UNIVERSITAT POLITÈCNICA DE VALÈNCIA

**Tesi doctoral**

# ASR and MT with deep neural networks for open educational resources, parliamentary contents and broadcast media

**Programa de Doctorat en Informàtica**

Gonçal Garcés Díaz-Munío

Directors: Dr. Alfons Juan Ciscar, Dr. Jorge Civera Saiz

15 d'octubre de 2024

# Contents


1 Motivation and objectives 3

2 Preliminary matters 5

3 OER: Evaluation & post-editing of ASR+MT w/ intell. interaction 7

4 OER: The transition from phrase-based MT to neural MT 11

5 Parliamentary contents: Speech data curation for streaming ASR 18

6 Broadcast media: Live & offline computer assisted subtitling 26

7 Achievements and conclusions 38

Publications 40

# 1 Motivation and objectives

## Trends in the 2010s–2020s

- Audiovisual contents: Push to increase accessibility and multilingualism
  - Open educational resources
  - Parliamentary debates
  - Broadcast media

- ASR and MT: Advances with deep neural networks
  - Output quality and robustness ↑↑
  - Complex applications ↑↑
  - Streaming ASR & MT, with high quality, low latency

**Proposal:** Speech recognition + Machine translation (+ Text-to-speech) for subtitling, transcription, translation, dubbing... live and offline

# 1 Motivation and objectives
## Objectives

1. Contributions to evaluation & post-editing of ASR+MT for OER
   - ▶ Context: EU project transLectures (2012–2014)
2. Developing state-of-the-art neural MT systems for OER
   - ▶ Context: EU project X5gon (2017–2020)
3. Contributions to streaming ASR data curation for parliamentary contents
   - ▶ Context: Spanish project Multisub (2019–2021), European Parliament
4. Contributions to live & offline assisted subtitling of broadcast media
   - ▶ Context: 2020–2023 À Punt Mèdia-UPV R&D agreement

# 2 Preliminary matters

## ASR: Automatic speech recognition

- Input: Audio.
  Output: Text.
  - Automatic transcriptions.
  - Same-language subtitles.
    Subtitles for the deaf and hard-of-hearing (SDH).

- State of the art:
  - 2010–: Hybrid neural-HMM systems.
  - 2014–: Neural “end-to-end” systems.
  - 2020s: Streaming ASR.

- Evaluation:
  - Automatic: Word Error Rate (WER).
  - Human: Post-editing time (RTF).

## MT: Machine translation

- Input: Text.
  Output: Text.
  - Automatic translations.
  - Translated subtitles (ASR+MT).
- State of the art:
  - 2000–2016: Phrase-based MT.
  - 2016–: Neural MT.
  - 2020s: Streaming MT.
- Evaluation:
  - Automatic: BLEU, neural metrics.
  - Human: Direct assessment, post-editing time (RTF).

# 3 OER: Evaluation & post-editing of ASR+MT with intelligent interaction

- **Objective 1:**
  Evaluation & post-editing of ASR+MT for OER, with intelligent interaction

- Open Educational Resources:
  - 2012 UNESCO Paris OER Declaration, 2013 EC Opening Up Education...
  - 2012 MOOC explosion (Coursera, edX...)

- The project: transLectures (EU FP7, **2012–2014**)
  Transcription and Translation of Video Lectures
  - Massive adaptation (to the domain, the speaker...)
  - Intelligent interaction

- The pilot: UPV Mèdia
  - Datasets: Spanish (114 hours); Spanish→English (15 hours).

# *3 OER: Evaluation & post-editing of ASR+MT*
## Types of evaluation performed

1. Automatic evaluations
   - ▶ Transcriptions: WER.
   - ▶ Translations: BLEU, TER.

2. Human evaluations by language experts
   - ▶ Transcriptions: RTF, comments.
   - ▶ Translations: Direct assessment, RTF, comments.

3. Human evaluations by users (UPV lecturers)
   - ▶ Transcriptions: RTF.
   - ▶ Translations: RTF.
   - ▶ Research on **post-editing with intelligent interaction**: impact on RTF.

# 3 OER: Evaluation & post-editing of ASR+MT
## 3) Post-editing ASR with intelligent interaction

- Human evaluations by users (UPV lecturers)
  - Research on post-editing ASR with intelligent interaction:
    - Phase 1: Complete post-editing
    - Phase 2: Intelligent interaction
    - Phase 3: Two-round post-editing (hybrid)
  - Compared to transcribing video lectures from scratch

- Results:
  - Post-editing: transcription time ↓40%
  - Intelligent interaction:
    - Transcription time ↓↓75%, reducing WER by 50% (residual WER 8%)

# 3 OER: Evaluation & post-editing of ASR+MT
## Conclusions

- Post-editing: transcription time ↓40%
  - Intelligent interaction: even more cost-effective
- ASR quality improved hugely thanks to DNN models
- We are offering this auto subtitling service for all UPV Mèdia since 2014
- Thus, focus for next research:
  - Improving ASR and MT technology with DNN
  - Increasing amounts of training data

# 4 OER: The transition from phrase-based MT to neural MT

- **Objective 2:** Developing state-of-the-art neural MT systems for OER

- The project: X5gon (EU H2020, **2017–2020**)
  Cross-Modal, Cross-Cultural, Cross-Lingual, Cross-Domain, and Cross-Site Global OER Network
  - Pilots
    - UPV Mèdia (UPV, Spain)
    - VideoLectures.net (Inst. Jožef Stefan, EU/Slovenia)
    - virtUOS (Univ. Osnabrück, Germany)
  - Languages
    - ASR (4): English, Spanish, German, Slovene.
    - MT (14): **English**↔{Spanish, French, **German**, Italian, Slovene}, French↔German, Spanish↔Portuguese.

# 4 OER: The transition from phrase-based MT to neural MT
## The MLLP-UPV German-English MT System for WMT18

- DE→EN neural MT system, based on Transformer architecture
- WMT18 bilingual corpus: cleaner data 6M; noisy data 36M
- Results, with techniques applied:

| System | newstest2018 BLEU |
|---|---|
| Base (6M cleaner pairs) | 39.1 |
| Filtered corpus (10M) | 42.2 |
| + Backtranslations (+20M) | 44.7 |
| Ensemble (×4) | **45.1** |

# 4 OER: The transition from phrase-based MT to neural MT
## The MLLP-UPV German-English MT System for WMT18

### WMT18 official results DE→EN

Human evaluation (official)

| German→English | | | |
|---|---|---|---|
| | Ave. % | Ave. z | System |
| 1 | 79.9 | 0.413 | RWTH |
| | 79.4 | 0.395 | UCAM |
| | 78.2 | 0.359 | NTT |
| | 77.3 | 0.346 | ONLINE-B |
| | 77.4 | 0.321 | MLLP-UPV |
| | 77.0 | 0.317 | JHU |
| | 76.9 | 0.315 | UBIQUS-NMT |
| | 76.7 | 0.310 | ONLINE-Y |
| | 75.7 | 0.268 | ONLINE-A |
| | 75.4 | 0.261 | UEDIN |
| 11 | 72.5 | 0.162 | LMU-NMT |
| | 72.2 | 0.149 | NJUNMT-PRIVATE |
| 13 | 65.2 | −0.074 | ONLINE-G |
| 14 | 58.5 | −0.296 | ONLINE-F |

Automatic evaluation

| System | BLEU (newstest2018) |
|---|---|
| RWTH | 48.4 |
| UCAM | 48.0 |
| NTT | 46.8 |
| JHU | 45.3 |
| **MLLP-UPV** | **45.1** |
| UBIQUS-NMT | 44.1 |
| UEDIN | 43.9 |
| LMU-NMT | 40.9 |
| NJUNMT-PRIVATE | 38.3 |

# Comparing phrase-based and neural MT systems

| System (DE→EN) | newstest2018 BLEU |
|---|---|
| **PBMT** base (6M pairs) | 28.0 |
| + Monolingual data (+210M sents.) | **30.6** |
| **NMT** base (6M pairs) | 39.1 |
| Filtered corpus (10M pairs) | 42.2 |
| + Backtranslations (+20M pairs) | 44.7 |
| NMT ensemble (×4) | **45.1** |

PBMT base → NMT base: **+40%**

PBMT best → NMT best: **+47%**

# 4 OER: The transition from phrase-based MT to neural MT
## Impact of NMT on real OER scenarios

- New NMT systems for X5gon, developed from the experience of WMT

| Language pair | UPV Mèdia (BLEU) | | | |
|---|---|---|---|---|
| | 2018 | 2019 | 2020 | △ |
| ES-EN | 24 | 30 | 34 | **+40%** |

# 4 OER: The transition from phrase-based MT to neural MT
## Impact of NMT on real OER scenarios

- New NMT systems for X5gon, developed from the experience of WMT

| Language pair | VideoLectures.net (BLEU) | | | WMT (BLEU) | | | Δ% |
|---|---|---|---|---|---|---|---|
| | Google | X5gon | Δ% | Google | X5gon | Δ% | average |
| ES-PT | - | - | - | 43.4 | **70.7** | 62.9 | +62.9 |
| PT-ES | - | - | - | 47.6 | **72.4** | 52.1 | +52.1 |
| SL-EN | 15.0 | **26.4** | 76.0 | 29.2 | **34.3** | 17.4 | +46.7 |
| EN-SL | 16.5 | **22.9** | 38.8 | 23.6 | **29.4** | 24.6 | +31.7 |
| DE-EN | 25.7 | **27.0** | 5.1 | 43.9 | **48.0** | 9.4 | +7.3 |
| EN-ES | 41.3 | **44.4** | 7.5 | **35.3** | 34.6 | -2.0 | +2.8 |
| ES-EN | 37.8 | **40.3** | 6.6 | 34.4 | **35.9** | -2.0 | +2.3 |
| FR-EN | **30.3** | 30.1 | -0.7 | 38.6 | **39.7** | 2.8 | +1.1 |
| DE-FR | **19.6** | 18.6 | -5.1 | 32.2 | **34.4** | 6.8 | +0.9 |
| EN-FR | **29.4** | 29.2 | -0.7 | 40.4 | **41.1** | 1.7 | +0.5 |
| IT-EN | - | - | - | **35.7** | 35.2 | -1.4 | −1.4 |
| FR-DE | **18.6** | 17.2 | -7.5 | 26.6 | **26.9** | 1.1 | −3.2 |
| EN-IT | - | - | - | **32.1** | 29.8 | -7.2 | −7.2 |
| EN-DE | **24.7** | 21.5 | -13.0 | **47.0** | 45.7 | -2.8 | −7.9 |

# 4 OER: The transition from phrase-based MT to neural MT
## Conclusions

- Developed state-of-the-art NMT systems for OER
  - 1st rank in WMT18
  - +47% BLEU over previous PBMT system
  - Applying new key techniques (corpus filtering, backtranslations)
- Showed the impact of NMT on real OER scenarios
  - UPV Mèdia: NMT Spanish→English +40% BLEU (2018–2020)
  - High-quality results, on par with Google

# 5 Parliamentary contents: Speech data curation for streaming ASR

- **Objective 3:**
  Speech data curation for streaming ASR of parliamentary contents

- Parliamentary contents: 2012 Declaration on Parliamentary Openness

- Context:
  - The project: Multisub (Spanish government, **2019–2021**)
    Multilingual subtitling of classrooms and plenary sessions
  - The data: European Parliament videos, transcriptions and translations
  - Language: English (and other 23 official EU languages)
  - Focus: Monolingual data

# 5 Parliaments: Speech data curation for streaming ASR
## Motivation

- Quality ASR requires **thousands of hours of speech data**
- Frequently, **existing transcriptions** are **not 100% verbatim**
  - Most large public corpora are non-verbatim
- We need techniques to **make the most of speech data**
  - Speech data curation
  - Filter out inaccurately transcribed parts
    $\rightarrow$ Filtering
  - What about **auto-improving** inaccurate transcriptions?
    $\rightarrow$ Verbatimization
- Also: **Lack of realistic tasks for streaming** ASR evaluation

# 5 Parliaments: Speech data curation for streaming ASR
## Goal: The Europarl-ASR corpus

- A new large speech corpus for:
  - Training & benchmarking (streaming) ASR
  - Benchmarking speech data curation techniques

- **1300 hours** of EN transcribed speech data
  - 3 full sets of timed transcriptions:
    official non-verbatim; auto noise-filtered; auto verbatimized

- Dev/test: 18 hours of **manually revised** transcriptions
  - 2 transcription sets: official non-verbatim; revised verbatim
  - 2 independent dev/test partitions for 2 realistic ASR tasks

`https://www.mllp.upv.es/europarl-asr`

# 5 Parliaments: Speech data curation for streaming ASR
## Data gathering and selection

- European Parliament “reports of proceedings” (1999–2020)
  - English speech recordings
  - English transcriptions
  - Translations into English

- 1263 hours of EN transcribed speech (non-verbatim)

# 5 Parliaments: Speech data curation for streaming ASR
## Definition of tasks and evaluations sets

- 2 realistic streaming ASR **tasks**:
  - Speaker-independent (*Guest*): no prior knowledge of guests
  - Speaker-dependent (*MEP*): prior knowledge about MEPs

| Set → | MEP-dev | MEP-test | Guest-dev | Guest-test | train |
|---|---|---|---|---|---|
| Speakers | 21 | 21 | 6 | 6 | 1034 |
| Length (h) | 4.6 | 4.7 | 4.3 | 3.9 | 1230 |

- Manual revision of dev/test: verbatim & non-verbatim texts ($\triangle$11% WER)
- **Text data** for language modelling (in-domain):

| Data source | Tokens |
|---|---|
| EP training set transcripts | 10M |
| EP transcripts without audio | 6M |
| EP translations into English | 42M |
| Europarl-v10 EN no overlap | 11M |
| DCEP English | 104M |
| Total | 173M |

# 5 Parliaments: Speech data curation for streaming ASR
## Speech data filtering and verbatimization

- **Goal**: To filter/verbatimize training data to improve ASR results.
- Speech data **filtering** (“discard”):
  - Force-align audio and official transcription.
    Accept/reject at word level (phoneme duration, alignment score).
  - 33% of the speech data is filtered out.
- Speech data **verbatimization** (“repair”):
  - For each speech, new automatic transcription
    with LM adapted to the non-verbatim transcription.
  - No speech data is discarded. Verbatimized text $\triangle$7% WER.

| Data set | Duration (h) |
|---|---|
| *raw* | 1007 |
| *filtered* | 672 |
| *verbatimized* | 1054 |

# 5 Parliaments: Speech data curation for streaming ASR
## Europarl-ASR baseline experiments and results

- Experimental setup:
  - 3 streaming ASR systems: raw, filtered, verbatimized
  - Training: Europarl-ASR audio & text
  - Evaluation: MEP & Guest tasks

- Strong baseline ASR results, streaming latency 0.65 s:

| Training set | WER | | | |
|---|---|---|---|---|
| | MEP-test | | Guest-test | |
| | Offline | Streaming | Offline | Streaming |
| *raw* | 8.6 | 8.8 | 7.6 | 7.8 |
| *filtered* | **7.8** | **7.9** | 7.4 | 7.5 |
| *verbatimized* | 8.2 | 8.3 | **7.0** | **7.3** |

# 5 Parliaments: Speech data curation for streaming ASR
## Conclusions

- Created and released the Europarl-ASR corpus:
  `https://www.mllp.upv.es/europarl-asr`
  - Goals: Benchmarking streaming ASR and speech data curation
  - 1300h EN speech data
  - 18h dev/test, manually revised; 2 tasks
  - In use for training SOTA end-to-end ASR models
  - Referenced in recent parliament-based corpora (Euskadi, JA, EL, PT)
- Strong ASR results, offline and streaming (lat. 0.65 s): 7.0–7.9 WER
- Speech data filtering & verbatimization improved WER by 9%
  - Verbatimization keeps all speech data for training

# 6 Broadcast media: Live & offline computer assisted subtitling

- **Objective 4:** Live & offline computer assisted subtitling of broadcast media
- Broadcast media: increasing legal requirements; Spanish Law 13/2022.
  - Subtitling for the deaf (SDH): Minimum 90% of the programming.
- The project: **2020–2023** À Punt Mèdia-UPV R&D agreement for real-time computer assisted subtitling of media contents.
- The pilot: À Punt Mèdia (TV, radio, Internet; live and non-live programming).
  - Languages (ASR): Catalan, Spanish.
  - Datasets: À Punt 2019 (CA 4 hours); À Punt 2020 (CA 37 hours); RTVE 2018–2020 (ES 624 hours).

# 6 Broadcast: Live & offline computer assisted subtitling
## Motivation: À Punt accessibility in 2019

- À Punt accessibility team in 2019: 6 people
  - For live subtitling, pre-recorded subtitling & audio description
- Subtitling for the deaf (SDH) workflow in 2019:
  - Pre-recorded programming: typing and timing from scratch
  - Live programming: typing in real time (in pairs); no respeaking
- Measures to increase subtitle coverage from 2020 on
  - Enlarge accessibility team from 6 to 8 people
  - Use language technologies

**Proposal:** Streaming ASR for live & offline computer assisted subtitling as a tool for the accessibility team to cover more programming without compromising quality

# 6 Broadcast: Live & offline computer assisted subtitling
## Background: MLLP systems for IberSpeech-RTVE 2020

- The benchmark for automatic subtitling of media in Spanish

| System | Configuration | Train. data (h) | WER | |
|---|---|---|---|---|
| | | | test-2018 | test-2020 |
| MLLP 2021 closed-set | Streaming (lat. 0.8 s) | 205 | 13.7 | 20.4 |
| MLLP 2021 open-set | Streaming (lat. 0.8 s) | 3924 | **12.3** | **16.9** |
| 2nd best 2021 (open-set) | Offline | 743 | n/a | 19.3 |

- Top results, with state-of-the-art latencies (0.8 s) for high-quality automatic live subtitling

# 6 Broadcast: Live & offline computer assisted subtitling
## Initial ASR system description

- System architecture:
  - Streaming-ready (latency 0.8 s)
  - Acoustic model: hybrid BLSTM-HMM
  - Language model: neural + n-gram
- Training data:
  - News & entertainment, parliamentary contents, video lectures...

| System | Audio (h) | Text (M words) |
|---|---|---|
| ASR CA Nov19 | 2900 | 320 |
| ASR ES Oct20 | 3900 | 3400 |

- For on-premises deployment in À Punt servers: streaming + offline

# 6 Broadcast: Live & offline computer assisted subtitling
## Initial ASR results & comparison with commercial provider

- Catalan ASR: WER over À Punt 2019, and by programme type

| System | WER<br>À Punt 2019 test |
|---|---|
| MLLP ASR CA Nov19 (offline) | **22** |
| Google S2T CA (Oct 19) (offline) | 49 |

| Dev | | | Test | | |
|---|---|---|---|---|---|
| Programme | Type | WER Nov19 | Programme | Type | WER Nov19 |
| Notícies matí | **TV news** | **8.4** | La qüestió | **TV news** | **11.9** |
| Al ras | Radio | 13.4 | Plaerdemavida | Radio | 12.1 |
| Línia de fons | Radio | 14.4 | Tot futbol | TV | 21.0 |
| La forastera | TV | 18.3 | L'alque. blanca | TV | 22.6 |
| Punt docs | TV | 23.1 | Trau la llengua | TV | 27.7 |
| Twist al tuit | TV | 26.4 | Valenc. al món | TV | 32.0 |
| Assumptes i. | TV | 31.7 | Comediants | TV | 34.3 |
| TOTAL<br>À Punt 2019 | | 18.7 | **TOTAL**<br>**À Punt 2019** | | **22.5** |

# 6 Broadcast: Live & offline computer assisted subtitling
## Initial ASR results & comparison with commercial provider

- Catalan ASR:

| System | WER<br>À Punt 2019 test |
|---|---|
| MLLP ASR CA Nov19 (offline) | **22** |
| Google S2T CA (Oct 19) (offline) | 49 |

- Spanish ASR:

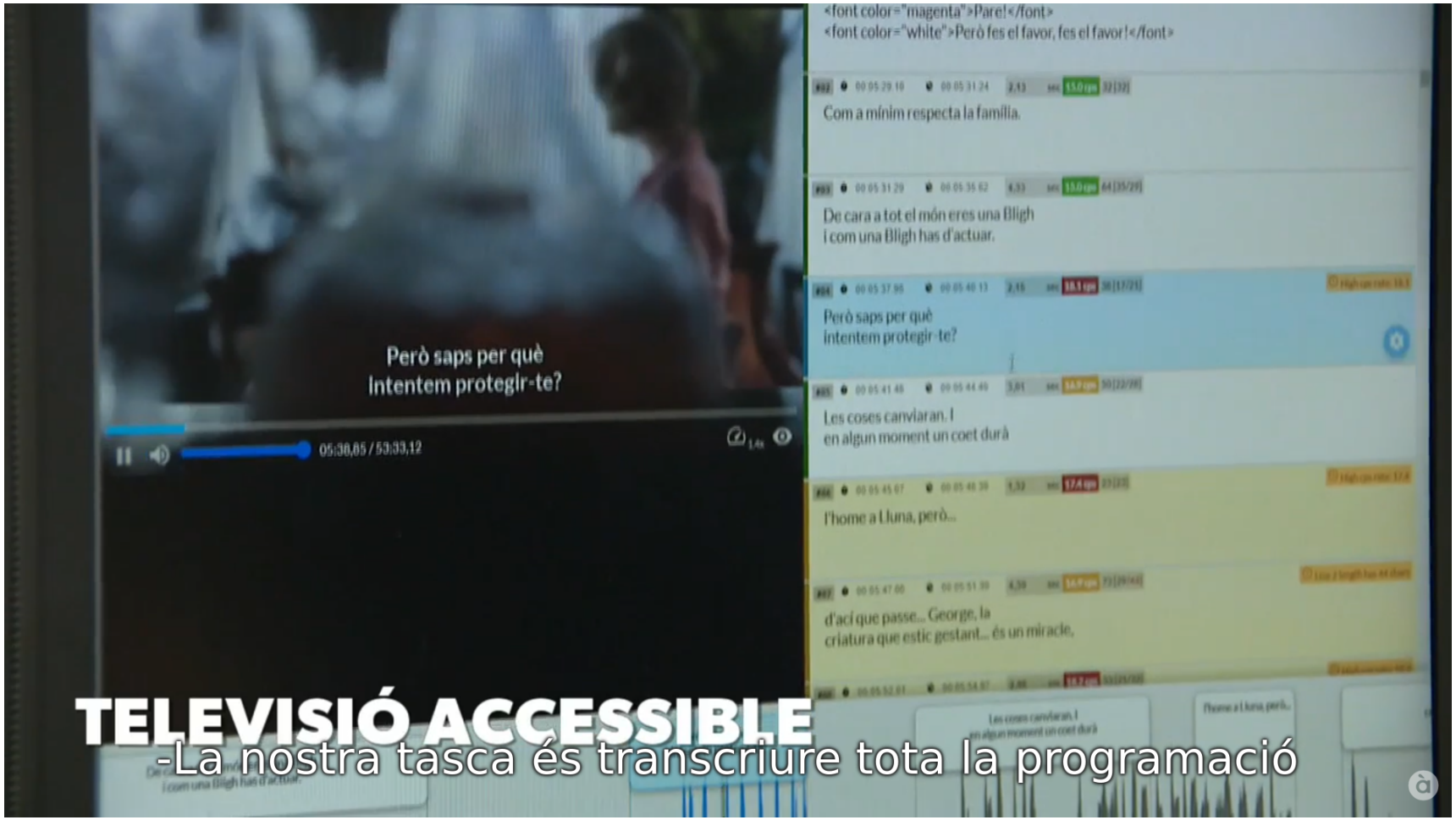
| System | WER<br>RTVE 2018 test |
|---|---|
| MLLP ASR ES Oct20 (offline) | **11** |
| MLLP ASR ES Oct20 (streaming) | **12** |
| Google S2T ES (Feb 20) (offline) | 45 |

# 6 Broadcast: Live & offline computer assisted subtitling
## Integration of ASR systems in À Punt

- On-premises deployment in À Punt servers
  - Offline computer assisted subtitling
  - Live computer assisted subtitling

# 6 Broadcast: Live & offline computer assisted subtitling

- Informatiu migdia À Punt, 28/9/2024, dia de les persones sordes.
  - Offline computer assisted subtitling



# 6 Broadcast: Live & offline computer assisted subtitling

- Informatiu migdia À Punt, 28/9/2024, dia de les persones sordes.
  - Live computer assisted subtitling

# 6 Broadcast: Live & offline computer assisted subtitling
## Improvements derived from actual use in À Punt

- Goal: Saving post-editing time
- Automatic ASR post-processing modules:
  - Automatic recasing (correct use of capital letters)
  - Automatic punctuation
  - Text-to-numbers transliteration ("tres mil" $\rightarrow$ "3.000")
  - Text substitution:
    new AVL2018 orthography; Valencian dialect; À Punt linguistic criteria
- Improving ASR systems
  - Focusing on language model:
    More training data. Recent vocabulary (e.g., COVID-19). New LM tech.

# 6 Broadcast: Live & offline computer assisted subtitling
## Improved ASR systems

- Training data:

| | Audio (h) | | Text (M words) | |
|---|---|---|---|---|
| | Initial | Improved | Initial | Improved |
| ASR CA | 2 900 | | 320 | **2 700** |
| ASR ES | 3 900 | | 3 400 | **17 000** |

- Catalan ASR:

| | WER À Punt 2019 | | |
|---|---|---|---|
| System config. \ System | Nov19 | Jul21 | $\Delta$ |
| Offline config. | 22.5 | **19.6** | $-13\%$ |
| Streaming config. | – | **20.4** | – |

- Spanish ASR:

| | WER RTVE 2018 | | |
|---|---|---|---|
| System config. \ System | Oct20 | Sep22 | $\Delta$ |
| Offline config. | 11.7 | **11.3** | $-3\%$ |
| Streaming config. | 12.8 | **11.7** | $-9\%$ |

# 6 Broadcast: Live & offline computer assisted subtitling
## Conclusions

- Implemented high-quality, low-latency streaming ASR in real broadcasting.
  - For a less-spoken and a widely spoken language (Catalan, Spanish).
  - Improved and adapted ASR systems for broadcast media needs.
  - Different scenarios of live and pre-recorded subtitling, different programmes.
- Strong ASR systems, streaming and offline.
  - WER: Catalan ~20% global, ~10% news; Spanish ~12% global.
- ASR system improvements of 3–13% WER.
  - Updating the language models with more training data.
  - Focusing on recognition of new terms and names in current events.
- Integrated into workflows for live and pre-recorded media subtitling.
  - Currently in actual use at À Punt.

# 7 Achievements and conclusions

1. Contributed to evaluation & post-editing of ASR+MT for OER
2. Developed state-of-the-art neural MT systems for OER
3. Contributed to streaming ASR data curation for parliamentary contents
4. Contributed to live & offline assisted subtitling of broadcast media

# *7 Achievements and conclusions*
## Future work

- [Ch3] Intelligent interaction: MT quality estimation
- [Ch4] New NMT SOTA: Decoder-only neural LLMs
  Streaming MT & Speech Translation for OER
- [Ch5] Europarl-ASR: Expanding with more English and other EU languages
- [Ch6] New lines of research in follow-up À Punt–UPV agreement

# Publications (1/2)

## Ch3: Evaluation & post-editing of ASR+MT w/ intell. interaction for OER

- J. A. Silvestre-Cerdà, M. Á. Del Agua, **G. Garcés**, G. Gascó, A. Giménez, A. Martínez, A. Pérez, I. Sánchez, N. Serrano, R. Spencer, J. D. Valor, J. Andrés-Ferrer, J. Civera, A. Sanchis, and A. Juan. transLectures. In *Proc. VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop (IberSpeech 2012)*, pages 345–351, Madrid (Spain), 2012.

- J. D. Valor Miró, R. N. Spencer, A. Pérez González de Martos, **G. Garcés Díaz-Munío**, C. Turró, J. Civera, and A. Juan. Evaluación del proceso de revisión de transcripciones automáticas para vídeos Polimedia. In *Proc. I Jornades d'Innovació Educativa i Docència en Xarxa (IN-RED 2014)*, pages 272–278, València (Spain), 2014.

- *Indexed international journal. Scimago **SJR Q2** ("Education", 2014):*

  J. D. Valor Miró, R. N. Spencer, A. Pérez González de Martos, **G. Garcés Díaz-Munío**, C. Turró, J. Civera, and A. Juan. Evaluating intelligent interfaces for post-editing automatic transcriptions of online video lectures. *Open Learning*, 29(1):72–85, 2014.

# Publications (2/2)

## Ch4: The transition from phrase-based MT to neural MT of OER

- J. Iranzo-Sánchez, P. Baquero-Arnal, **G. V. Garcés Díaz-Munío**, A. Martínez-Villaronga, J. Civera, and A. Juan. The MLLP-UPV German-English Machine Translation System for WMT18. In *Proc. 3rd Conf. on Machine Translation (WMT18)*, pages 422–428, Brussels (Belgium), 2018.

## Ch5: Streaming ASR data curation for parliamentary contents

- *Indexed international conference.* ***CORE 2021 A****:*

  **G. V. Garcés Díaz-Munío**, J. A. Silvestre-Cerdà, J. Jorge, A. Giménez Pastor, J. Iranzo-Sánchez, P. Baquero-Arnal, N. Roselló, A. Pérez-González de Martos, J. Civera, A. Sanchis, and A. Juan. Europarl-ASR: A Large Corpus of Parliamentary Debates for Streaming ASR Benchmarking and Speech Data Filtering/Verbatimization. In *Proc. Interspeech 2021*, pages 3695–3699, Brno (Czech Republic), 2021.

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

**Tesi doctoral**

# ASR and MT with deep neural networks for open educational resources, parliamentary contents and broadcast media

**Programa de Doctorat en Informàtica**

Gonçal Garcés Díaz-Munío

Directors: Dr. Alfons Juan Ciscar, Dr. Jorge Civera Saiz

15 d'octubre de 2024

# 5 Parliaments: Speech data curation for streaming ASR
## Motivation: ASR corpora in 2020

- Lack of realistic tasks for streaming ASR evaluation
- Speech corpora usually based on non-verbatim transcriptions

| | LibriSpeech | TED-LIUM | Europarl-ST | VoxPopuli | Europarl-ASR |
|---|---|---|---|---|---|
| Release date | 2015 | 2012–2018 | 2020 | 2021 | 2021 |
| Domain | Audiobooks | TED talks | European Parliament | European Parliament | European Parliament |
| Focus | ASR | ASR | Speech translation | Speech-to-speech, unsupervised | Streaming ASR, data curation |
| EN transcribed speech (h) | 1000 | 452 | 186 | 543 (unsup 24K) | 1263 |
| dev-test sets | Verbatim, segmented | Verbatim, segmented | Non-verbatim | Non-verbatim | Verbatim & non-verbatim, long-form |