From Conceptual Model to Data Model in Multimedia & Multimodal Corpus Integration

The Institute of Linguistics
The Chinese Academy of Social Sciences
Yeuguo.gu@gmail.com
zhyongwei@qq.com

Main Headings

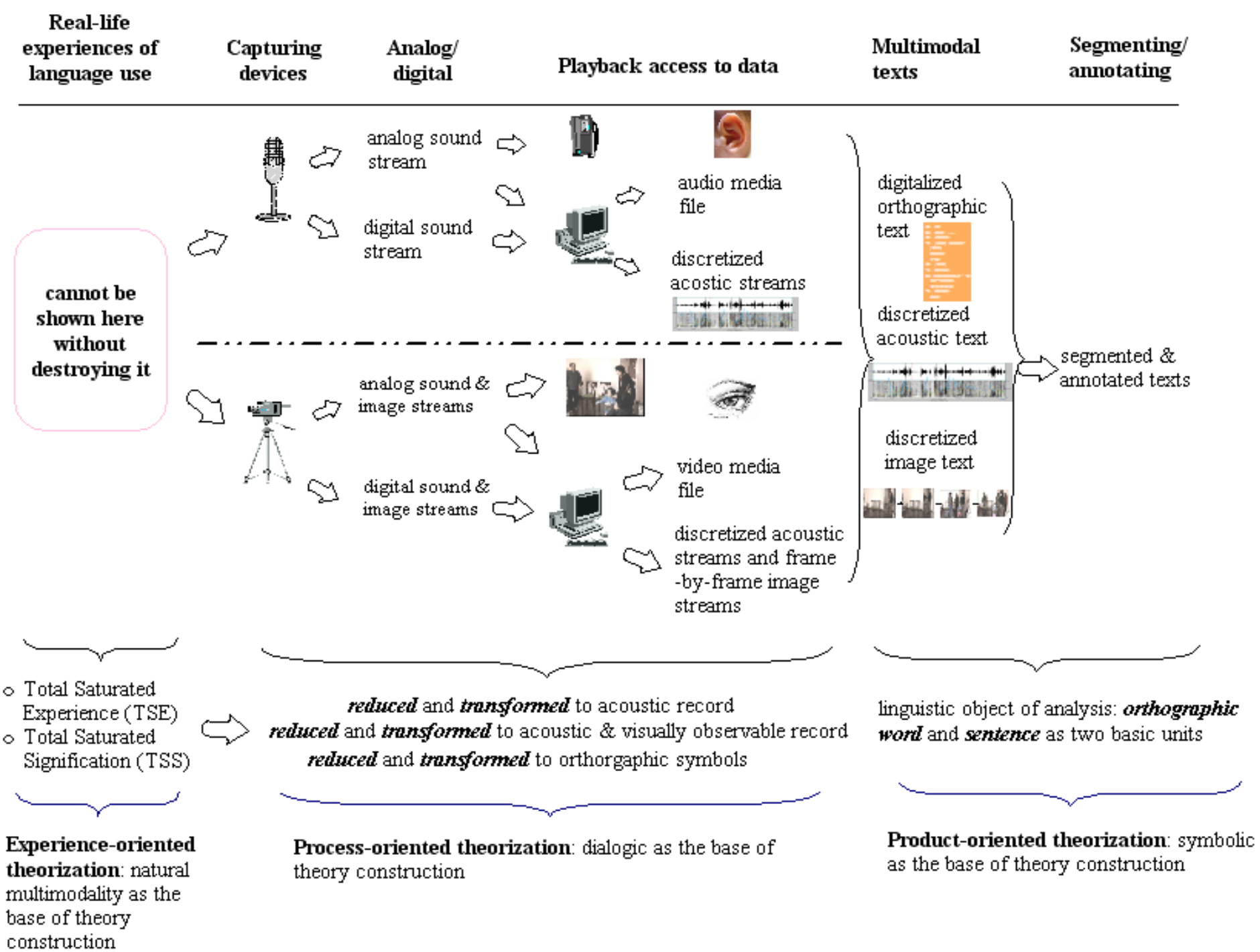
- What is a multimedia corpus?
- What is a multimodal corpus?
- An introduction to Spoken Chinese Corpus of Situated Discourse (SCCSD for short)
- How to do modeling the real-life activity recorded in multimedia & multimodal corpora
- Two application examples: two research projects which we are undergoing based on SCCSD

What is a multimedia corpus?

- Corpus
  - A collection of writings, conversations, speeches, etc., that people use to study and describe a language
- In view of media, the content of a corpus can be recorded by orthographic texts, audio streams, static images, video streams and other media files.
- A corpus which contains orthographic texts, audio files, static images, video files and other media files can be regarded as a multimedia corpus.

What is a multimodal corpus?

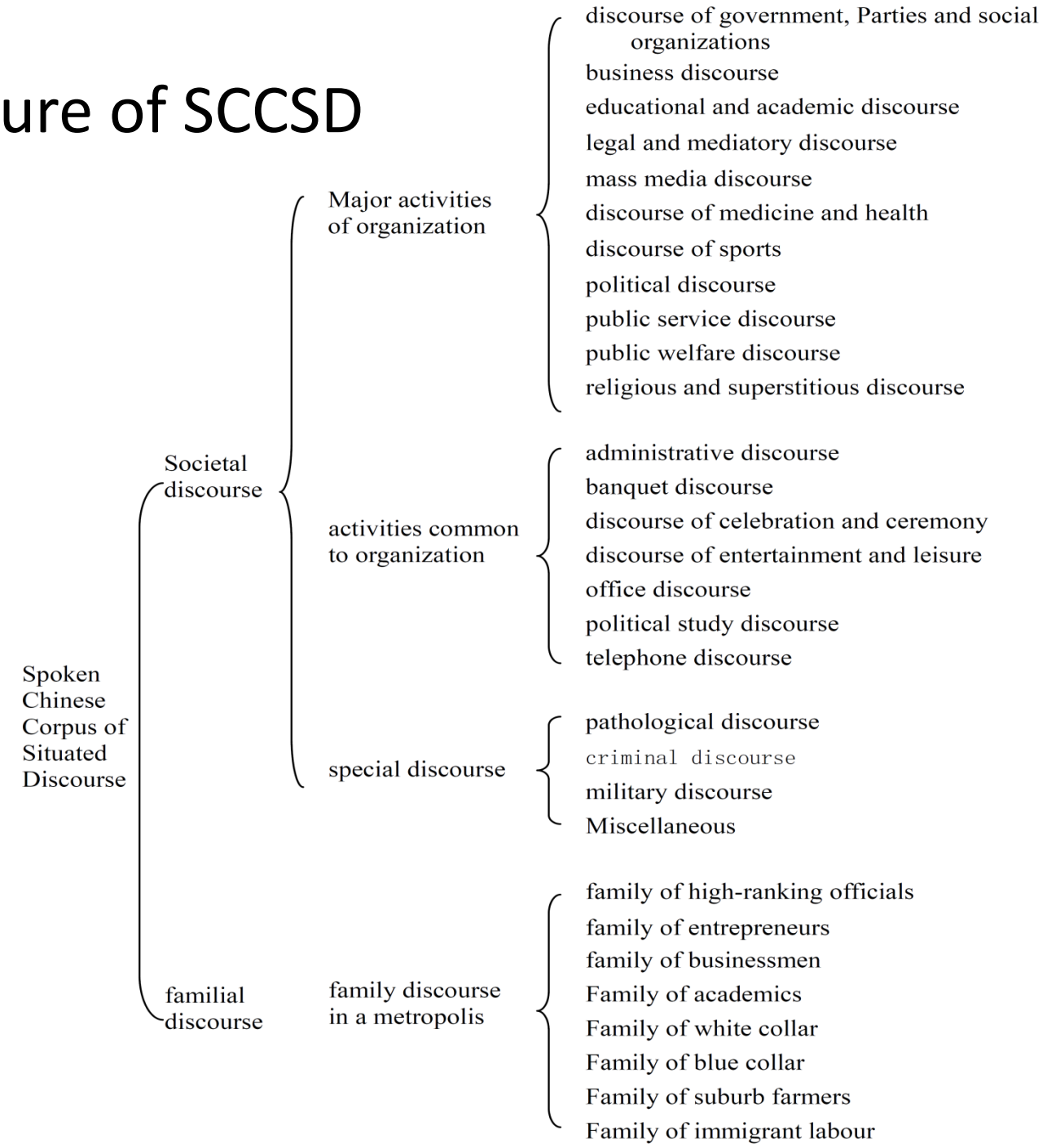
- We communicate not only via verbal language, but also through our use of intonation, gaze, hand gestures, body gestures, and facial expressions. (Gibbon D, Mertins I, Moore R. 2000)
- Each modality is one way of communication between humans.
- Communication between humans uses many modalities.
- A corpus which is annotated by more than one communicate modality can be regarded as a multimodal corpus.
- Real-life experiences of language use is Total Saturated Experience (TSE for short)


Real-life experiences of language use
Capturing devices
Analog/ digital
Playback access to data
Multimodal texts
Segmenting/ annotating
cannot be shown here without destroying it
analog sound stream
digital sound stream
audio media file
discretized acostic streams
digitalized orthographic text
discretized acoustic text
analog sound & image streams
digital sound & image streams
video media file
discretized acoustic streams and frame -by-frame image streams
discretized image text
segmented & annotated texts
Total Saturated Experience (TSE)
Total Saturated Signification (TSS)
reduced and transformed to acoustic record
reduced and transformed to acoustic & visually observable record
reduced and transformed to orthorgaphic symbols
linguistic object of analysis: orthographic word and sentence as two basic units
Experience-oriented theorization: natural multimodality as the base of theory construction
Process-oriented theorization: dialogic as the base of theory construction
Product-oriented theorization: symbolic as the base of theory construction

An introduction to SCCSD

- **S**poken **C**hinese **C**orpus of **S**ituated **D**iscourse
  - SCCSD for short
  - Gu Yueguo's group has spent more than 20 years on building SCCSD
  - It was first trialed in 1993
  - It was started from 1998 until today
  - Contains 1,000 hours audio records(WAV format, stored in 1,000 CDs);
  - Contains 1,000 hours video records (MPG format, stored in 1,000 DVDs)
  - More than 18 million words transcription

The architecture of SCCSD



Research Methodology

- In View of Multimodal corpus Research, we use simulative modeling as our research methodology
- Two Questions:
  - What is a model?
  - What is Simulative modeling?

What is a model?

- Model
  - Productive model
    - The object doesn’t exist;
    - From a concept to a product;
    - What it should be;
    - What function should it has;
  - Simulative model
    - There exists an object;
    - We want to talk about it;
    - We want to understand it;
    - We want to handle it;

What is a model?

- What's the difference between the productive model and the simulative model?
  - The productive model: to model an object that does not exist at the time of modeling;
  - The simulative model: to model an object or phenomenon that exists already at the time of modeling

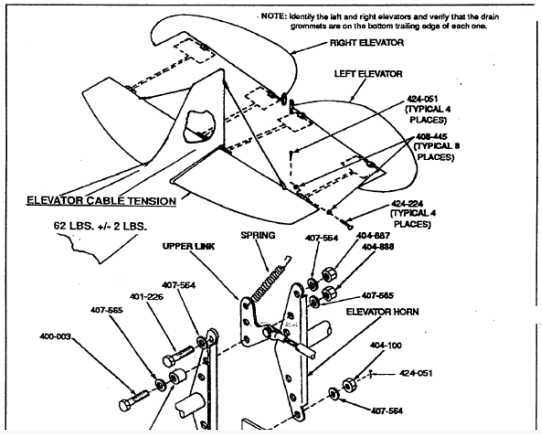
PIPER AIRCRAFT CORPORATION
LOCK HAVEN, PENNA.

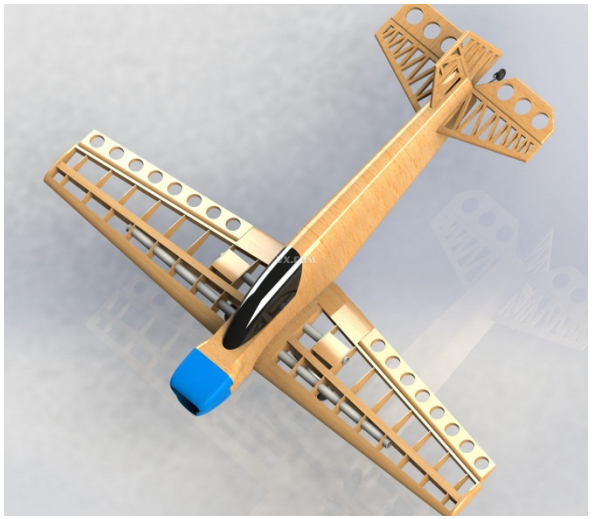
REPORT 699
PAGE 1
MODEL L-21A

SPECIFICATION FOR AIRPLANE
U. S. AIR FORCE MODEL L-21A
PIPER MODEL PA-18 (MODIFIED)
SHORT RANGE OBSERVATION
LIAISON AIRPLANE
CONTRACT NO. AF33(038)-24719

DATE: 24 APRIL 1951.

PAGE 8 OF 11 SERVICE BULLETIN NO. 966





Productive model

Simulative model

In our research project

- we are concerned with simulative modeling;
- the object or phenomenon already exists.
- Take, for example, Alzheimer's disease patients' discourse.
- If it is the phenomenon we want to study by way of analyzing it, we should use simulative modeling as our research methodology.

Simulative modeling

– Three steps:
  - conceptual modeling
  - data modeling
  - Implementation and verification

Conceptual modeling

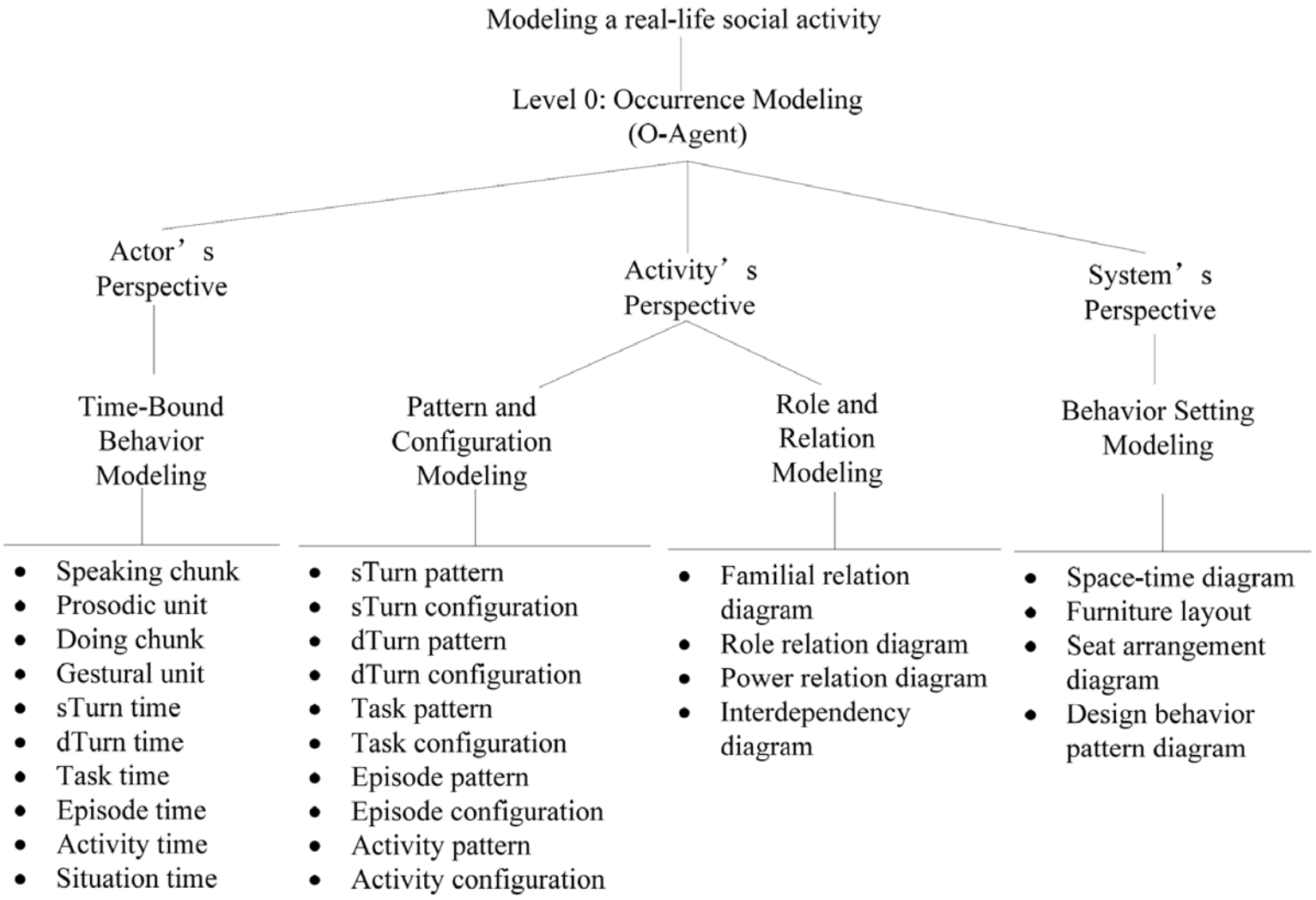
- Basic Principle: Multiple-perspective
- It is impossible to describe the whole activity at one time.
- Each perspective represents a particular view of the activity what we are concerned at one time.
- Multiple-perspective may simulate the whole view of the real-life activity

Multiple-perspective

There is statue standing on a stone.

- The linguistic behavior is modeled from a range of perspectives

The conceptual model of real-life social activity



Yueguo Gu 2009:453-6

Data modeling

- What does Data modeling do?
  - Build the data model according to conceptual model.
  - Convert the understanding of the phenomenon (Conceptual model of real-life activity) to data which can be stored in computer and can be used to retrieval and do statistics.
- Basic Principle: Multiple-layer (According to Multiple-perspective)
  - The relationship of perspectives and layers is not only one-to-one mapping, but also one to multi mapping.

Processing model of multimedia corpus

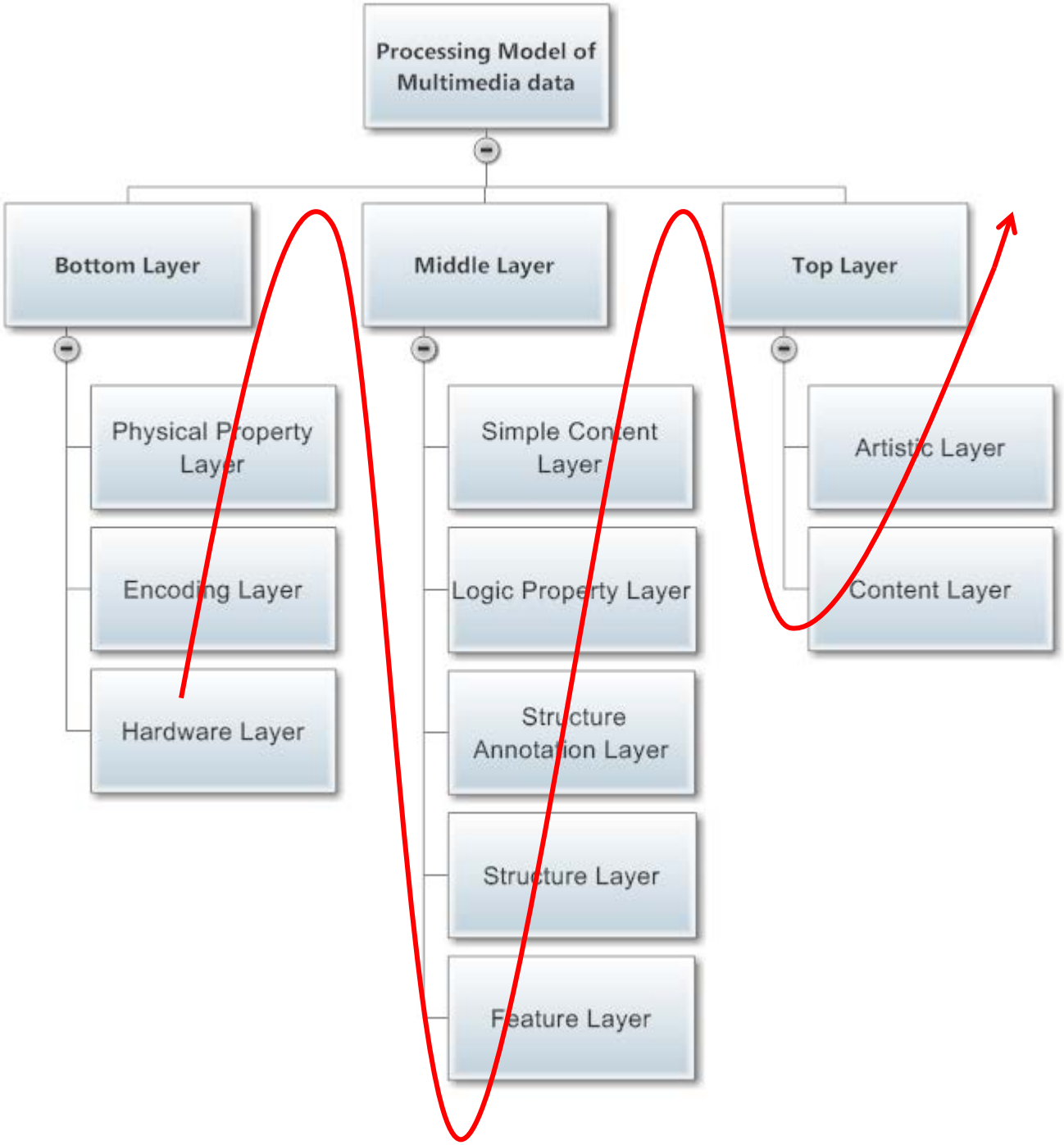
- Use a hierarchical processing model to process multimedia files.
- Generally, the processing model of multimedia corpus can be divided into the bottom layer, middle layer and top layer.
- Among the three layers, the bottom layer is closely related to the computer hardware, so the computer processing is relatively easy.
- The top layer is closely related to the advanced semantics and artistic appreciation, and the computer processing is the most difficult in this part.

Processing model of multimedia corpus

- The middle layer consists of many sub-layers; from the bottom up, the difficulties of conducting the computer processing to each sub-layer increase in sequence, and more and more manual interventions are needed when conducting the corpus segmentation and annotation.
- Different medias have different processing models which can instruct the procedure of processing multimedia data

Take Static Image as An Example

- The bottom layer of the processing model contains hardware layer, encoding layer and physical property layer.
- The middle layer of the processing model contains feature layer, structure layer, structure annotation layer, logic property layer and simple content layer.
- The top layer of the processing model contains content layer and artistic layer.

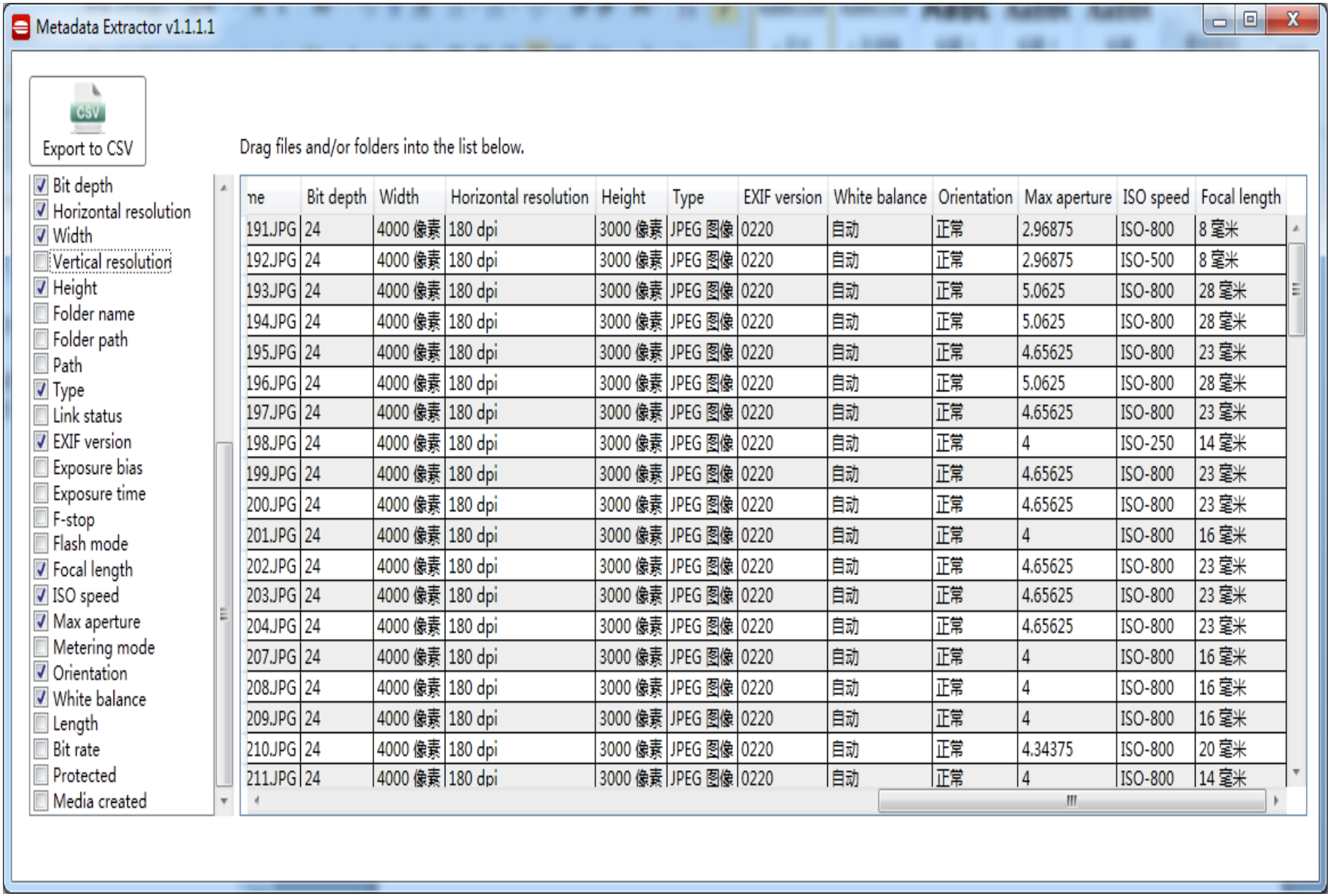

Processing Model of Multimedia data
Bottom Layer
Middle Layer
Top Layer
Physical Property Layer
Encoding Layer
Hardware Layer
Simple Content Layer
Logic Property Layer
Structure Annotation Layer
Structure Layer
Feature Layer
Artistic Layer
Content Layer

Hardware Layer & Encoding Layer

- Hardware layer and encoding layer concern
  - how hardware devices capture the multimedia data
  - how to store the multimedia data
  - How to show the multimedia data
- We don’t concern much about hardware layer and encoding layer

Physical Property Layer

- Describe properties of the media files which are generated by the media file capture devices. (Camera, recorder etc.)
- Take documentary photography (static image) as an example
  - Camera write date and time information, manufacturer information, exposure time, ISO speed information to static image
  - Exif is an standard to record these data, and almost all camera manufacturers use it. (Besides Exif standard, there are also other related standards: XMP、IPTC、JFIF、TIFF)
  - Computer can view and edit these data easily.


Metadata Extractor v1.1.1.1
Export to CSV
Drag files and/or folders into the list below.
Bit depth
Horizontal resolution
Width
Vertical resolution
Height
Folder name
Folder path
Path
Type
Link status
EXIF version
Exposure bias
Exposure time
F-stop
Flash mode
Focal length
ISO speed
Max aperture
Metering mode
Orientation
White balance
Length
Bit rate
Protected
Media created
ne | Bit depth | Width | Horizontal resolution | Height | Type | EXIF version | White balance | Orientation | Max aperture | ISO speed | Focal length
191.JPG | 24 | 4000 像素 | 180 dpi | 3000 像素 | JPEG 图像 | 0220 | 自动 | 正常 | 2.96875 | ISO-800 | 8 毫米
192.JPG | 24 | 4000 像素 | 180 dpi | 3000 像素 | JPEG 图像 | 0220 | 自动 | 正常 | 2.96875 | ISO-500 | 8 毫米
193.JPG | 24 | 4000 像素 | 180 dpi | 3000 像素 | JPEG 图像 | 0220 | 自动 | 正常 | 5.0625 | ISO-800 | 28 毫米
194.JPG | 24 | 4000 像素 | 180 dpi | 3000 像素 | JPEG 图像 | 0220 | 自动 | 正常 | 5.0625 | ISO-800 | 28 毫米
195.JPG | 24 | 4000 像素 | 180 dpi | 3000 像素 | JPEG 图像 | 0220 | 自动 | 正常 | 4.65625 | ISO-800 | 23 毫米
196.JPG | 24 | 4000 像素 | 180 dpi | 3000 像素 | JPEG 图像 | 0220 | 自动 | 正常 | 5.0625 | ISO-800 | 28 毫米
197.JPG | 24 | 4000 像素 | 180 dpi | 3000 像素 | JPEG 图像 | 0220 | 自动 | 正常 | 4.65625 | ISO-800 | 23 毫米
198.JPG | 24 | 4000 像素 | 180 dpi | 3000 像素 | JPEG 图像 | 0220 | 自动 | 正常 | 4 | ISO-250 | 14 毫米
199.JPG | 24 | 4000 像素 | 180 dpi | 3000 像素 | JPEG 图像 | 0220 | 自动 | 正常 | 4.65625 | ISO-800 | 23 毫米
200.JPG | 24 | 4000 像素 | 180 dpi | 3000 像素 | JPEG 图像 | 0220 | 自动 | 正常 | 4.65625 | ISO-800 | 23 毫米
201.JPG | 24 | 4000 像素 | 180 dpi | 3000 像素 | JPEG 图像 | 0220 | 自动 | 正常 | 4 | ISO-800 | 16 毫米
202.JPG | 24 | 4000 像素 | 180 dpi | 3000 像素 | JPEG 图像 | 0220 | 自动 | 正常 | 4.65625 | ISO-800 | 23 毫米
203.JPG | 24 | 4000 像素 | 180 dpi | 3000 像素 | JPEG 图像 | 0220 | 自动 | 正常 | 4.65625 | ISO-800 | 23 毫米
204.JPG | 24 | 4000 像素 | 180 dpi | 3000 像素 | JPEG 图像 | 0220 | 自动 | 正常 | 4.65625 | ISO-800 | 23 毫米
207.JPG | 24 | 4000 像素 | 180 dpi | 3000 像素 | JPEG 图像 | 0220 | 自动 | 正常 | 4 | ISO-800 | 16 毫米
208.JPG | 24 | 4000 像素 | 180 dpi | 3000 像素 | JPEG 图像 | 0220 | 自动 | 正常 | 4 | ISO-800 | 16 毫米
209.JPG | 24 | 4000 像素 | 180 dpi | 3000 像素 | JPEG 图像 | 0220 | 自动 | 正常 | 4 | ISO-800 | 16 毫米
210.JPG | 24 | 4000 像素 | 180 dpi | 3000 像素 | JPEG 图像 | 0220 | 自动 | 正常 | 4.34375 | ISO-800 | 20 毫米
211.JPG | 24 | 4000 像素 | 180 dpi | 3000 像素 | JPEG 图像 | 0220 | 自动 | 正常 | 4 | ISO-800 | 14 毫米

Feature Layer

- Extract features from the multimedia files
- Written texts: Term Frequency (TF), Document Frequency (DF), TFIDF=TF/DF, etc.
- Static Images: Color, Texture, Shape, Scale-Invariant Feature Transform(SIFT) ,etc.
- Audio streams: Waveform, Power, Spectrum Envelope, Linear Predictive Cepstral Coefficient(LPCC), Mel Frequency Cepstral Coefficients(MFCC), etc.
- Video streams: Color, Texture, Shape, Motion, etc.
- Computer can extract most of the features from multimedia files

Structure Layer

- Describe the structure of the media file.
- Segmentation is the basic operation.
- Written texts: paragraphs, sentences, phrases, words, characters.
- Audio streams: phonemes, syllables, discourse units, turns and other meaningful segments of audio streams.

Structure Layer

- Static Images:

Original Image

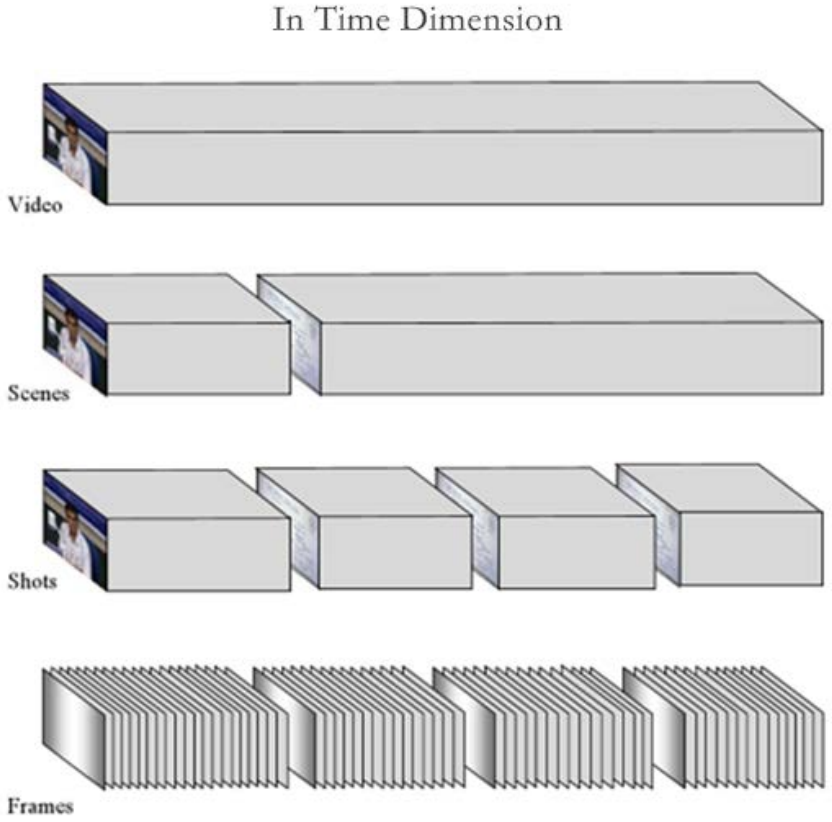
Segment the Image using Specific Graph

Circular, rectangular ellipse, and polygon.

Segment the Image According to the Boundry of The Object In the Image

the purpose and degree of granularity

- Video Streams:



In Space Dimension

Video streams can also be segmented in time and space dimension simultaneously.

Structure Annotation Layer

- Structure annotation layer contains annotations of the segmentation

A = At|As|Ats|An
At = <start,end,Avalue>
An = <key,Avalue>
As = <space_description,Avalue>
Ats = <start,end, space_description, Avalue>
Avalue = Avalue,
Avalue|free_text|structured_text|keyword_text|dependency_text |enum_text

A: Annotation
At: Annotation with time description
As: Annotation with space description
Ats: Annotation with time and space description
An: Annotation without time or space description
start: The start media time of the segment
end: The end media time of the segment
Avalue: Annotation value
space_description: Space Description

free_text: Annotation in free text format
structured_text: Annotation in structured text format
keyword_text: Annotaion in keyword text format
dependency_text : Annotation in dependeny text format
enum_text: Annotation value can be choosed in a set
|: or

Annotations can be categorized into tiers.

T = Tt|Ts|Tts|Tn
Tt = <{At},type,{An}>
Ts= <{As},type,{An}>
Tts = <{Ats},type,{An}>
Tn= <{An},type>

T: Tier, Layer
Tt: Tier with time description
Ts: Tier with space description
Tts: Tier with time description and space description
Tn: Tier without time or space description
At: Annotation with time description
As: Annotation with space description
Ats: Annotation with time and space description
An: Annotation without time or space description
type: Category which this tire belongs to
{X}: Repeat X for many times

R=<{T},{An}>

R: The integration result

Logic Property Layer

- Describe properties of the media file which are **not** generated by the media file capture devices. (Camera, recorder etc.)
- For example:
  - Who generate the media file
  - The copyright of the media file
  - The usage information of the media file etc.

Simple Content Layer & Content Layer

- We separate content layer into a simple one and a normal one.
- In simple content layer, we usually describe the content of the media through WHO, WHAT, WHERE, WHEN, WHY and HOW.
- If there are more information need to describe, we describe them in (normal) content layer.
- Both simple content layer and content layer need human work to annotate.

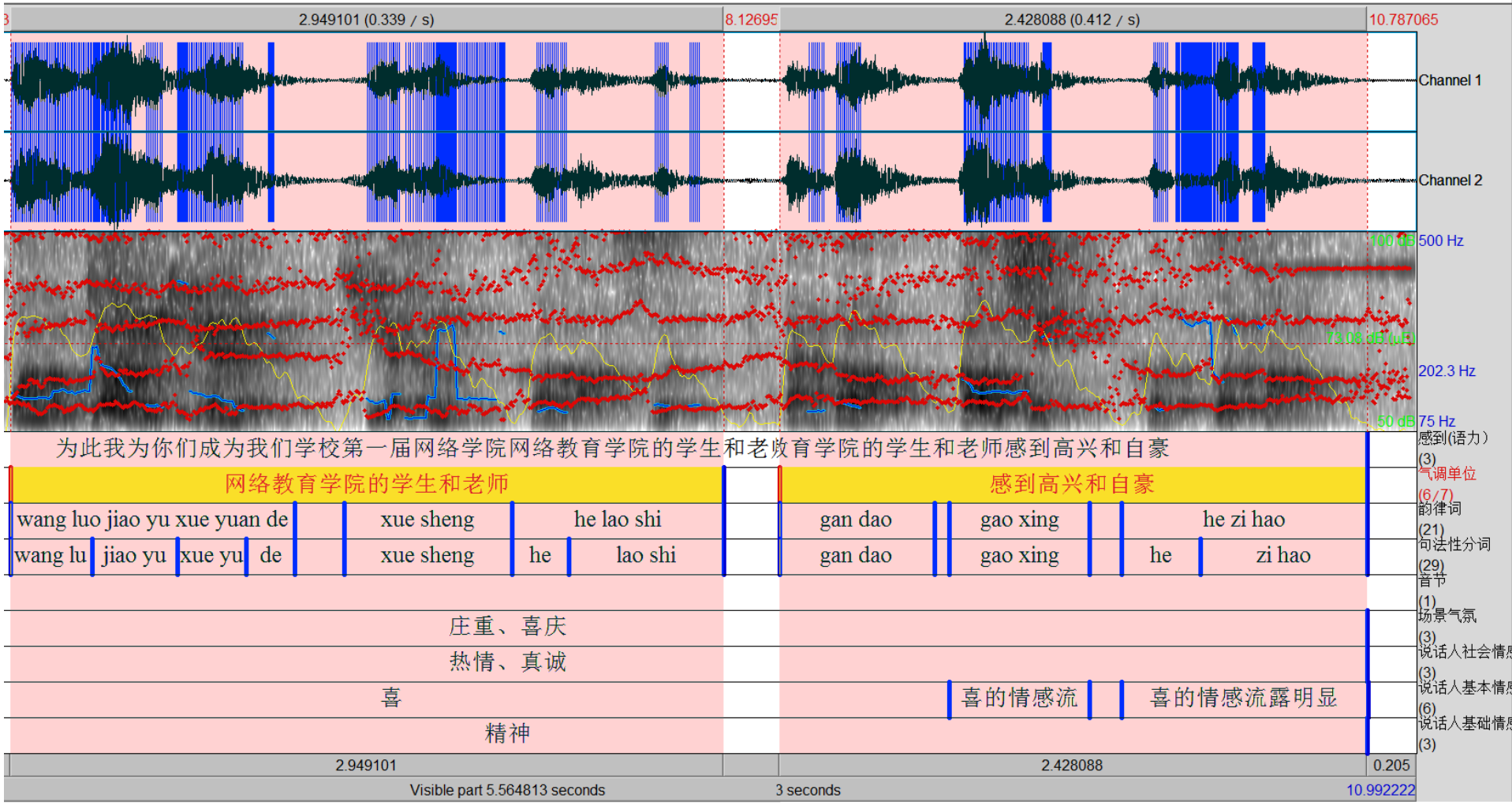
Artistic Layer

- Artistic layer describes the deeper understanding of the media.
- Different people may have different understandings of the media.
- So, it is unpractical to let computer to describe this layer.

Synchronization & Integration

- There are several software using different meta-language to synchronize and integrate different kinds of multimedia files: ELAN, Anvil, C-BAS, EXMARaLDA Editor, MacVisSTa, Transformer, Theme, etc.
- We use MPEG-7 as meta-language to describe different layers' content.
- MPEG-7, formally known as Multimedia Content Description Interface which is an international standard

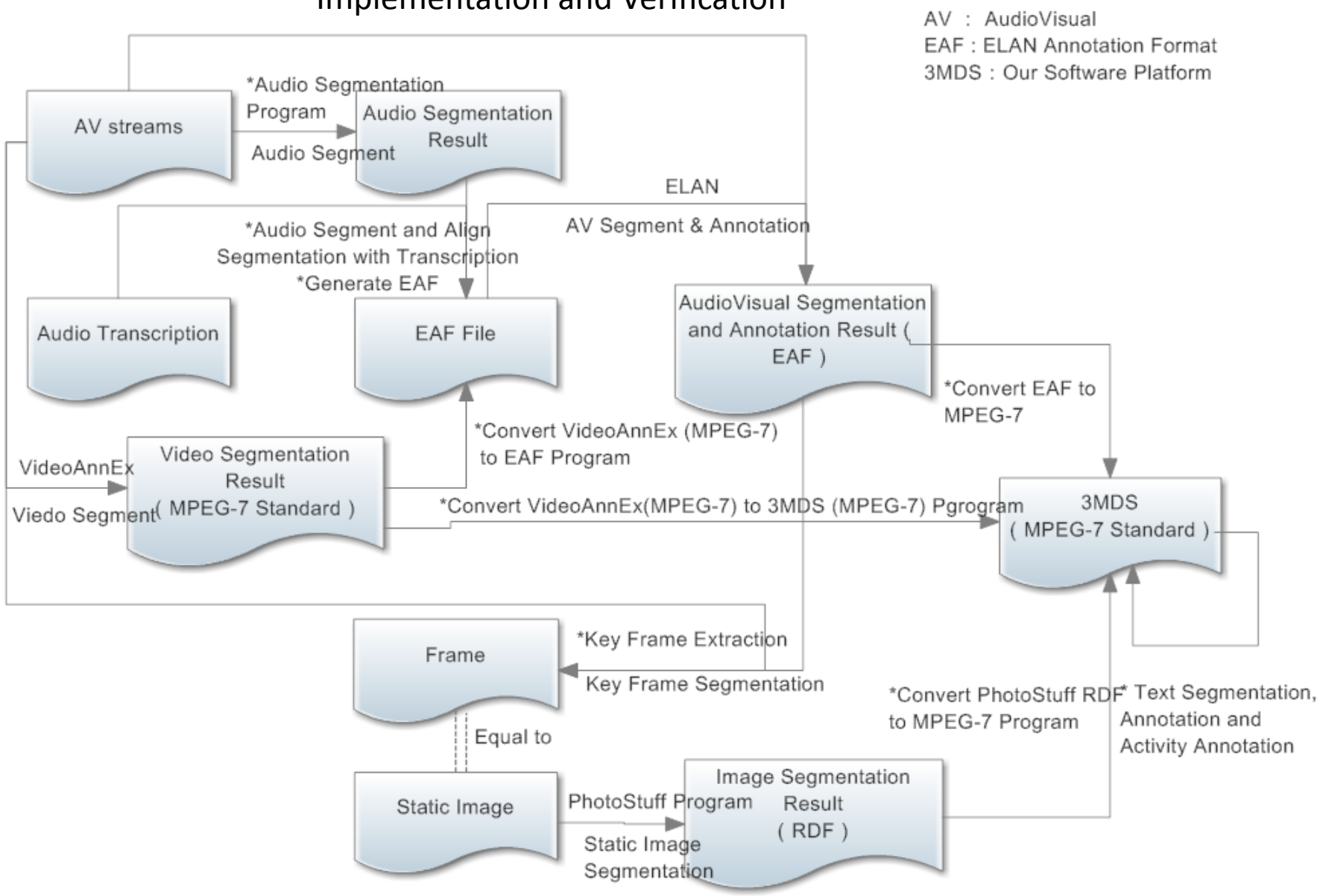
Multi-layered segmentation and annotation



Simulative modeling

– Three steps:
  - conceptual modeling
  - data modeling
  - Implementation and verification

Implementation and Verification



Implementation and Verification

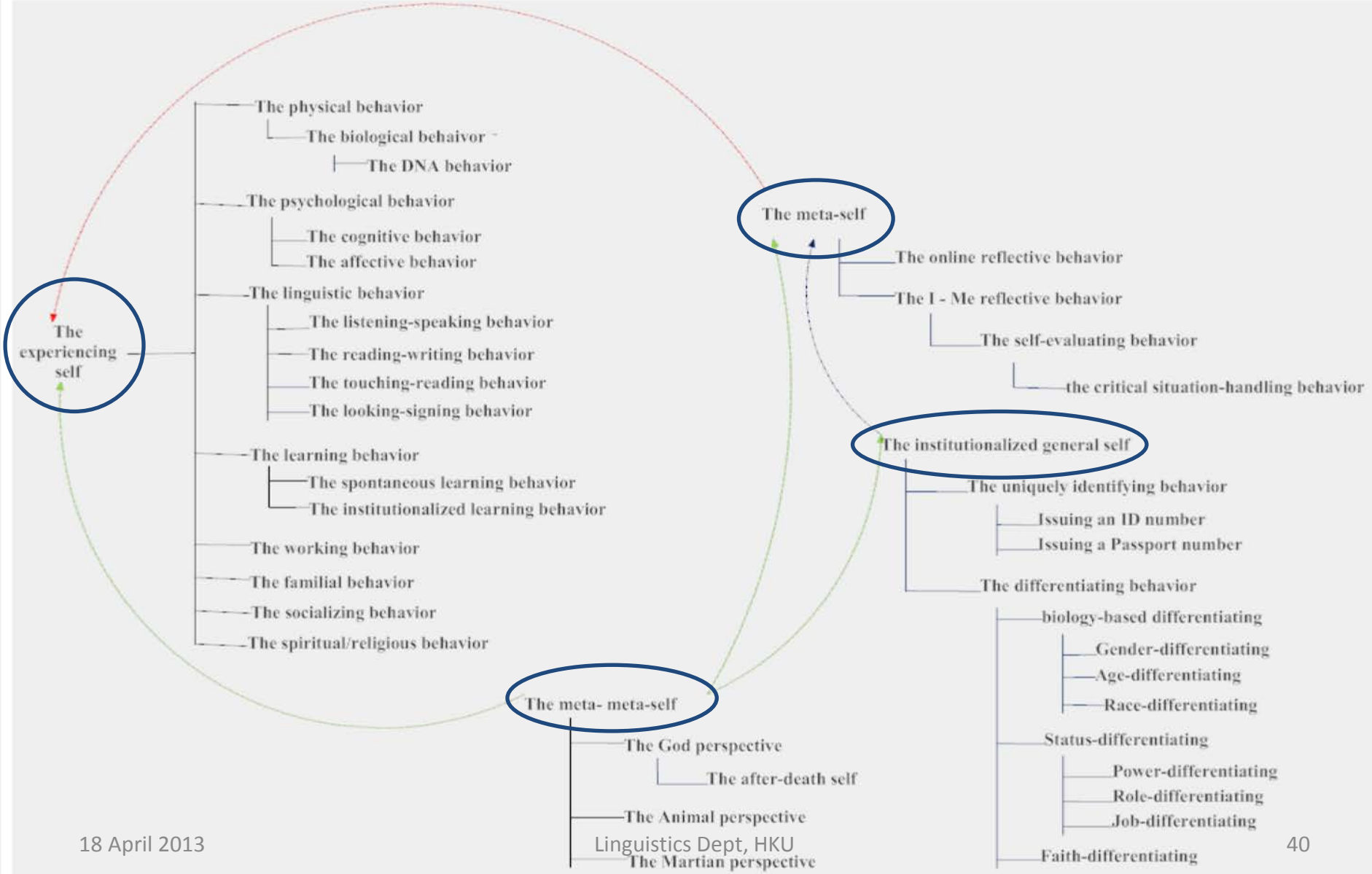
Application Example (1)

- We have s
  Analysis L
  Polytechn
- The under
  difference
  patients' d
  discourse

Conceptual modeling

- Modeling your understanding of the phenomenon;
- Understanding: Dementia affects every aspects of a person's life;
- Gu's (2013) model of human agency

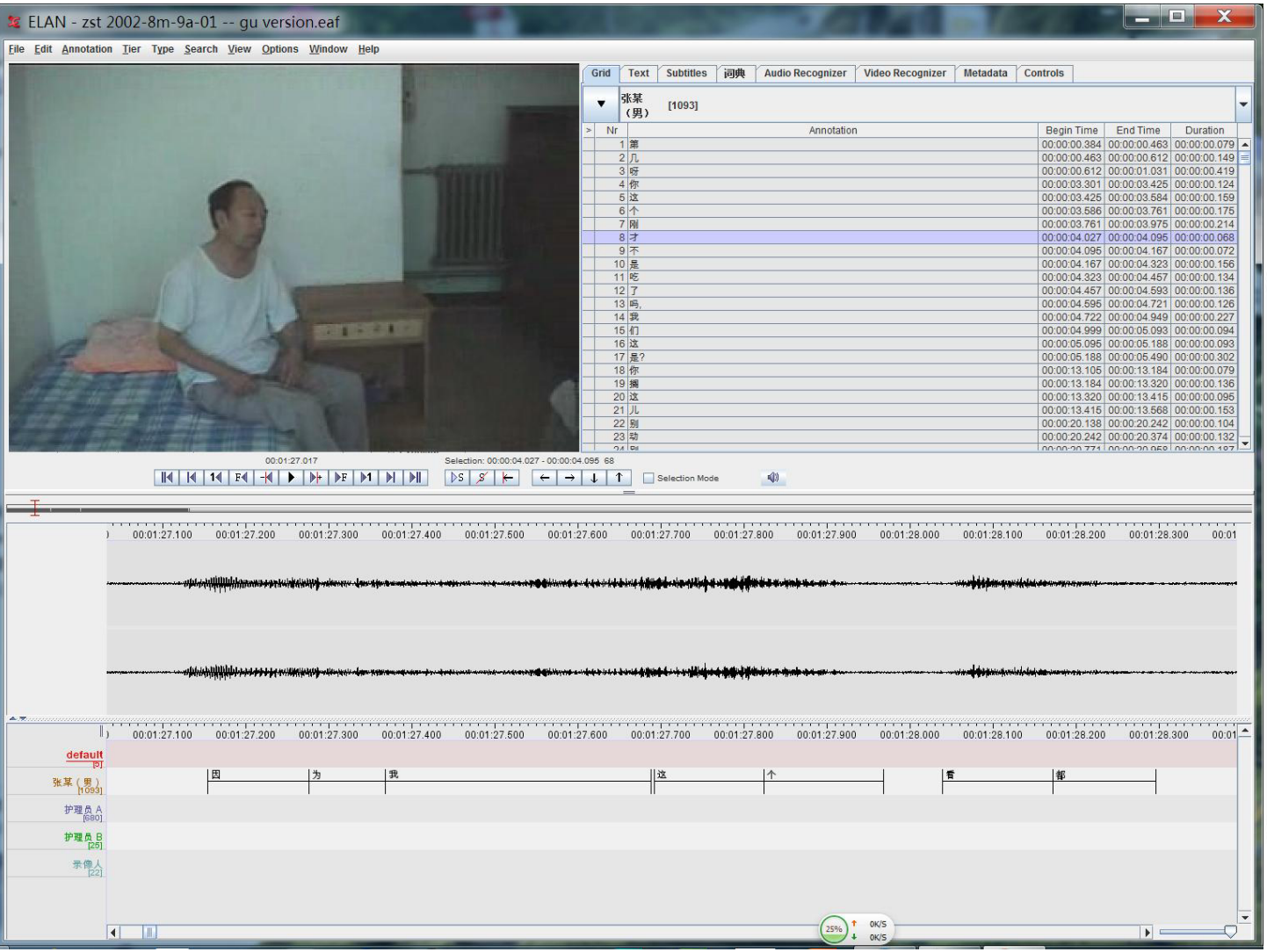
The whole man is modeled with a set of behaviors

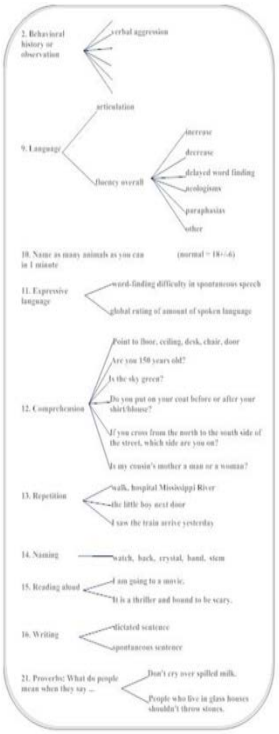


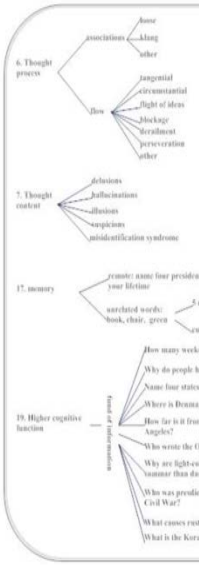
Self-identity lost

This video was captured by a
student of professor Gu Yueguo

A tier each for all perspectives, and parameters







Contrasted parameters

Normal ageing

Alzheimer's disease

- Verbal behavior (audio + video)
  a) Articulation
  b) Fluency
  c) Pragmatics
  d) Discourse coherence
- Paralinguistic behavior (video)
  a) Hand gestures
- Doing-behavior (video)
  a) Orientation
  b) Self-caring
  c) Household chores
- Emotional state (audio + video)

Our hope to contribute

1. Mental state examination is clinically handled through interviews;
2. Information for many parameters depends on the clinician's intuitive judgments made on the spot, and under the time pressure;
3. We use audio-, video-taped data and can look at the data, assisted by the toosl such as Praat and Elan, in the way as physicians look at their data through microscopes.

Ultimate Goal

- Train robots fixed with audio and video sensors to do automatic analysis of audio, and video streams;
- The automated analyzed data helps the clinician make better informed diagnosis.

Application Example (2)

- Intuitively
  - What is said （言）
    - is connected with
  - what is thought of （思）
    - is connected with
  - what is felt （情）
    - is connected with
  - What is embodied（貌）

The STFE-Match Assumption

- There is a perfect match between what is said, what is thought of, what is felt, and what is embodied i.e.
- the STFE-Match Assumption;
- The Assumption is generally upheld in child discourse, but subject to flouting and manipulating in adult discourse.

- The Integrity Person is modeled from four perspectives (i.e., STFE) in three phases, conceptual modeling, data modeling and implementation/critical evaluation.
- The data, mainly from audio and video recordings of everyday activities, are segmented and annotated using ELAN.

- We have collected a lot of audio and video recordings of prisoners.
- It will be helpful to judge whether a person is lying or not.
- We think there must be some other fields in which multimedia & multimodal corpus can be used.

- Brugman H, Russel A. 2004. Annotating multimedia/multi-modal resources with ELAN: Proceedings of the Fourth International Conference on Language Resources and Evaluation. Citeseer.
- Chang, S. F., Sikora, T., & Purl, A. 2001. Overview of the MPEG-7 standard. Circuits and Systems for Video Technology, IEEE Transactions on, 11(6), 688-695.
- Geoffrey Leech, The State of The Art in Corpus Linguistics, 1991, In Aijmar, K. and Altenberg, B. , eds. , English Corpus Linguistics: Studies in Honor of Jan Svartvik, London: Longman, 1991.
- Gibbon D, Mertins I, Moore R. 2000. Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology, and Product Evaluation. Kluwer Academic Publishers.

- Gibbon D, Moore R, Winski R. 1997. Handbook of standards and resources for spoken language systems. Mouton de Gruyter, Berlin.
- Manjunath B S, Salembier P, Sikora T. 2002. Introduction to MPEG-7: multimedia content description interface. John Wiley & Sons Inc.
- Schmidt, T., Duncan, S., Ehmer, O., Hoyt, J., Kipp, M., Loehr, D. Kipp M. et al. 2009. An Exchange Format for Multimodal Annotations. In Multimodal corpora (pp. 207-221). Springer Berlin Heidelberg.
- Yueguo G. 2006. Multimodal text analysis: A corpus linguistic approach to situated discourse. Text & Talk, 26, 2:127-167.

- Gu, Yueguo. 2006. "Sampling situated discourse for spoken Chinese corpus." Retrieved June 15 (2006):
- Yueguo G. 2009. From real-life situated discourse to video-stream data-mining: An argument for agent-oriented modeling for multimodal corpus compilation. International journal of corpus linguistics,14(4):433-466.
- Yueguo G. 2012. A conceptual model for segmenting and annotating a documentary photograph corpus (DPC). In xxx, xxx.
- 顾曰国，2002，北京地区现场即席话语语料库的取样与代表性问题，见中国社会科学院世界经济研究中心编，《全球化与21世纪——首届“中法学术论坛”论文集》。北京：社会科学文献出版社。484-490页。
- 顾曰国，2007，多媒体、多模态学习剖析。《外语电化教学》第2期，3-12页。
- 顾曰国，2011，当代语言学的波形发展主题三:语言、媒介载体与技术。《当代语言学》第01期，22-48页。

Thanks to:

- Aijun Li
- Andrea Deme
- Danqing Liu
- Huba Bartos
- Jianhua Hu
- Katalin Mády
- Kenesei István
- Yueguo Gu
- Zhengda Tang
- The Research Institute of Linguistics of the Hungarian Academy of Sciences
- Institute of Linguistics of Chinese Academy of Social Sciences