LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE

Machine Learning and Open Courseware

Colin de la Higuera cdlh@univ-nantes.fr

Région PAYS DE LA LOIRE

CominLabs

K4A knowledge for all

CO CO

UNIVERSITÉ DE NANTES

ECOLE DES MINES DE NANTES

Outline

- What is open courseware?
- Is there still research to be done?
- Recommendation
- Translation and Transcription
- Annotation
- Navigation
- Traces

CC BY

Opencourseware=Open Educational Resources

- Organization: OCWC
  - 30 000+ Modules
  - 280+ Organizations
  - 40 Countries
  - 29 Languages

CC BY

Some reasons (1)

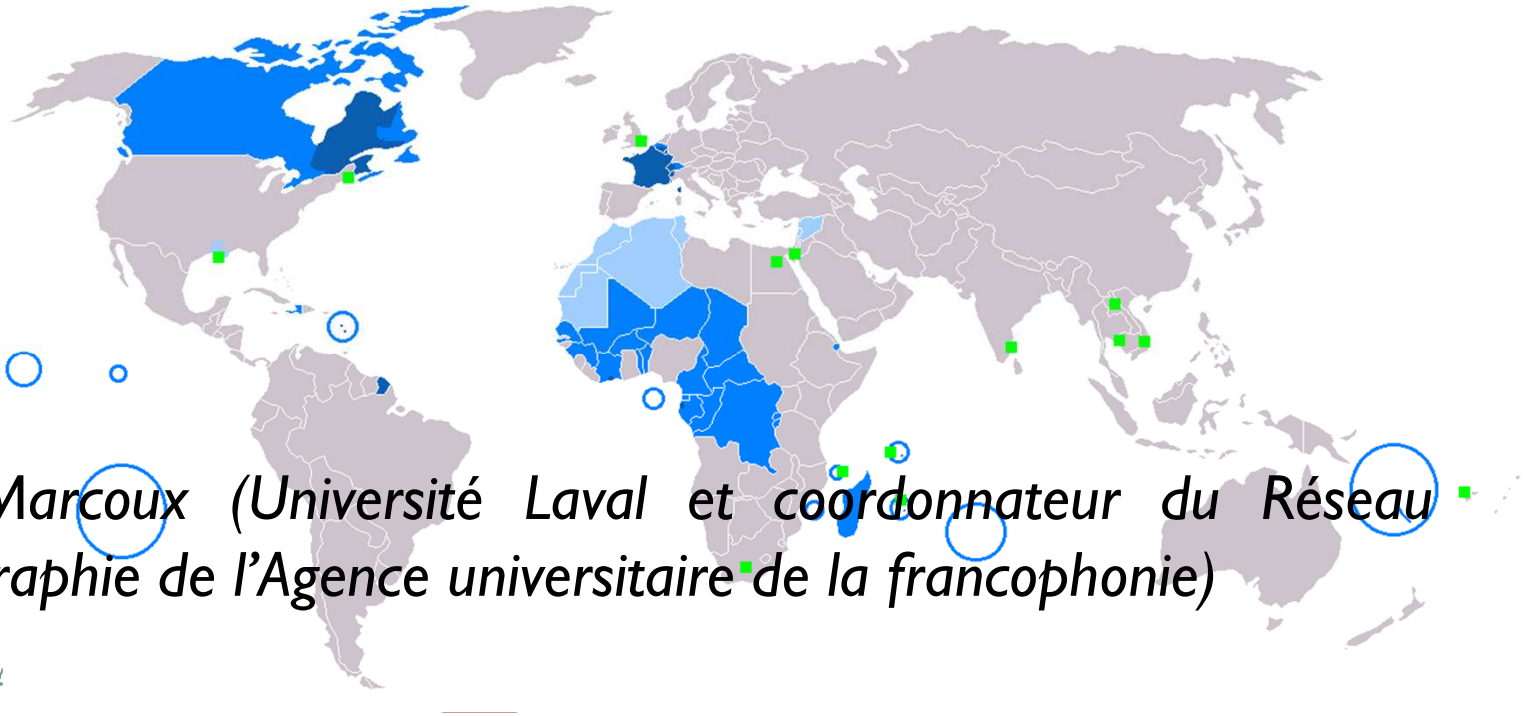
- *By 2025, the global demand for higher education will double to ~200M per year, mostly from emerging economies (NAFSA 2010)*
- 100 M/15 years… 80 000 new students/week!

- Diana Laurillard. http://www.ucd.ie/teaching/u21conference2013/
- Anka Mulder http://ankamulder.weblog.tudelft.nl/

CC BY

Some reasons (2)

- *220 M French Speakers in 2010… 700 M in 2050*
- *3% (2012) → more than 7% in 2050*
- *By 2030, more French Speakers than English Speakers (only 5% in 2050)*



- *René Marcoux (Université Laval et coordonnateur du Réseau Démographie de l'Agence universitaire de la francophonie)*

CC BY

How does this relate to MOOCs?

- Historically, appeared before MOOCs
- Most MOOC players acknowledge inheritance
- Also, through MOOCs only the upper layer of what we teach is being shared
- Question is: can we do something with the courseware that has been produced so far?

- Some believe we can do better

CC BY

OpenCourseware and Pascal

- The Pascal network ran 2003-2012
- One key asset was videolectures.net
- PASCAL 2 → Knowledge 4 All foundation
- Goal is to build upon the Pascal legacy

CC BY

VideoLectures current stats

Content

Events : 800
Authors: 11,000
Lectures: 15,000
Videos: 17,000
Organisations: 7,000
Categories : 600
Comments: 8500

Website

Video views: 6 million
Page views: 27 million
Signed in users: 25,000
Average time: 36 min
Attachments: 620,000
Files: 1,3 million
8 servers
New Visitor: 60.83%
Licenses: Creative Commons

Users

Top countries: United States, India, Slovenia, UK, Germany, China, Canada

Tutorials: 350
Keynote: 800
Interviews:250

Cdlh 2014

http://conference.ocwconsortium.org/2014/

CC BY

ARE THERE REALLY SOME RESEARCH QUESTIONS THERE?

CC BY

Research is e-learning?

- In part. Lots of new scenarios
- Much more learning data which needs analysing

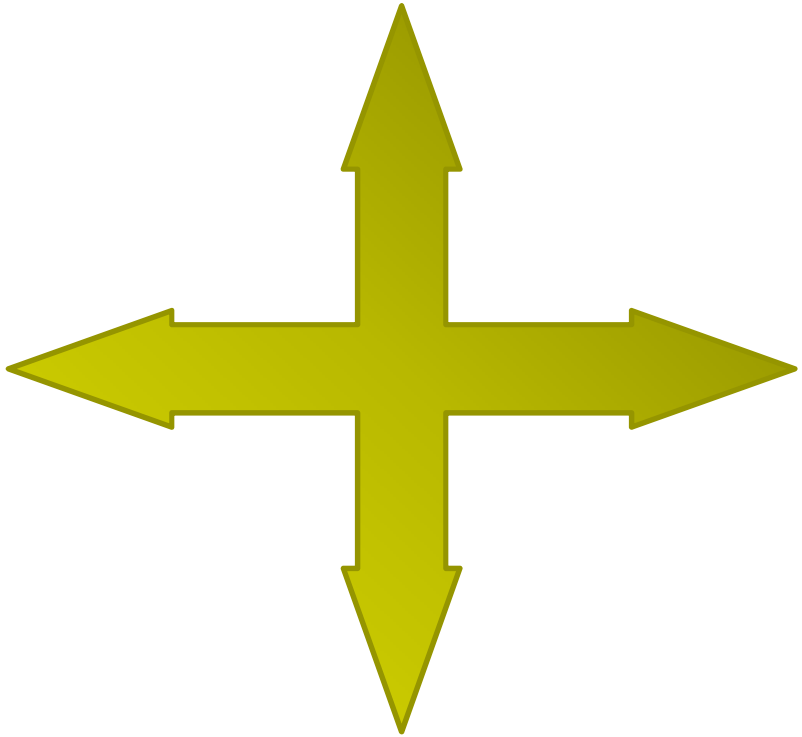
- But not only

CC BY

A simplified picture

The learner and his experience (e-learning)

Sharing open educational resources



Knowledge as a merchandise

Helping the teacher deal with the masses of information (machine learning)

Cdlh 2014

LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE

Clue: Think about MOOCs

- Coursera: Andrew Ng and Daphne Koller
- Eduacity,
- Keywords: automatic evaluation, learning analytics, big data

CC BY

Clue: Some titles of talks from the Internet of Education conference, Slovenia, November 2013

http://www.k4all.org/Internet_of_Education/

- TransLectures: cost-effective transcription and translation of video lectures
- Knowledge Building Through Collaborative Hypervideo Creation
- Annotations, a key asset for video-based e-learning
- A MediaMixer for online learning? – making learning materials more valuable for their owner and more useful for their consumer

cc BY

Clue: OCWC 2014 Global

- Stats from the accepted papers
- 4 Tracks

Track	# papers	# posters
Open Educational Policies	16	
Research and Technology	20	13
Pedagogical Impact	20	4
Project dissemination	10	

CC BY

Slides based on those by Matjaz Rihtar <matjaz.rihtar@ijs.si>

MACHINE LEARNING AND RECOMMENDATION

CC BY

Why recommendation?

- Simplest question is: after this video, which one should I consider?
- Basic system uses history of transactions: *most viewers who watched A also watched B*
- Can we use?
  - Individual history
  - Thematic similarity
  - History over what the users really do

CC BY

Machine learning approach

- Extract many different features
- Let the learning algorithm put different weights on these depending on their usefulness
- Think carefully about the validation issues in order to verify what you have learnt

Obviously a slide improper for SML

CC BY

Experience from *LaVie*

- Were used for recommendation
  - Transcriptions
  - Graphs of authors
  - Related pdfs
  - Other ontologies (Wikipedia, DBLP and Google)

CC BY

Topic and user modeling

- **7 features:**
  1. Lecture popularity
     - Number of visits
  2. Content similarity
     - $BoW(L_c) \cdot BoW(L_p)$
  3. Category similarity
     - $BoC(L_c) \cdot BoC(L_p)$
  4. User content similarity (computed on the fly)
     - $BoW(Hist(U)) \cdot BoW(L_p)$
  5. User category similarity (computed on the fly)
     - $BoC(Hist(U)) \cdot BoC(L_p)$
  6. Co-visits
     - Number of times of L_c and L_p viewed in the same browsing session
  7. User similarity
     - Number of users who have watched both L_c and L_p

L_c = current lecture

L_p = proposed lecture

U = user

1 table has approx. 70 million entries (for features 2,3,6 and 7)

CC BY

Recommendation (1)

- Using SVM classifier for training:
  - Positive samples: two months of clicks using current recommender
  - Resulting feature weights:

Feature	Weight
Lecture popularity	-0.00003
Content similarity	0.00452
Category similarity	0.00148
User content similarity	0.02724
User category similarity	0.04167
Co-visits	0.00187
User similarity	0.01519

CC BY

Recommendation (2)

- Final recommendation
  - A linear SVM classifier was used to rank all possible recommendation links:

$Given\ L_c\ and\ U$:

$For\ all\ Lp \neq L_c$:

$\vec{x}$ ... *feature vector computed for the triplet (Lc, L_p, U)*

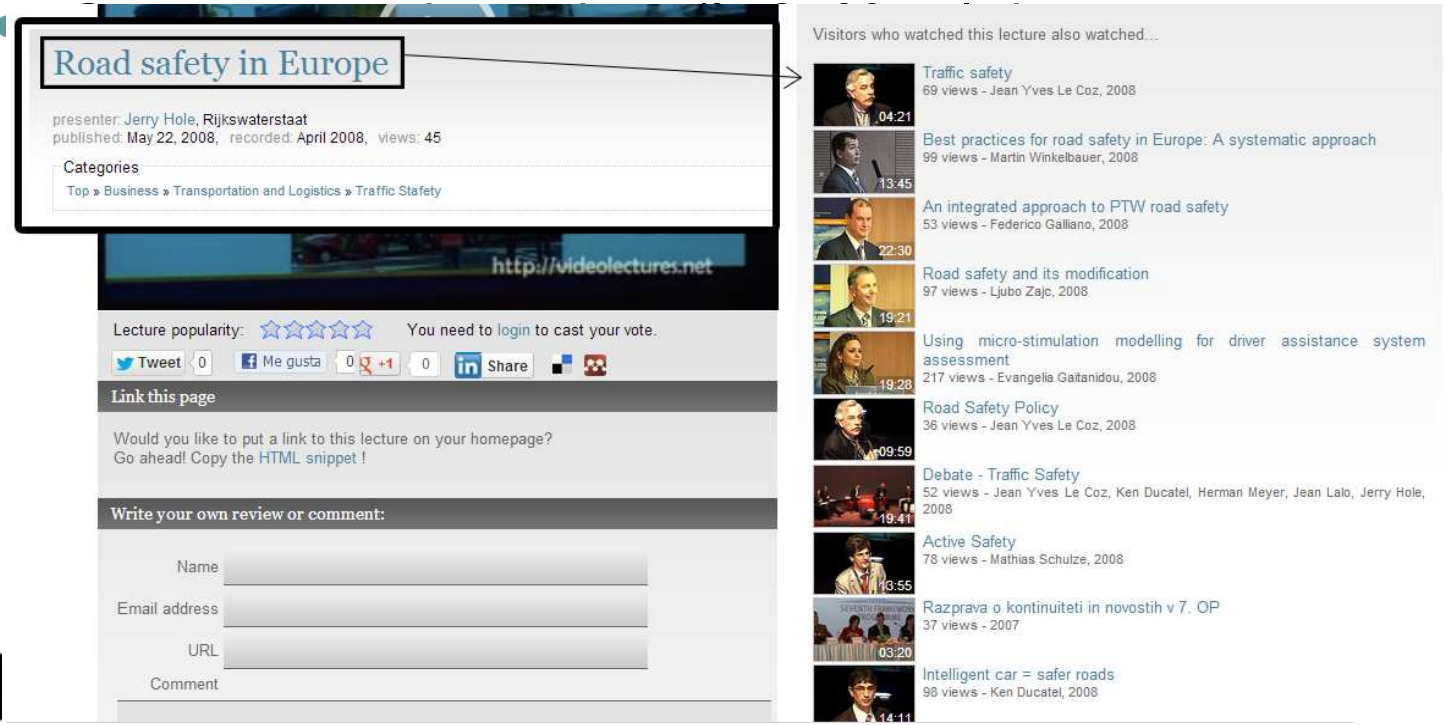
$$\text{score}(\vec{x}) = \vec{w} \cdot \vec{x} = \sum_{n=1}^{7} w_n \cdot x_n$$

  - Lectures with top 10 scores are recommended

CC BY

Evaluation

- Evaluation
  - Using coin flipping between old and new recommender
  - Counting the number of clicks
- Try http://dev.videolectures.net/



TRANSCRIPTION AND TRANSLATION

CC BY

LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE

transLectures

Transcription and Translation of Video Lectures

Slides by Gonçal Garcés ggarces@dsic.upv.es
Project coordinator: Alfons Juan-Ciscar ajuan@dsic.upv.es

cnrs

UNIVERSITÉ DE NANTES

ECOLE DES MINES DE NANTES

The transLectures partners

	Name	Country
1	Universitat Politècnica de València	Spain
2	Xerox SAS	France
3	Institut Jožef Stefan	Slovenia
3+	Knowledge for All Foundation	UK
4	RWTH Aachen University	Germany
5	EML – European Media Laboratory	Germany
6	DDS – Deluxe Digital Studios	UK

CC BY

The transLectures approach

1. Automatic Speech Recognition (ASR) and Machine Translation (MT)
   - Adaptation: Taking advantage of the characteristics of video lecture repositories
   - High-quality automatic transcriptions and translations
2. Interactive postediting: intelligent interaction for reduced effort

CC BY

Goals

- Massive adaptation
- Intelligent interaction

- Implementation
  - Case studies: Videolectures.NET & Polimedia
  - Real-life evaluation
- Integration into Opencast Matterhorn

http://opencast.org/matterhorn/

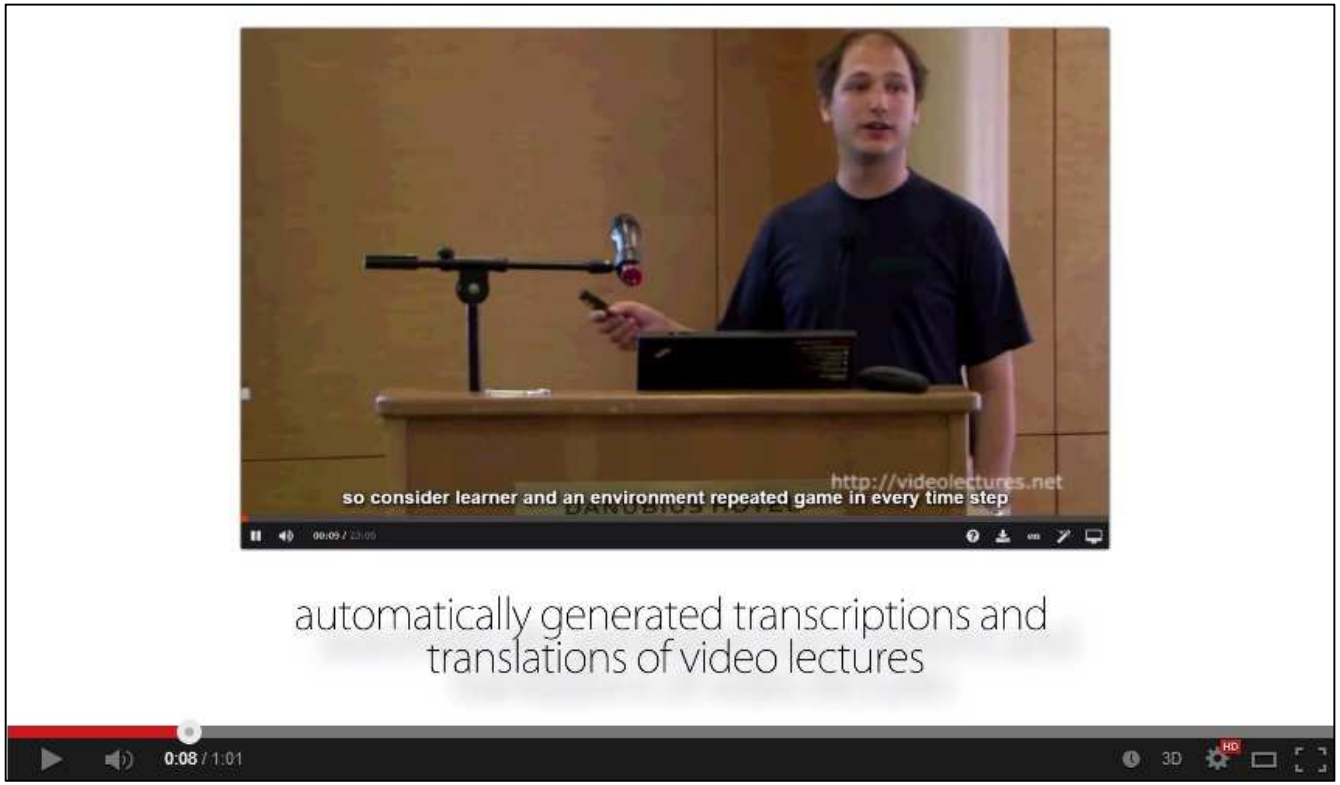
CC BY

Languages

- Transcription (ASR)
  - EN
  - SL
  - ES
- Translation (MT)
  - EN>SL , SL>EN
  - EN>ES , ES>EN
  - EN>FR
  - EN>DE

CC BY

transLectures: video demo

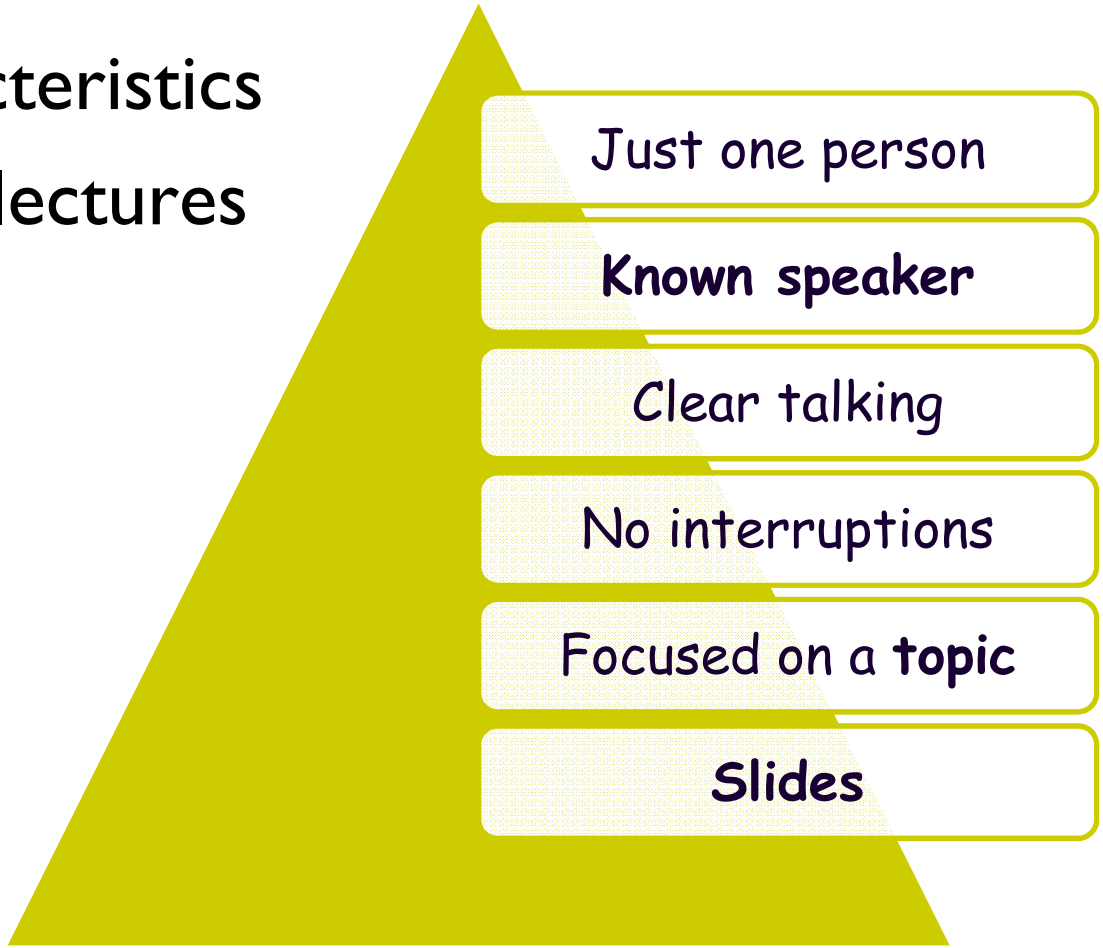


CC BY

Massive adaptation

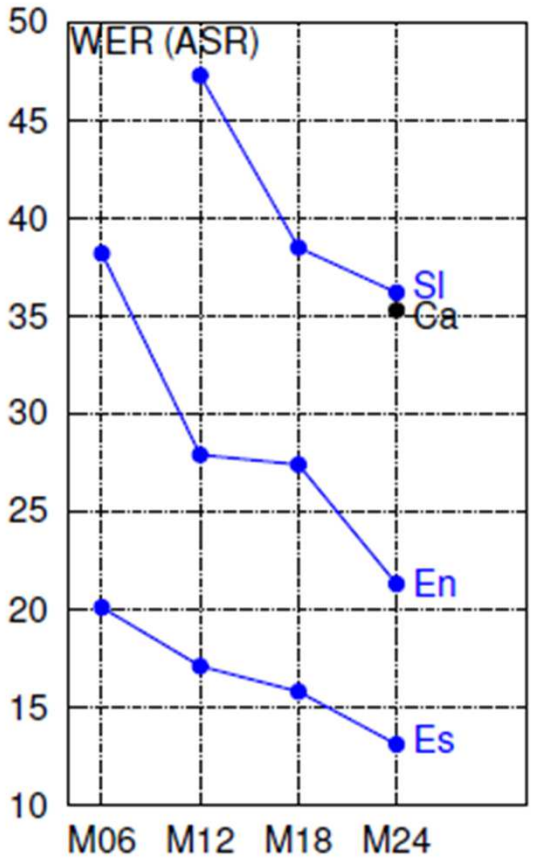
- Characteristics

of video lectures



Scientific evaluations

- Transcription results
  - WER: Word Error Rate (%)
  - Goal: WER < 25%
  - EN, SL, ES



Worse

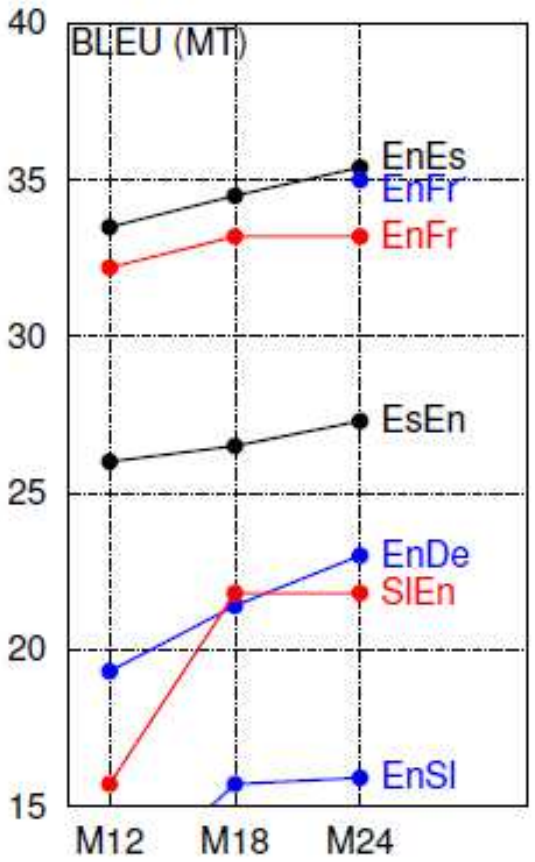
Better

CC BY

Scientific evaluations

- Translation results
  - BLEU
  - Goal: BLEU > 30

  - EN>SL , SL>EN
  - EN>ES , ES>EN
  - EN>FR
  - EN>DE



Better

Worse

CC BY

Y1 results and comparison

- Videolectures.NET:

WER	tL	Google
English	**27.9**	45.4
Slovenian	44.4	N/A

- poliMedia:

WER	tL	Google
Spanish	**17.1**	34.3

CC BY

Y1 results and comparison

- Videolectures.NET:

WER	tL	Google
English	**27.9**	45.4
Slovenian	44.4	N/A

		BLEU	
		tL	Google
English	→ Spanish	33.5	**34.1**
English	→ French	**32.2**	32.0
English	→ German	**20.6**	18.6
English	→ Slovenian	**15.5**	10.5
Slovenian	→ English	15.7	**17.5**

- poliMedia:

WER	tL	Google
Spanish	**17.1**	34.3

		BLEU	
		tL	Google
Spanish	→ English	26.0	**27.6**

CC BY

Intelligent interaction

- Post-editing automatic transcriptions/translations
  - The user invests the least possible effort
  - The system learns the most from it

- Confidence measures
- Fast constrained search

cc BY

Intelligent interaction

Cdlh 2014

LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE

User evaluations

2. Intelligent interaction

CC BY

User evaluations

- User evaluations at UPV: results

Table 2: Comparison between all the interaction models evaluated.

Supervision mode	Initial WER	Final WER	Lecturer RTF	SS-score
1st - Manual supervision	16.9	0.0	5.6	9.1
2nd - Intelligent interaction	14.5	8.0	2.2	7.2
3rd - Two-step supervision	28.4	0.0	5.3	7.8

CC BY

transLectures: Open source tools

- The tL player (& editor)
  - Coming soon (www.translectures.eu)
- The transLectures-UPV Toolkit (TLK) for ASR
  - www.translectures.eu/tlk
- RWTH Aachen: rASR, Jane (MT)
  - http://www-i6.informatik.rwth-aachen.de/web/Software/

CC BY

Multilingualism

- Multilingualism is a crucial issue
  - If English is the unique language we have many deprived users
  - But also we are missing a lot of excellent material
- It is a highly sensitive issue
- Answers
  - Being able to navigate between languages
  - Being able to translate: translectures project

CC BY

Olivier Aubert - @Olivier_Aubert
Yannick Prié - @yprie

ANNOTATION ISSUES

CC BY

Key questions

- Obtaining the participation of the learner can be vital (at least if the main purpose of the resource is to allow a learner to learn)
- Examining the MOOC's success helps
- Why would someone want to annotate our videos?
- What tools can we provide in order for quality annotation to take place ?

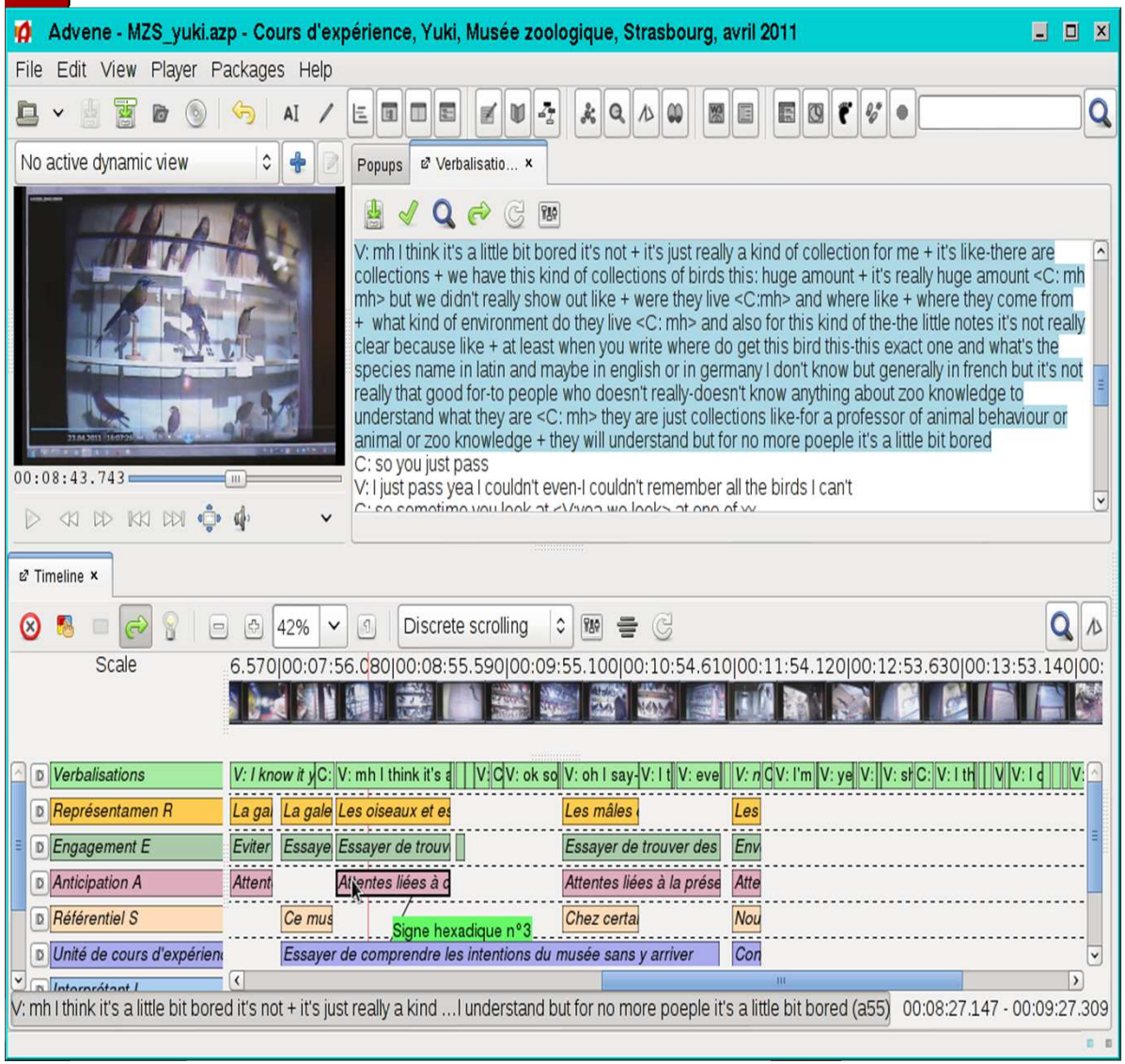
CC BY

Video-based e-learning activities

- Different activities based on
  - the nature of the video document
  - the status of the annotator
  - the status of the recipient

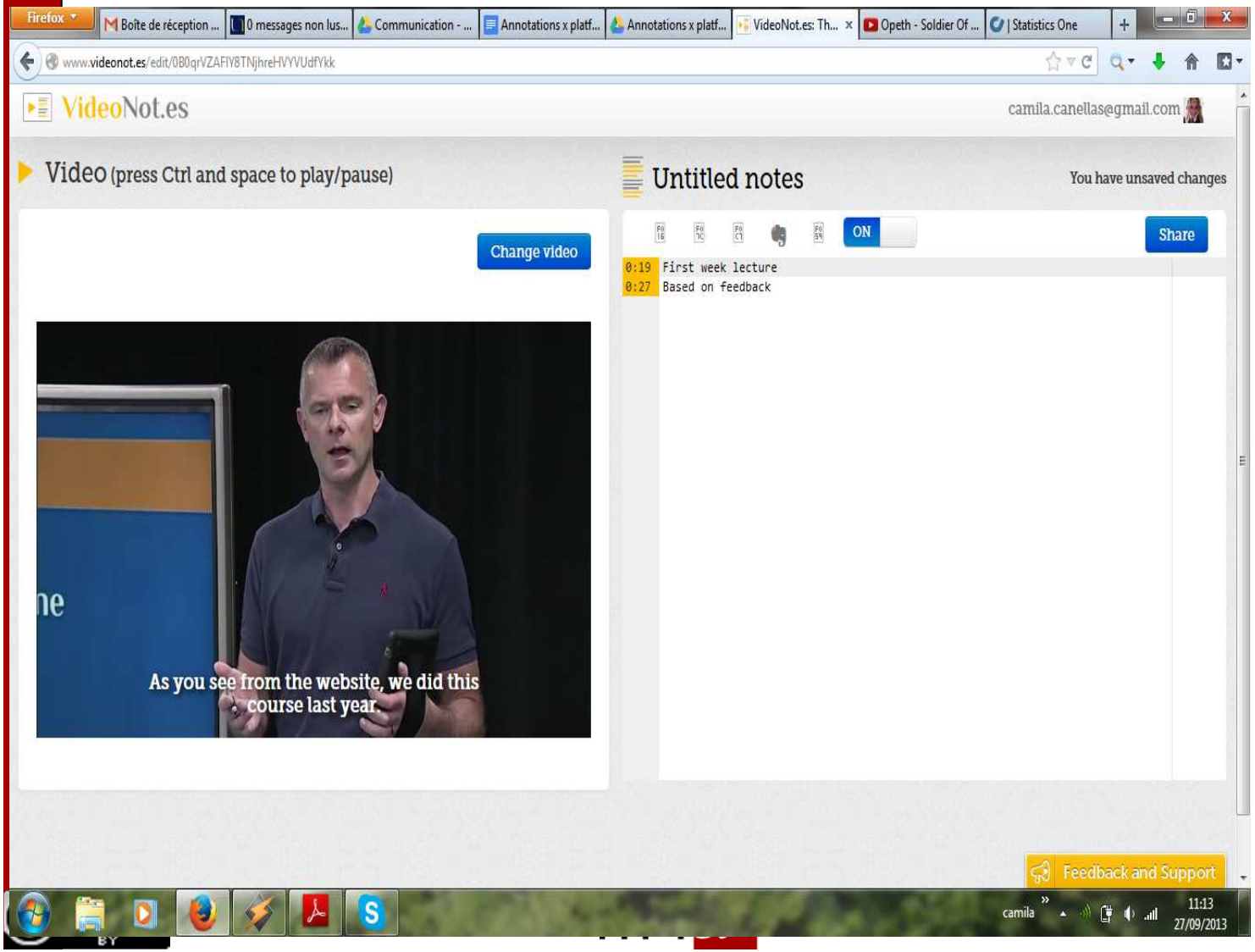
CC BY

Assimilation - example



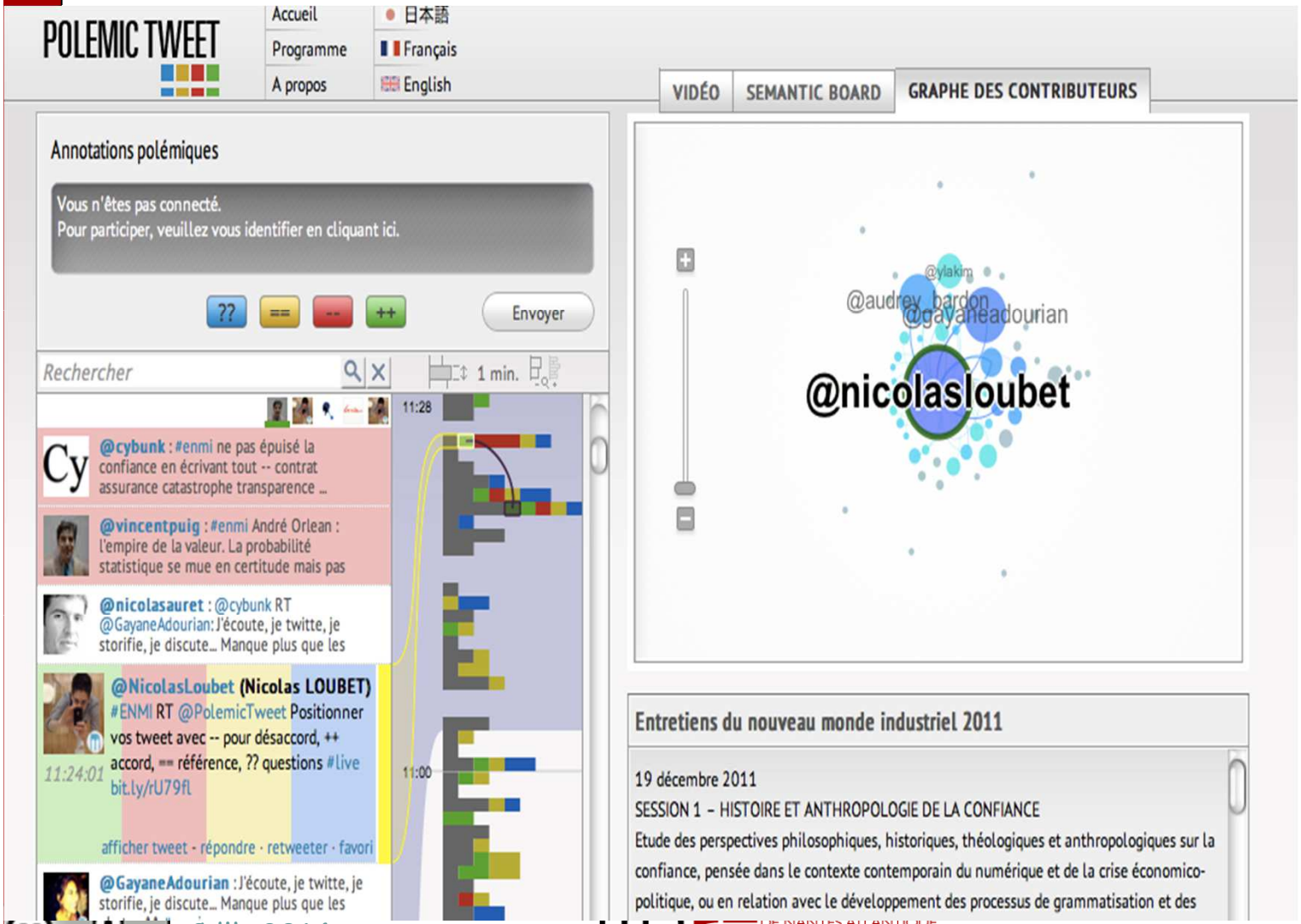
Advene: an example of an offline anotator

Collaborative assimilation example



VideoNot.es

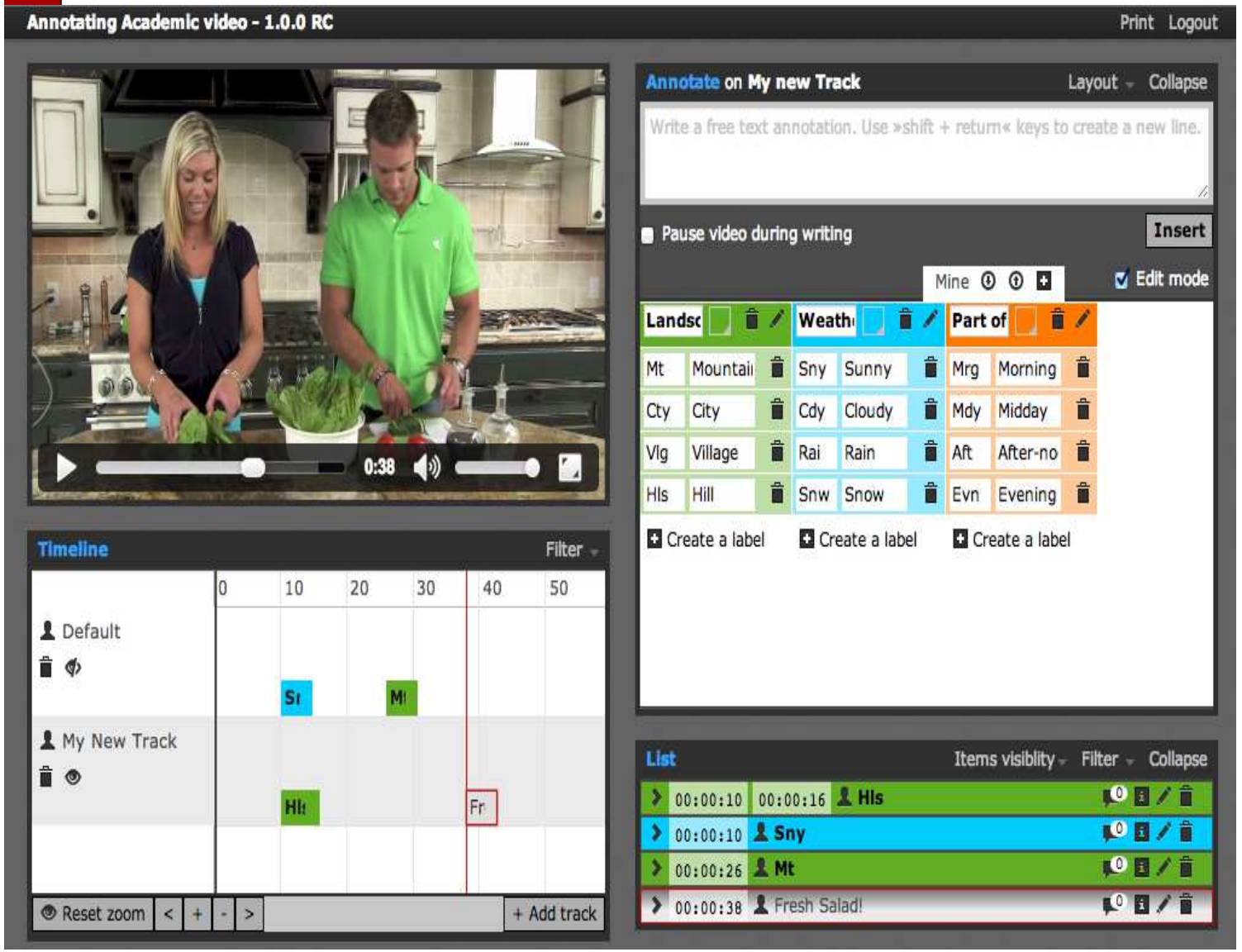
Feedback - example



PolemicTweet

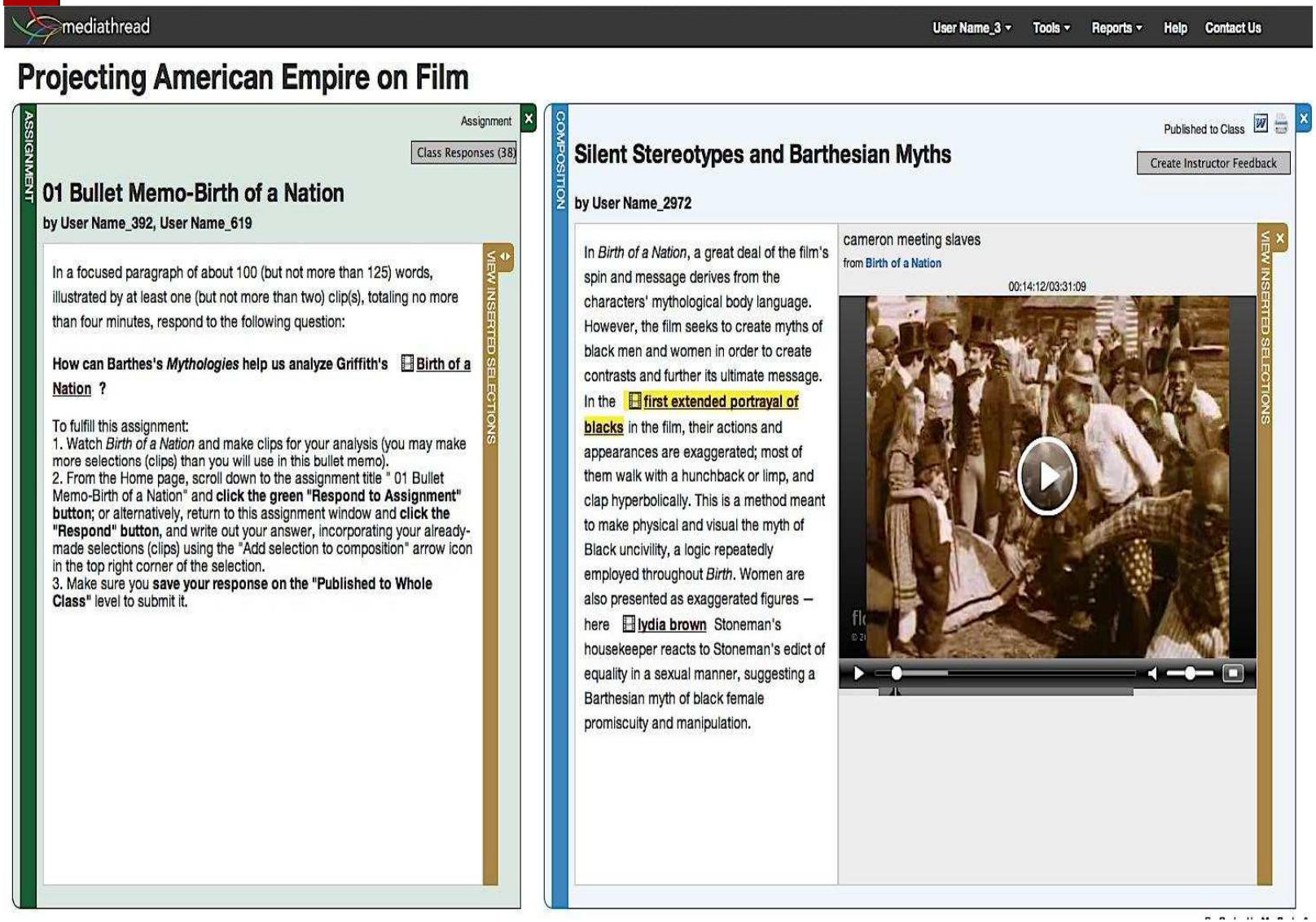
BY

App. of an analysis grid - example



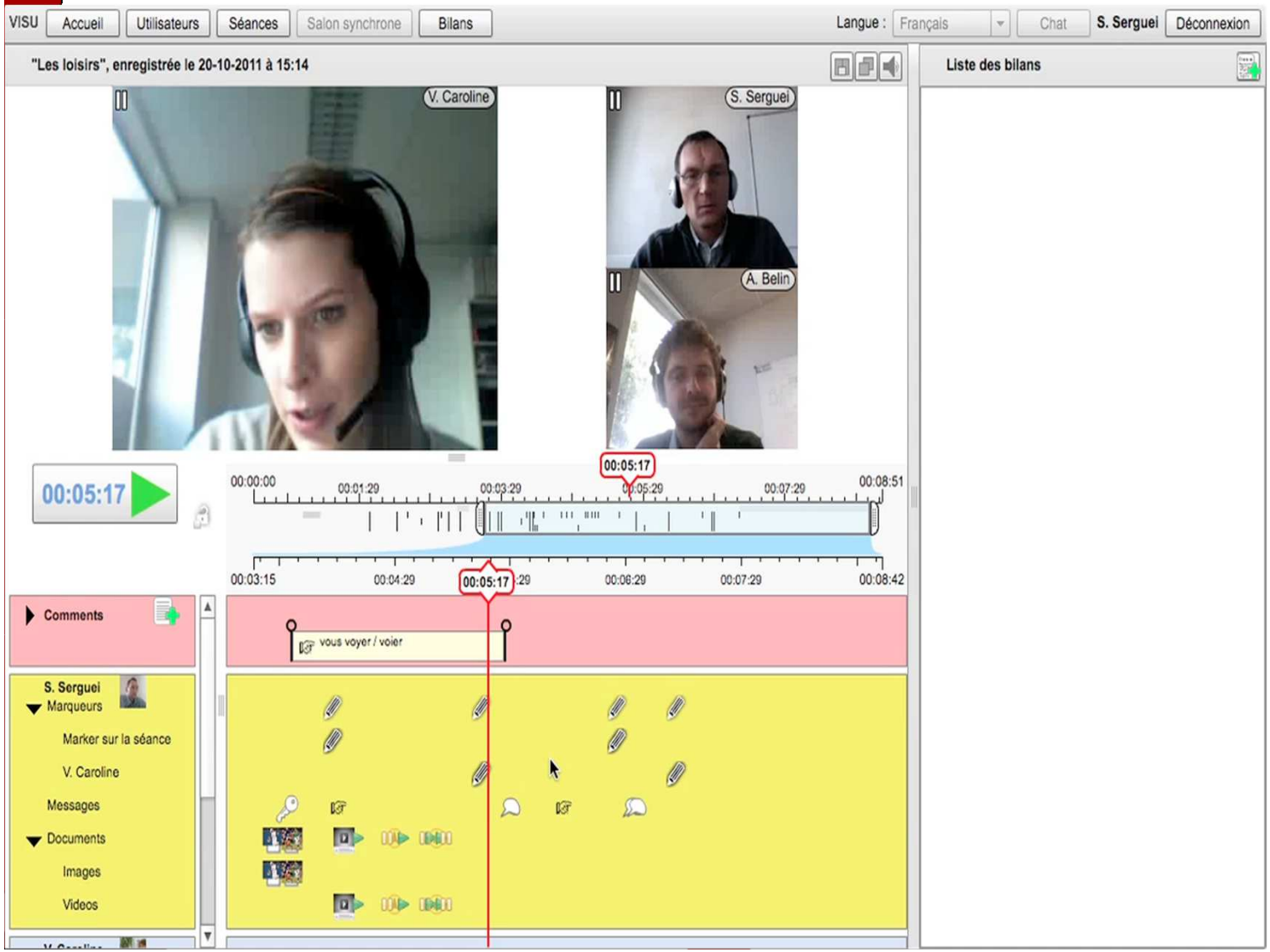
Matterhorn Engage player

Analysis - example

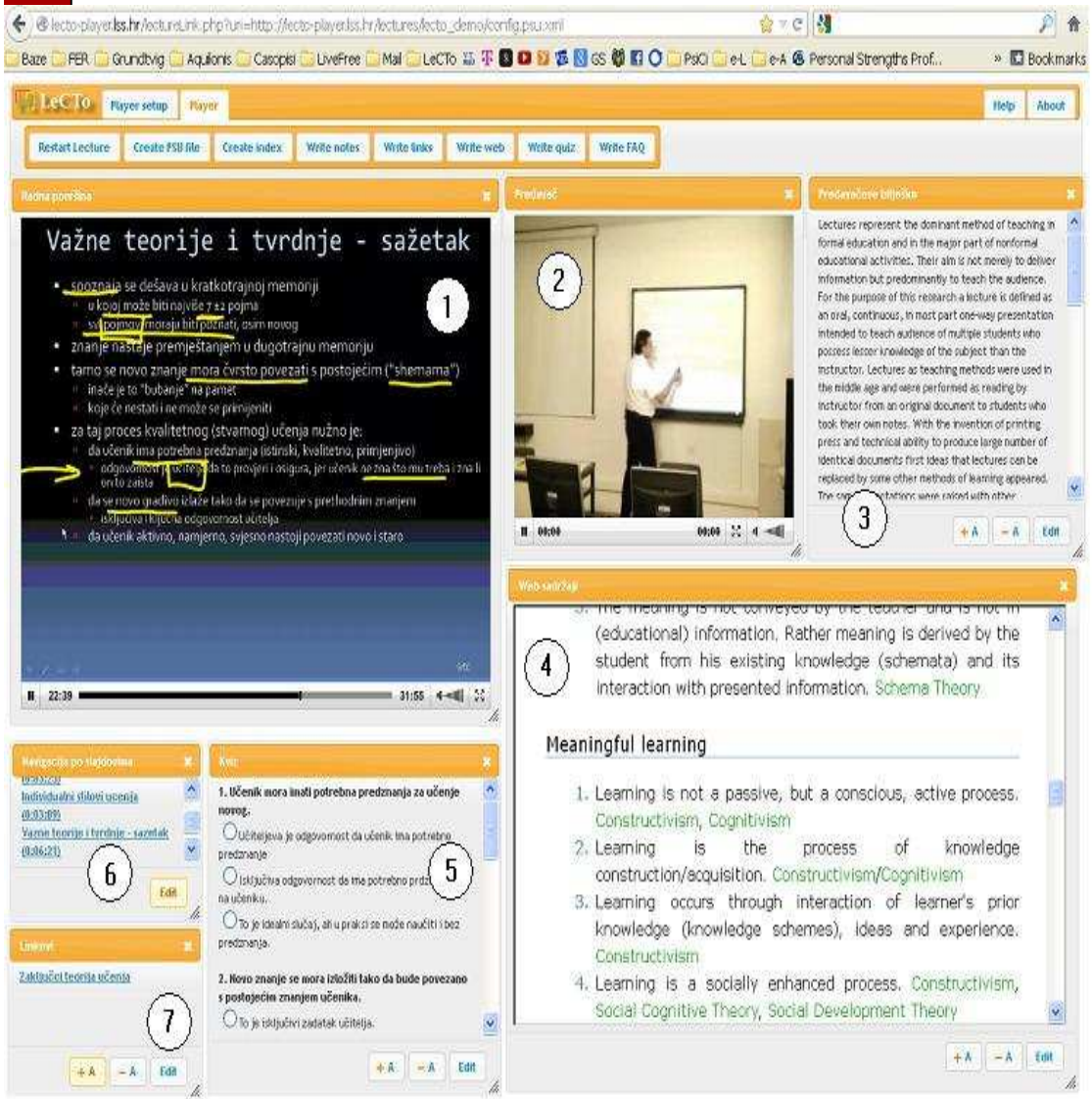


MediaThread

Reflexivity/Feedback - example



Course enrichment - example



Annotation challenges

- Goals
  - Ensure interoperability
  - Ensure durability
- Support
  - Anchoring now normalized (MediaFragment)
  - From unstructured free-text annotations to semantic annotations

CC BY

Semi-automatic annotation challenges

- Many efforts to do automatic generation (Translectures, linkedTV) but not perfect yet
- Provide tools that combine automatic algorithms and correction interfaces
- Being able to cope with hundreds of annotations

CC BY

NAVIGATION (AND INDEXING)

CC BY

Linking the segments

- As videos are becoming increasingly basic objects, how do we navigate inside a video?
- And from video to video?

MediaMixer:

Community set-up and networking for the reMIXing of online MEDIA fragments

The main rationale of MediaMixer is to set up and sustain a community of video producers, hosters, and redistributors who will be supported in the adoption of semantic multimedia technology in their systems and workflows to build a European market for media fragment re-purposing and re-selling.

CC BY

Multimodality

- In a first approach, having to deal with material coming from different sources is a (technical) problem
- But it is also an asset!
  - Pdfs give us a lot of information towards language modelling
  - Slides allow us to envisage segmentation

CC BY

COCo project (in Nantes)

CominOpenCourseware

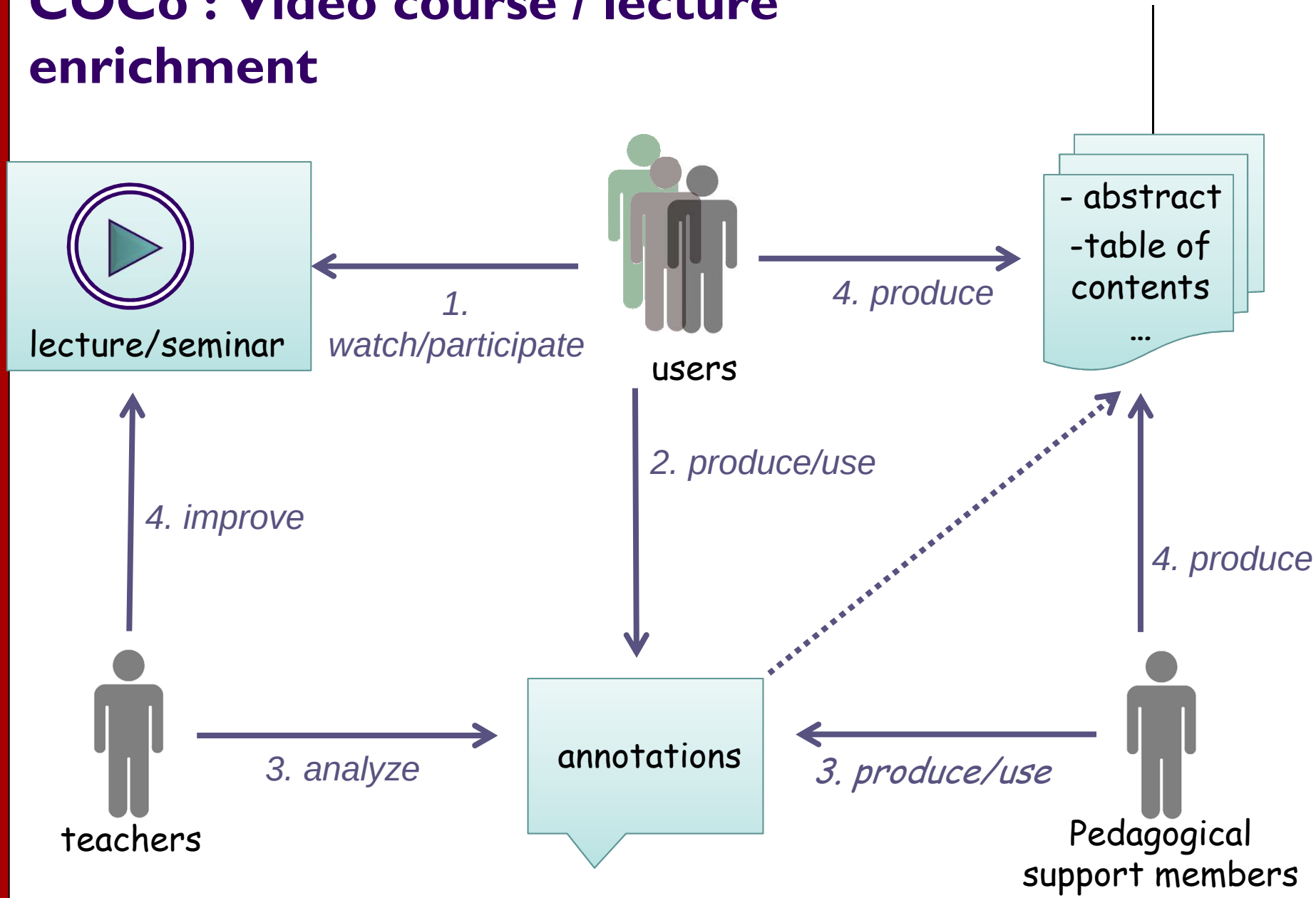
Région PAYS de la LOIRE

ComInLabs

CC BY

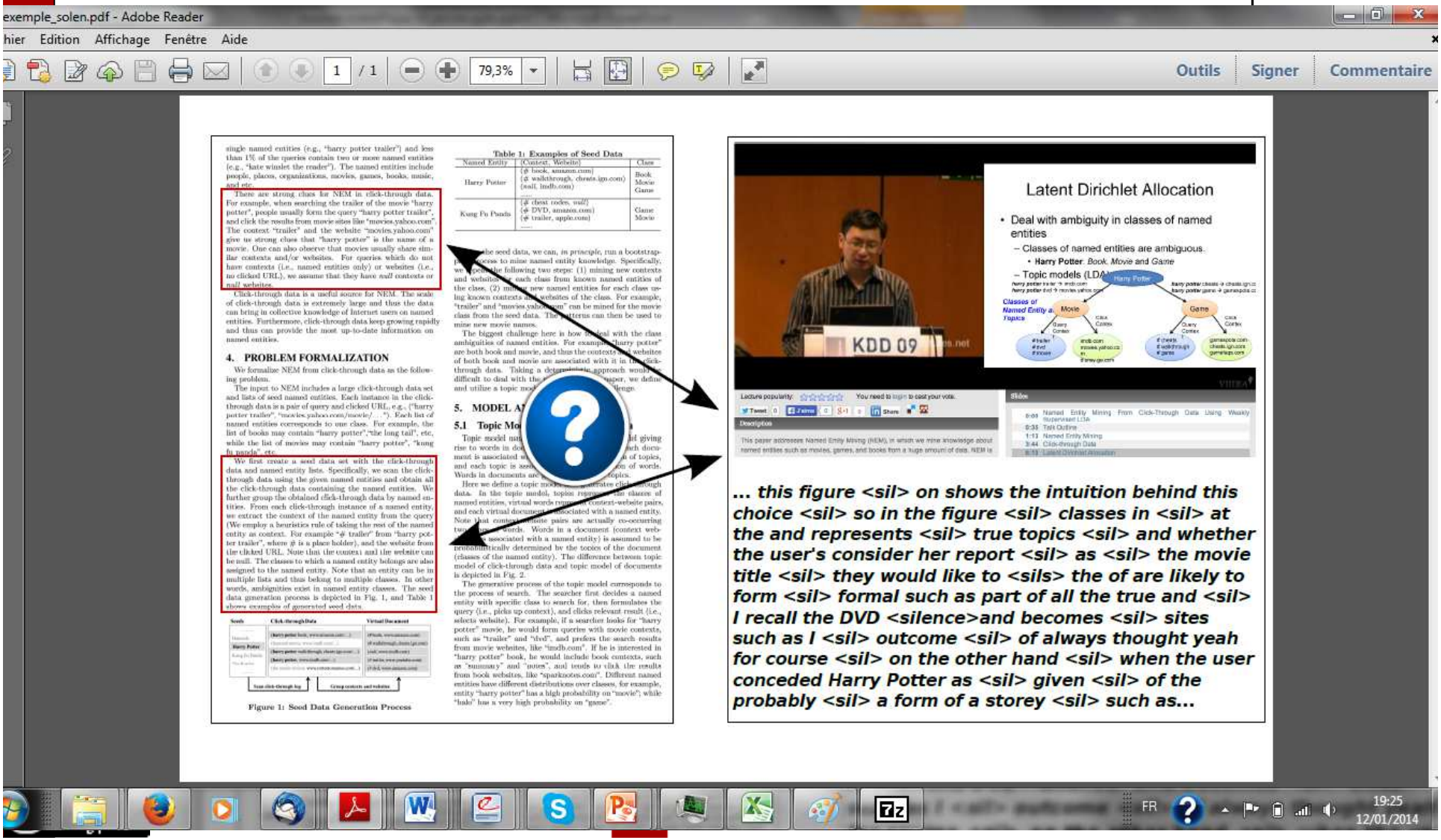
lina LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE

COCo : Video course / lecture enrichment



Cdlh 2014

Automatic alignment



TRACKS (LEARNING ANALYTICS)

CC BY

Daphne Koller

- You can turn the study of learning from the hypothesis driven mode to the data driven mode… a transformation that has for example revolutionized biology

- http://www.ted.com/talks/daphne_koller_what_we_re_learning_from_online_education

CC BY

A very small conclusion

- There is a lot of research aiming to improve the technologies linked with online education
- Machine learning is a key (but not unique) technology
- More at OCWC 2014 Global!

CC BY

LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE

Thank you.

WEBSITES:

http://www.k4all.org/

http://videolectures.net/

http://www.translectures.eu/

Knowledge 4All Foundation Ltd

UNIVERSITÉ DE NANTES

ECOLE DES MINES DE NANTES

Sugata Mitra

A teacher who can be replaced by a computer, should be.
http://sfltdu.blogspot.fr/2012/11/a-teacher-who-can-be-replaced-by.html

CC BY