معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

The AMARA Corpus: Building Parallel Language Resources for the Educational Domain

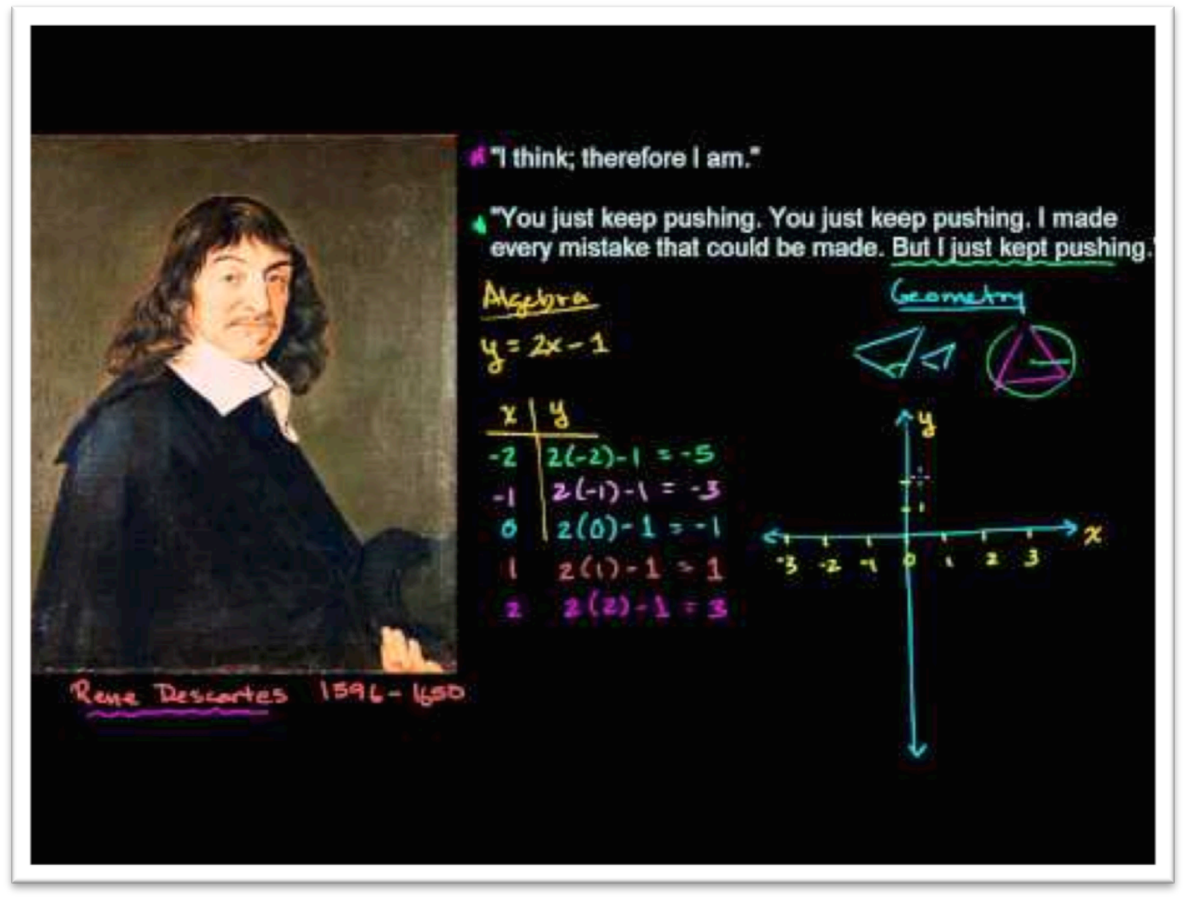
Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, Stephan Vogel

Motivation

Massive Online Open Courses (MOOCs) bridge financial/geographical gap

Thousands of lectures are available (70% English)

TED UDACITY KHAN ACADEMY coursera


"I think; therefore I am."
"You just keep pushing. You just keep pushing. I made every mistake that could be made. But I just kept pushing."
Algebra
Geometry
y = 2x − 1
Rene Descartes 1596–1650

Source:Khan Academy

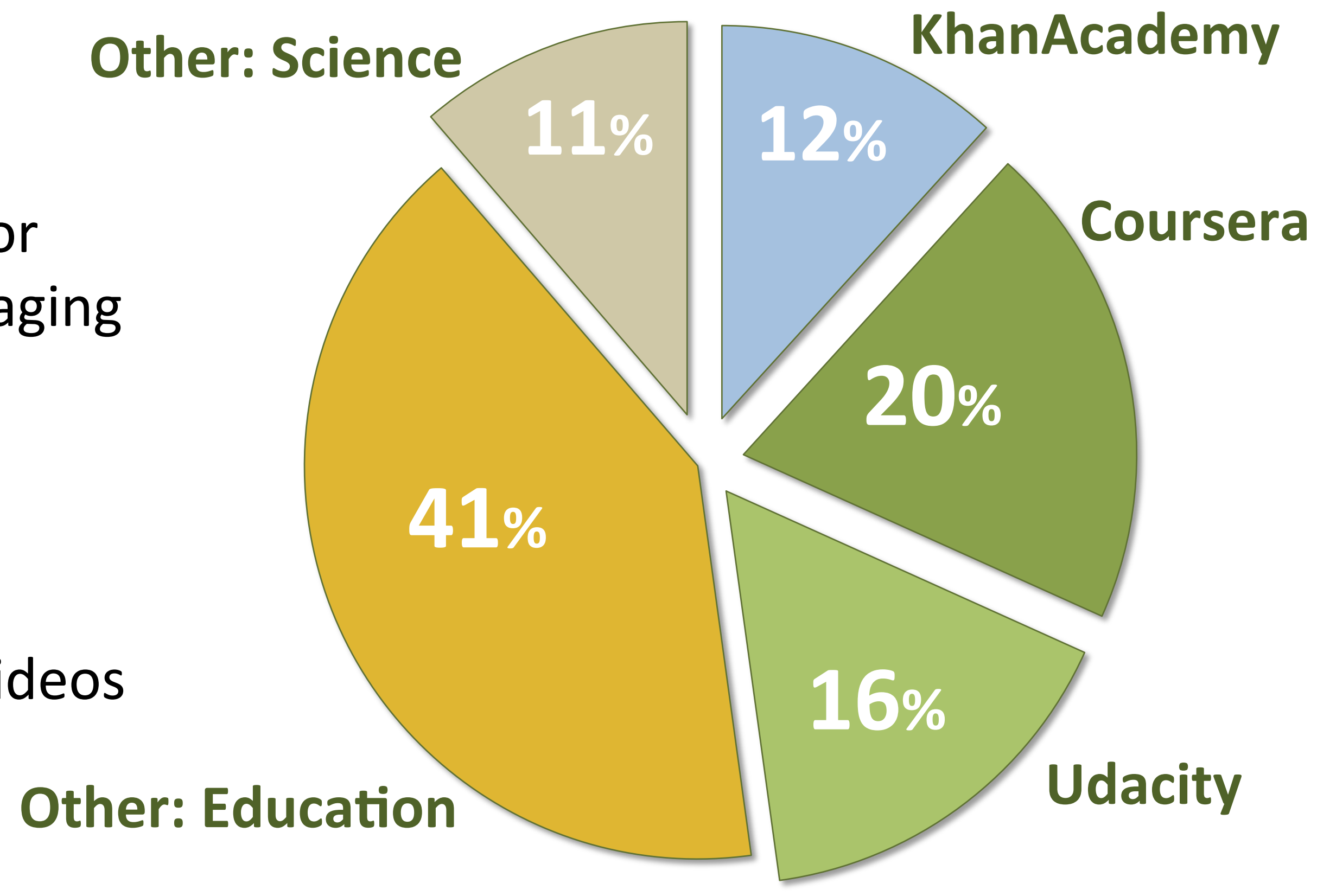
Source

Amara

Is a web-based platform for creating, editing and managing video subtitles

Is driven by volunteers

Hosts many educational videos subtitled/translated



Building Parallel Resources for Education

Crawl

(a) download subtitles

- ➔43K videos
- ➔160 language
- ➔121K translations/transcripts

(b) keep educational videos

Validate

Content of the subtitles

(a) completeness:
Discard empty or incomplete

(b) correct language:
Using Cybozu language detector

- ➔12.2K videos
- ➔ 20 language
- ➔34K translation/transcript

Align

Parallel segments

(a) strict synchronization:
Accept only parallel segments w/identical ds and timestamps

(b) cascaded synchronization:
Re-align the discarded segments using length statistics and information from a bilingual dictionary

	en	sp	pt	zhs	zht	tr	pl	ar	fr	cz	ru	it	ja	kr	bg	de	th	nl	da	hi
en	**2.5M**																			
sp	335K	**479K**																		
pt	231K	117K	**291K**																	
zhs	139K	52K	46K	**281K**																
zht	117K	49K	48K	191K	**231K**															
tr	169K	72K	67K	47K	51K	**205K**														
pl	151K	88K	72K	47K	52K	65K	**197K**													
ar	158K	83K	73K	58K	59K	90K	69K	**185K**												
fr	125K	63K	58K	29K	29K	36K	59K	48K	**161K**											
cz	132K	61K	65K	55K	56K	66K	69K	56K	40K	**158K**										
ru	77K	39K	36K	22K	20K	18K	30K	29K	29K	25K	**146K**									
it	97K	52K	49K	29K	29K	38K	43K	44K	39K	38K	23K	**124K**								
ja	98K	44K	44K	46K	42K	47K	39K	48K	31K	43K	23K	29K	**113K**							
kr	83K	36K	36K	30K	32K	28K	32K	37K	25K	30K	14K	28K	26K	**109K**						
bg	79K	39K	44K	28K	33K	45K	44K	39K	33K	42K	15K	27K	21K	19K	**100K**					
de	77K	49K	45K	22K	25K	30K	45K	36K	42K	35K	22K	30K	24K	23K	28K	**99K**				
th	85K	50K	40K	31K	29K	56K	38K	45K	21K	29K	14K	22K	27K	15K	20K	20K	**95K**			
nl	73K	43K	42K	25K	29K	33K	41K	40K	33K	41K	19K	31K	25K	22K	25K	30K	19K	**85K**		
da	48K	27K	34K	18K	21K	29K	25K	30K	23K	32K	12K	21K	18K	16K	25K	16K	10K	21K	**58K**	
hi	43K	26K	22K	14K	14K	17K	24K	25K	16K	17K	8K	26K	16K	14K	13K	13K	13K	17K	15K	**48K**

download

The AMARA corpus from:
http://amaracorpus.alt.qcri.org/

Segment Alignment	Spanish	Arabic	Russian
Strict Synchronization	241K	128K	56.3K
Cascade Sync	691K	318K	157K
Improvements	**287%**	**249%**	**279%**

Evaluate

Run translation experiments

	BLEU NIST v13	
Source Lang.	**tst2014a**	**tst2014b**
Spanish (sp)	48.2	41.4
Portuguese (pt)	52.1	46.6
Arabic (ar)	38.0	34.4
Polish (pl)	34.7	29.4
Czech (cz)	33.7	32.9
French (fr)	31.5	35.1
German (de)	35.2	34.0
Russian (ru)	34.3	38.6
Dutch (nl)	39.8	45.6
Danish (da)	40.5	35.3

Sample Translations

Spanish		
	Src	Luego el brazo volverá a su posición original
	Trans	Then the arm will return to its original position
	Src	Entonces esto es igual a 4x al cuadrado menos 2x más 8.
	Trans	So this is equal to 4x squared minus 2x plus 8.

Arabic		
	Src	نطرح ٣ من كلا الطرفين
	Trans	Subtract 3 from both sides.
	Src	ماهو الرسم البياني لـ y=x^2؟
	Trans	What is the graph of y is equal to minus x squared?

High quality translations

- In-domain data
- Short segments
- Limited reordering

acknowledgements We thank the Amara.org staff, especially Nicholas Reville, Aleli Alcala and Dean Jansen for their support in the development of this corpus.