MLLP | Machine Learning and Language Processing

Language technologies for Education: recent results by the MLLP group

Alfons Juan
2nd Internet of Education Conference 2015
18 September 2015, Sarajevo

Contents

MLLP | Machine Learning and Language Processing

- Research group at the *Univ. Politècnica de València* (Spain)

- **Research areas:**
  - Machine Learning and Applications
  - Natural Language Processing
  - Educational Technologies and Big Data

- **Recent research projects:**
  - trans**Lectures**: Transcription and translation of video lectures
  - EMMA: European Multiple MOOC Aggregator
  - Active Interaction for Speech Transcription and Translation

trans**Lectures** (Nov 2011 – Oct 2014)

Main goal: to develop innovative, cost-effective (quasi-automatic) solutions to produce accurate subtitles for educational videos

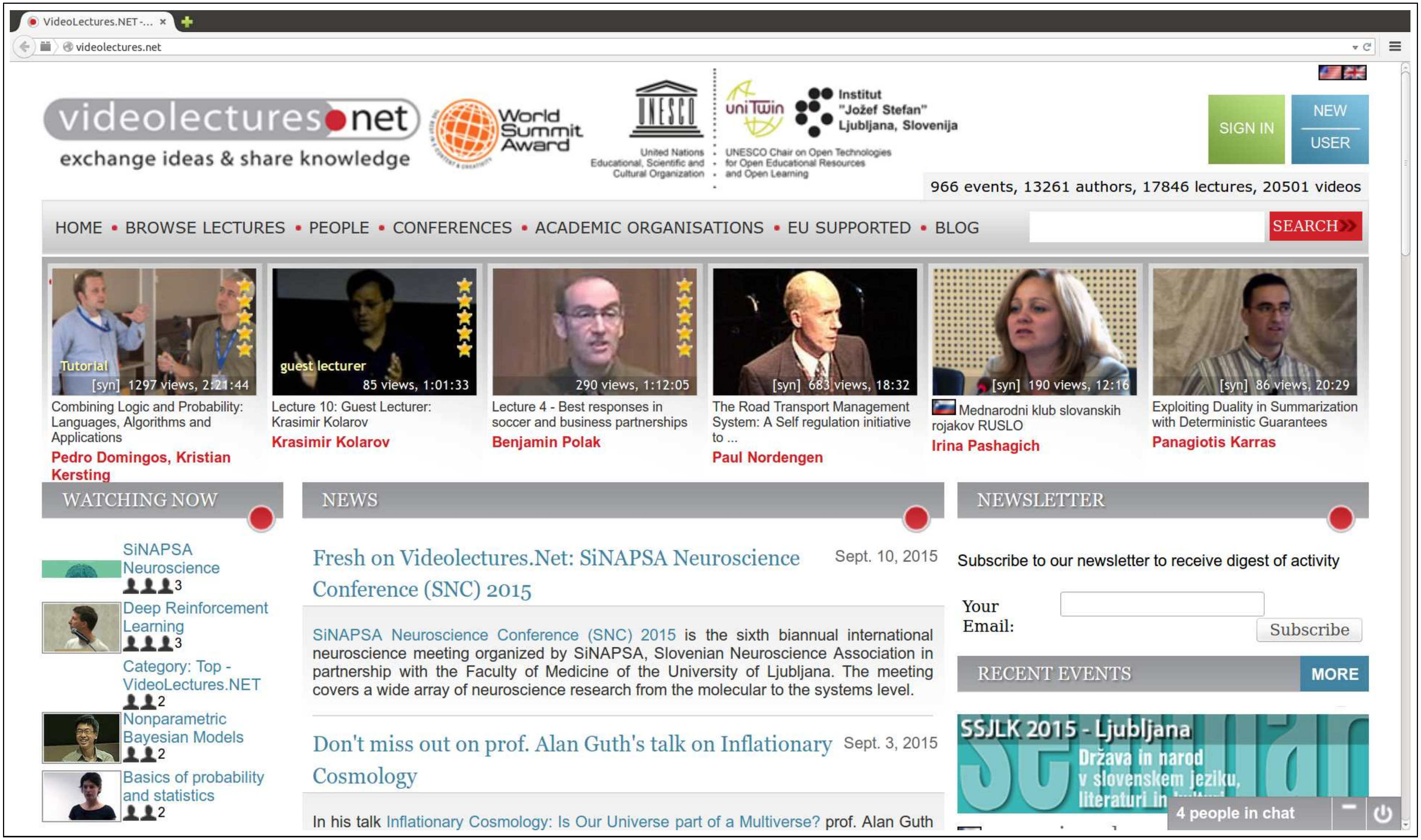
Two pilots:

- ***VideoLectures.NET:*** En, Sl $\quad$ En→{De,Sl,Fr,Es} $\quad$ Sl→En
- ***poliMedia:*** Es, Ca $\quad$ Es↔En $\quad$ Ca↔{Es,En}

Three scientific and technological objectives:

- ***Massive adaptation*** to improve subtitling quality
- ***Intelligent interaction*** to improve subtitling quality
- ***Integration into Opencast*** to enable real-life evaluation

transLectures: VideoLectures.NET


VideoLectures.NET -...
videolectures.net
videolectures.net
exchange ideas & share knowledge
World Summit Award
UNESCO
United Nations Educational, Scientific and Cultural Organization
uniTwin
UNESCO Chair on Open Technologies for Open Educational Resources and Open Learning
Institut "Jožef Stefan" Ljubljana, Slovenija
SIGN IN
NEW USER
HOME • BROWSE LECTURES • PEOPLE • CONFERENCES • ACADEMIC ORGANISATIONS • EU SUPPORTED • BLOG
SEARCH
Tutorial
guest lecturer
WATCHING NOW
NEWS
NEWSLETTER
Subscribe
RECENT EVENTS
MORE
SSJLK 2015 - Ljubljana
Država in narod v slovenskem jeziku, literaturi in
4 people in chat

966 events, 13261 authors, 17846 lectures, 20501 videos

[syn] 1297 views, 2:21:44
Combining Logic and Probability: Languages, Algorithms and Applications
Pedro Domingos, Kristian Kersting

85 views, 1:01:33
Lecture 10: Guest Lecturer: Krasimir Kolarov
Krasimir Kolarov

290 views, 1:12:05
Lecture 4 - Best responses in soccer and business partnerships
Benjamin Polak

[syn] 683 views, 18:32
The Road Transport Management System: A Self regulation initiative to ...
Paul Nordengen

[syn] 190 views, 12:16
Mednarodni klub slovanskih rojakov RUSLO
Irina Pashagich

[syn] 86 views, 20:29
Exploiting Duality in Summarization with Deterministic Guarantees
Panagiotis Karras

WATCHING NOW

SiNAPSA Neuroscience 3
Deep Reinforcement Learning 3
Category: Top - VideoLectures.NET 2
Nonparametric Bayesian Models 2
Basics of probability and statistics 2

NEWS

Fresh on Videolectures.Net: SiNAPSA Neuroscience Conference (SNC) 2015 — Sept. 10, 2015

SiNAPSA Neuroscience Conference (SNC) 2015 is the sixth biannual international neuroscience meeting organized by SiNAPSA, Slovenian Neuroscience Association in partnership with the Faculty of Medicine of the University of Ljubljana. The meeting covers a wide array of neuroscience research from the molecular to the systems level.

Don't miss out on prof. Alan Guth's talk on Inflationary Cosmology — Sept. 3, 2015

In his talk Inflationary Cosmology: Is Our Universe part of a Multiverse? prof. Alan Guth

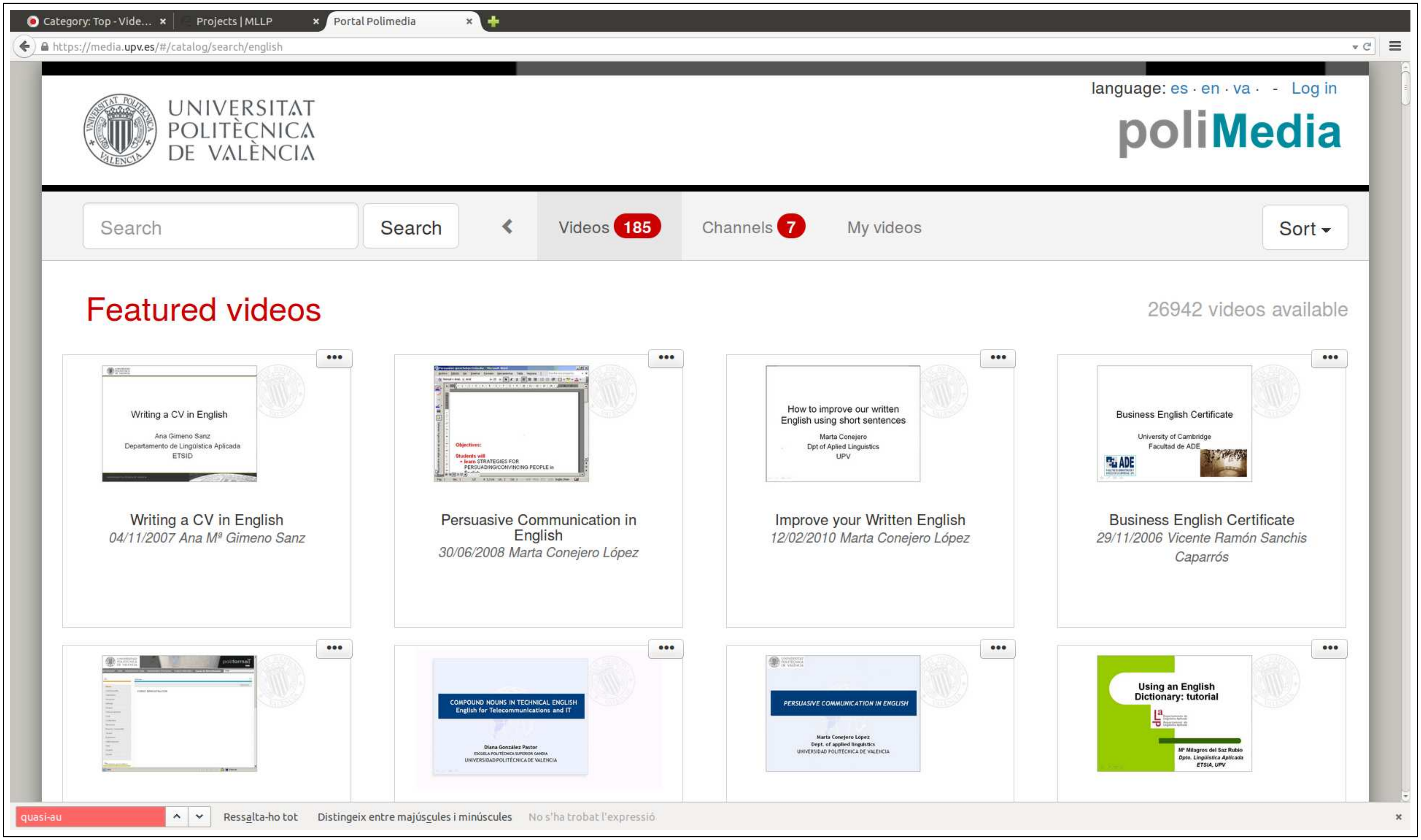
NEWSLETTER

Subscribe to our newsletter to receive digest of activity

Your Email:

>20000 videos (45 min. on avg.): 85% English, 13% Slovenian, . . .

transLectures: poliMedia



Category: Top - Vide...
Projects | MLLP
Portal Polimedia
https://media.upv.es/#/catalog/search/english
UNIVERSITAT POLITÈCNICA DE VALÈNCIA
language: es · en · va · - Log in
poliMedia
Search
Search
Videos 185
Channels 7
My videos
Sort
quasi-au
Ressalta-ho tot
Distingeix entre majúscules i minúscules
No s'ha trobat l'expressió

Featured videos

26942 videos available

Writing a CV in English
04/11/2007 Ana Mª Gimeno Sanz

Persuasive Communication in English
30/06/2008 Marta Conejero López

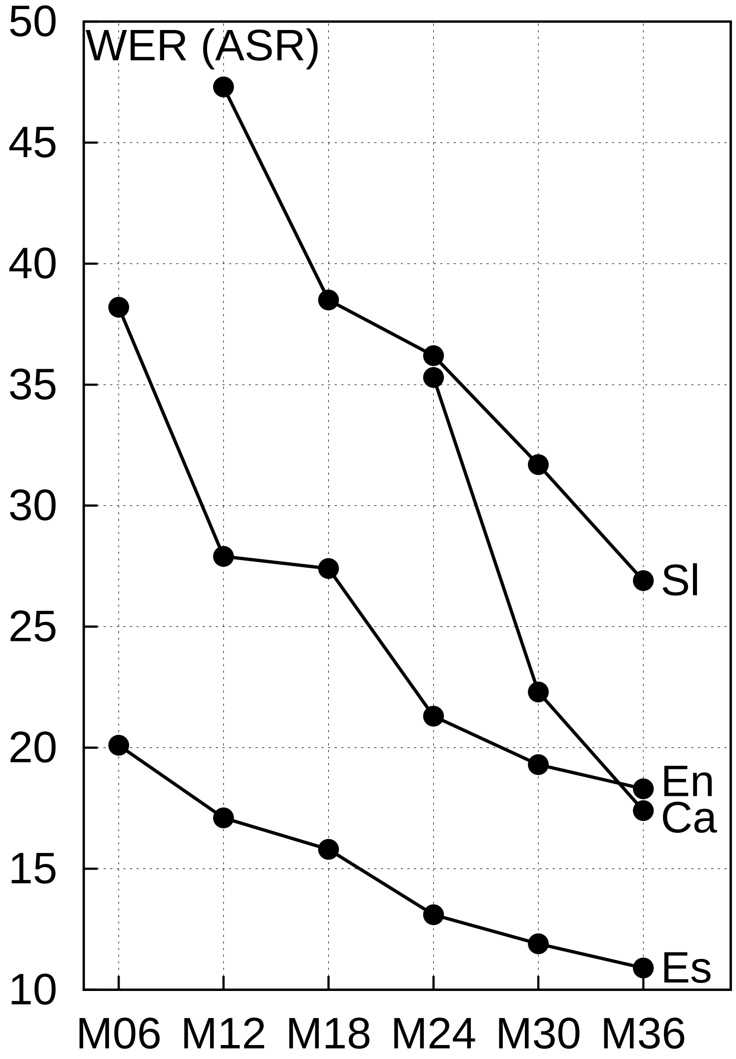
Improve your Written English
12/02/2010 Marta Conejero López

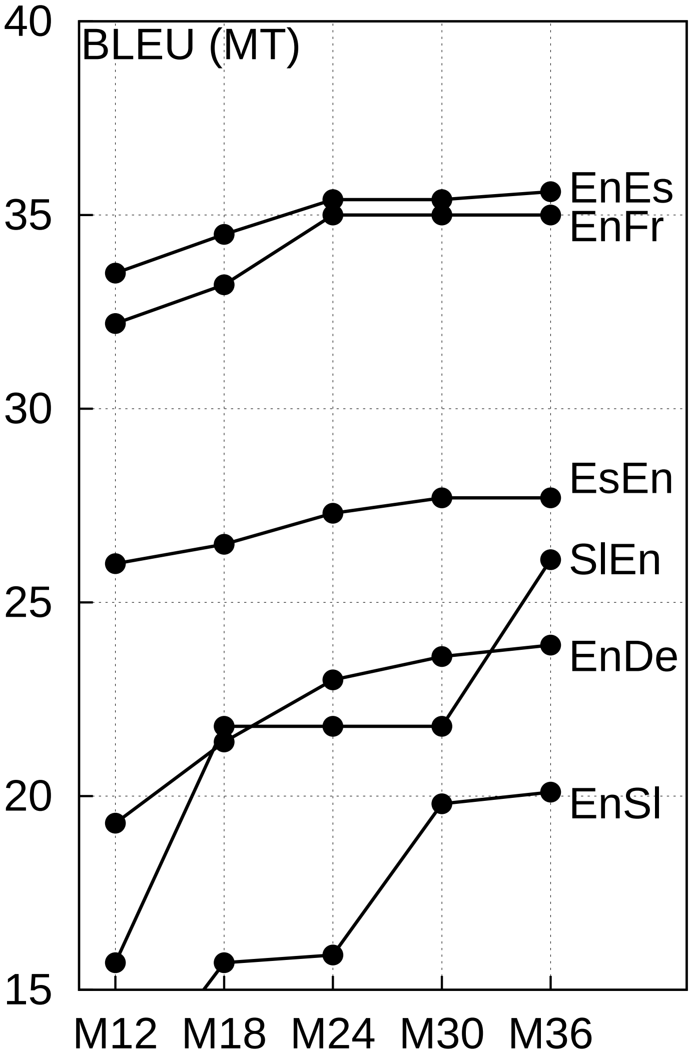
Business English Certificate
29/11/2006 Vicente Ramón Sanchis Caparrós

Writing a CV in English
Ana Gimeno Sanz
Departamento de Lingüística Aplicada
ETSID
How to improve our written English using short sentences
Marta Conejero
Dpt of Aplied Linguistics
UPV
Business English Certificate
University of Cambridge
Facultad de ADE
COMPOUND NOUNS IN TECHNICAL ENGLISH
English for Telecommunications and IT
Diana González Pastor
PERSUASIVE COMMUNICATION IN ENGLISH
Marta Conejero López
Using an English Dictionary: tutorial

27000 videos (2-10 min.): 88% Spanish, 3% Catalan, . . .

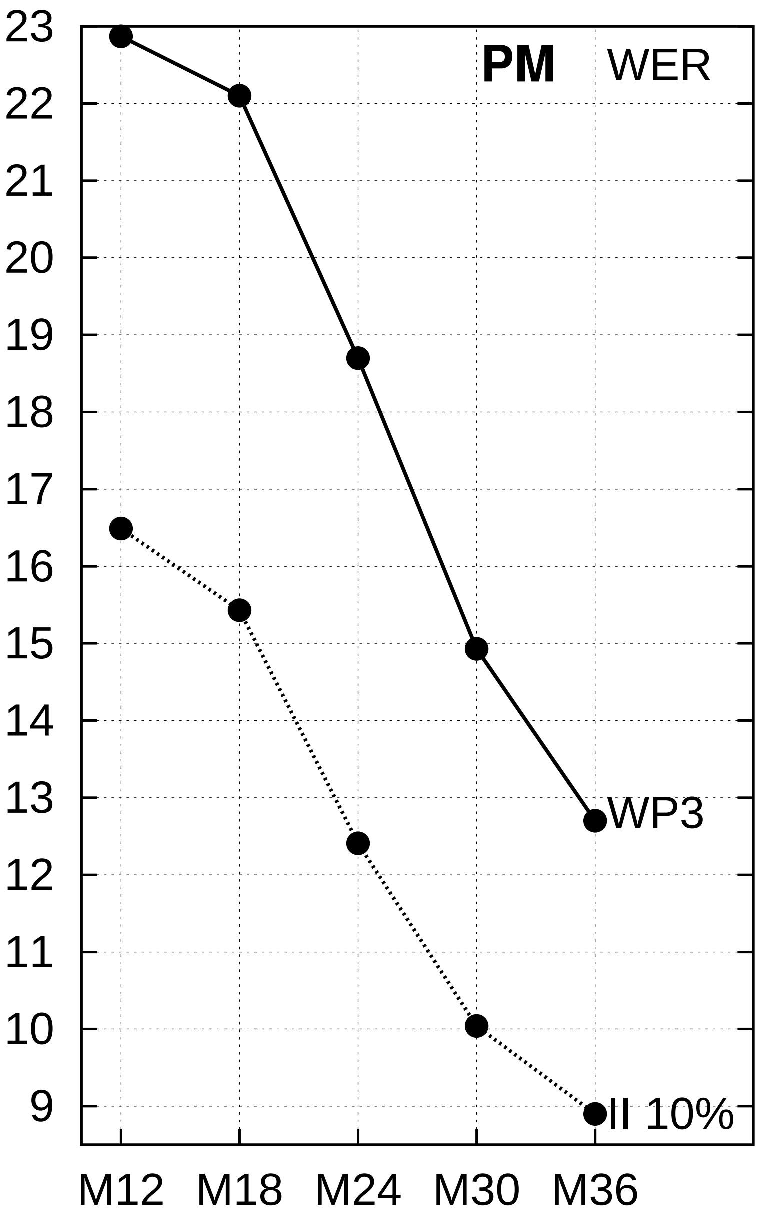
transLectures: Massive adaptation





- Key advances: neural networks and “extra” resources (slides)
- ASR: all error rates below 30 % (Sl: 26.9 %, others below 20 %)
- SMT: all BLEU scores above 20 (En→Sl: 20.1, mostly above 25)

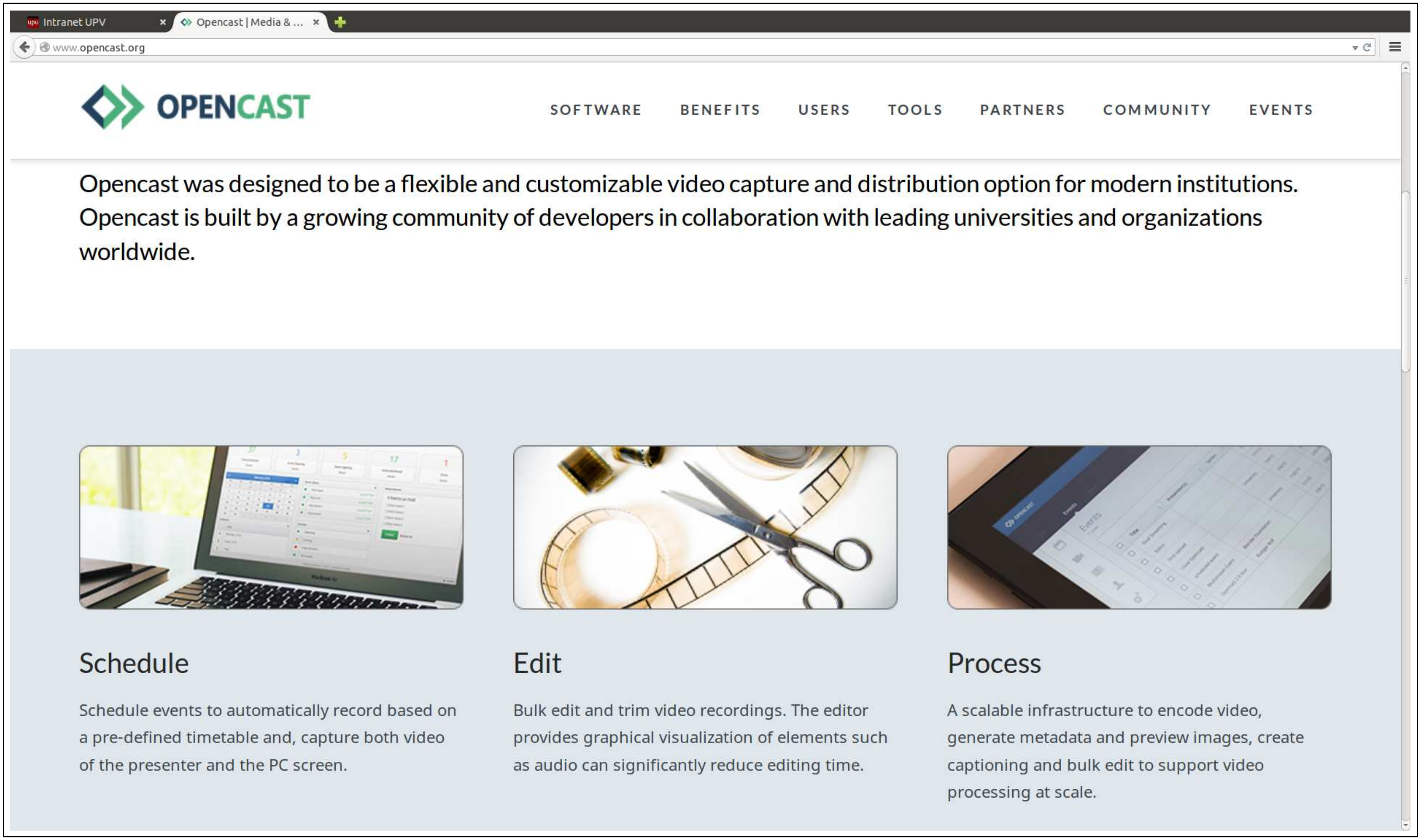
transLectures: Intelligent interaction



poliMedia transcription

The 10% of words recognized with minimum confidence include 30% of the actual word recognition errors

transLectures: Integration into Opencast



Opencast was designed to be a flexible and customizable video capture and distribution option for modern institutions. Opencast is built by a growing community of developers in collaboration with leading universities and organizations worldwide.

Schedule

Schedule events to automatically record based on a pre-defined timetable and, capture both video of the presenter and the PC screen.

Edit

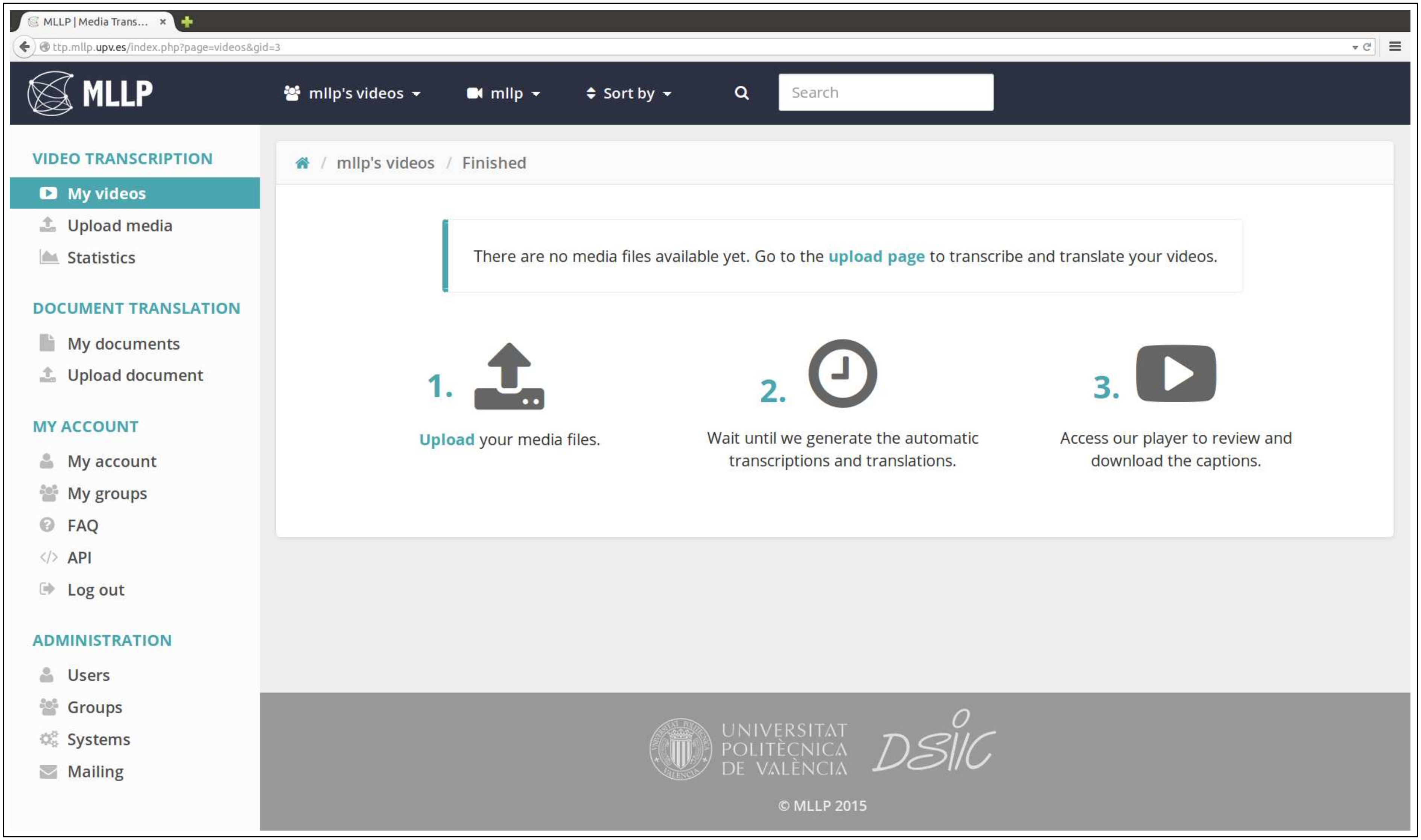
Bulk edit and trim video recordings. The editor provides graphical visualization of elements such as audio can significantly reduce editing time.

Process

A scalable infrastructure to encode video, generate metadata and preview images, create captioning and bulk edit to support video processing at scale.

>300 different institutions worldwide

transLectures: try our tools



There are no media files available yet. Go to the upload page to transcribe and translate your videos.

1. Upload your media files.

2. Wait until we generate the automatic transcriptions and translations.

3. Access our player to review and download the captions.

Try our tools at `ttp.mllp.upv.es`

190 users and 1197 videos (264 hours) since May 2014

trans**Lectures**: try our tools (cont.)

Institution	Country	Language
Univ. Politècnica de València	Spain	ES, CA, EN
Univ. Carlos III de Madrid	Spain	ES, EN
Universidade Aberta	Portugal	PT, EN
University of Naples Federico II	Italy	IT, EN
Open Universiteit Nederland	Netherlands	NL, EN
University of Leicester	UK	EN, ES

Examples:

- Lecture recorded at poliMedia studios. [link]
- Lee Rubenstein, edX VP of Business Development: “Reinventing education”, 30 June 2015, València. [link]
- Text-to-Speech demo [link]

EMMA (Feb 2014 – Jul 2016)

European Multiple MOOC Aggregator (EMMA)

Main goal: to provide multilingual access to European MOOCs

Motivation:

- Most MOOCs are offered in few languages (En, Es, Fr)
- Language barrier is keeping many learners from taking MOOCs
- MOOC components: texts, images, videos, forums
- EMMA uses trans**Lectures** tools to translate videos and texts
  - Few hours of video in 7 En, Es, It, Nl, Et, Pt and Fr
  - Source language is the national language of the MOOC provider
  - Target languages: En, Es and It

EMMA: cost of manually translating MOOCs

Texts:

- Manual translation rate is approximately 2500 words per day
- A 6-week course with 75000 words takes 1.5 PM

Videos:

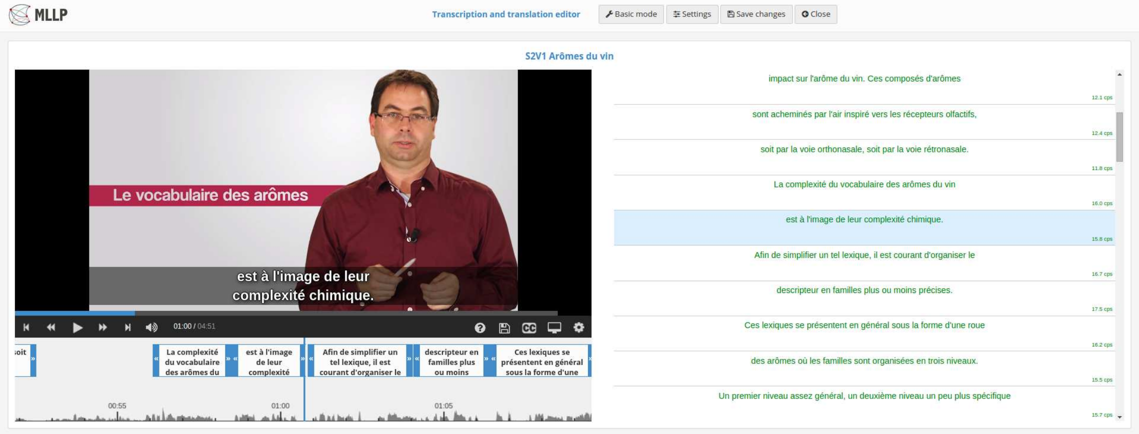
- Before translating, videos are manually transcribed (10 RTF)
- Then, transcriptions are translated (30 RTF)
- A course including 2 hours of video takes 0.5 PM

Solutions to lower costs:

- Crowdsourcing (e.g. TED talks)
- ASR and MT: user effort is reduced to 30% (2 $\rightarrow$ 0.6 PM)

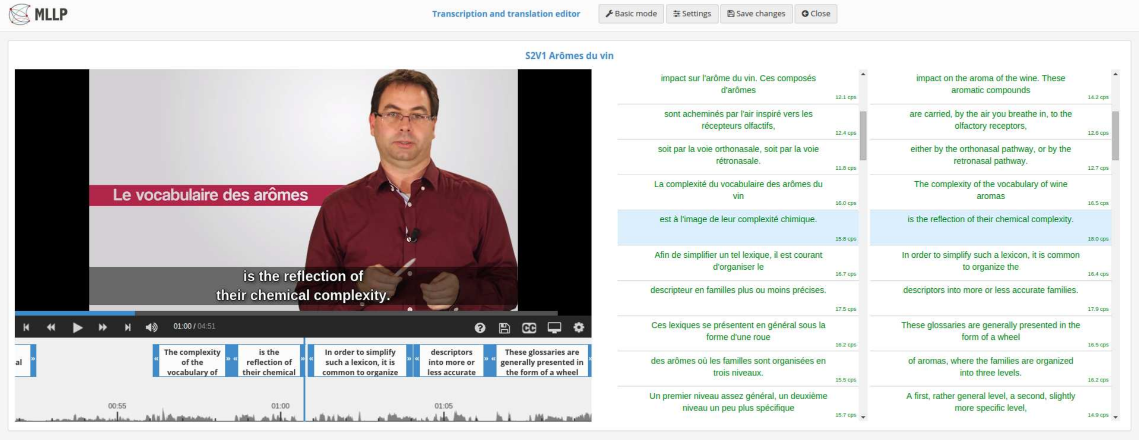
EMMA: automatic video subtitling

1. Generation of ***automatic transcriptions*** from video
2. ***Manual review*** of automatic transcriptions


MLLP
Transcription and translation editor
Basic mode
Settings
Save changes
Close
S2V1 Arômes du vin
Le vocabulaire des arômes
est à l'image de leur complexité chimique.
01:00 / 04:51
soit
La complexité du vocabulaire des arômes du
est à l'image de leur complexité
Afin de simplifier un tel lexique, il est courant d'organiser le
descripteur en familles plus ou moins
Ces lexiques se présentent en général sous la forme d'une
00:55
01:00
01:05

impact sur l'arôme du vin. Ces composés d'arômes	12.1 cps
sont acheminés par l'air inspiré vers les récepteurs olfactifs,	12.4 cps
soit par la voie orthonasale, soit par la voie rétronasale.	11.8 cps
La complexité du vocabulaire des arômes du vin	16.0 cps
est à l'image de leur complexité chimique.	15.8 cps
Afin de simplifier un tel lexique, il est courant d'organiser le	16.7 cps
descripteur en familles plus ou moins précises.	17.5 cps
Ces lexiques se présentent en général sous la forme d'une roue	16.2 cps
des arômes où les familles sont organisées en trois niveaux.	15.5 cps
Un premier niveau assez général, un deuxième niveau un peu plus spécifique	15.7 cps

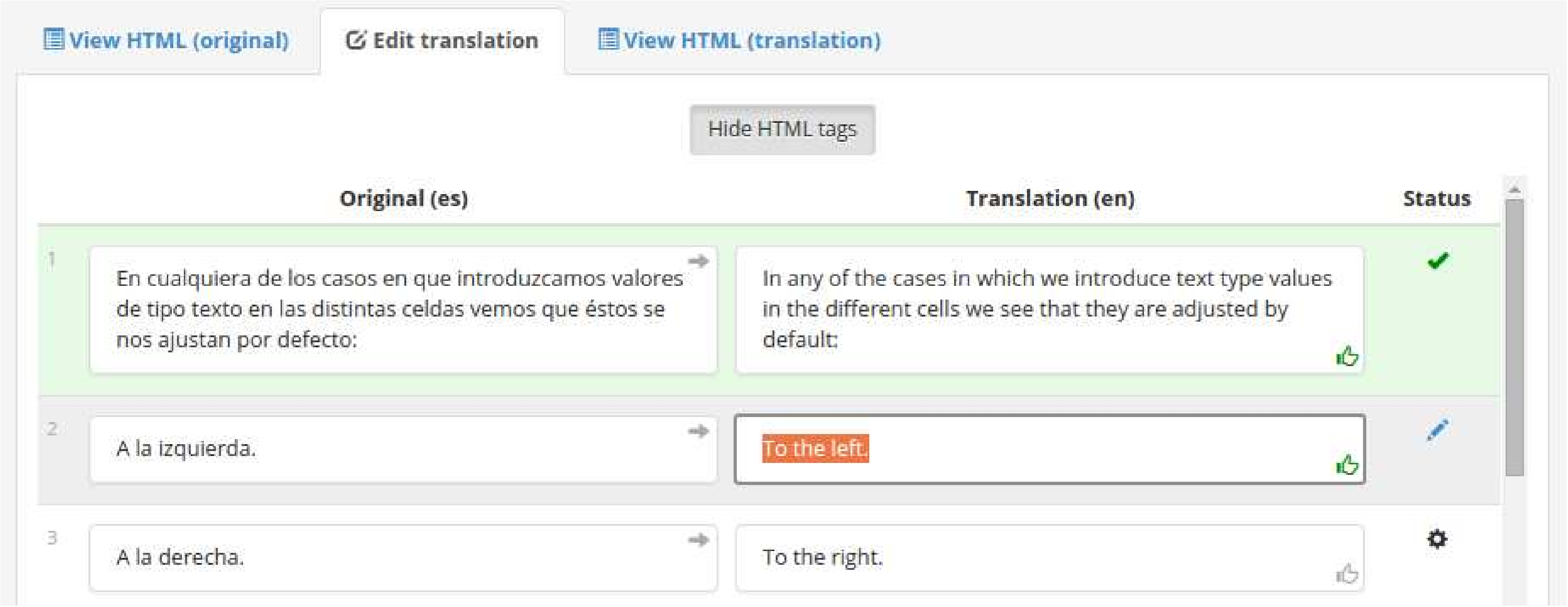
3. Generation of ***automatic translations*** from transcriptions
4. ***Manual review*** of automatic translations


MLLP
Transcription and translation editor
Basic mode
Settings
Save changes
Close
S2V1 Arômes du vin
Le vocabulaire des arômes
is the reflection of their chemical complexity.
01:00 / 04:51
al
The complexity of the vocabulary of
is the reflection of their chemical
In order to simplify such a lexicon, it is common to organize
descriptors into more or less accurate
These glossaries are generally presented in the form of a wheel
00:55
01:00
01:05

impact sur l'arôme du vin. Ces composés d'arômes	12.1 cps	impact on the aroma of the wine. These aromatic compounds	14.2 cps
sont acheminés par l'air inspiré vers les récepteurs olfactifs,	12.4 cps	are carried, by the air you breathe in, to the olfactory receptors,	12.6 cps
soit par la voie orthonasale, soit par la voie rétronasale.	11.8 cps	either by the orthonasal pathway, or by the retronasal pathway.	12.7 cps
La complexité du vocabulaire des arômes du vin	16.0 cps	The complexity of the vocabulary of wine aromas	16.5 cps
est à l'image de leur complexité chimique.	15.8 cps	is the reflection of their chemical complexity.	18.0 cps
Afin de simplifier un tel lexique, il est courant d'organiser le	16.7 cps	In order to simplify such a lexicon, it is common to organize the	16.4 cps
descripteur en familles plus ou moins précises.	17.5 cps	descriptors into more or less accurate families.	17.9 cps
Ces lexiques se présentent en général sous la forme d'une roue	16.2 cps	These glossaries are generally presented in the form of a wheel	16.5 cps
des arômes où les familles sont organisées en trois niveaux.	15.5 cps	of aromas, where the families are organized into three levels.	16.2 cps
Un premier niveau assez général, un deuxième niveau un peu plus spécifique	15.7 cps	A first, rather general level, a second, slightly more specific level,	14.9 cps

EMMA: automatic document translation

- Course text is ingested into the translation system
- Source and target texts are reviewed in parallel
- Preview of source and target texts also available
- Translated text is imported back into the EMMA platform



	Original (es)	Translation (en)	Status
1	En cualquiera de los casos en que introduzcamos valores de tipo texto en las distintas celdas vemos que éstos se nos ajustan por defecto:	In any of the cases in which we introduce text type values in the different cells we see that they are adjusted by default:	✓
2	A la izquierda.	To the left.	✎
3	A la derecha.	To the right.	⚙

EMMA: evaluations

Video subtitling

Language pairs	Transcription RTF (10)	Translation RTF (30)	Total RTF (40)
Spanish → English	3	7	10
English → Spanish	6	17	23

Document translation

Language pairs	Translation RTF
Spanish → English	7
English → Spanish	17

EMMA: conclusions

- Multilingual access to your course boosts visibility
- The cost of manually translating your course is high (2 PM)
- Automatic translation can reduce the temporal cost to 30%
- Accuracy of automatic translation depends on several factors:
  - Languages involved
  - Availability of annotated data resources related to your course
  - Specificity of the course
- Designing a multilingual MOOC should also take into account: slides, images, application interfaces (demos), bibliography