



# TraMOOC

Translation for Massive Open Online Courses

@enetCollect #Cost

---



- TraMOOC in a nutshell
- Project Motivation
- Project Objectives
- Work Description
- The TraMOOC Platform
- The TraMOOC Consortium

SUMMARY

WHY?

WHAT?

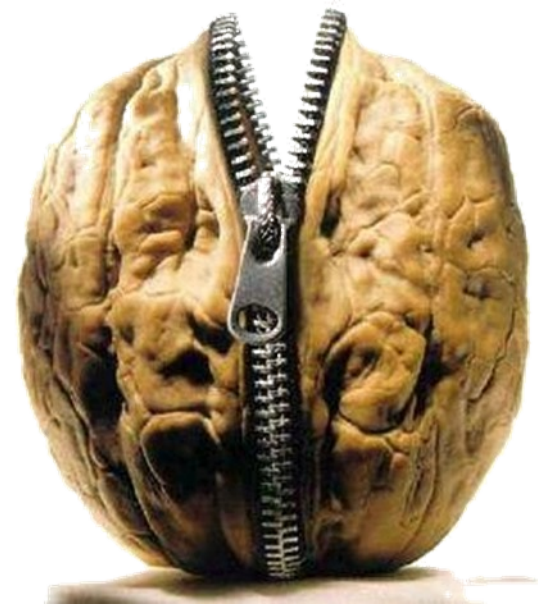
HOW?

RESULT?

WHO?

# Table of contents

- TraMOOC in a nutshell
- Project Motivation
- Project Objectives
- Work Description
- The TraMOOC Platform
- The TraMOOC Consortium



SUMMARY  
WHY?  
WHAT?  
HOW?  
RESULT?  
WHO?



15/03/2019,  
#enetCollect  
#cost #Lisbon

*Valia Kordoni (UBER,  
TraMOOC Coordinator)*

TraMOOC  
Confidential





## Project details

<b><i>Partially funded</i></b>	European Union's Horizon 2020 research and innovation programme under grant agreement No 644333
<b><i>Thematic priority</i></b>	Information and Communications Technologies (ICT)
<b><i>Topic</i></b>	Approved under 1 <sup>st</sup> call of the ICT priority in the strategic objective "Cracking the language barrier"
<b><i>Duration</i></b>	36 months (From 2015-02-01 to 2018-02-01)
<b><i>Budget</i></b>	Approximately 3M€
<b><i>Coordinated by</i></b>	Humboldt-Universität zu Berlin (UBER)
<b><i>Consortium</i></b>	9 organizations from 6 European countries

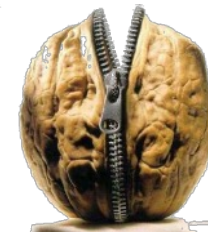


15/03/2019,  
#enetCollect  
#cost #Lisbon

**Valia Kordoni (UBER,  
TraMOOC Coordinator)**

TraMOOC  
Confidential





## *Objectives & expected impacts*

- TraMOOC has made existing monolingual educational material available to speakers of other languages.
- The project's vision has been to tear down language barriers, thus providing previously excluded groups of people with new educational chances.
- The project results has been showcased and tested on the openHPI platform and on the VideoLectures.Net digital video lecture library.
- The core of the service is open-source, with some premium add-on services which will be commercialised.
- The translation methodology is automatic and language-independent in nature and showcased for 11 indicative language pairs - 9 EU (DE, IT, PT, DU, BG, EL, PL, CS and CR), and 2 BRIC languages (RU and ZH).

15/03/2019,  
#enetCollect  
#cost #Lisbon

**Valia Kordoni (UBER,  
TraMOOC Coordinator)**

TraMOOC  
Confidential





## *Main novelties*

- Novel research in online and open education
  - Novel translation evaluation schemata
  - Added value to existing tools and resources in linguistics, natural language processing, text analytics, data mining and machine translation scientific communities
  - Topic identification of the source and translated text
  - Sentiment analysis on users' posts on fora and social websites has been used for extracting users' opinion on the translated material



15/03/2019,  
#enetCollect  
#cost #Lisbon

*Valia Kordoni (UBER,  
TraMOOC Coordinator)*

TraMOOC  
Confidential



# Table of contents

- TraMOOC in a nutshell
- **Project Motivation**
- Project Objectives
- Work Description
- The TraMOOC Platform
- The TraMOOC Consortium



SUMMARY

**WHY?**

WHAT?

HOW?

RESULT?

WHO?

15/03/2019,  
#enetCollect  
#cost #Lisbon

*Valia Kordoni (UBER,  
TraMOOC Coordinator)*

TraMOOC  
Confidential

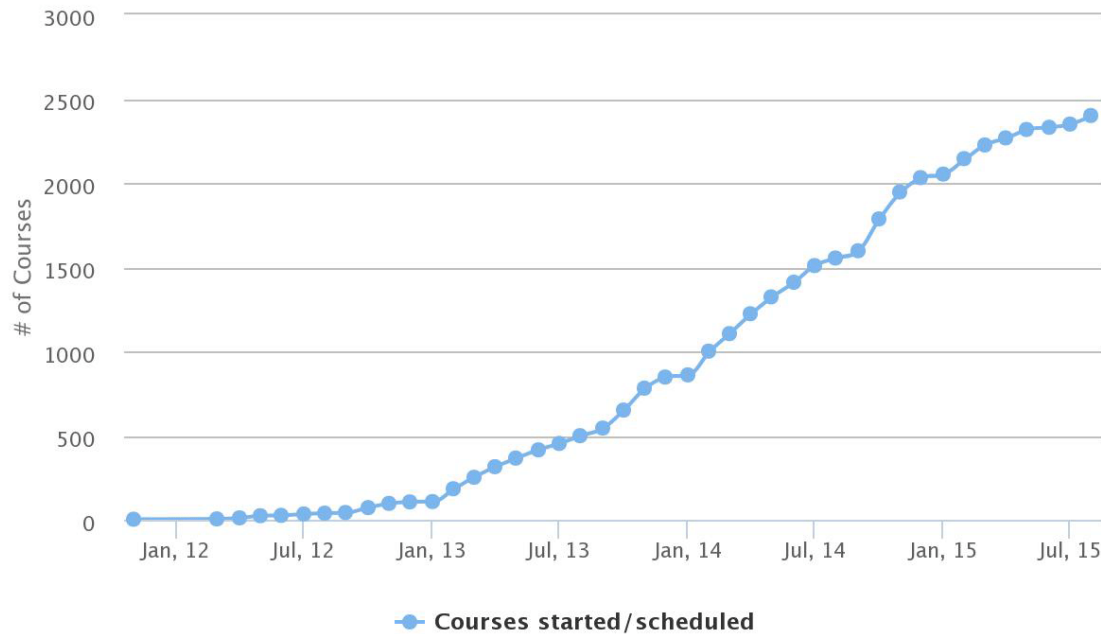




- MOOCs have been growing rapidly in size and impact

## Growth of MOOCs

Cumulative number of courses started/scheduled



*This year, the number of universities offering MOOCs has doubled to exceed 400 universities, with a doubling of the number of cumulative courses offered, to 2400. It is estimated that 16-18 million students attend MOOCs worldwide\*.*

\* Source: <https://www.edsurge.com/n/2014-12-26-moocs-in-2014-breaking-down-the-numbers>



15/03/2019,  
#enetCollect  
#cost #Lisbon

**Valia Kordoni (UBER,  
TraMOOC Coordinator)**

TraMOOC  
Confidential





# Table of contents



- TraMOOC in a nutshell
- Project Motivation
- Project Objectives
- **Work Description**
- The TraMOOC Platform
- The TraMOOC Consortium

- SUMMARY
- WHY?
- WHAT?
- HOW?**
- RESULT?
- WHO?



15/03/2019,  
#enetCollect  
#cost #Lisbon

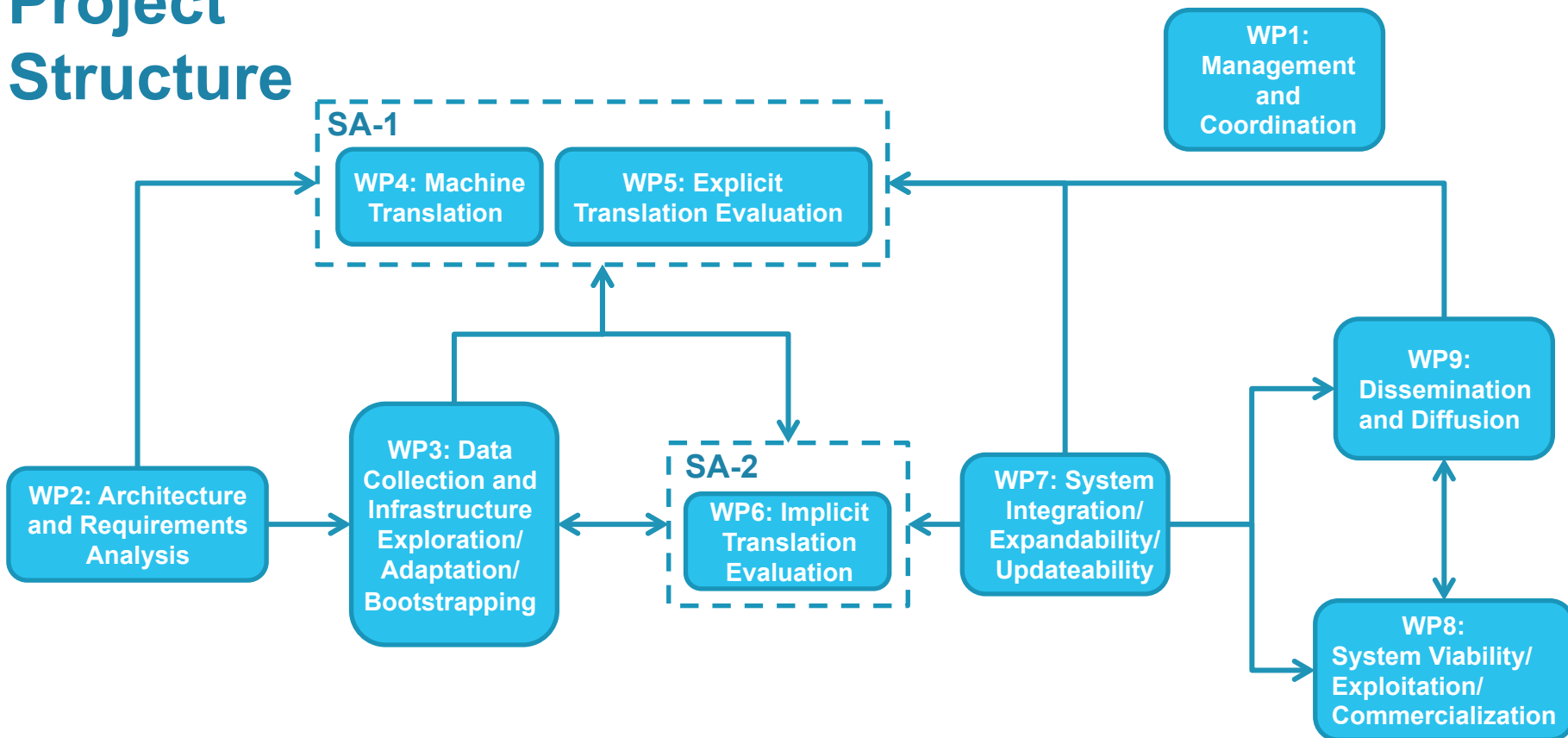
*Valia Kordoni (UBER,  
TraMOOC Coordinator)*

TraMOOC  
Confidential





## Project Structure



15/03/2019,  
#enetCollect  
#cost #Lisbon

*Valia Kordoni (UBER,  
TraMOOC Coordinator)*

TraMOOC  
Confidential



# Table of contents



SUMMARY  
WHY?  
WHAT?  
HOW?  
RESULT?  
WHO?

- TraMOOC in a nutshell
- Project Motivation
- Project Objectives
- Work Description
- The TraMOOC Platform
- The TraMOOC Consortium

15/03/2019,  
#enetCollect  
#cost #Lisbon

*Valia Kordoni (UBER,  
TraMOOC Coordinator)*

TraMOOC  
Confidential



# TraMOOC

Translation for Massive Open Online Courses

# The TraMOOC Consortium



- TraMOOC brings together a consortium of leading researchers, highly relevant industry organizations and leading use-case partners.
- The partners' diverse interests in machine translation, linguistics, text mining web analytics and crowdsourcing

methodologies-related areas make the consortium ideally placed to tackle the challenges associated with TraMOOC.

- The design of the scientific areas and the associated work packages have been arranged carefully to ensure maximum efficiency of input from each partner while maintaining a suitable distribution of responsibilities.



15/03/2019,  
#enetCollect  
#cost #Lisbon

**Valia Kordoni (UBER,  
TraMOOC Coordinator)**

TraMOOC  
Confidential



# Crowdsourcing Platform Selection

Multiple platforms have been researched and ranked.

CrowdFlower was ranked second best after Amazon Mechanical Turk, the latter being rejected due to its inflexible USA-based payment process.

CrowdFlower was selected due to its:

- configurability,
- robust infrastructure,
- densely populated crowd channels and the evaluation and ranking process they undergo,
- convenient payment options,
- high reception and popularity level in the microtasking field

## Crowdsourcing Activities:

1. CA1: Translation
2. CA2: Translation Evaluation
3. CA3: Sentiment/Topic Annotation

Activity: Parallel translation of EN segments to 11 target languages, 11M segments.

Data Sources: iversity, Coursera, QED Corpus

**Translate This Sentence in Greek!**

Sentence	Translation
Each time you do a web search on Google or Bing, that works so well because their machine learning software has figured out how to rank web pages.	

EL (146 participants) – NL (36 participants) with Gold Standard translations

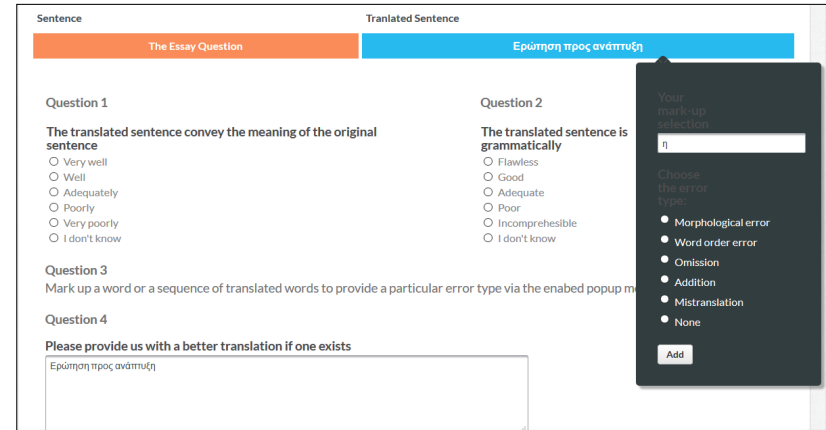
- Design refinements and settings calibration, e.g.
  - Time frame allocated to completing a job
  - Number of segments/job etc
- Quality Control Approach refinement (evaluation), e.g.
  - Number of test questions
  - Type of test questions

English sentence	Translated Sentence
Mergers and acquisitions	Choose best Greek translation from the options below: <ul style="list-style-type: none"><li><input type="radio"/> Συγωνεύσεις κι εξαγορές</li><li><input type="radio"/> Συγχωνεύσεις και εξαγορές</li><li><input type="radio"/> συγχωνεύσεις και εξαγορές</li></ul>

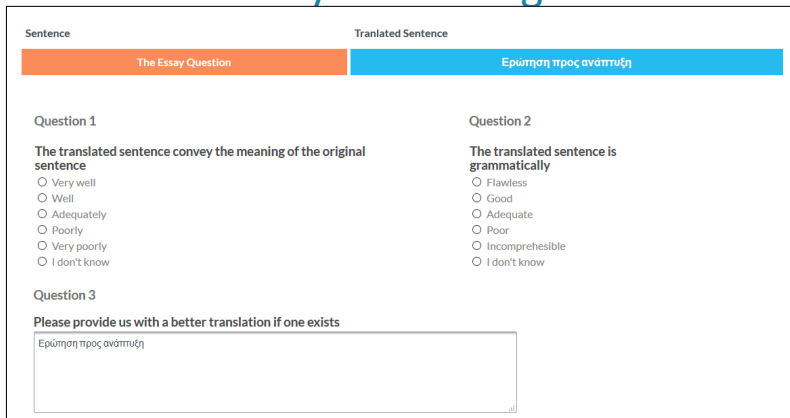


### Activity:

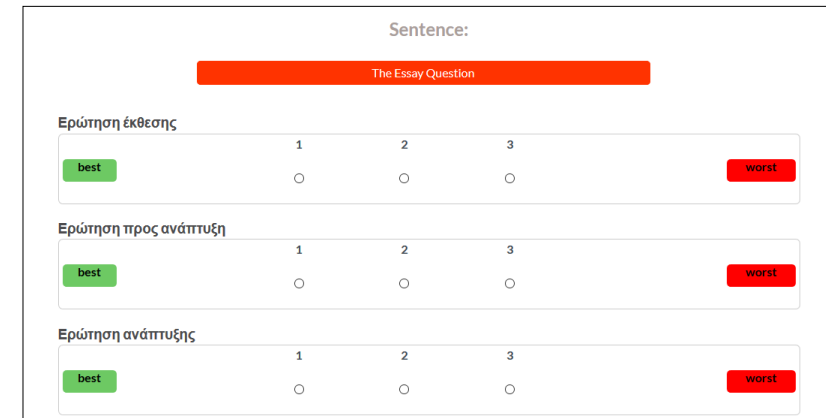
- Sub-activity 1: Adequacy/Fluency Markup, Error Type Markup, and Post-editing
- Sub-activity 2: Adequacy/Fluency Markup, and Post-editing
- Sub-activity 3: Ranking



The screenshot shows a user interface for evaluating a translation. It has two columns: 'Sentence' and 'Translated Sentence'. The 'Sentence' column contains 'The Essay Question' and 'Question 1: The translated sentence convey the meaning of the original sentence'. The 'Translated Sentence' column contains 'Ερώτηση προς ανάπτυξη' and 'Question 2: The translated sentence is grammatically'. A pop-up menu is open on the right, titled 'Your mark-up selection', with a search bar containing the Greek letter η. Below the search bar, it says 'Choose the error type:' and lists options: Morphological error, Word order error, Omission, Addition, Mistranslation, and None. An 'Add' button is at the bottom of the pop-up.



This screenshot shows the same interface as above, but with a different question in the 'Translated Sentence' column: 'Question 3: Please provide us with a better translation if one exists'. Below this question is a text input field with the placeholder text 'Ερώτηση προς ανάπτυξη'.



This screenshot shows a ranking task. At the top, it says 'Sentence:'. Below that is a red bar labeled 'The Essay Question'. There are three rows of ranking questions, each with three columns labeled '1', '2', and '3'. The first row is 'Ερώτηση έκθεσης', the second is 'Ερώτηση προς ανάπτυξη', and the third is 'Ερώτηση ανάπτυξης'. Each row has a 'best' button on the left and a 'worst' button on the right, with radio buttons in between.

### Activity:

- 1.Sub-activity 1: Sentiment Annotation
- 2.Sub-activity2: Topic/Entity Annotation

Sentence:

Social Innovation - just a buzzword or a substantial concept?

What is the author's sentiment of the previous sentence?

Positive-happy,excited,enthusiastic,complimentary

Neutral-questions,unrelated comments,mixed feelings

Negative-hate, anger, sadness, frustration

Positive **0** Neutral **0** Negative **0**

Mark up any entities that appear in the following sentence

Course : "Critical Thinking, Chapter 1.1"

English Context

"Hello, everybody, and welcome to this Massive Online Open Course in Critical Thinking! My name is Radu Atanasiu and I teach Critical Thinking for the Executive MBA at Maastricht School of Management Romania, here, in Bucharest."

Greek Context

"Γεια σας και καλωσορίσατε σε αυτό το Ανοιχτό Μαζικό Διαδίκτυο Μάθημα για την Κριτική Σκέψη! Ονομάζομαι Radu Atanasiu και διδάσκω Κριτική Σκέψη για το Μεταπτυχιακό Διοίκησης Επιχειρήσεων για Στελέχη στη Σχολή Διοίκησης του"

Mark up the English Entities

"This is the first Massive Online Open Course that is produced in Romania and I'm very, very proud to be part of this exciting project."

Align the Greek Entities

Αυτό είναι το πρώτο Ανοιχτό Μαζικό Διαδίκτυο Μάθημα που γίνεται εδώ στη Ρουμανία και είμαι πάρα πολύ περήφανος που είμαι μέλος αυτού του συναρπαστικού εγχειρήματος.

1. Search the entity in wikipedia.      2. Mark up and Add the entity.      3. Paste Wikipedia URL      4. Add Wikipedia URL

English Wikipedia Homepage	Massive Online Open Course	/wiki/Massive_open_online_course
Greek Wikipedia Homepage	Ανοιχτό Μαζικό Διαδίκτυο Μάθημα	NONE <input type="button" value="Add URL"/>



- Tokenization
- Sentence segmentation
- Truecasing
- Assure proper sentence alignment
- Marking of URLs
- Other steps specific to each data source (e.g., conversion from PDF into plain text)
- Goal: well-tokenized and as much as possible well-segmented grammatical parallel data
- Generally performed using a pipeline of available sentence splitters/aligners, data source tailored Python and shell scripts
- All data stored in a protected data repository provided by UBER

- Need for a uniform in-domain test set throughout the project
- 80,000 words extracted from MOOC materials provided by IVE and DME
- Manual translation into Greek, Italian, and Portuguese performed by the respective partners
- 3 of the 4 language pairs in MT prototype 1 covered
- The translations for the remaining 8 languages were produced via crowdsourcing



- Challenging crawling: often complicated structure of the web resource; did not allow for large-scale automatic crawling
- Challenging data extraction and alignment: most materials in PDF, possible misalignments during conversion into plain text
- Representativeness: slides, notes, assignments are rarely translated
- Copyright issues



Language pair	Size (million words)
EN-DE	2.7
EN-BG	1.5
EN-PT	4.8
EN-EL	2.4
EN-NL	1.3
EN-CZ	1.5
EN-RU	1.4
EN-CR	0.2
EN-PL	1.7
EN-IT	2.3
EN-ZH	8.7

- A case study with Croatian and Serbian
- Vanilla Moses trained on Coursera data in Croatian and Serbian shown to outperform a system trained only on Croatian in- and out-of-domain data
- High-quality MT of Serbian Coursera data into Croatian
- The Croatian data resulting from (rule-based) MT was added to the “normal” Croatian Coursera in-domain training corpus: best performing system

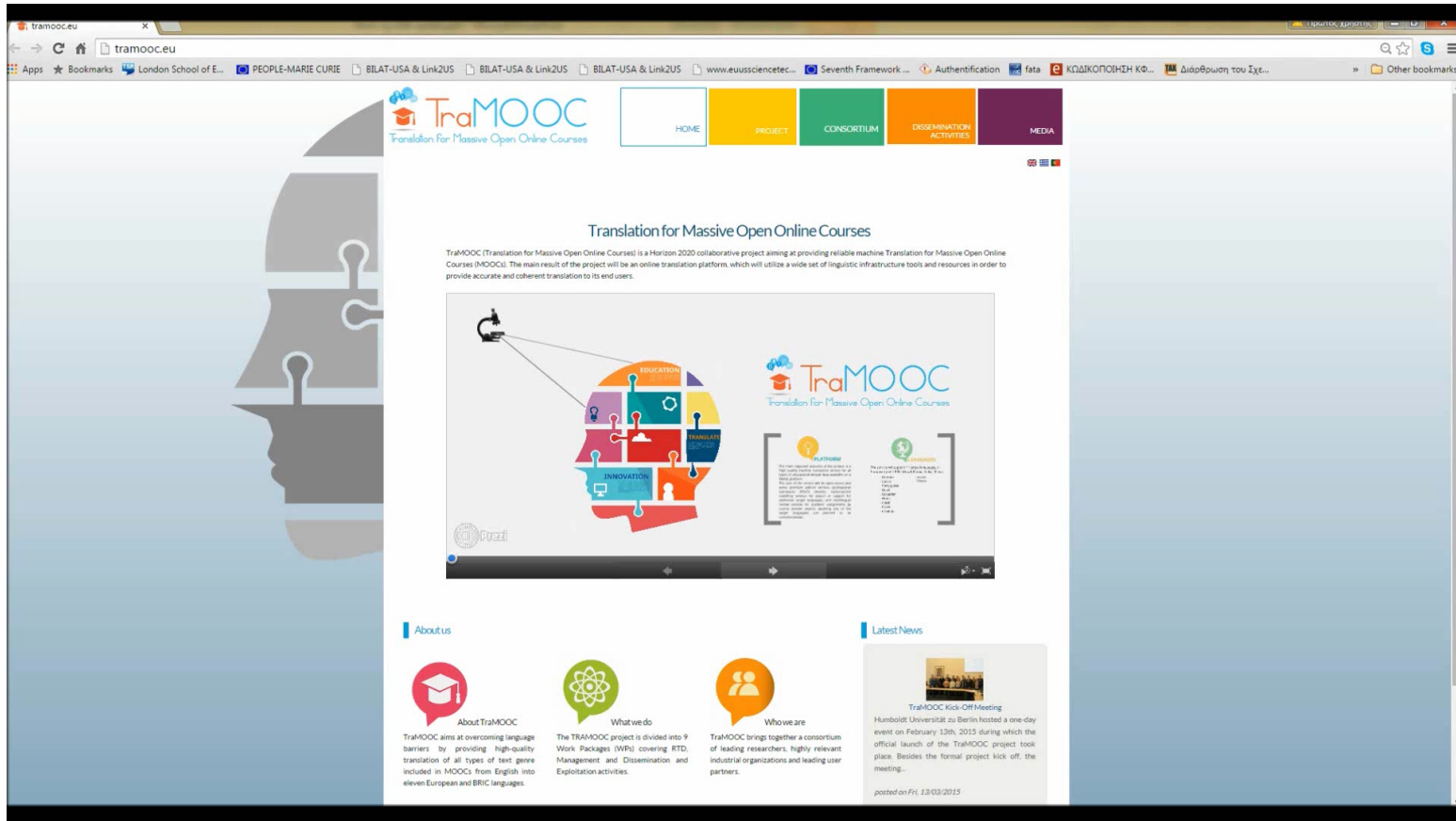


# TraMOOC The TraMOOC website

Translation for Massive Open Online Courses



Find out more about the TraMOOC project and platform at



15/03/2019,  
#enetCollect  
#cost #Lisbon

**Valia Kordoni (UBER,  
TraMOOC Coordinator)**

TraMOOC  
Confidential



SUMMARY

WHY

WHAT

HOW

RESULT

WHO





*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644333.*

*This document and all information contained herein is the sole property of the TraMOOC Consortium or the company referred to in the slides. It may contain information subject to intellectual property rights. No intellectual property rights are granted by the delivery of this document or the disclosure of its content. Reproduction or circulation of this document to any third party is prohibited without the consent of the author(s).*

*The statements made herein do not necessarily have the consent or agreement of the TraMOOC consortium and represent the opinion and findings of the author(s).*

*All rights reserved.*