

Blending **academic** and **entrepreneurial** knowledge  
in technology enhanced learning

# TOWARDS TRANSLATION OF EDUCATIONAL RESOURCES USING GIZA++



# BAEKTEL

*Ivan Obradović*  
*Dalibor Vorkapić*  
*Ranka Stanković*  
*Nikola Vulović*  
*Miladin Kotorčević*

11/8/2016

*University of Belgrade, Faculty of Mining and Geology*

# Content

MOOCs language barrier

Translation for Massive Open Online Courses (TraMOOC)

Related work – Coursera translated lectures

Current approaches for translation of educational resources

Environment for text alignment

Towards machine translation for Serbian

# MOOCs language barrier

## World statistics

500 Universities

4200 Courses

35 million users

### **Problem:**

- The language barrier is the biggest obstacle that stands in the way of broader usage of online courses as the majority of such courses are offered in English

### **So far solutions:**

- The solutions provided so far have been fragmentary, human-based, and implemented off-line by the majority of course providers

# Translation for Massive Open Online Courses (TraMOOC)

- Horizon 2020, EU project
- The main result of the project will be an online translation platform
- It constitutes a solution to online course content translation that aims at eleven target languages
- It is based on statistical machine translation (SMT) techniques
- Serbian not included

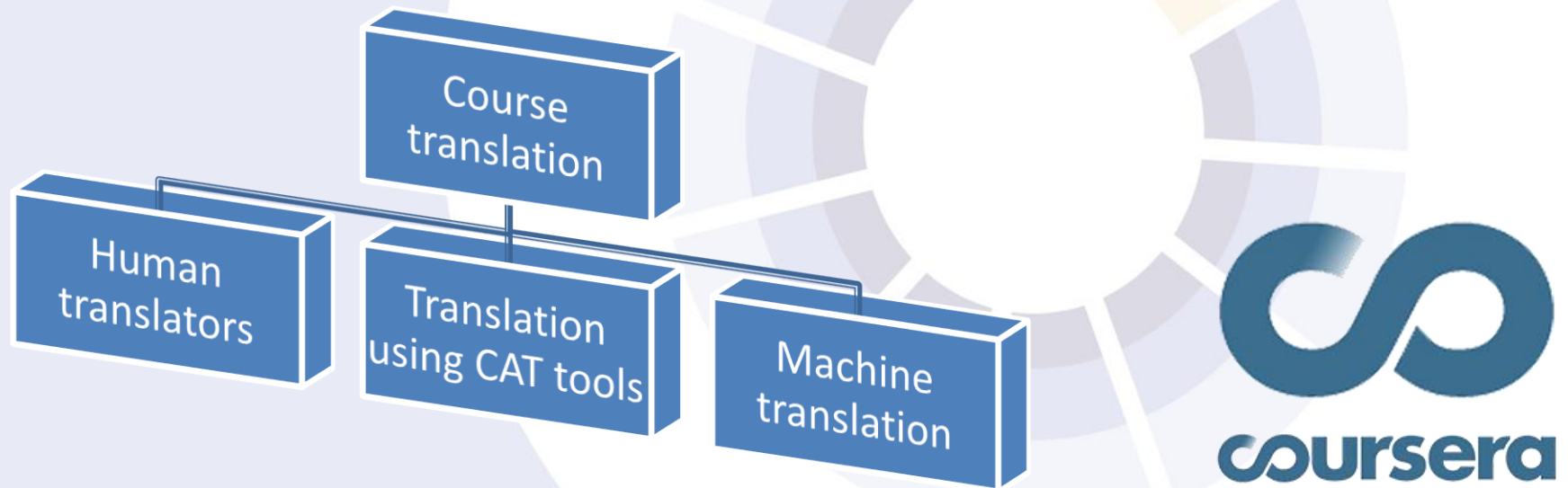


# Translation for Massive Open Online Courses (TraMOOC)

- TraMOOC translation is aimed at all types of text genre included in MOOCs assignments, tests, presentations, lecture subtitles, forum text, from English into eleven languages:
  - German, Italian, Portuguese, Greek, Dutch, Czech, Bulgarian, Croatian, Polish, Russian, Chinese
- The results will be showcased and tested on the Iversity MOOC platform and on the VideoLectures.NET digital video lecture library
- The translation engine employed in TraMOOC is Moses, the most widely used SMT toolkit available in academia and commercial environments

# Related work - Coursera

- Leading MOOC provider
- For multilingual support Coursera uses **Transifex**
- Coursera interface is available in 5 different languages



# Translation of educational resources - current approaches

For translation of eLearning resources both language translation and eLearning skills are necessary

The translation needs several reviews before publishing or preparation for voice recording

## Computer Aided Translation (CAT) Tool

- segments the source text in segments, usually sentences
- the source text and translation of each segment are saved together as a TU (translation unit)
- translation memory (TM) is a database of TUs
- CAT tool has support for terminology look-up, display and insertion of the search results into the text being translated

# Environment for text alignment – Step 1

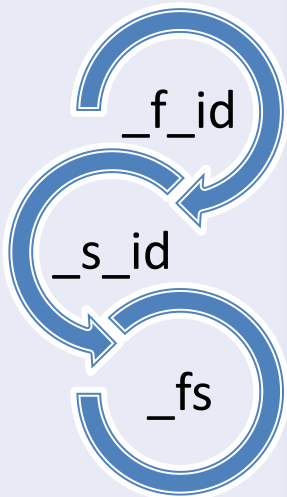
First step in text alignment:  
XML document preparation  
according to TEI guidelines

- Mark-up the divisions, titles, paragraphs and segments using text or XML editor
- Support for DTD scheme validation and well-formedness check desirable
- This part can be partly automated using finite-state transducers (manual intervention is still necessary)



# Environment for text alignment – Step 2

- Step 2 = **PARALLELIZATION**
- The task is thus to establish the connection between originals and their translations
- Parallelization can be performed using ACIDE software

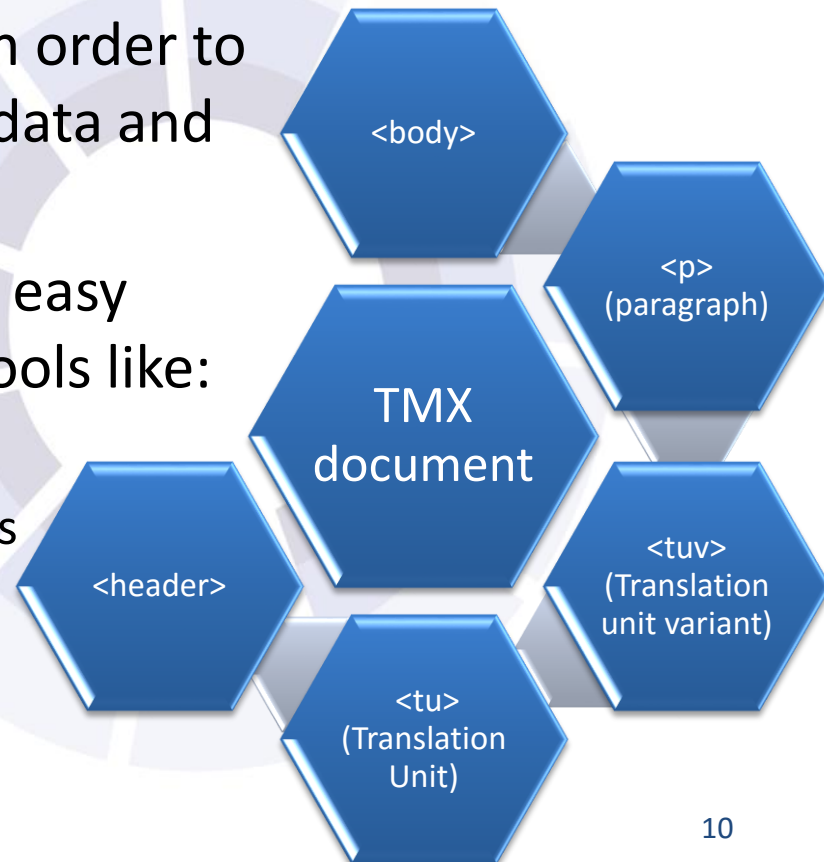


```
<div>
  <p>
    <seg id="n15">1. An Ocean of Digital Words</seg>
    <seg id="n16">A society of information offers almost
      a limitless amount of information to everyone.</seg>
    <seg id="n17">Without the usage of intelligent,
      efficient applications for information extraction,
      which are based on highly advanced techniques and
      methods, one can benefit only from the smallest part
      of potential offered by new technology (Piskorski
      1999).</seg>
    <seg id="n18">If we define information as a result of
      collecting, processing, manipulating and organizing
      data in order to present new knowledge to the
      recipient, than it can be said that a piece of data
```

```
<div>
  <p>
    <seg id="n15">1. Okean digitalnih reči</seg>
    <seg id="n16">Informatičko društvo stavlja na
      raspolaganje gotovo neograničenu količinu informacija
      svakom pojedincu.</seg>
    <seg id="n17">Bez upotrebe inteligentnih, efikasnih
      aplikacija za ekstrakciju informacija koje se zasnivaju
      na vrlo naprednim tehnikama i metodama, pojedinac nije u
      mogućnosti da iskoristi ni delić nesagledivog
      potencijala koji nude nove tehnologije (Piskorski 1999).
    </seg>
    <seg id="n18">Ako informaciju definišemo kao rezultat
      sakupljanja, obrade, manipulacije i organizovanja
      podataka sa ciljem da se primaocu predstavi novo znanje
```

# Environment for text alignment – Step 3

- Production of a TMX document
- Metadata code (element `<prop>`) is attached to each aligned sentence (element `<tu>`) in order to establish a direct relation to metadata and the original
- From aligned TMX documents it is easy to produce parallel text form for tools like:
  - Giza++
  - JSON format suitable for web services
  - Mongo and other NoSQL databases



# Aligned segments in TMX

```
<tu>
  <prop type="Domain">Gucul-Milojević, 2010, vol. XI:1, ID: 1.2010.1.4</prop>
  <tuv xml:lang="en" creationid="n15" creationdate="20110513T151548Z">
    <seg>1. An Ocean of Digital Words </seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n15" creationdate="20110513T151548Z">
    <seg>1. Okean digitalnih reči </seg>
  </tuv>
</tu>
<tu>
  <prop type="Domain">Gucul-Milojević, 2010, vol. XI:1, ID: 1.2010.1.4</prop>
  <tuv xml:lang="en" creationid="n16" creationdate="20110513T151548Z">
    <seg>A society of information offers almost a limitless amount of information to everyone. </seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n16" creationdate="20110513T151548Z">
    <seg>Informatičko društvo stavlja na raspolaganje gotovo neograničenu količinu informacija svakom pojedincu. </seg>
  </tuv>
</tu>
<tu>
  <prop type="Domain">Gucul-Milojević, 2010, vol. XI:1, ID: 1.2010.1.4</prop>
  <tuv xml:lang="en" creationid="n17" creationdate="20110513T151548Z">
    <seg>Without the usage of intelligent, efficient applications for information extraction, which are based on highly advanced techniques and methods, one can benefit only from the smallest part of potential offered by new technology (Piskorski 1999). </seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n17" creationdate="20110513T151548Z">
    <seg>Bez upotrebe inteligentnih, efikasnih aplikacija za ekstrakciju informacija koje se zasnivaju na vrlo naprednim tehnikama i metodama, pojedinac nije u mogućnosti da iskoristi ni delić nesagledivog potencijala koji nude nove tehnologije (Piskorski 1999). </seg>
  </tuv>
</tu>
```

# GIZA ++ Basic facts

- GIZA++ is an extension of the program GIZA
- It was developed by the Statistical Machine Translation team in 1999 at the Center for Language and Speech Processing at Johns-Hopkins University (CLSP/JHU)
- The extensions of GIZA++ were designed and written by Franz Josef Och
- GIZA ++ is installed on the Faculty of Mining and Geology as part of Moses
- GIZA is quite a demanding tool, and it therefore requires extra resources (Linux OS, larger amount of RAM (16GB) )

# GIZA ++ Corpus preparation

## Tokenisation

- Spaces have to be inserted between words and punctuation

## Truecasing

- The initial words in each sentence are converted to their most probable casing - this helps reduce data sparsity

## Cleaning

- Long sentences and empty sentences are removed as they can cause problems with the training pipeline, and obviously mis-aligned sentences are also removed

# Tokenisation

```
~/mosesdecoder/scripts/tokenizer/tokenizer.perl -l en \  
  < ~/corpus/training/edX.en \  
  > ~/corpus/edX.tok.en
```

```
~/mosesdecoder/scripts/tokenizer/tokenizer.perl -l sr \  
  < ~/corpus/training/edX.sr \  
  > ~/corpus/edX.tok.sr
```

# Truecaser

```
~/mosesdecoder/scripts/recaser/train-truecaser.perl \  
  --model ~/corpus/truecase-model.en --corpus \  
  ~/corpus/edX.tok.en
```

```
~/mosesdecoder/scripts/recaser/train-truecaser.perl \  
  --model ~/corpus/truecase-model.sr --corpus \  
  ~/corpus/edX.tok.sr|
```

# Cleaning

```
~/mosesdecoder/scripts/recaser/truecase.perl \  
  --model ~/corpus/truecase-model.en \  
  < ~/corpus/edX.tok.en \  
  > ~/corpus/edX.true.en
```

```
~/mosesdecoder/scripts/recaser/truecase.perl \  
  --model ~/corpus/truecase-model.sr \  
  < ~/corpus/edX.tok.sr \  
  > ~/corpus/edX.true.sr
```

```
~/mosesdecoder/scripts/training/clean-corpus-n.perl \  
  ~/corpus/edX.true sr en \  
  'corpus/edX.clean 1 80
```


# GIZA ++ Language Model Training

- A language model (LM) is used to ensure fluent output (target language - English)
- The following script creates a *lm* folder, selects this folder as the output folder and executes a command that will build a 3-gram language model

```
mkdir ~/lm
cd ~/lm
~/mosesdecoder/bin/lmplz -o 3 <~/corpus/edX.true.en >
edX.arpa.en
~/mosesdecoder/bin/build_binary \
edX.arpa.en \
edX.blm.en
```



# GIZA ++ Language Model Training



## Running word- alignment (using GIZA++)

- phrase extraction and scoring
- lexicalised reordering tables creation
- using Moses configuration file (all with a single command)
- duration: 90min

## Result:

- aligned Serbian and English words (and phrases)
- with a factor of accuracy for translation
  - from Serbian to English and
  - from English into Serbian



# The result of machine translation using GIZA++ tool

Word_sr	Word_en	P_sr_en	P_en_sr	P_su
skladu	accordance	0,977011	0,977011	1,9540
,u skladu	, in accordance	0,904762	0,968254	1,8730
1. ovog	1 of this	0,935484	0,935484	1,8709
centar	the Centre	0,935484	0,935484	1,8709
daljem tekstu	further text	0,935484	0,935484	1,8709
daljem	further	0,935484	0,935484	1,8709
tekstu	text	0,935484	0,935484	1,8709
u daljem tekstu	in further text	0,935484	0,935484	1,8709
u daljem	in further	0,935484	0,935484	1,8709
,u	, in	0,913043	0,942029	1,8550
zakonom.	law .	0,925926	0,925926	1,8518
iz stava	from paragraph	0,886792	0,962264	1,8490
izuzetno od stava	Exceptionally from paragraph	0,818182	0,818182	1,6363
izuzetno od	Exceptionally from	0,818182	0,818182	1,6363
Republika, autonomna	the Republic, autonomous	0,818182	0,818182	1,6363
aktom o osnivanju	the establishment act	0,818182	0,818182	1,6363
skladu sa ovim i posebnim zakonom	accordance with this law and separate laws	0,818182	0,818182	1,6363
predmeta	subjects	0,866667	0,733333	1,6
za naknadu	employee about the	0,2	0,6	0,8
za naknadu	the employee about the	0,2	0,6	0,8
za naknadu	the	0,2	0,6	0,8

# Conclusion

- There is a growing need for translating MOOCs into other languages
- GIZZA++ is a suitable tool for that, but it needs a parallel corpus
- For a corpus we need to:
  - prepare input text in both languages
  - perform parallelization
  - Perform tokenisation, truecasing and cleaning
  - finally, use the language model
- The presented method yielded promising results, but bigger a corpus is needed for better results
- Great efforts are being made for additional text alignment and augmentation of the Biblisha library of aligned texts
- Detailed evaluation will be performed when we reach at least 100.000 sentence pairs



# BAEKTEL

Blending academic and entrepreneurial knowledge  
in technology enhanced learning



Hvala za vašo pozornost

Thank you for your attention

Grazie per la vostra attenzione

Vă mulțumesc pentru atenție

Хвала на пажњи

Hvala na pažnji

