(Semi-)automatisch ondertitelen en vertalen van leermateriaal

Prof. Dr. Marco Kalz
Chair for Open Educaton
Faculty Management, Science and Technology &
Welten Institute
Open University of the Netherlands

Dr. Jorge Civera, Gonçal Garcés Díaz-Munío, Prof. Dr. Alfons Juan
Machine Learning and Language Processing research group
Universitat Politècnica de València

SURF, 26/09/2016

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MLLP

Open Universiteit
www.ou.nl

Achtergrond

EUROPEAN MULTIPLE MOOC AGGREGATOR

http://europeanmoocs.eu

- 30 month H2020 policy support project
- Net afgesloten (review 23/09/16)
- Resultaten:
  - Open MOOC platform
  - Cross-cultural, cross-language
  - Aggregator: Productie in moedertaal, publicatie in EN (+ES/IT) op EMMA
  - OUNL & UPV samenwerking voor NL

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MLLP

Open Universiteit
www.ou.nl

Technologie: MLLP platform for transcription and translation

- Main features:
  - Automatic multilingual subtitling (transcription + translations)
  - Text-to-Speech synthesis (voice-to-voice translation)
  - Document translation
  - Multilingual website translation
- Advanced editors:
  - TLP Media Player: Post-editing of multilingual subtitles
  - TLP Text Translation Editor: Side-by-side text translation post-editing

http://ttp.mllp.upv.es

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MLLP

Open Universiteit
www.ou.nl

Talen

- Automatic **Speech Recognition** (automatic subtitling):
  - English (En), Spanish (Es), Catalan (Ca), French (Fr), Estonian (Et), Italian (It), **Dutch (Nl)**, Portuguese (Pt), German (De), Slovene (Sl)

- **Machine Translation** (for subtitles, documents and websites):
  - {Es, Ca, Fr, Et, It, **Nl**, Pt} $\rightarrow$ En
  - En $\rightarrow$ {Es, Ca, Fr, It, Pt, Sl}

- **Text-to-Speech Synthesis** (voice-to-voice translation):
  - English (En), Spanish (Es), Catalan (Ca)
  - E.g., **Dutch (Nl)** $\rightarrow$ {En, Es, Ca} voice-to-voice translation

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MLLP

Open Universiteit
www.ou.nl

Resultaten

- Transcription accuracy
- Time saving with respect to transcribing from scratch (10 RTF)

Language	Accuracy	Time saving
Spanish	Excellent (83%)	69%
English	Good (78%)	47%
Italian	Good (74%)	61%
French	Good (77%)	34%
Dutch	**Good (71%)**	**42%**
Catalan	Average (63%)	44%

- Translation accuracy
- Time saving with respect to translating from scratch (30 RTF)

Language pairs	Accuracy	Time saving
Dutch → English	**Good (69%)**	**68%**
English → Spanish	Good (67%)	70%
Spanish → English	Good (63%)	74%
English → Italian	Good (64%)	68%
Italian → English	Good (53%)	58%
French → English	Average (41%)	23%

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MLLP

Open Universiteit
www.ou.nl

Resultaten

- Transcription error (WER)

Language	EMMA	YouTube
Spanish	14.8	22.5 (+52%)
Italian	17.1	31.6 (+85%)
French	20.6	32.0 (+55%)
Estonian	27.1	—
Dutch	25.8	48.1 (+86%)
English	39.2	65.9 (+41%)
Portuguese	43.0	62.3 (+31%)

- Translation quality (BLEU)

Language pair		EMMA	Google
English	→ Spanish	42.5	38.3 (-10%)
English	→ Italian	52.6	44.1 (-16%)
Spanish	→ English	40.9	34.6 (-15%)
Estonian	→ English	27.8	26.3 (- 5%)
French	→ English	33.1	30.8 (- 7%)
Italian	→ English	47.8	37.5 (-22%)
Dutch	→ English	41.6	34.0 (-18%)
Portuguese	→ English	55.5	45.4 (-18%)

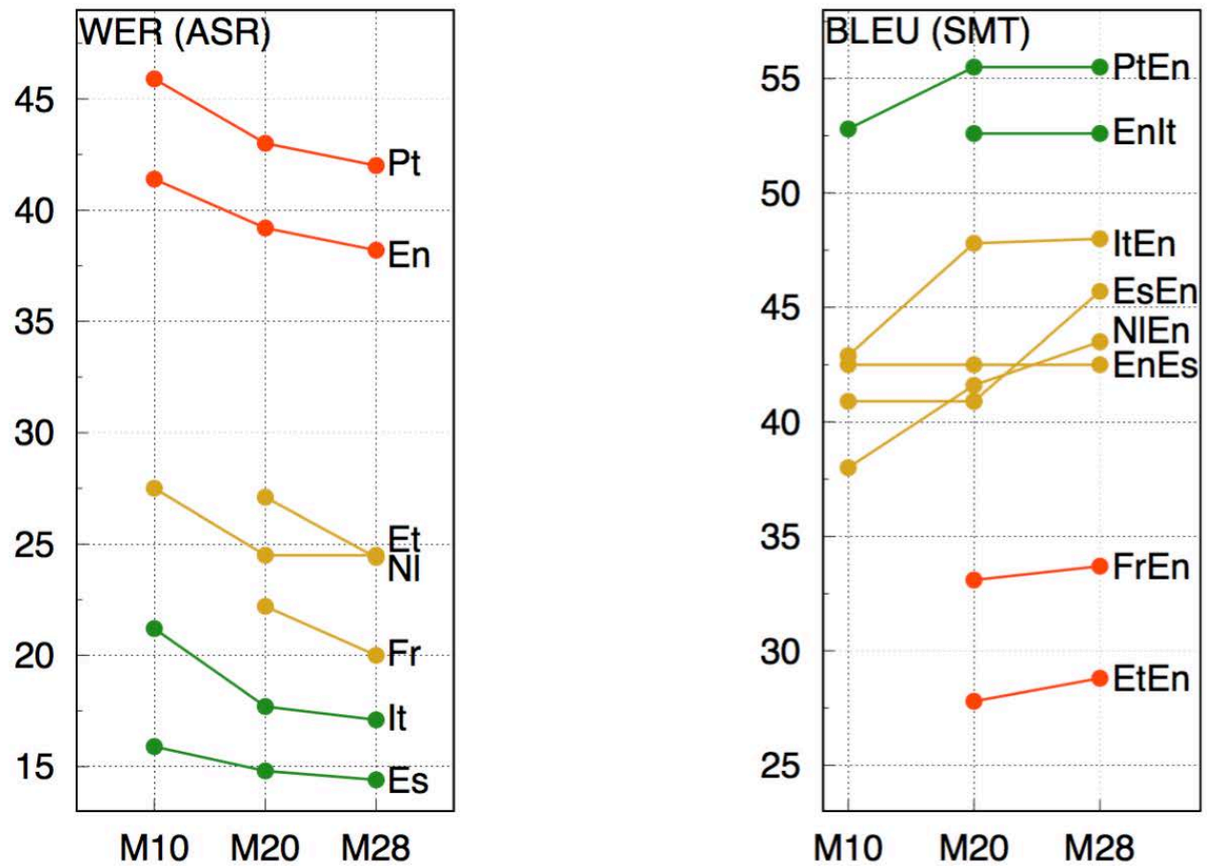
(Brouns, Serrano Martínez-Santos, Civera, Kalz, & Juan, 2015)

MLLP

Open Universiteit
www.ou.nl

Resultaten

Transcription and translation quality



- Quality: Major revision (orange), worth post-editing (yellow) and minor revision (green)
- Time saving with respect to transcribing and translating from scratch: 50% on average[1]

[1] Papers: Speech Communication; EC-TEL 2015; Submission to Learning, Media and Technology

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MLLP

Open Universiteit
www.ou.nl

Uitdagingen & Kansen

Uitdagingen

- Onvoldoende capaciteit voor training van het systeem
- Ingewikkelde video's (1 uur, discussie, non-native sprekers)

Kansen

- Verhoging van visibility van open cursussen op EMMA +76%
- Semi-automatisch ondertitelen en vertalen kan kosten voor meerdere talen 30% - 50% verminderen
- Betere indexering van video-content

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MLLP

Open Universiteit
www.ou.nl

Voorbeelden

1. Open Universiteit EMMA (16.1 Word Error Rate = 84% accuracy) https://ttp.mllp.upv.es/s/adsAy

2. Aankondiging MOOC Formatief toetsen (17.4 WER = 83% accuracy) https://ttp.mllp.upv.es/s/blHgE

3. Feedback als effectieve leer- en instructiestrategie (21.5 WER = 78% accuracy) https://ttp.mllp.upv.es/s/aC4Q6

4. Leren en gedragsverandering de context van gezondheidsgedrag (25.2 WER = 75% accuracy) https://ttp.mllp.upv.es/s/bKTwM

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MLLP

Open Universiteit
www.ou.nl

Bedankt

marco.kalz@ou.nl

jcivera@dsic.upv.es

ggarces@dsic.upv.es

ajuan@dsic.upv.es

@mkalz

@mllpresearch

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MLLP

Open Universiteit
www.ou.nl