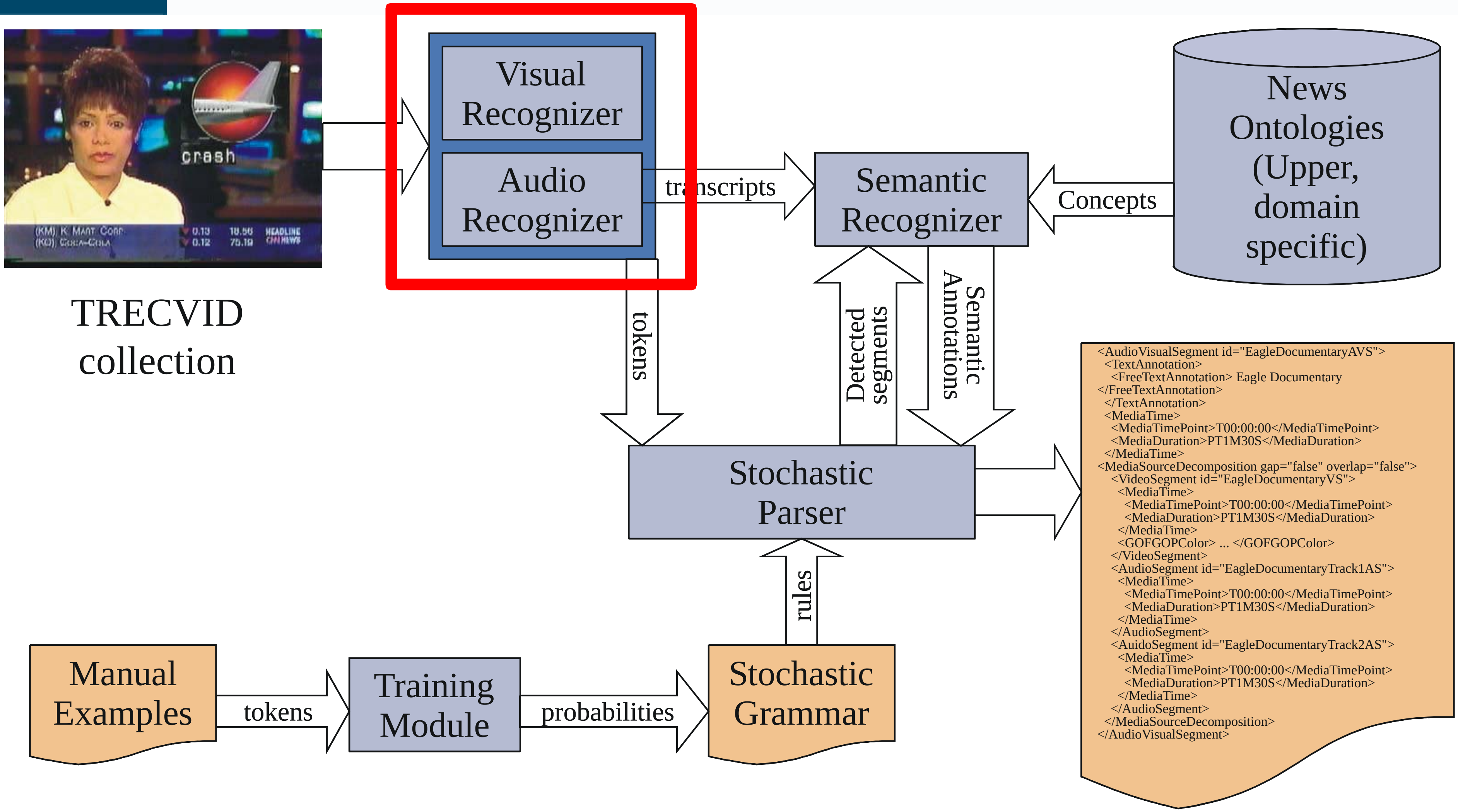
```
Intro --> MainTitle HeadlinesPresentation[0.8]
Intro --> MainTitle                      [0.2]
```

- When presented a sequence of detected tokens, the parser finds the most probable generating tree

Stochastic parser (4)

- Implementation
  - Almost finished
  - Based on the *JavaChart* open source parser (http://nlpfarm.sourceforge.net/javachart/) that has been extended to handle probabilistic grammar rules

System architecture

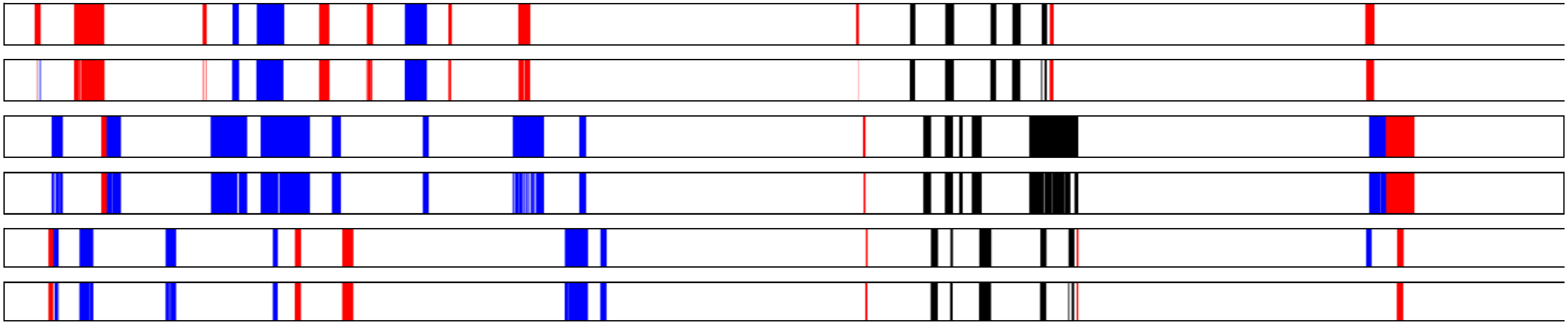


```
<AudioVisualSegment id="EagleDocumentaryAVS">
  <TextAnnotation>
    <FreeTextAnnotation> Eagle Documentary
</FreeTextAnnotation>
  </TextAnnotation>
  <MediaTime>
    <MediaTimePoint>T00:00:00</MediaTimePoint>
    <MediaDuration>PT1M30S</MediaDuration>
  </MediaTime>
<MediaSourceDecomposition gap="false" overlap="false">
    <VideoSegment id="EagleDocumentaryVS">
      <MediaTime>
        <MediaTimePoint>T00:00:00</MediaTimePoint>
        <MediaDuration>PT1M30S</MediaDuration>
      </MediaTime>
      <GOFGOPColor> ... </GOFGOPColor>
    </VideoSegment>
    <AudioSegment id="EagleDocumentaryTrack1AS">
      <MediaTime>
        <MediaTimePoint>T00:00:00</MediaTimePoint>
        <MediaDuration>PT1M30S</MediaDuration>
      </MediaTime>
    </AudioSegment>
    <AuidoSegment id="EagleDocumentaryTrack2AS">
      <MediaTime>
        <MediaTimePoint>T00:00:00</MediaTimePoint>
        <MediaDuration>PT1M30S</MediaDuration>
      </MediaTime>
    </AudioSegment>
  </MediaSourceDecomposition>
</AudioVisualSegment>
```

Visual recognizer

- Two modules currently:
  - Unsupervised anchor shot detection
    - Anchor shots differ between instances of one series
    - Anchor shots are very similar inside one instance
    - Only parameter: Nr. of expected anchor shot types
  - Supervised token classification
    - Visual tokens may be shown only once in an instance
    - Tokens are very similar between instances
    - Supervised training of token models

Anchor shot detection



Average total error: 1,09%
Average error regarding presentations: 8,42%

Token classification

- Three stage process:
  - Classify single frames, based on
    - Visual characteristics (color, texture)
    - Eigenface features
  - Do relaxation labeling on frames with temporal coherence constraint
  - Do relaxation labeling on contiguous segments using model-based constraints

Token classification (2)

- Relaxation labeling constraints based on the following rules:
  - After the “Intro” there is always a “Presentation”.
  - After a “Presentation” there may follow a “Map”, a “Report”, or the “Credits”, where “Presentation” followed by “Report” occurs less often.
  - A “Map” is always followed by a “Report”.
  - After a “Report”, there are always the “Credits”.
  - The “Credits” are always followed by a “Presentation”.

	“Intro”	“Presentation”	“Map”	“Report”	“Credits”
“Intro”	0.5	0.5	-1	-1	-1
“Presentation”	-1	0.5	0.5	0.25	0.5
“Map”	-1	-1	0.5	0.5	-1
“Report”	-1	-1	-1	0.5	0.5
“Credits”	-1	0.5	-1	-1	-1

Class	Precision	Recall
“Intro”	1	1
“Presentation”	0.958	0.953
“Map”	0.861	0.912
“Report”	0.993	1
“Credits”	1	1

Audio recognizer

- Speech-/Non-speech segmentation
- Speaker segmentation
- Speaker clustering
  - Will be used for token detection
- Keyword spotting
  - Keyword selection based on the Context-of-Capture model
  - Used for topic categorization

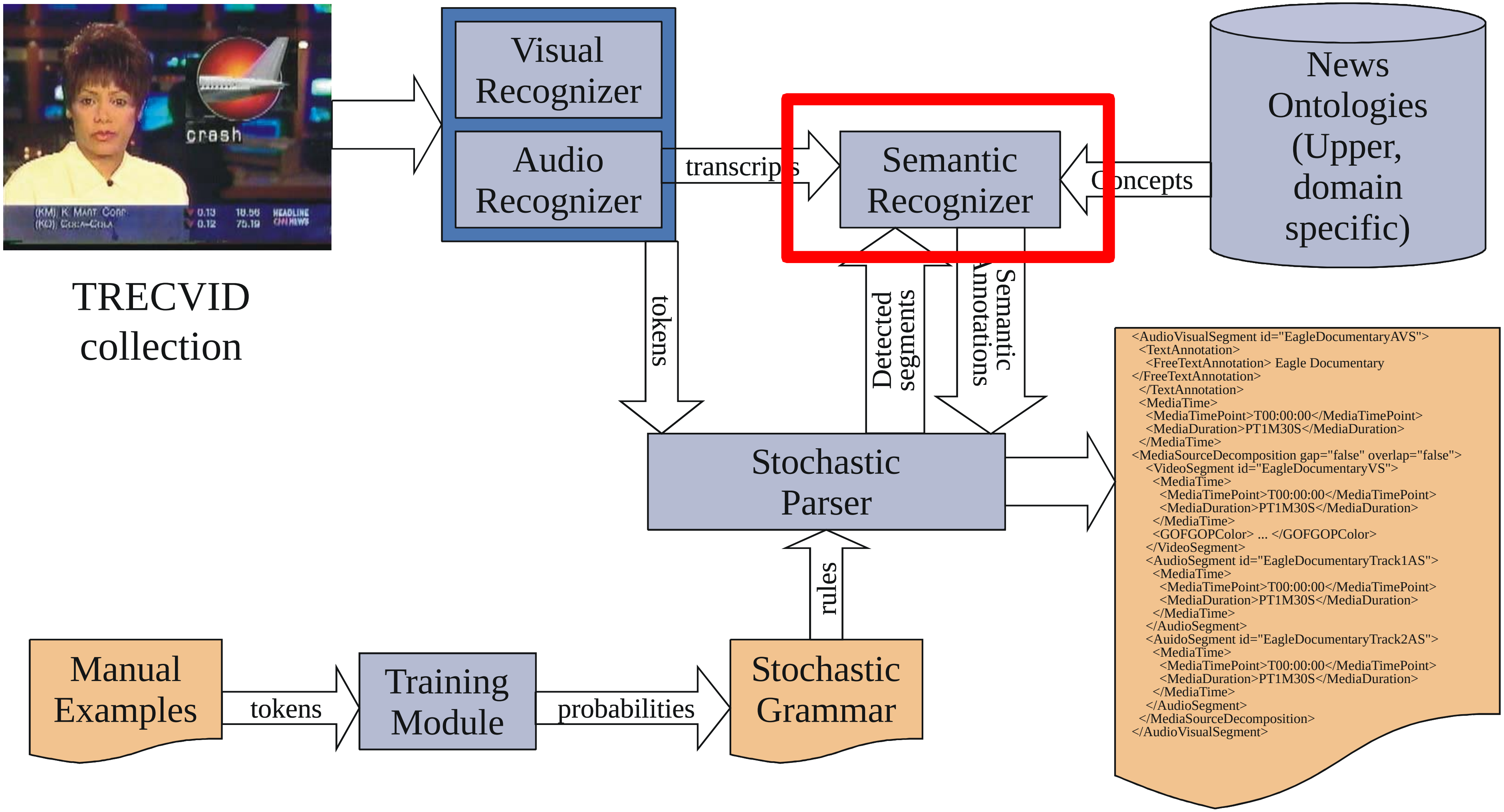
Audio recognizer (2)

- Work in progress:
  - Combination of the robustness of word-based speech recognition with the flexibility of syllable-based speech recognition (syllable-based recognition is not constrained to a pre-defined vocabulary)
  - Determination of the optimal balance between reliance on phonotactic information and reliance on acoustic information for keyword spotting in challenging audio conditions

Audio recognizer (3)

- Work in progress (2):
  - Experimentation with compositions of keyword clusters to detect topics (appreciable advantages are to be gained from creating clusters that include longer keywords, which can be robustly recognized)
  - Experimentation with different keyword lists, facilitated by a web interface to a server-based keyword spotter

System architecture



TRECVID collection

Visual Recognizer

Audio Recognizer

Semantic Recognizer

News Ontologies (Upper, domain specific)

Stochastic Parser

Manual Examples

Training Module

Stochastic Grammar

```
<AudioVisualSegment id="EagleDocumentaryAVS">
  <TextAnnotation>
    <FreeTextAnnotation> Eagle Documentary
</FreeTextAnnotation>
  </TextAnnotation>
  <MediaTime>
    <MediaTimePoint>T00:00:00</MediaTimePoint>
    <MediaDuration>PT1M30S</MediaDuration>
  </MediaTime>
<MediaSourceDecomposition gap="false" overlap="false">
    <VideoSegment id="EagleDocumentaryVS">
      <MediaTime>
        <MediaTimePoint>T00:00:00</MediaTimePoint>
        <MediaDuration>PT1M30S</MediaDuration>
      </MediaTime>
      <GOFGOPColor> ... </GOFGOPColor>
    </VideoSegment>
    <AudioSegment id="EagleDocumentaryTrack1AS">
      <MediaTime>
        <MediaTimePoint>T00:00:00</MediaTimePoint>
        <MediaDuration>PT1M30S</MediaDuration>
      </MediaTime>
    </AudioSegment>
    <AuidoSegment id="EagleDocumentaryTrack2AS">
      <MediaTime>
        <MediaTimePoint>T00:00:00</MediaTimePoint>
        <MediaDuration>PT1M30S</MediaDuration>
      </MediaTime>
    </AudioSegment>
  </MediaSourceDecomposition>
</AudioVisualSegment>
```

Semantic recognizer

- Receives the generating tree from the parser
- Semantically annotates news segments
  - News story categorization:
    Sports, Weather, Politics, Economics, Social
- Upper ontology based on News-ML
  - Using OWL and an OWL/MPEG-7 interoperability framework
- Specialized domain ontologies for different topic classes

Semantic recognizer (2)

- But: News story boundaries do not necessarily coincide with boundaries of parsed segments
  - Topic change can then only be detected by textual means
  - Need for topic segmentation based on lexical cohesion

Semantic recognizer (3)

- Current implementation
  - Topic modeling based on lexical chains
  - Exploitation of news ontologies and semantic relationships of WordNet.
  - News story segmentation based on lexical chaining of the news telecast text.

Overview

- Motivation
- Goals
- System Architecture
  - Stochastic parser
  - Audio-/visual recognizers
  - Semantic recognizer
- Plans for JPA3

Plans for JPA3

- Provide browsing and access capabilities based on analysis results
- Detection of non-structuring tokens
  - Interview, debate, correspondent
- Account for more topic classes
- Automatically identify token classes
- Enhancement of keyword spotting
- Large-vocabulary speech recognition

Questions?