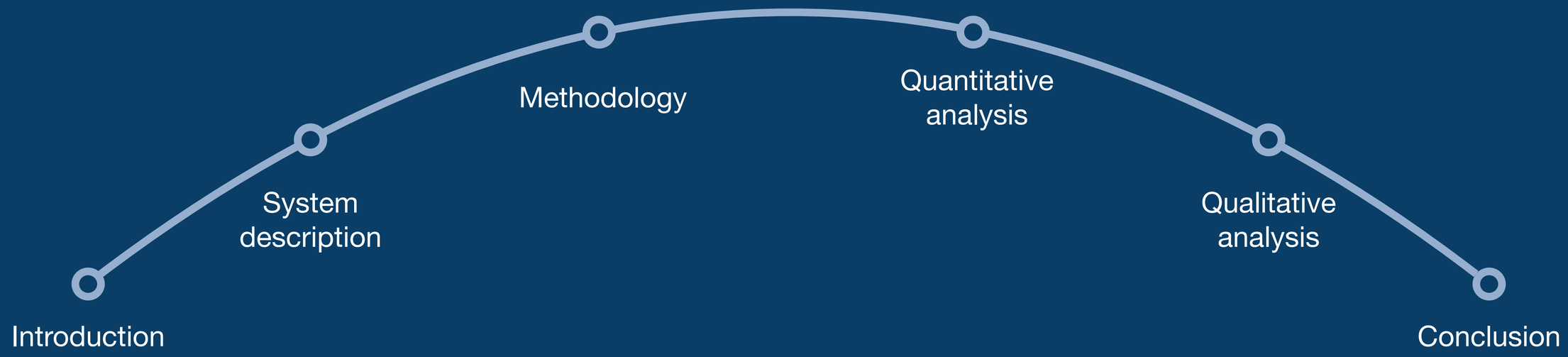
An experiment in error analysis of real-time speech machine translation using the example of the European Parliament's Innovation Partnership

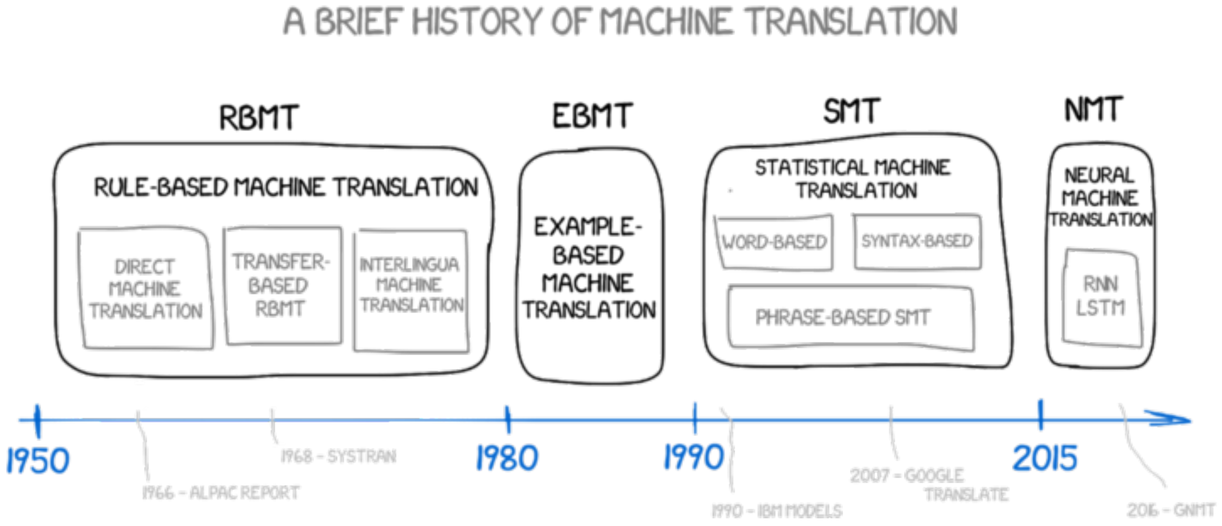
Elisa Di Nuovo, PhD
Directorate General for Translation
Directorate for Citizens' Language
Speech to Text Unit

European Parliament

Overview

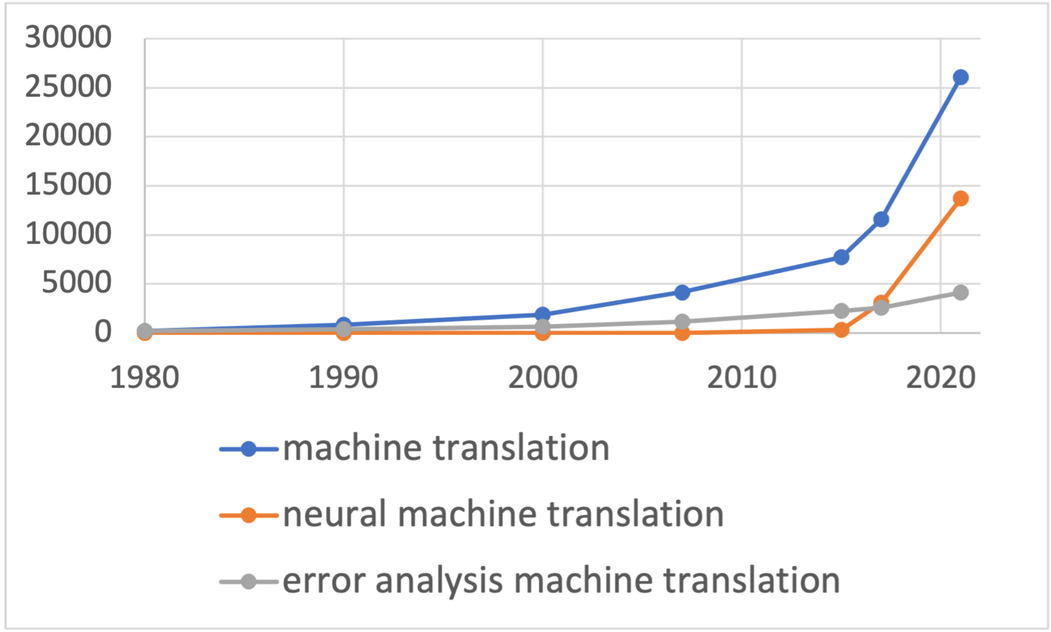


European Parliament



https://www.freecodecamp.org/news/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5/

PUBLICATIONS ABOUT…



Machine translation is difficult to evaluate

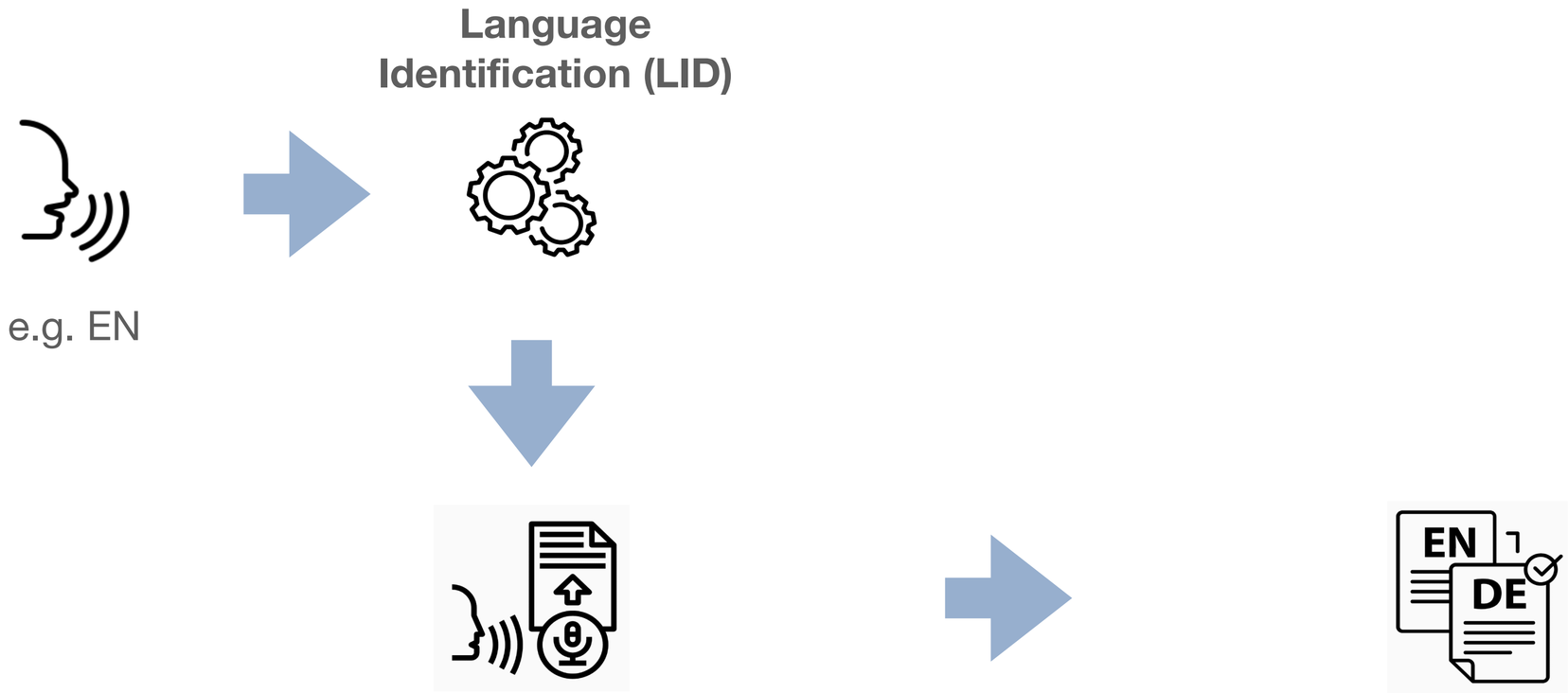
There is more than one correct translation

Quality evaluation:

- **Human / Manual**
  - Directly expressed judgement (DEJ) evaluation methods
    - Adequacy / Accuracy, Fluency
  - Non-DEJ-based evaluation methods
    - Error classification, post-editing
- **Automatic**
  - Reference translation-base metrics
    - BLEU, TER…
  - Quality estimation metrics
  - Linguistic checkpoint-based evaluation

Used in shared tasks, e.g. IWSLT

Low-latency cascaded ASR + MT tool



To date, the tool is able to transcribe and translate automatically in 19 languages

2020/2021

- English (EN)
- French (FR)
- German (DE)
- Spanish (ES)
- Italian (IT)
- Polish (PL)
- Greek (EL)
- Romanian (RO)
- Dutch (NL)
- Portuguese (PT)

- Bulgarian (BG)
- Czech (CS)
- Slovak (SK)
- Slovenian (SL)
- Croatioan (HR)
- Lithuanian (LT)
- Finnish (FI)
- Hungarian (HU)
- Swedish (SV)

2021/2022

Quality factors:

- Architecture of the neural network
- Volume of data
- Features of data
  - Languages
  - Domain
- Data quality
  - Noise
  - Bias reduction

Challenges:

- Not segmented text
- High degree of multilingualism
- Non-native accents
- Large variety of domains

ASR trained on other data + European Parliament data

EN 864 h
FR 263 h
DE 296 h
ES 143 h
RO 85 h

MT trained on different parallel corpora (including the EUROPARL corpus)

Hundreds of millions of words

Quantitative and qualitative study on ASR and MT quality

- Automatic metrics (WER, word error rate, for ASR)
  - Test set consisting of 92 speeches delivered in March and May 2022 by Members of the European Parliament (2 h 31 m 57 s)

- non-DEJ human evaluation method (Multidimensional Quality Metrics, MQM, for both ASR and MT) involving 4 annotators
  - Test set consisting of
    - 18 automatically transcribed speeches
      - 3 speeches in EN, ES, DE, FR, IT, RO
    - 30 automatically translated speeches
      - 3 speeches in EN translated into DE, ES, FR, RO, IT
      - 3 speeches in DE, ES, FR, RO translated into IT
      - 3 speeches in IT translated into EN

Test set evaluated using WER metric

Language	Number of speeches	Time (hh:mm:ss)
EN	9	00:14:56
FR	3	00:04:52
DE	4	00:06:28
ES	4	00:05:03
IT	7	00:15:31
RO	4	00:04:37
PL	10	00:19:10
EL	1	00:01:05
NL	1	00:01:21
PT	1	00:01:10
BG	4	00:07:19
CS	4	00:07:58
SK	4	00:05:42
SL	4	00:06:10
HR	4	00:05:59
LT	4	00:06:13
FI	8	00:13:20
HU	9	00:13:03
SV	7	00:12:00
Total	92	02:31:57

Table 1: ASR automatic evaluation test set.

Test set using non-DEJ human evaluation method (MQM)

Annotator	Language	CEFR level
Ann A	RO	Native language
	EN	C2
	IT	C2
Ann B	IT	Native language
	RO	Native language
	EN	C2
	ES	C2
	DE	C2
Ann C	IT	Native language
	EN	C1
	FR	C1
	ES	B2
Ann D	IT	Native language
	EN	C2
	FR	C1

Table 2: Annotators and language knowledge.

Evaluated task	Number (source id)	Annotations
RO transcription	3 speeches	Ann A – Ann B
IT transcription	3 speeches	Ann B – Ann C
EN transcription	3 speeches	Ann B – Ann C
ES transcription	3 speeches	Ann B – Ann C
FR transcription	3 speeches	Ann C – Ann D
DE transcription	3 speeches	Ann B
MT EN-IT	3 speeches	Ann B – Ann C
MT EN-RO	3 speeches	Ann A – Ann B
MT EN-FR	3 speeches	Ann C – Ann D
MT EN-ES	3 speeches	Ann B
MT EN-DE	3 speeches	Ann B
MT RO-IT	3 speeches	Ann A – Ann B
MT IT-EN	3 speeches	Ann B – Ann C
MT ES-IT	3 speeches	Ann B – Ann C
MT FR-IT	3 speeches	Ann C – Ann D
MT DE-IT	3 speeches	Ann B

Table 3: Human evaluation test set.

Error categories

ASR

- Over-segmentation
- Under-segmentation
- Lexical substitution
- Lexical deletion
- Lexical addition
- Morpho-syntactic
- Terminology

MT

- Accuracy
- Punctuation
- Grammar
- Register
- Terminology
- Other
- Unintelligible

Error severity

Neutral → 0 points

Minor → 1 point

Major → 5 points

Results using WER metric

Language (source id)	LID	WER	Language (source id)	LID	WER
All 19 languages, 1 speech each	On	6.45	PL	Off	5.05
RO	On	2.77	PL	On	7.80
IT	On	3.22	HU	Off	9.18
IT	Off	5.58	HU	On	9.57
EN	On	8.98	CS	Off	4.03
ES	On	4.94	SK	Off	2.52
FR	On	8.91	SL	Off	5.02
DE	On	7.81	HR	Off	5.63
EN	Off	5.25	LT	Off	11.14
EN	On	5.48	FI	Off	5.48
BG	Off	5.83	SV	Off	10.78

Table 4: WER results.

Average WER: 6.34%

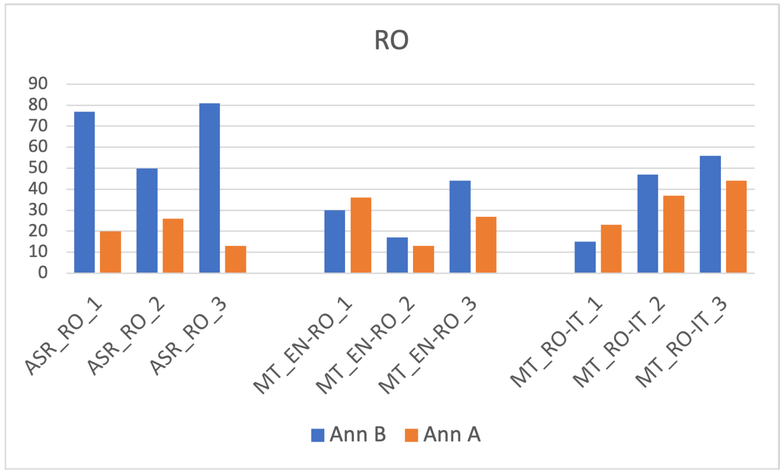
Results using WER metric

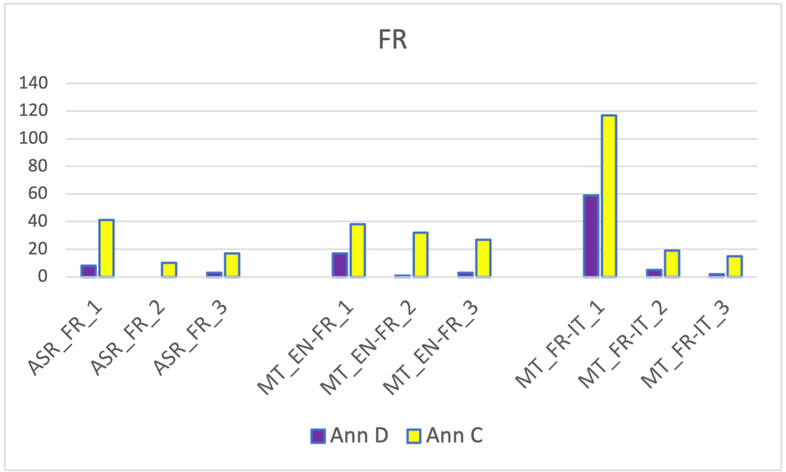
Language (source id)	LID	WER	Language (source id)	LID	WER
All 19 languages, 1 speech each	On	6.45	PL	Off	5.05
RO	On	2.77	PL	On	7.80
IT	On	3.22	HU	Off	9.18
IT	Off	5.58	HU	On	9.57
EN	On	8.98	CS	Off	4.03
ES	On	4.94	SK	Off	2.52
FR	On	8.91	SL	Off	5.02
DE	On	7.81	HR	Off	5.63
EN	Off	5.25	LT	Off	11.14
EN	On	5.48	FI	Off	5.48
BG	Off	5.83	SV	Off	10.78

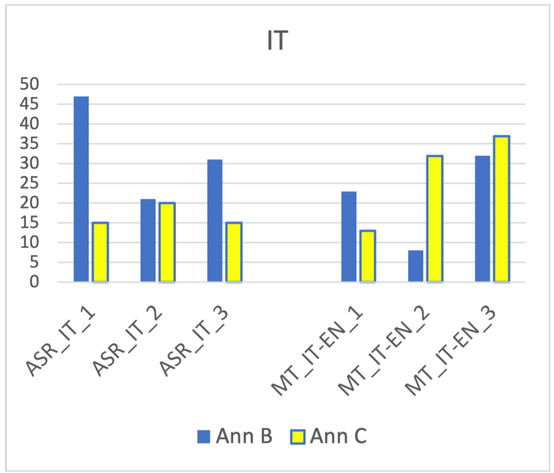
Table 4: WER results.

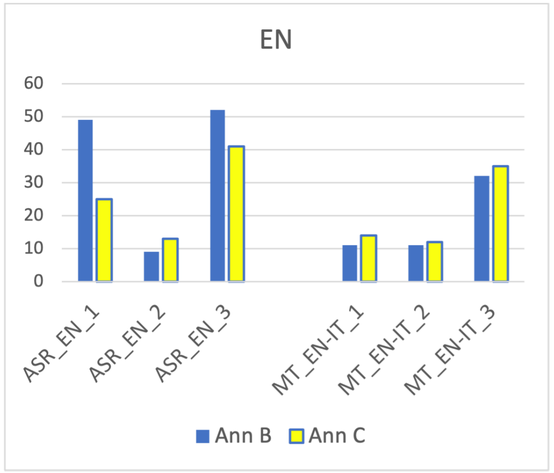
Average WER: 6.34%

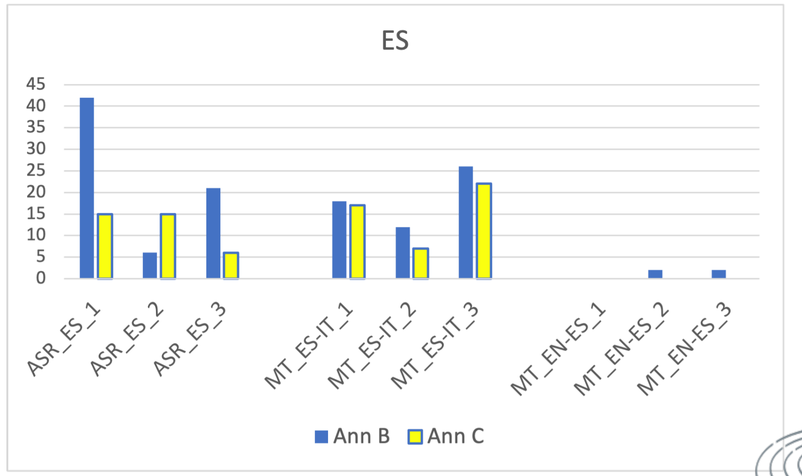
Same speeches, different annotators



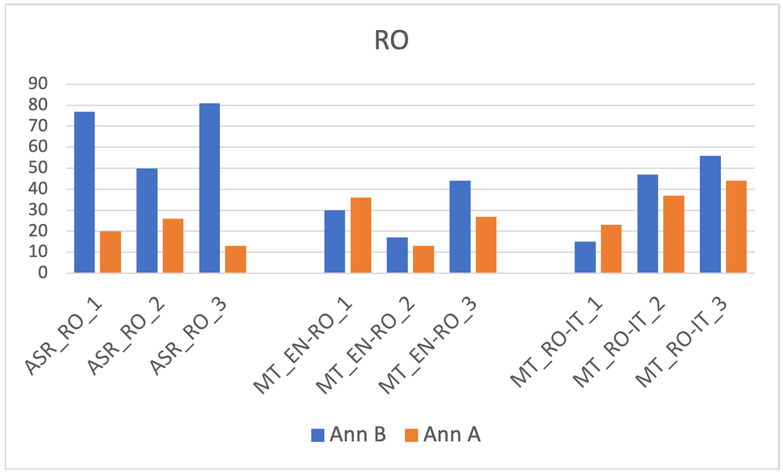


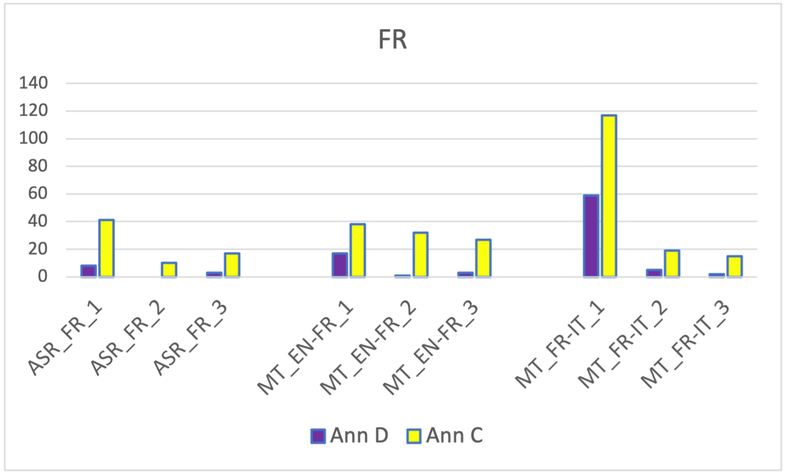


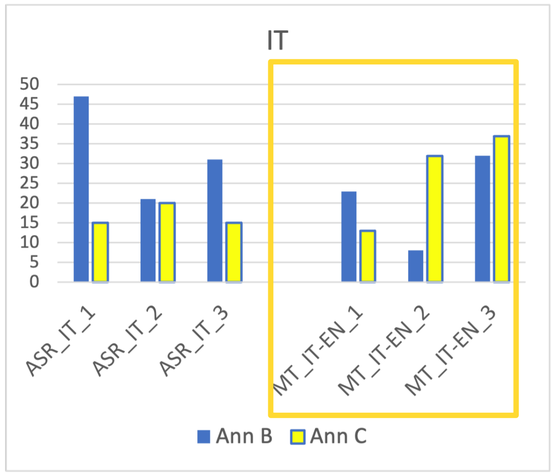


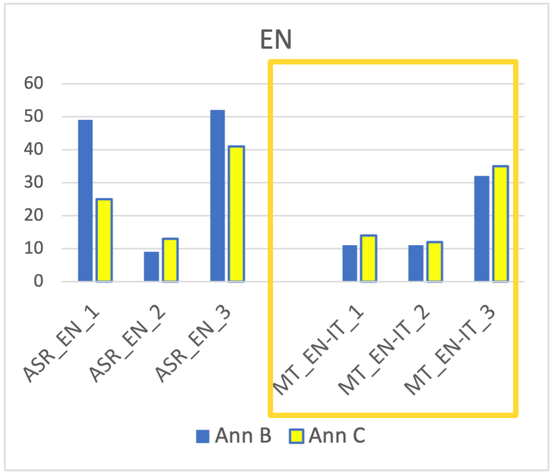


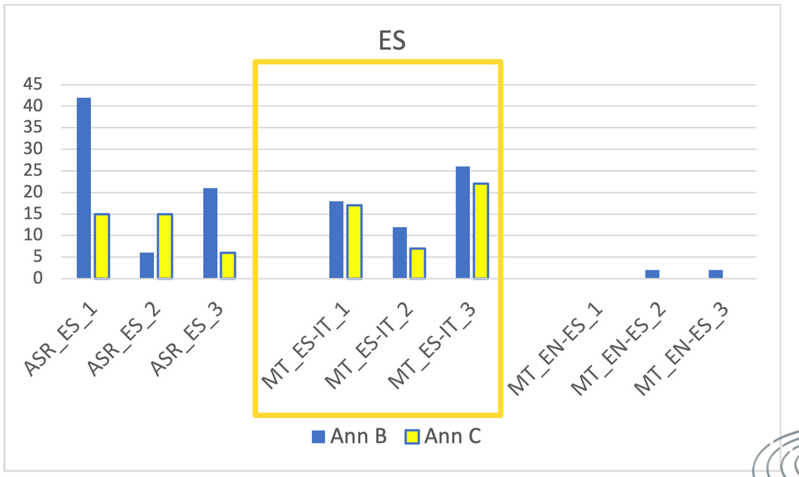
Same speeches, different annotators



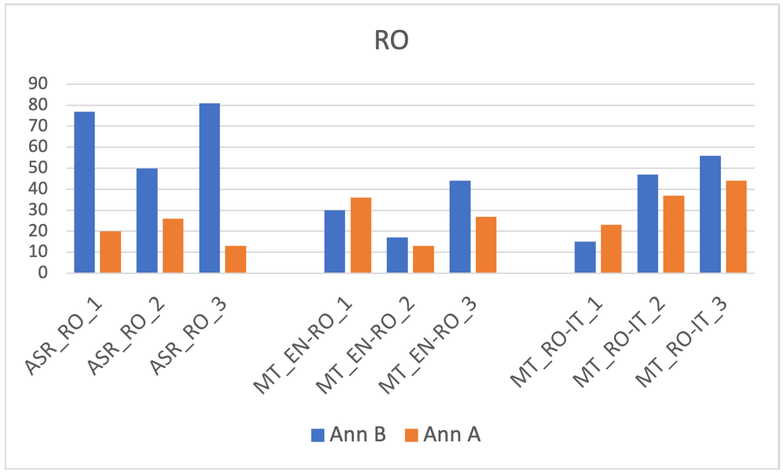


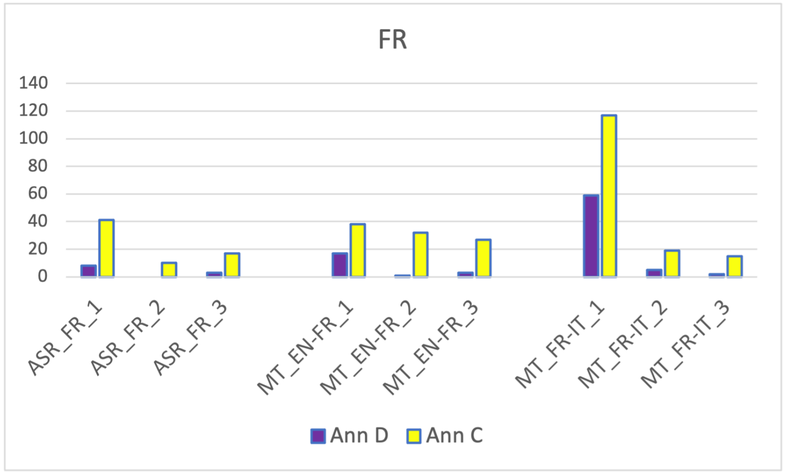


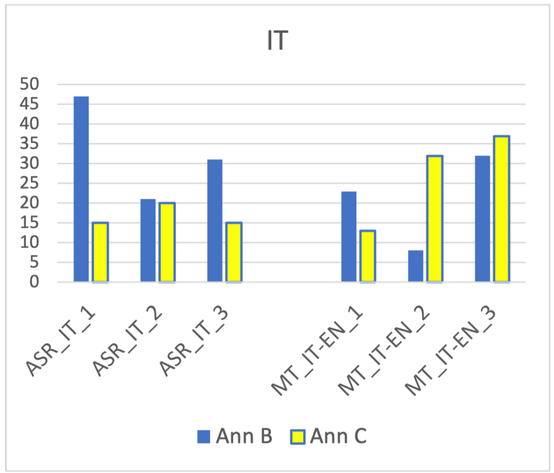


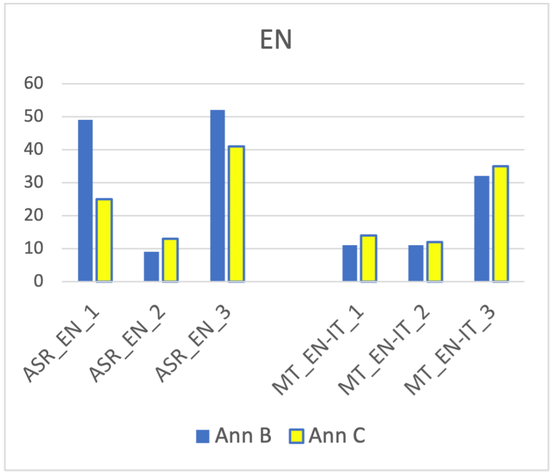


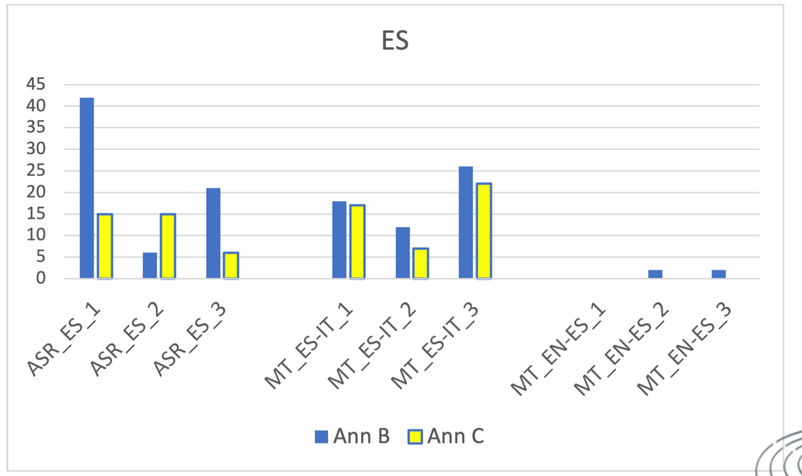
Same speeches, different annotators



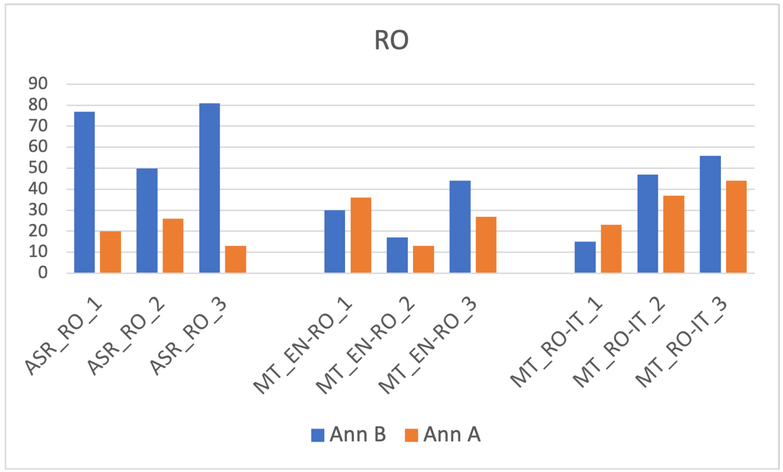


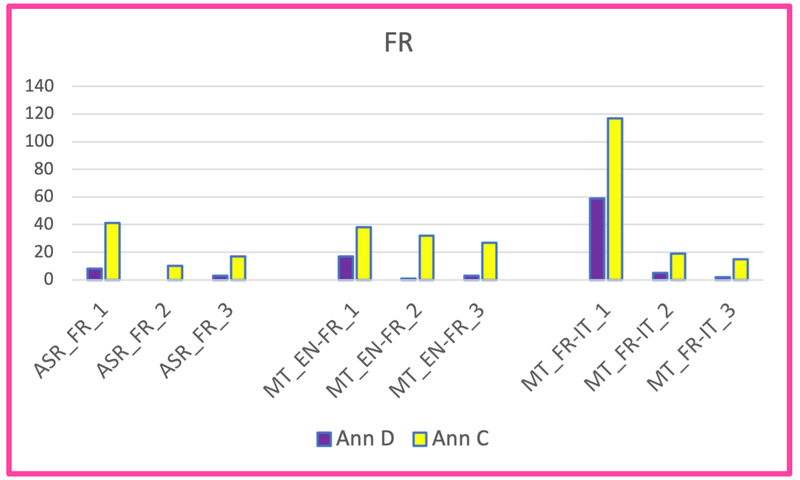


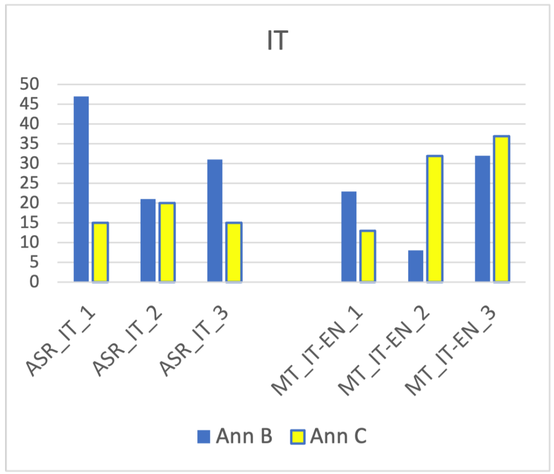


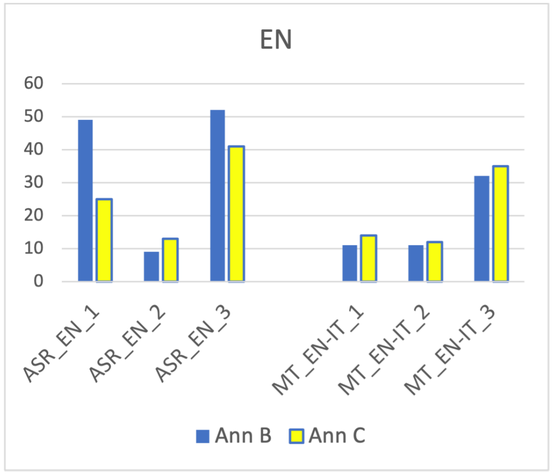


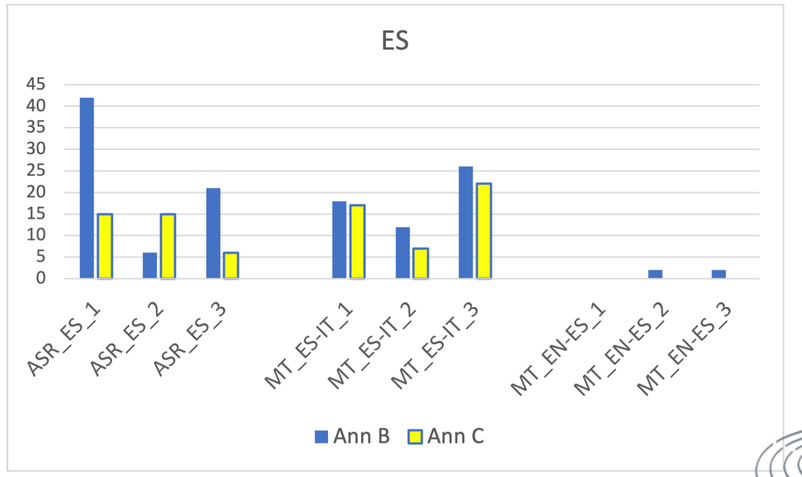
Same speeches, different annotators





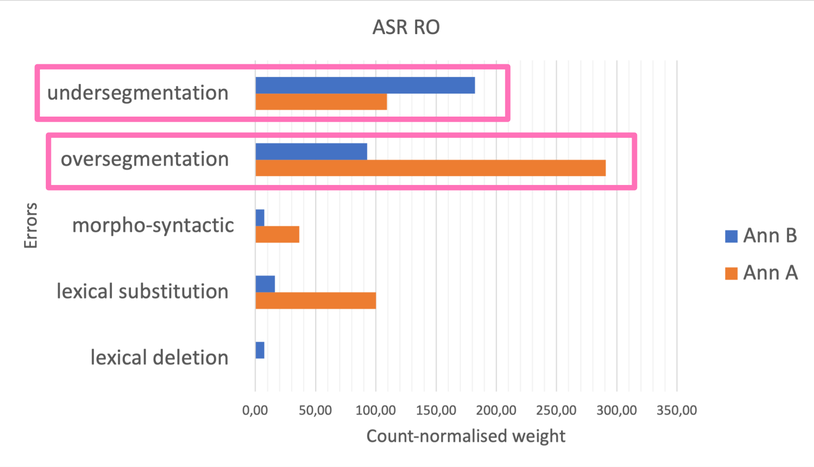


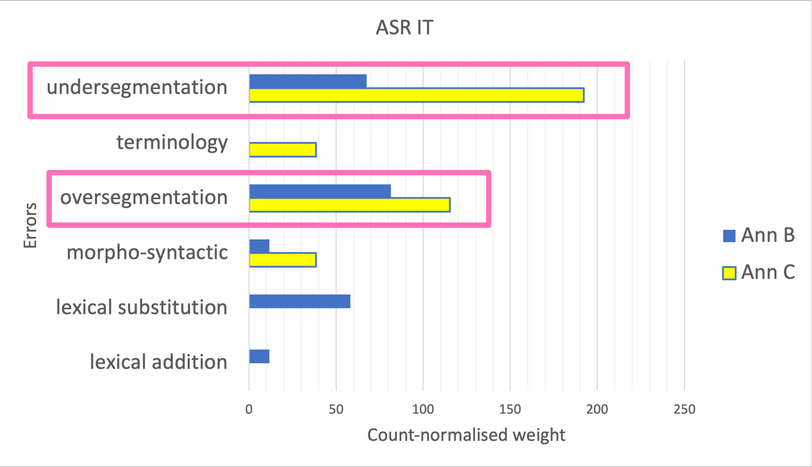


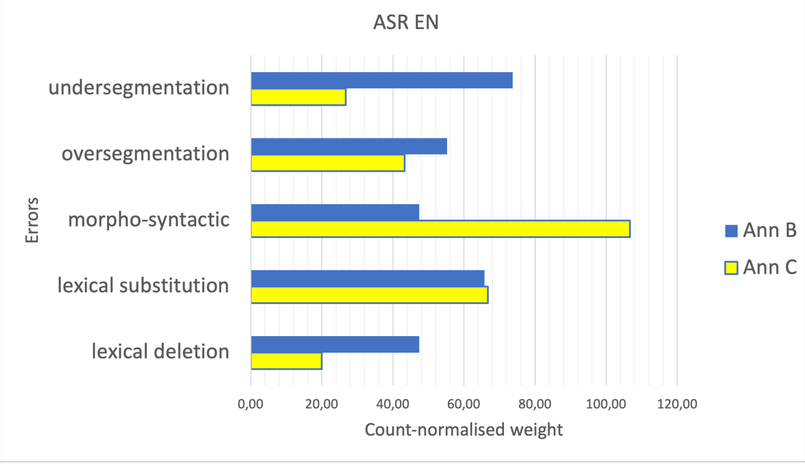


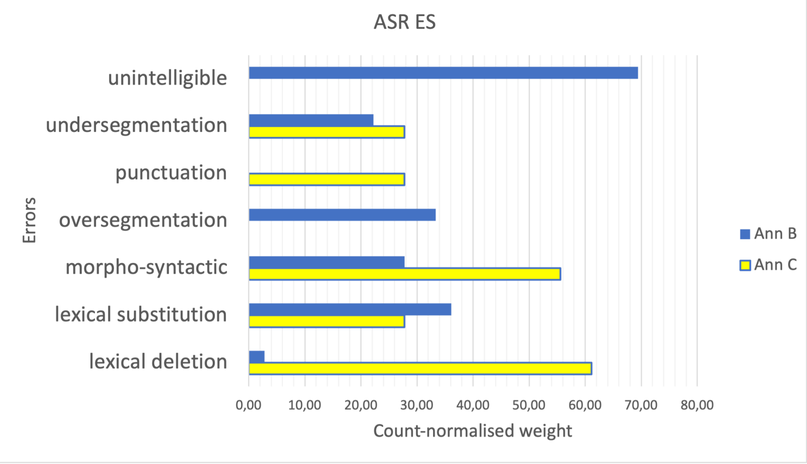
ASR error categories

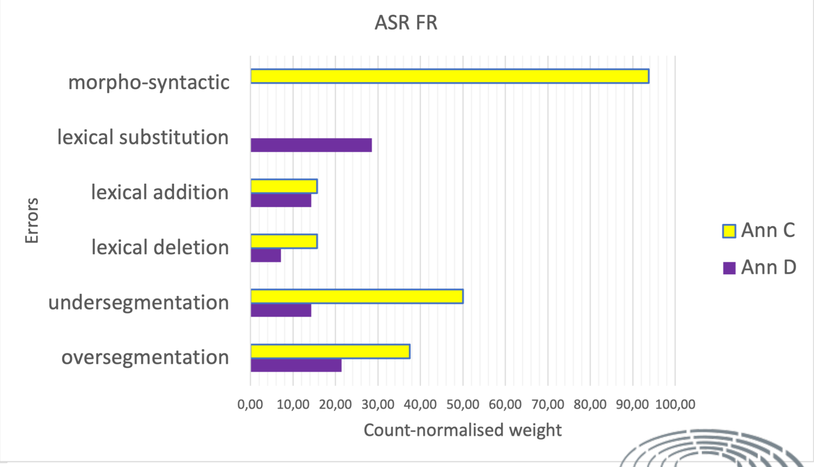
RO	On	2.77
IT	On	3.22
IT	Off	5.58
EN	On	8.98
ES	On	4.94
FR	On	8.91
DE	On	7.81





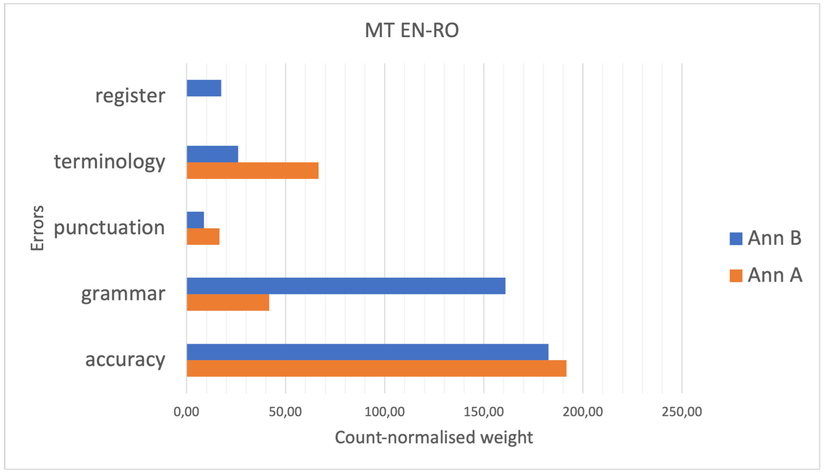


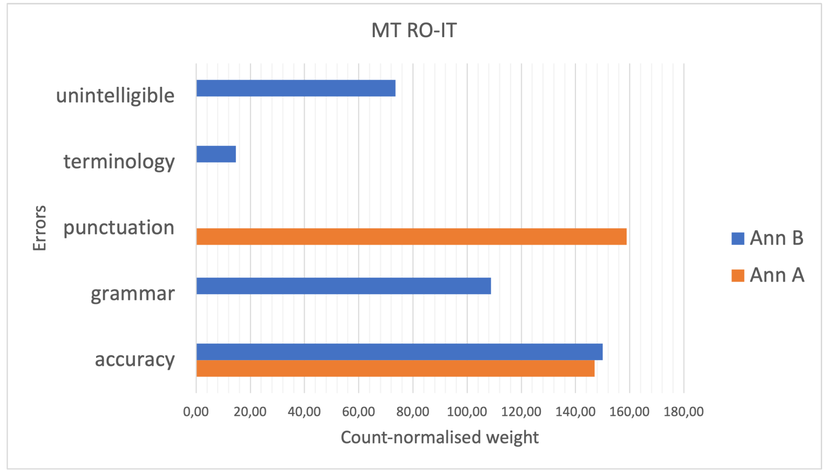




European Parliament

MT error categories



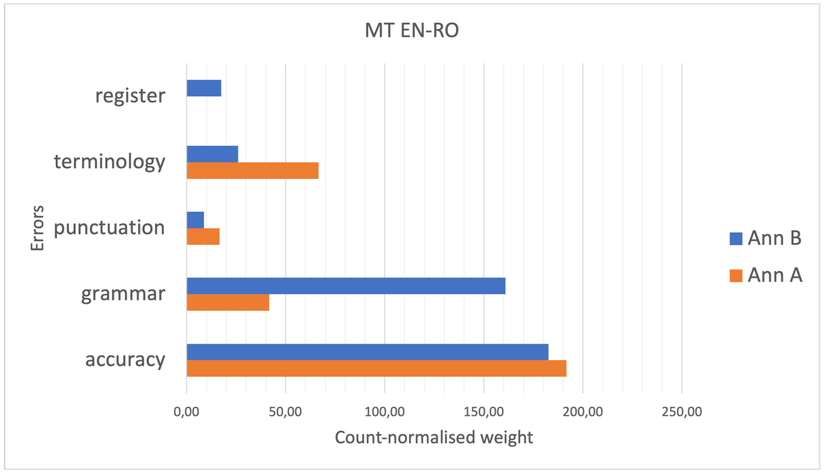


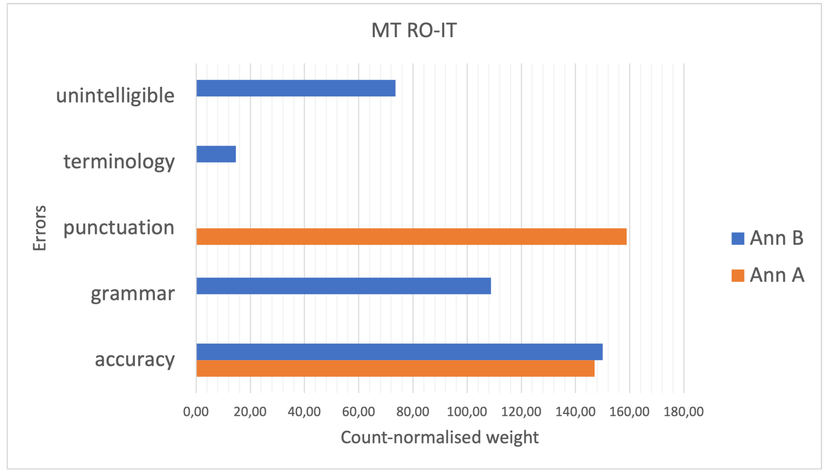
REF: [...] state precum Federația Rusă utilizează instrumentele moderne pentru **a ataca** state, pentru a ataca entități, pentru a pune în pericol democrația europeană**,** **acest** lucru necesită un răspuns **rapid** și unit.
*[...] countries like the Russian Federation use modern tools **to attack** states, to attack entities, to endanger European democracy**, this** requires a rapid and united response.*

ASR: State precum Federația Rusă utilizează instrumentele moderne pentru **a. Ataca** state pentru a ataca entități pentru a pune în pericol democrația europeană**.** **Acest** lucru necesită un răspuns**. Rapid** și unit, [...]

RO-IT: Paesi come la Federazione Russa usano strumenti moderni per **Attaccano** gli Stati per attaccare entità **che** mettono in pericolo la democrazia europea**.** **Ciò** richiede una risposta**. Veloce** e unito,

MT error categories



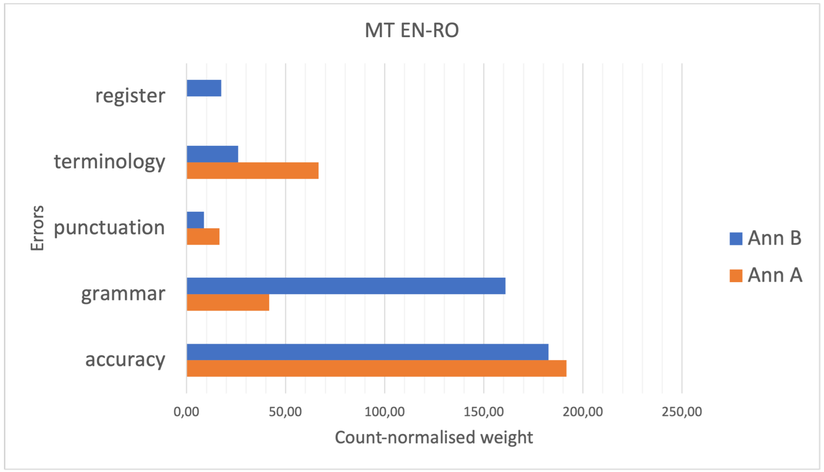


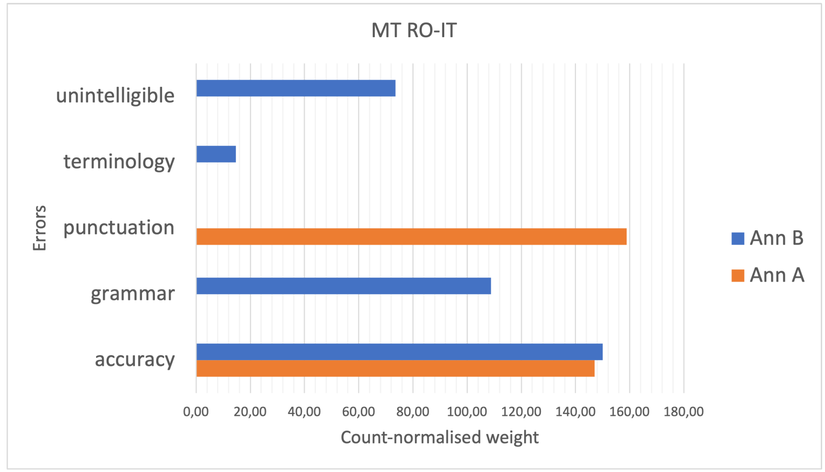
REF: [...] state precum Federația Rusă utilizează instrumentele moderne pentru **a ataca** state, pentru a ataca entități, pentru a pune în pericol democrația europeană**,** **acest** lucru necesită un răspuns **rapid** și unit.
*[...] countries like the Russian Federation use modern tools **to attack** states, to attack entities, to endanger European democracy**, this** requires a rapid and united response.*

ASR: State precum Federația Rusă utilizează instrumentele moderne pentru **a. Ataca** state pentru a ataca entități pentru a pune în pericol democrația europeană**.** **Acest** lucru necesită un răspuns**. Rapid** și unit, [...]

RO-IT: Paesi come la Federazione Russa usano strumenti moderni per **Attaccano** gli Stati per attaccare entità **che** mettono in pericolo la democrazia europea**.** **Ciò** richiede una risposta**. Veloce** e unito,

MT error categories



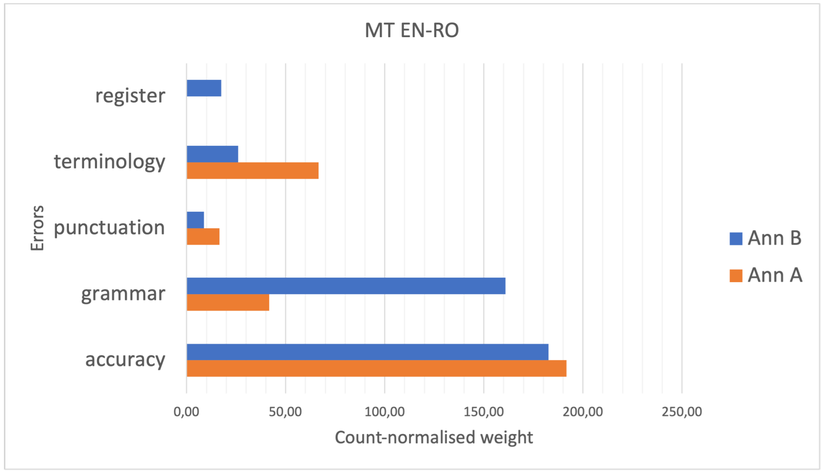


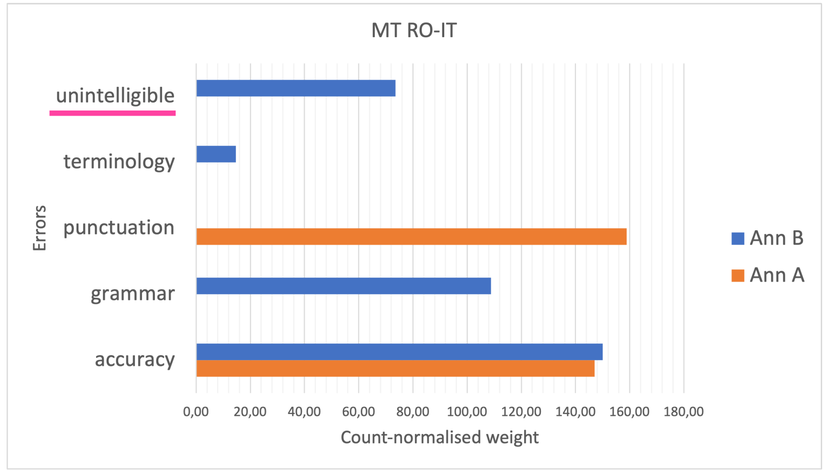
REF: [...] state precum Federația Rusă utilizează instrumentele moderne pentru **a ataca** state, pentru a ataca entități, pentru a pune în pericol democrația europeană**,** **acest** lucru necesită un răspuns **rapid** și unit.
*[...] countries like the Russian Federation use modern tools **to attack** states, to attack entities, to endanger European democracy**, this** requires a rapid and united response.*

ASR: State precum Federația Rusă utilizează instrumentele moderne pentru **a. Ataca** state pentru a ataca entități pentru a pune în pericol democrația europeană**.** **Acest** lucru necesită un răspuns**. Rapid** și unit, [...]

RO-IT: Paesi come la Federazione Russa usano strumenti moderni per **Attaccano** gli Stati per attaccare entità **che** mettono in pericolo la democrazia europea**.** **Ciò** richiede una risposta**. Veloce** e unito,

MT error categories



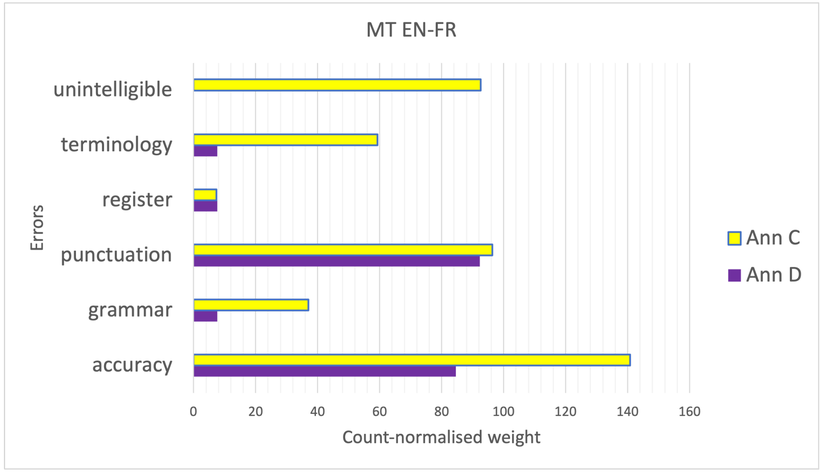


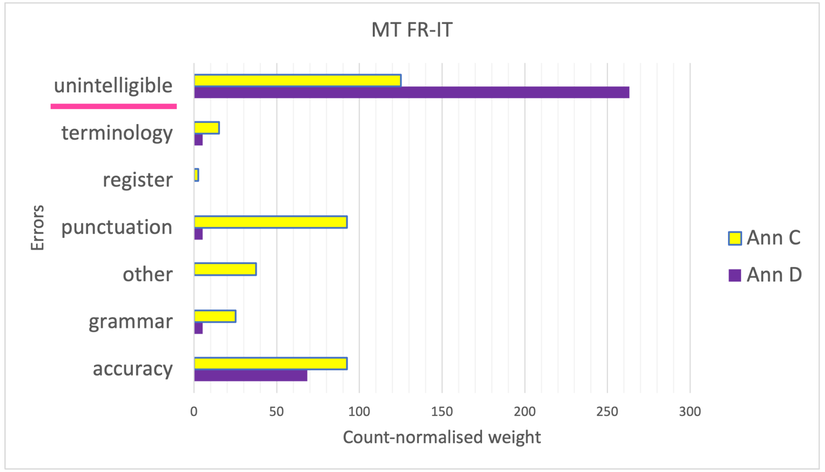
REF: [...] state precum Federația Rusă utilizează instrumentele moderne pentru **a ataca** state, pentru a ataca entități, pentru a pune în pericol democrația europeană**,** **acest** lucru necesită un răspuns **rapid** și unit.
*[...] countries like the Russian Federation use modern tools **to attack** states, to attack entities, to endanger European democracy**, this** requires a rapid and united response.*

ASR: State precum Federația Rusă utilizează instrumentele moderne pentru **a. Ataca** state pentru a ataca entități pentru a pune în pericol democrația europeană**.** **Acest** lucru necesită un răspuns**. Rapid** și unit, [...]

RO-IT: Paesi come la Federazione Russa usano strumenti moderni per **Attaccano** gli Stati per attaccare entità **che** mettono in pericolo la democrazia europea**.** **Ciò** richiede una risposta**. Veloce** e unito,

MT error categories





REF: Ceux qui ont harcelé et appelé au meurtre **sur Internet Samuel Paty, sont-ils**, étaient-ils, des vecteurs de liberté d'expression? Poser la question, c'est déjà y apporter une réponse.
Were those who harassed and called for the murder of Samuel Paty on the Internet vectors of freedom of expression? To ask the question is to answer it.

ASR: Ceux qui ont harcelé. **Su Internet. Internet. Jsem jej petic. Et** appelé au meurtre sur Internet, Samuel Paty**. Sont-ils**, étaient-ils des vecteurs de liberté d'expression. Poser la question c'est déjà y apporter une réponse.

FR-IT: Coloro che hanno molestato. **Su internet. Internet. Sono una petizione. E** ha chiesto omicidio su Internet, Samuel Paty**. Sono loro**, erano vettori della libertà di espressione. Fare la domanda è già fornire una risposta.
*Those who harassed. **On the Internet. Internet. They are a petition. And** called for murder on the internet, Samuel Paty**. It's them,** they were vectors of freedom of expression. To ask the question is already to provide an answer.*

- Human evaluation of ASR output can shed light on errors which are not considered in the commonly-used WER metric (i.e. sentence segmentation issues)
- Error annotation is a highly subjective task and subjectivity attains all categories, also those considered clear-cut categories (e.g. sentence segmentation)
- Annotator background has an influence on error severity perception and error identification, and should be investigated in detail
- Some annotators assigned multiple error categories to the same error, in the attempt to annotate also the consequences of the error
- Annotating errors in the transcription, our evaluators tended to rate the output as if it was a written text and not an oral text transposed into a written form
- A second round of annotations is necessary to obtain more reliable annotations
- Segment-level annotation could lead to higher severity error annotation than document-level error annotation

For the **official outline of the evaluation methodology used in the European Parliament**, please refer to the Call for Tender, available here:

https://etendering.ted.europa.eu/cft/cft-document.html?docId=58722

Thank you for your attention!

Elisa Di Nuovo, PhD
Directorate General for Translation
Directorate for Citizens' Language
Speech to Text Unit
elisa.dinuovo@europarl.europa.eu

European Parliament