TraMOOC

Translation for Massive Open Online Courses

@enetCollect #Cost

TraMOOC
Translation for Massive Open Online Courses

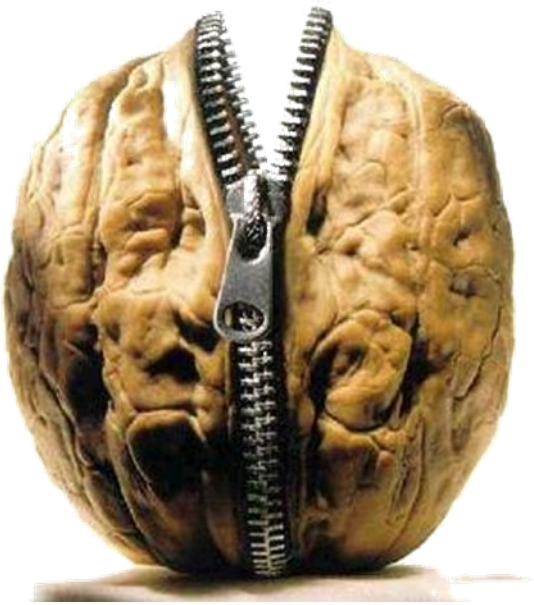
Table of contents

- TraMOOC in a nutshell — SUMMARY
- Project Motivation — WHY?
- Project Objectives — WHAT?
- Work Description — HOW?
- The TraMOOC Platform — RESULT?
- The TraMOOC Consortium — WHO?

TraMOOC
Translation for Massive Open Online Courses

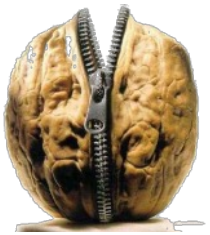
Table of contents

- TraMOOC in a nutshell — SUMMARY
- Project Motivation — WHY?
- Project Objectives — WHAT?
- Work Description — HOW?
- The TraMOOC Platform — RESULT?
- The TraMOOC Consortium — WHO?



TraMOOC
Translation for Massive Open Online Courses

TraMOOC in a nutshell

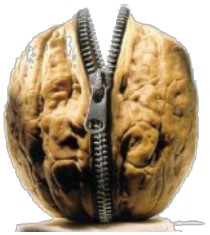


Project details

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

TraMOOC
Translation for Massive Open Online Courses

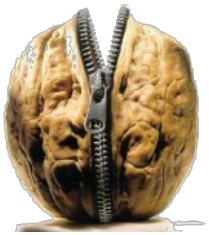
TraMOOC in a nutshell



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).

TraMOOC in a nutshell



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

Table of contents

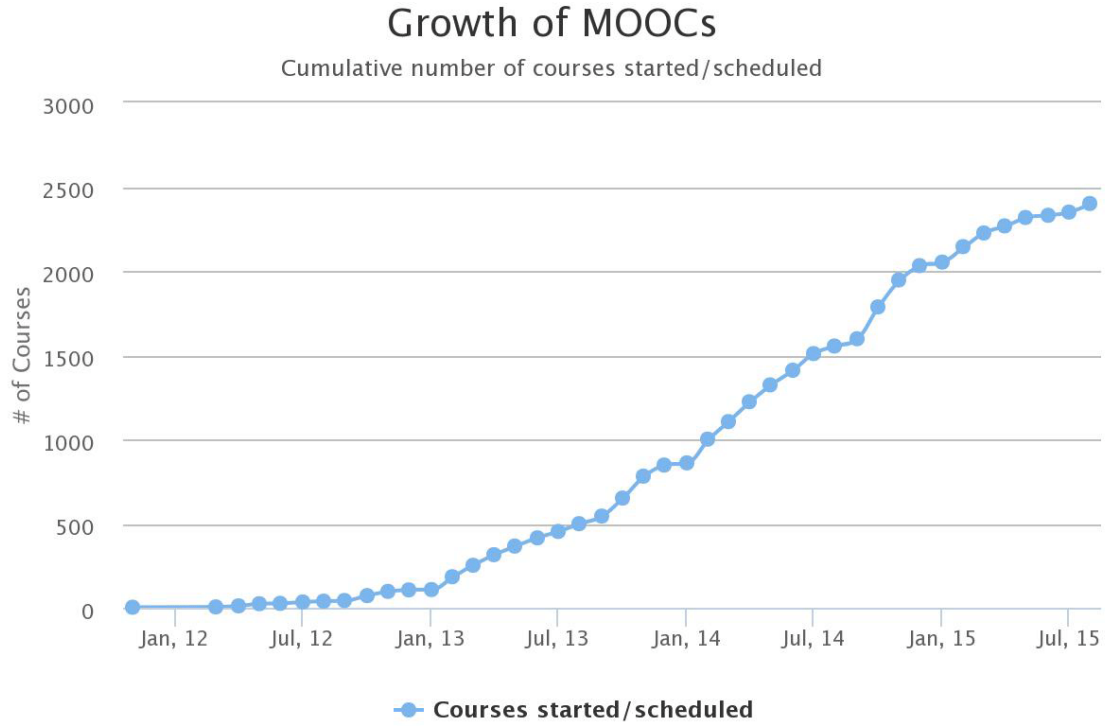
- TraMOOC in a nutshell
- Project Motivation
- Project Objectives
- Work Description
- The TraMOOC Platform
- The TraMOOC Consortium

SUMMARY
WHY?
WHAT?
HOW?
RESULT?
WHO?

TraMOOC
Translation for Massive Open Online Courses

Project Motivation

- MOOCs have been growing rapidly in size and impact



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

Table of contents

- TraMOOC in a nutshell — SUMMARY
- Project Motivation — WHY?
- Project Objectives — WHAT?
- Work Description — HOW?
- The TraMOOC Platform — RESULT?
- The TraMOOC Consortium — WHO?

Work Description

HOW?

Project Structure

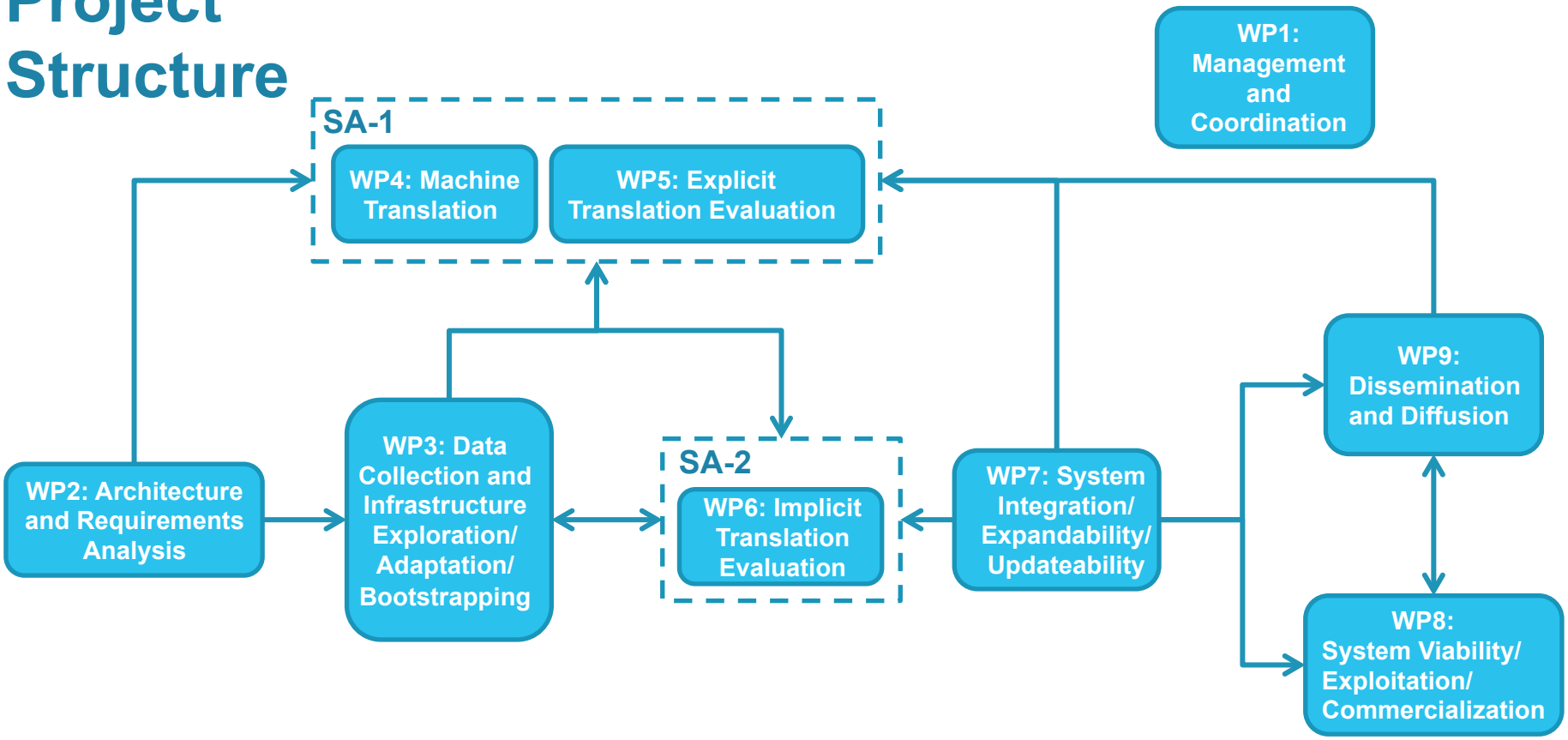


Table of contents

- TraMOOC in a nutshell — SUMMARY
- Project Motivation — WHY?
- Project Objectives — WHAT?
- Work Description — HOW?
- The TraMOOC Platform — RESULT?
- The TraMOOC Consortium — WHO?

TraMOOC
Translation for Massive Open Online Courses

The TraMOOC Consortium

Consortium Members

HUMBOLDT-UNIVERSITÄT ZU BERLIN · DCU · deluxe · THE UNIVERSITY of EDINBURGH · IONIAN UNIVERSITY · Radboud University · TILBURG UNIVERSITY · easn · K4A knowledge for all

- 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.

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 Platform Configuration

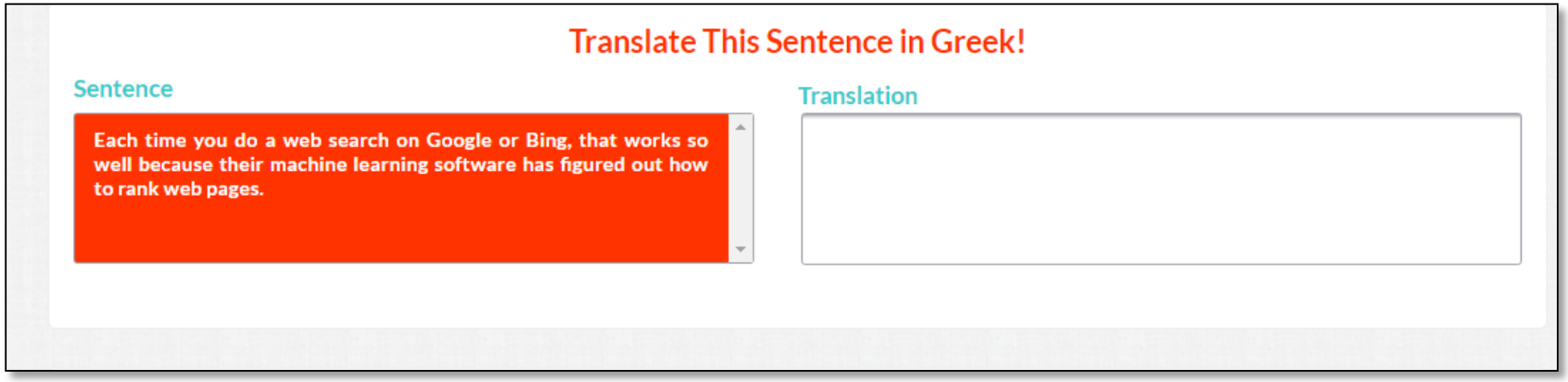
Crowdsourcing Activities:

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

Crowdsourcing Platform Configuration
CA1: Translation

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

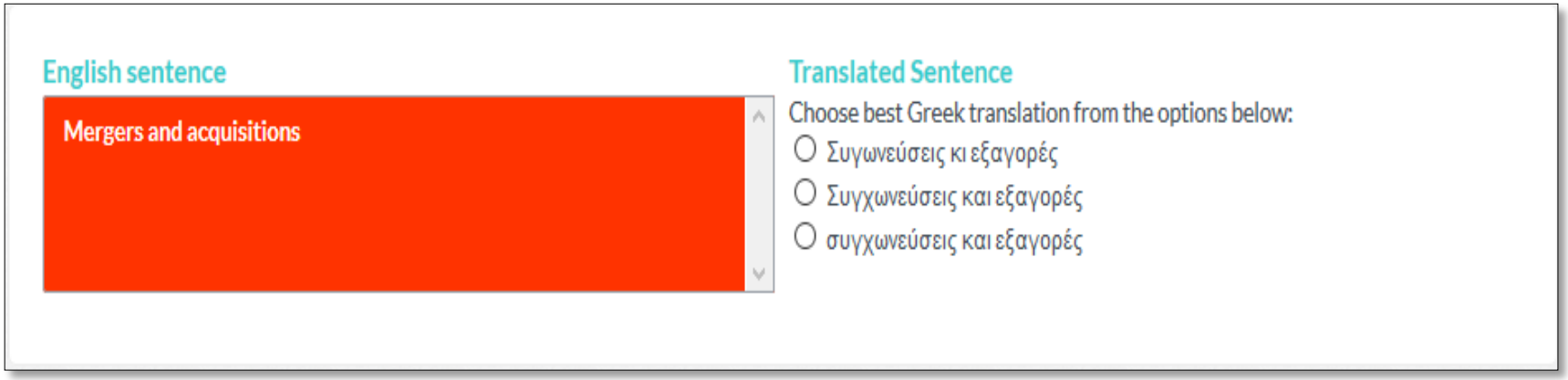
Data Sources: iversity, Coursera, QED Corpus



Crowdsourcing Platform Configuration
CA1: Translation (Trials)

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

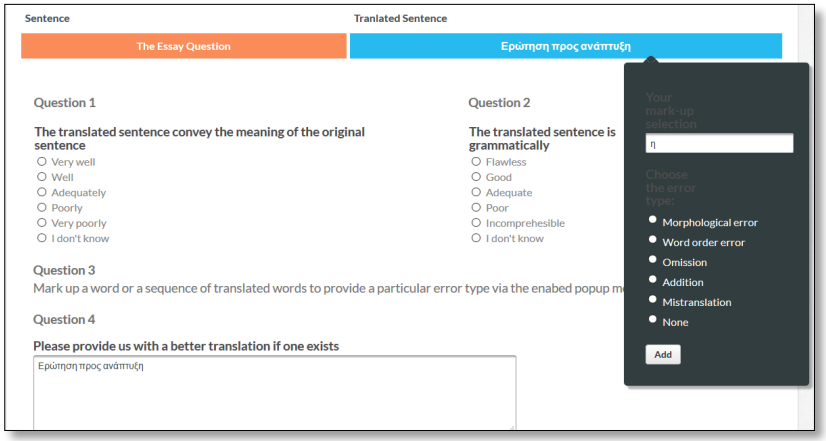


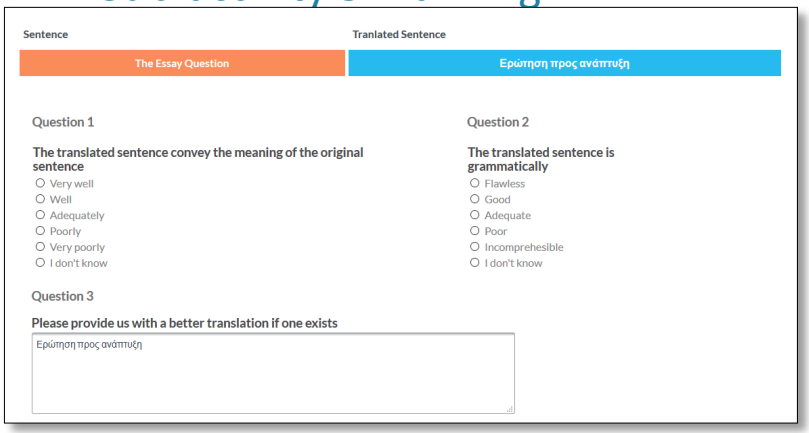
Crowdsourcing Platform Configuration
CA2: Translation Evaluation

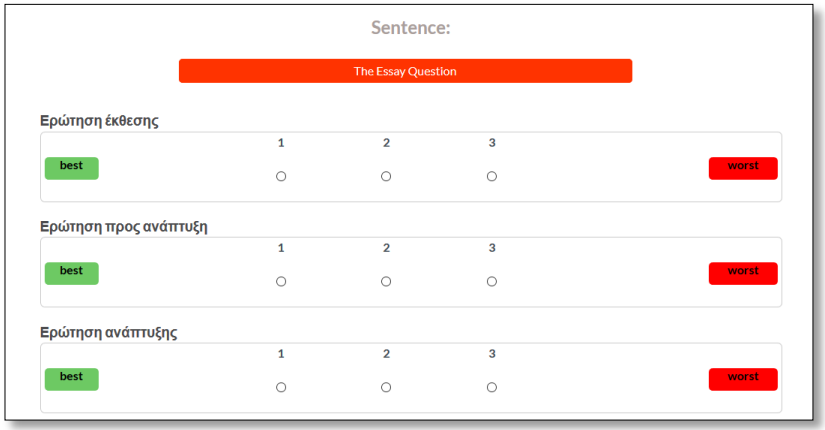
TraMOOC
Translation for Massive Open Online Courses

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



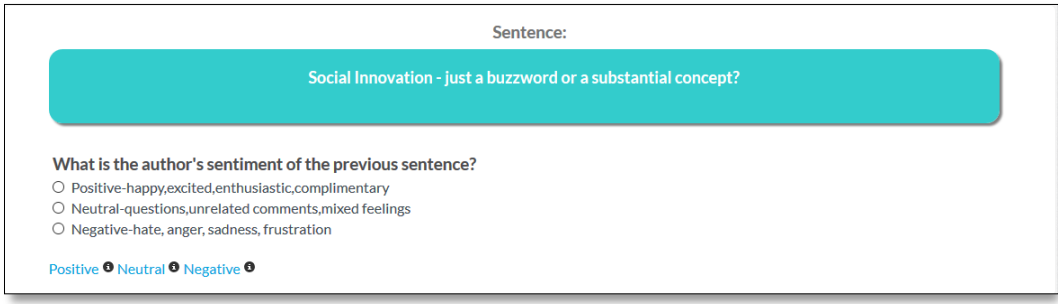


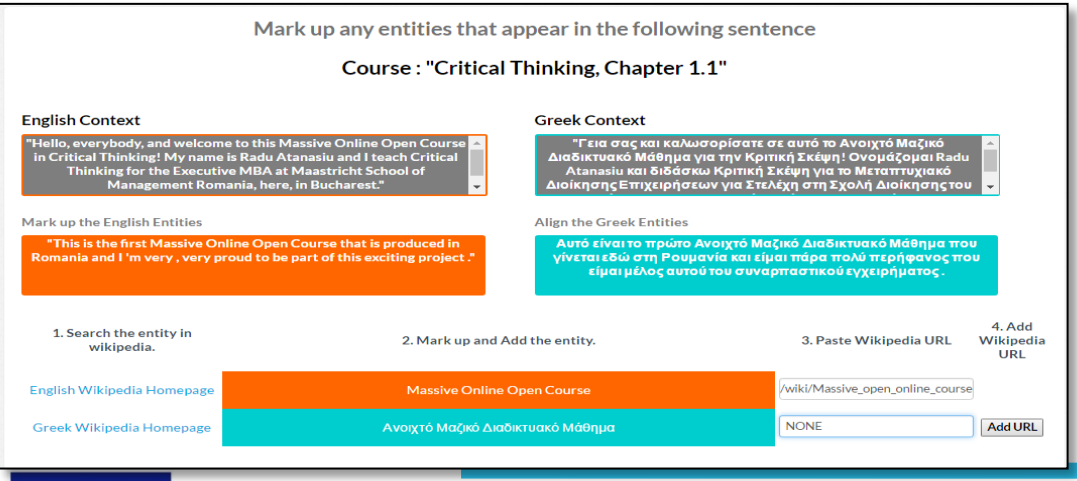


Crowdsourcing Platform Configuration
CA3: Sentiment/Topic Annotation

Activity:

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





Data Processing

- 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

Tuning and Test Sets

- 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

Issues with Data Collection

- 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

In-domain Data Summary

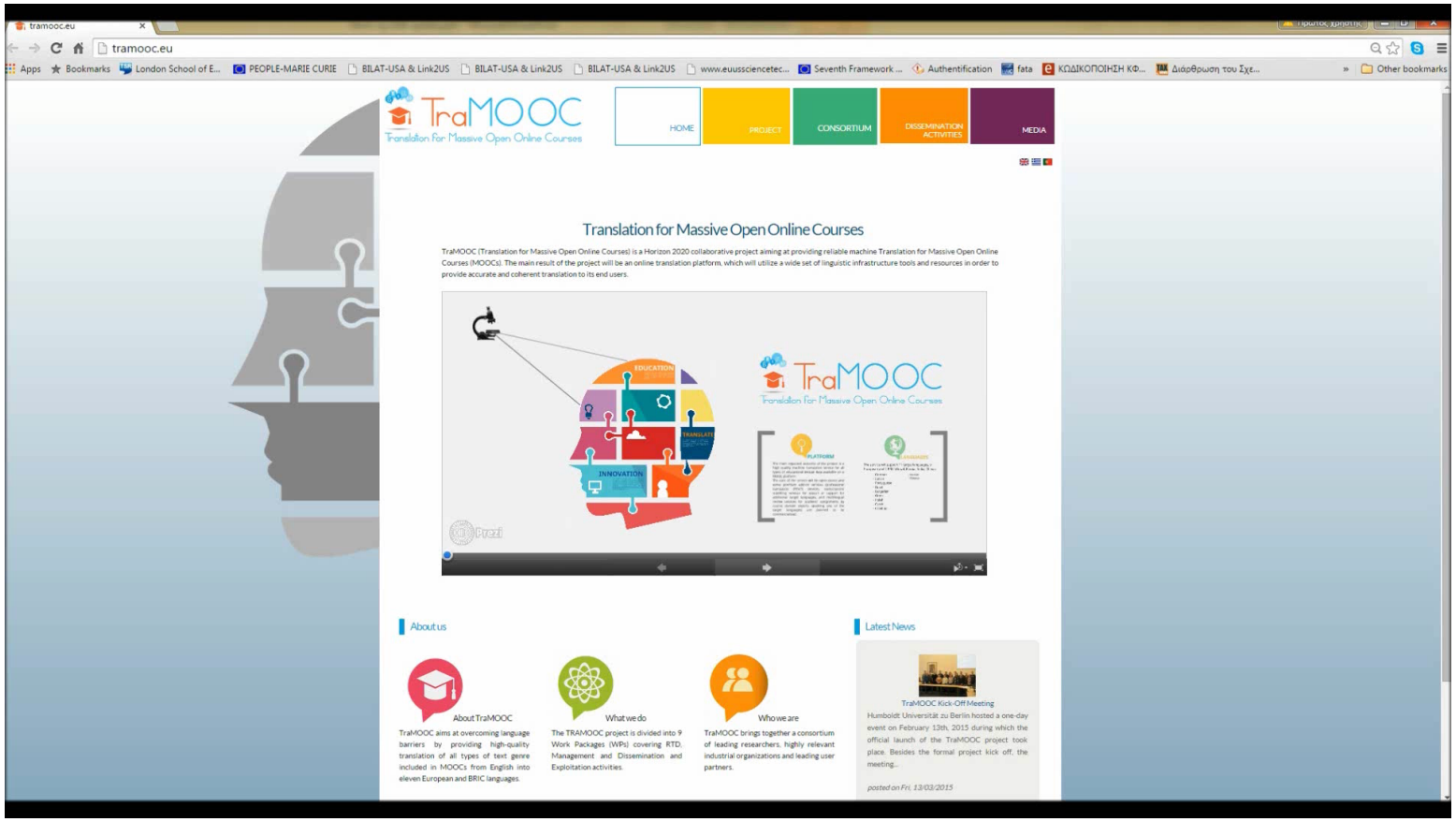
Language pair	Size (million words)
EN-DE	2.7
EN-BG	1.5
EN-PT	4.8
EN-EL	2.4
EN-NL	1.3
EN-CZ	1.5
EN-RU	1.4
EN-CR	0.2
EN-PL	1.7
EN-IT	2.3
EN-ZH	8.7

Bootstrapping of Data/Resources

- 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

The TraMOOC website

Find out more about the TraMOOC project and platform at



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