

# Evaluating Automatic Subtitling: Correlating Post-editing Effort and Automatic Metrics

Alina Karakanta, Mauro Cettolo, Matteo Negri, Luisa Bentivogli



Universiteit  
Leiden  
The Netherlands

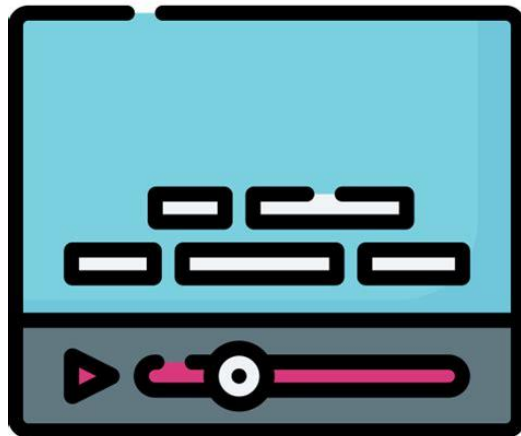
LREC-COLING 2024  
Torino, May 2024



# Meeting increasing subtitling demand



The Subtitling Boom!



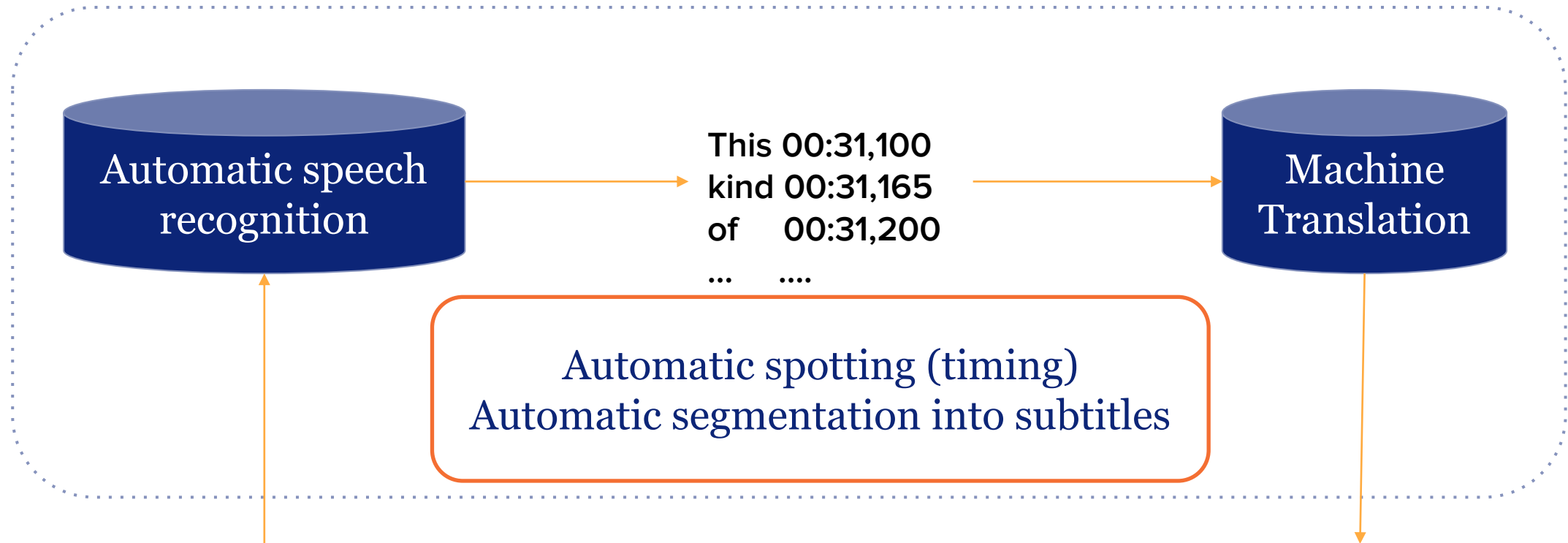
Automatic subtitling



Provide direct access to audiovisual products

As a tool to increase subtitlers' productivity

# Automatic subtitling



This 00:31,100  
kind 00:31,165  
of 00:31,200  
... ..



9  
00:00:31,100 --> 00:00:33,090  
Questo tipo di molestia impedisce  
10  
00:00:33,414 --> 00:00:36,191  
alle donne di accedere a Internet,  
ovvero dalla conoscenza.

# Post-editing in automatic subtitling

The screenshot displays the Matesub web application interface for post-editing subtitles. The browser address bar shows the URL: <https://app.matesub.com/editor/task1/ac5143cf-551b-4486-a7a2-3d73a5ca0990#09:08,09>. The application header includes the Matesub logo, the task name "task1", and a language selection dropdown set to "English US → German" with a 64% progress indicator. An "Export" button and a user profile icon labeled "AN" are also visible.

The main interface is divided into three primary sections:

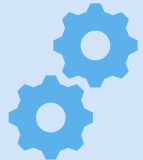
- Subtitles list (240):** A sidebar on the left displays a list of subtitle entries. Entry #164 is partially visible: "und ihren Teams zu vertrauen, dass... richtigen Entscheidungen treffen w...". Entry #165 is selected and shows the time range "09:08:04 → 09:09:05" and the text "auf dem Weg dorthin.". Entry #166 is partially visible: "Die Erstellung der Ausrichtung erfo... eine offene und transparente Kom...".
- Video Player:** The central area shows a video player with a speaker on stage. A subtitle "auf dem Weg dorthin." is overlaid on the video. The player includes a progress bar, playback controls, and a "Translated" button.
- Timeline:** A detailed timeline at the bottom shows the subtitle alignment. It includes a word cloud for the selected subtitle: "and", "trust", "their", "teams", "to", "make", "right", "decisions", "how", "to", "get", "there.", "now", "creating", "alignm". Below the word cloud, an audio waveform is shown with a red box indicating a speed of "28.98 cps" for the selected subtitle.

The Windows taskbar at the bottom shows the search bar with the text "Scrivi qui per eseguire la ricerca", the system tray with a temperature of 17°C, and the date/time "14:16 09/11/2021".

# Motivation

- The way subtitlers interact with automatically generated subtitles has not been yet explored. Previous studies have focused on **translation edits**, without studying the **effort of adjusting the timestamps and segmentation**.
- New **automatic subtitling metrics** for assessing the quality of automatically-generated subtitles
- **Neural-based metrics** with increased correlations with human ratings
- However, to date there exists no study on the usefulness of automatic metrics for assessing productivity and quality in automatic subtitling.

# Contributions



A correlation analysis of automatic MT metrics with human measures of post-editing effort in automatic subtitling



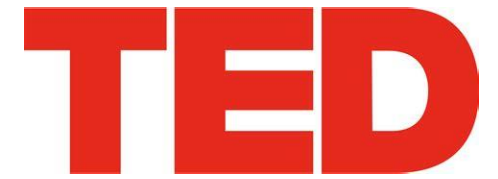
A new dataset in subtitling PE (en->it,de), containing:

product data (subtitles),  
process data (time, keystrokes)  
participant-based data (user  
experience ratings)

# Data collection

# Data

- MuST-Cinema test set (Karakanta et al. 2020)
- 9 English single-speaker TED Talks
- 1 hour video (545 sentences, 10K words)
- English into Italian x2 subtitles
- English into German x1 subtitle

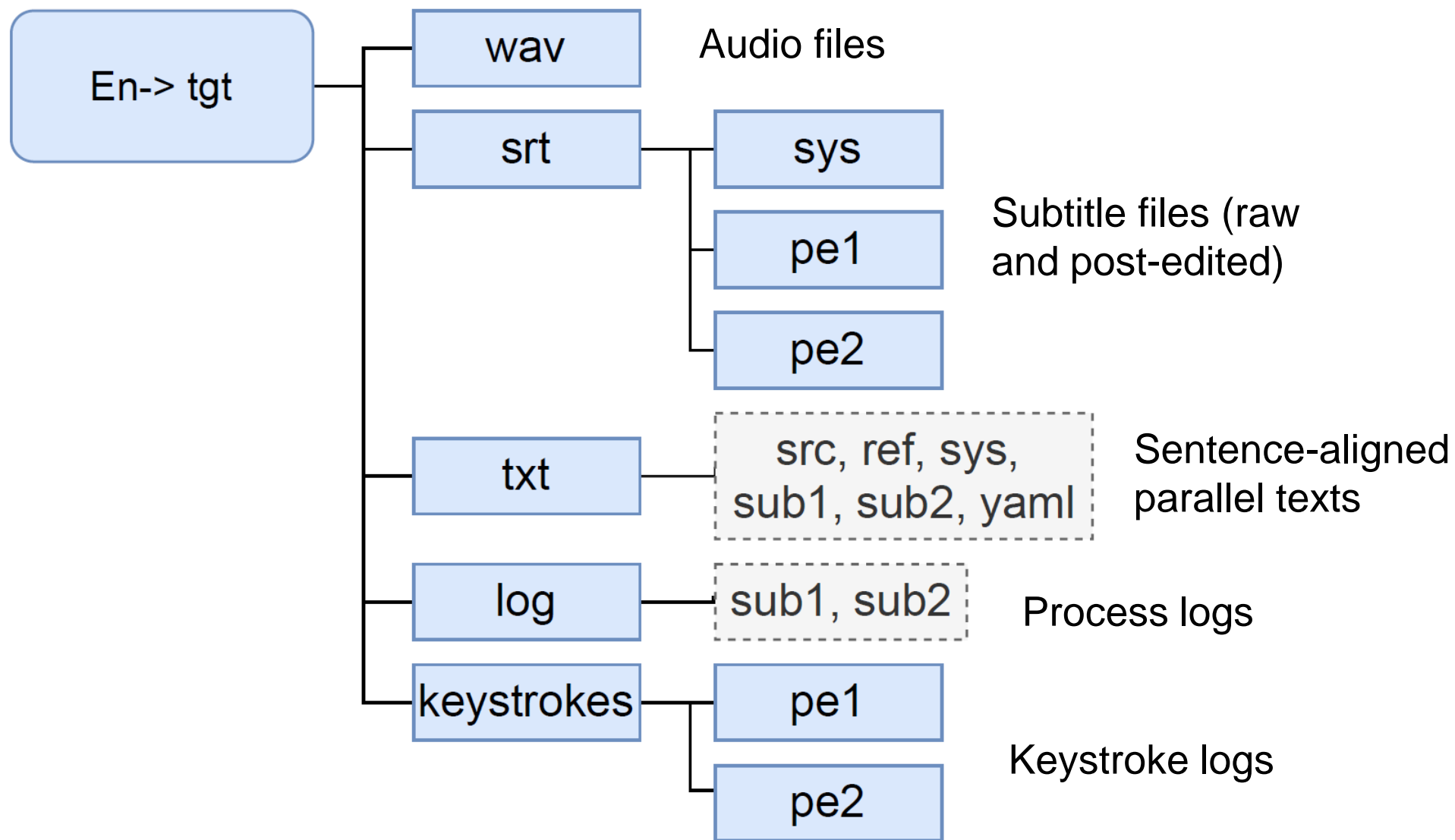


# Task

- Post-editing in Matesub
- Professional subtitlers with MTPE experience, users of Matesub
- Guidelines and test session
- 12 tasks (3 video minutes each)
- Process logs downloaded from Matesub
- Keystrokes using InputLog
- Screen recordings
- UX questionnaire

| text   | original_text                                     | start  | orig_start | end    | orig_end | time_activity |
|--|---|--------|------------|--------|----------|---------------|
| aber einander in den letzten 10 Jahren<br>höchstens E-Mails und Statusberichte | wir senden einander E-Mails<br>und Statusberichte | 375.92 | 375.93     | 379.64 | 378.32   | 144649        |
|  | in den letzten zehn Jahren                        |        | 378.41     |        | 380.05   |               |
| geschickt hatten.  |   | 379.72 |            | 380.85 |          | 12957         |

# Corpus structure



# Correlation analysis

# Effort measures

## temporal effort



### **Post-Editing Speed (PES):**

average number of edited words per minute

## technical effort



### **Total interaction events (Tot\_int):**

the sum of all keystrokes and mouse clicks, normalised by video length



### **Mouse clicks/interaction events (Mouse/Int):**

the percentage of mouse clicks over total interaction events

# Automatic metrics

## Subtitle metrics

- Subtitle Edit Rate (SubER)
- Timed-BLEU (T-BLEU)
- AS-BLEU
- Sigma

## Traditional MT metrics

- HTER
- BLEU
- charF

## Neural metrics

- COMET
- BERTScore
- BLEURT

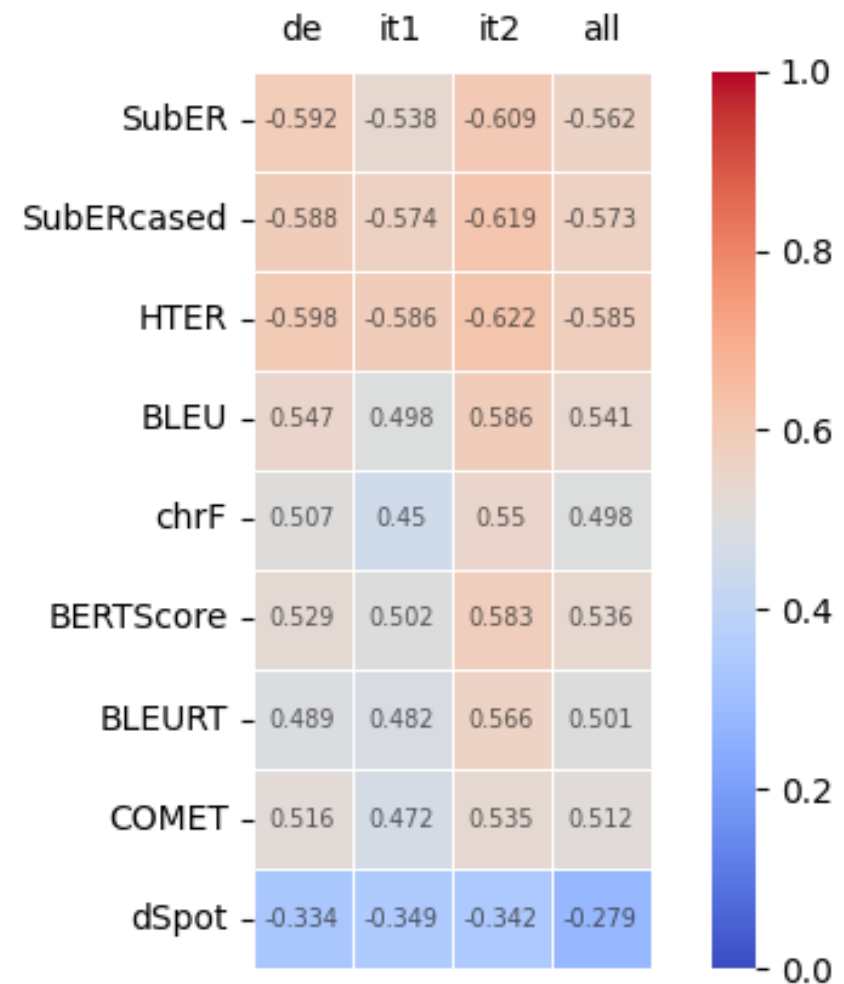
# Subtitle vs task

- Subtitle level
  - Correlations with **PES** based on process logs
  - **SubER, dSpot, traditional and neural MT metrics**
  - T-BLEU=AS-BLEU=BLEU
  - Per subtitler and aggregated for all subtitlers
- Task level
  - Correlations with **PES, Total interactions, Mouse/Int** based on keystroke logs
  - **All automatic metrics**
  - Aggregated for all subtitlers

# Results

# Subtitle-level correlations

## Moderate correlations

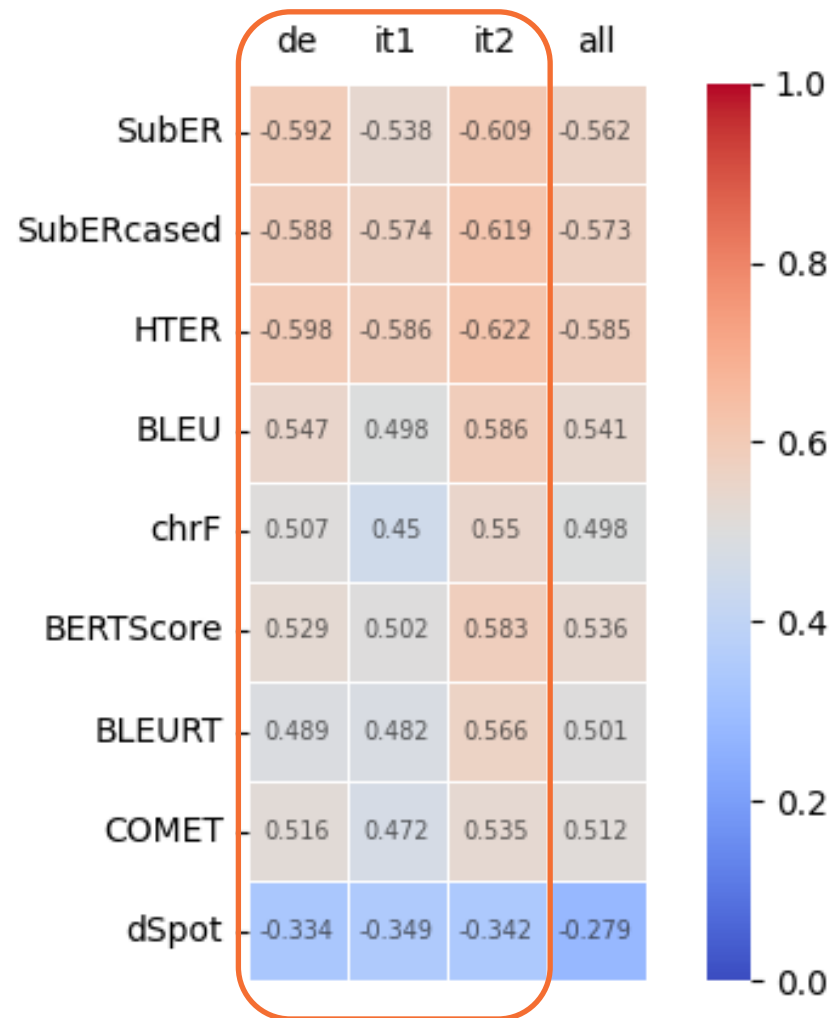


# Subtitle-level correlations

Moderate correlations

Individual differences

- Correlations:  $it2 > de > it1$
- Average PES:  $it1 > it2 > de$



# Subtitle-level correlations

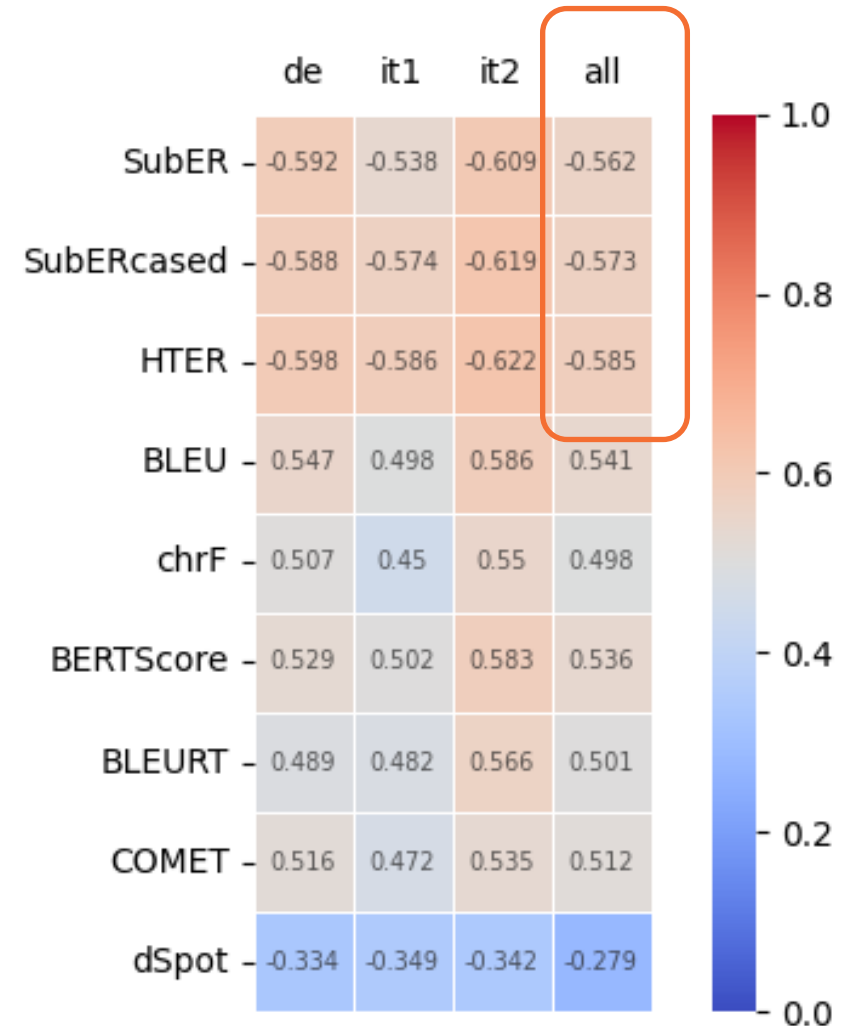
## Moderate correlations

## Individual differences

- Correlations:  $it2 > de > it1$
- Average PES:  $it1 > it2 > de$

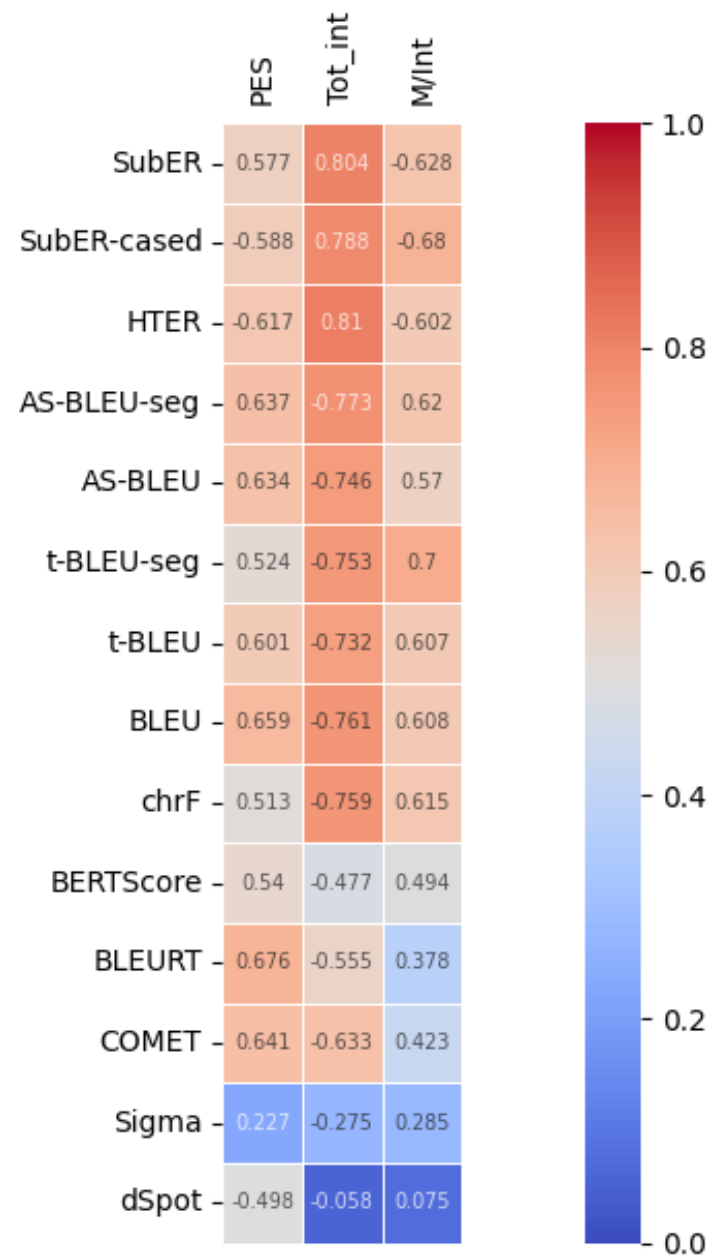
## All

- Edit distance metrics correlate the most
- Followed by BLEU and BERTScore



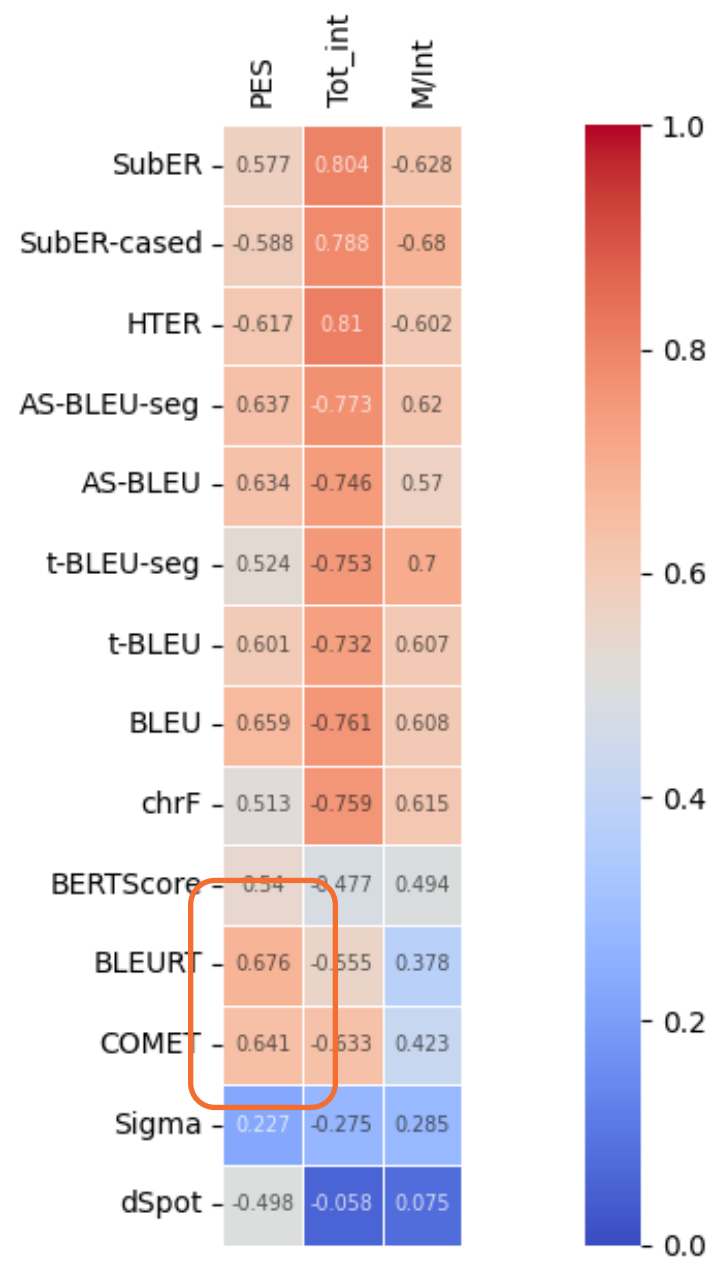
# Task-level correlations

- Stronger correlations



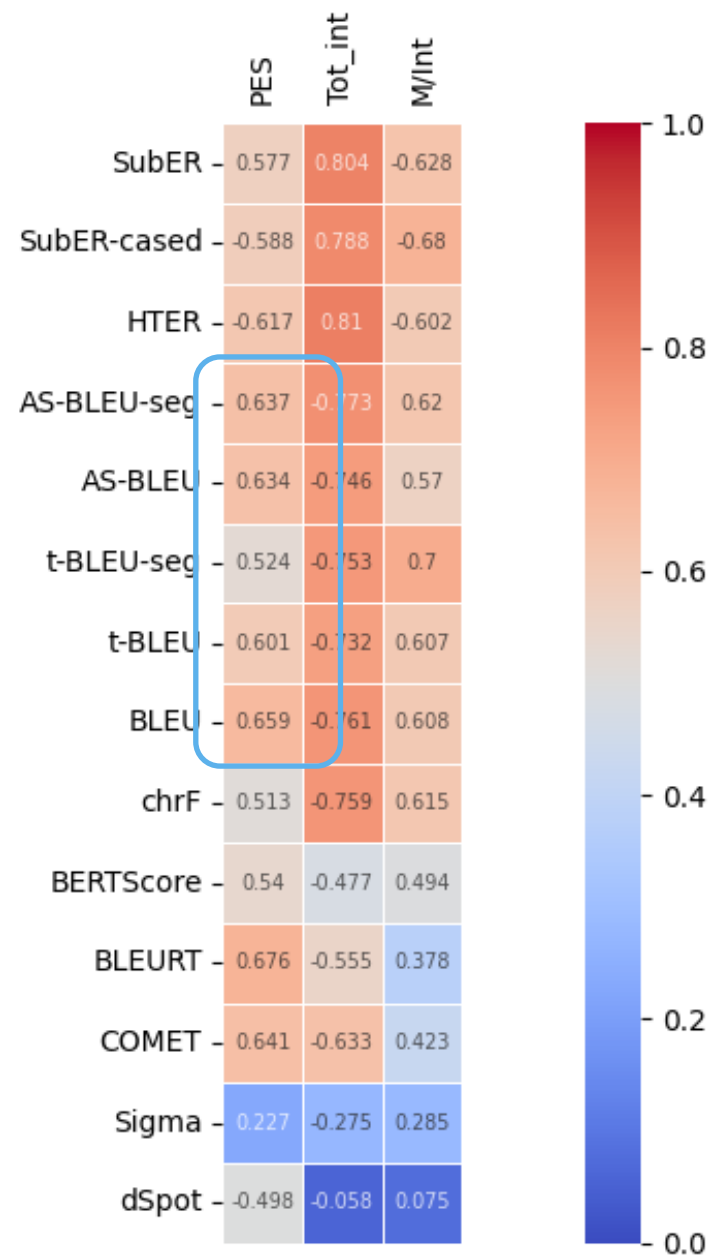
# Task-level correlations

- Stronger correlations
- PES
  - Neural metrics correlate best



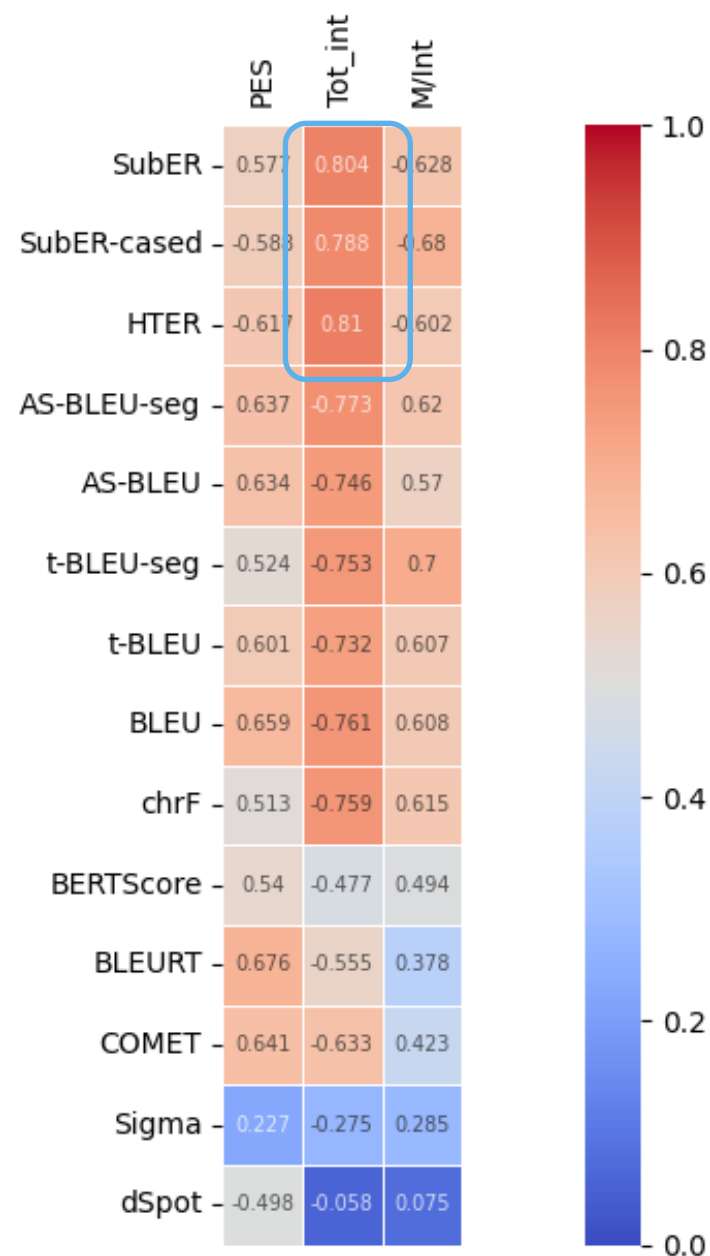
# Task-level correlations

- Stronger correlations
- PES
  - Neural metrics correlate best
  - BLEU (variants) come next



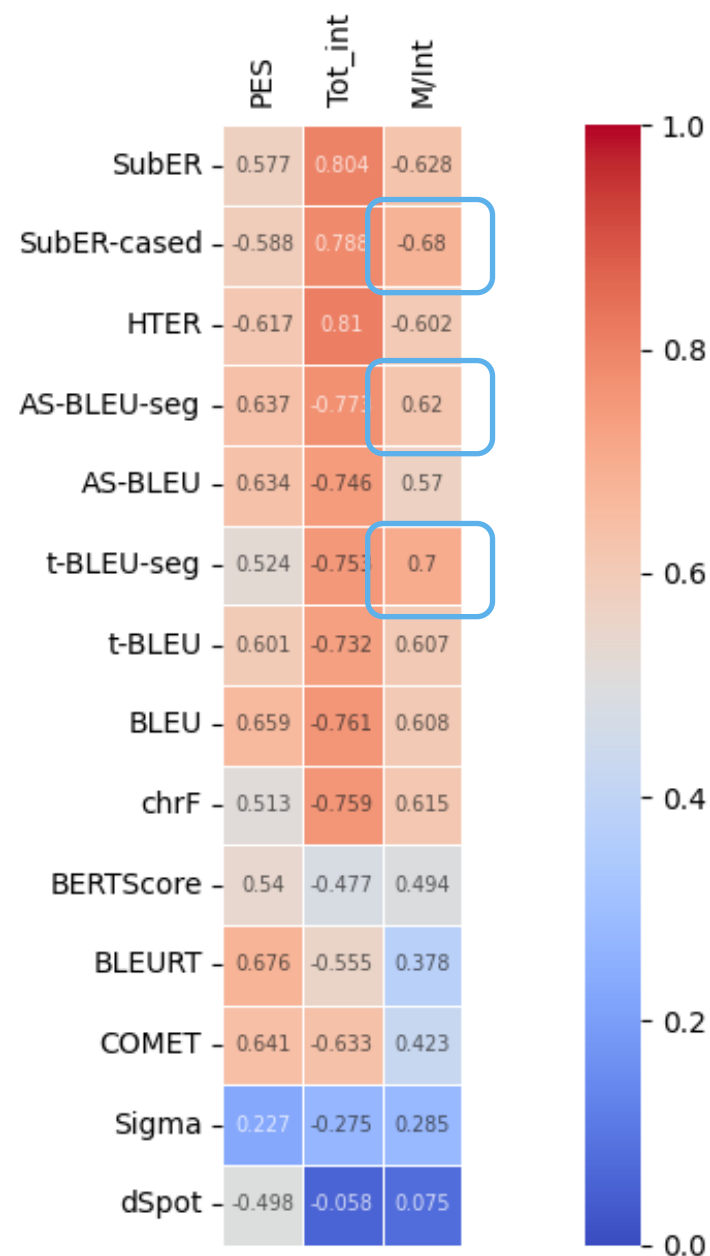
# Task-level correlations

- Stronger correlations
- PES
  - Neural metrics correlate best
  - BLEU (variants) come next
- Total interactions
  - Edit distance metrics > string-based > neural



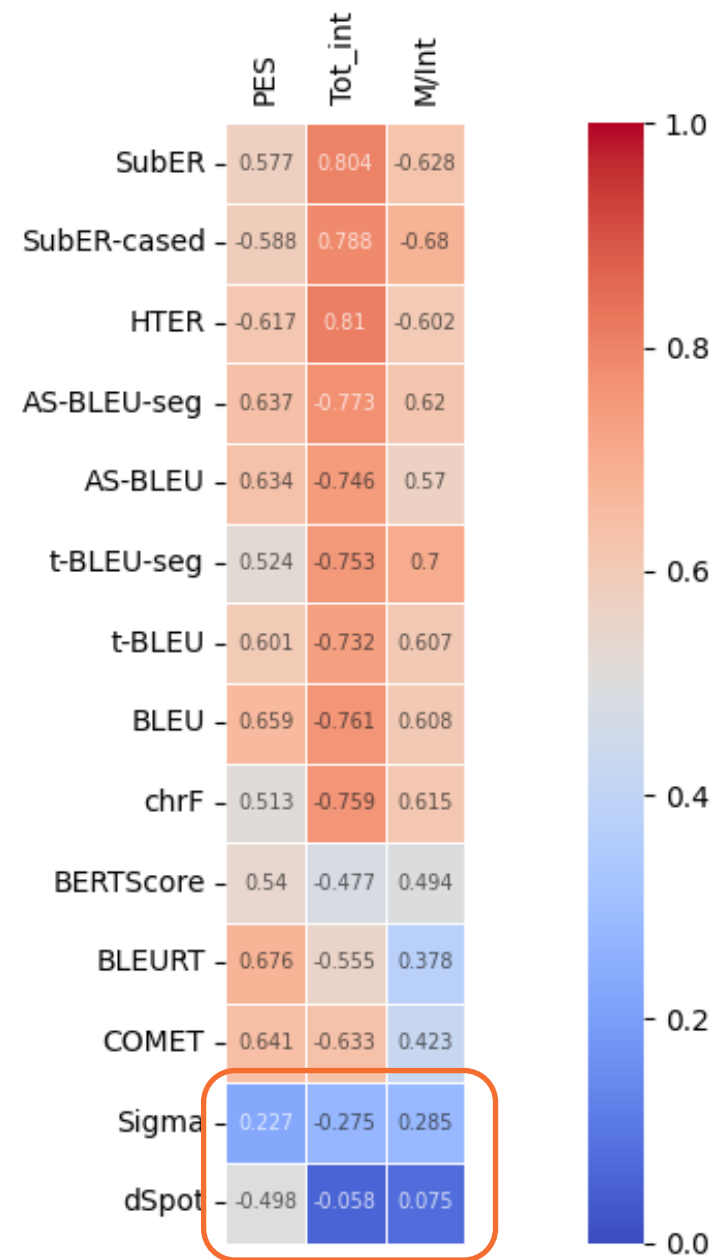
# Task-level correlations

- Stronger correlations
- PES
  - Neural metrics correlate best
  - BLEU (variants) come next
- Total interactions
  - Edit distance metrics > string-based > neural
- Mouse/Int ratio
  - Metrics considering case/segmentation have higher correlations than their unaware counterparts



# Task-level correlations

- Stronger correlations
- PES
  - Neural metrics correlate best
  - BLEU (variants) come next
- Total interactions
  - Edit distance metrics > string-based > neural
- Mouse/Int ratio
  - Metrics considering case/segmentation have higher correlations than their unaware counterparts
- Sigma and dSpot only capture one aspect of the subtitling process



# Conclusion

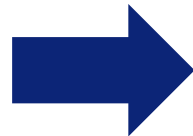
- A new **corpus** containing process-, product- and participant-based data of PE in automatic subtitling
- Correlations of **automatic MT quality metrics** with **technical** and **temporal PE effort**
- **Edit distance metrics** correlate extremely well with the total technical effort in editing automatic subtitles when considering an entire task (video)
- **Neural metrics**, when computed at the level of sentences, correlate well with PE speed
- Automatic metrics only moderately capture productivity and effort at the **subtitle level**

Corpus available through CLARIN <http://hdl.handle.net/10032/tm-a2-y2>

More details on the project can be found at <https://mt.fbk.eu/must-cinema-pe/>

# Conclusion

- A new **corpus** containing process-, product- and participant-based data of PE in automatic subtitling
- Correlations of **automatic MT quality metrics** with **technical** and **temporal PE effort**
- **Edit distance metrics** correlate extremely well with the total technical effort in editing automatic subtitles when considering an entire task (video)
- **Neural metrics**, when computed at the level of sentences, correlate well with PE speed
- Automatic metrics only moderately capture productivity and effort at the **subtitle level**



further investigations with more **languages**, **subtitlers** and **domains** will grant us a better understanding into the subtitle PE process, individual subtitler differences and the evaluation of automatic subtitling

Corpus available through CLARIN <http://hdl.handle.net/10032/tm-a2-y2>

More details on the project can be found at <https://mt.fbk.eu/must-cinema-pe/>

# Evaluating Automatic Subtitling: Correlating Post-editing Effort and Automatic Metrics

Alina Karakanta, Mauro Cettolo, Matteo Negri, Luisa Bentivogli



Universiteit  
Leiden  
The Netherlands

LREC-COLING 2024  
Torino, May 2024

