8th Live Subtitling and Accessibility Symposium
Barcelona, 19 April 2023

Automatic speech recognition in Spain: the Basque and Catalan case

Ana Tamayo (University of the Basque Country, UPV/EHU)

Irene de Higes Andino (Universitat Jaume I, UJI)

UNIVERSITY OF LEEDS

Universidade Vigo

SWISS TXT

TRANS MEDIA CATALONIA

QuaLiSub

The Quality of Live Subtitling

- Led by Universidade de Vigo and funded by Spanish Ministry of Science and Innovation (ref. PID2020-117738RB-I00)
- To measure the quality of live subtitles in programmes and events in all four official languages in Spain (Catalan, Basque, Galician and Spanish) as well as human and automatic subtitling in English in the UK, the US and Canada.
- To explore the automation of the NER model

Universidad del País Vasco Euskal Herriko Unibertsitatea

In charge of Basque subtitles

UNIVERSITAT JAUME I

In charge of Catalan subtitles

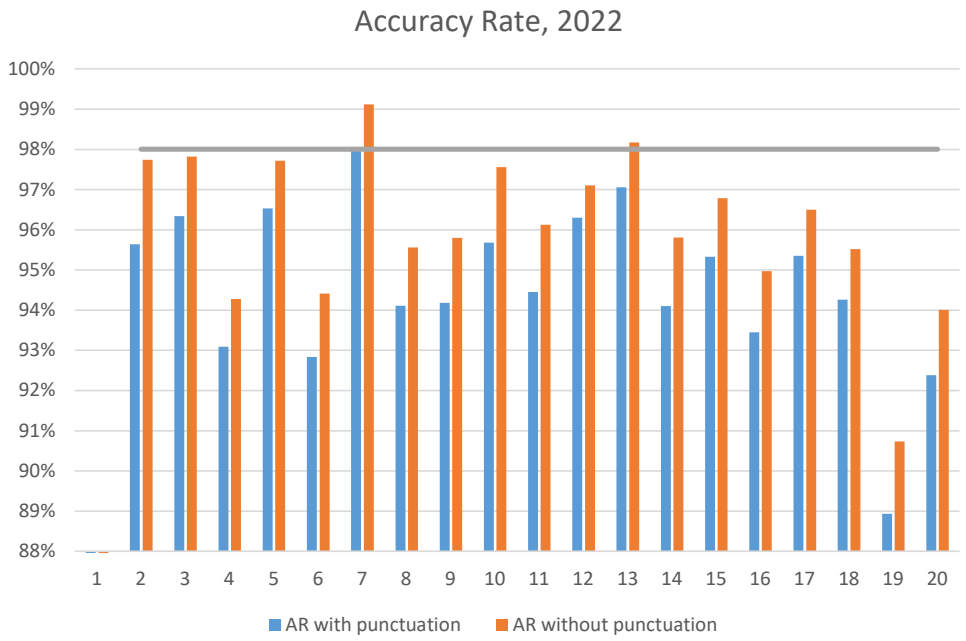
Basque, introduction

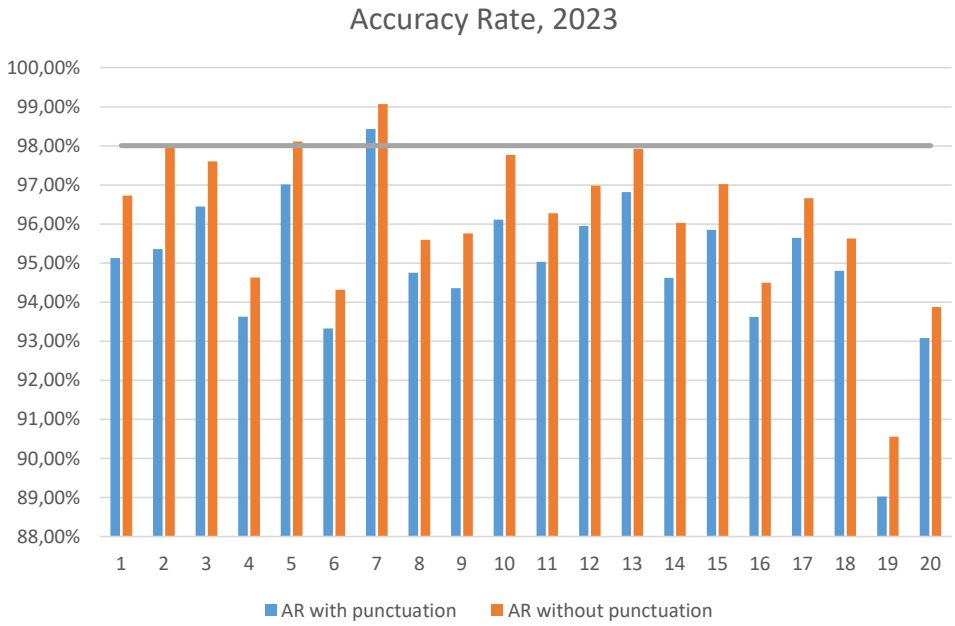
- Basque co-oficial status in 1978
- EiTB in 1982
  - In Basque: ETB1 and ETB3
  - In Spanish ETB2 and most of ETB4
- In Basque, only prerecorded subtitles using speech-to-text program (Idazle, by Vicomtech). Mostly with post edition (except for the weather forecast, withoug editing)

Basque, methodology

- 20 samples (5 minutes each)
- Recorded from ETB1 in May, 2022
- Automatic speech-to-text with ADITU (Elhuyar foundation) in two versions
  - June 2022 (one sample not recognised by the program)
  - March 2023
- 198 minutes analysed 3708 subtitles analysed
- Analysed using the NER model

Basque, accuracy rate

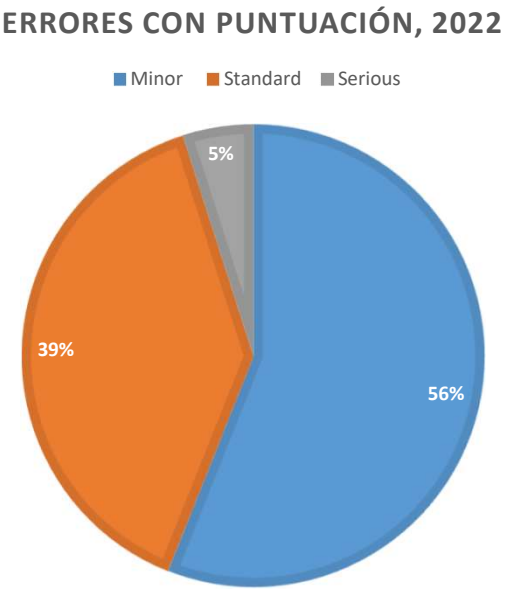


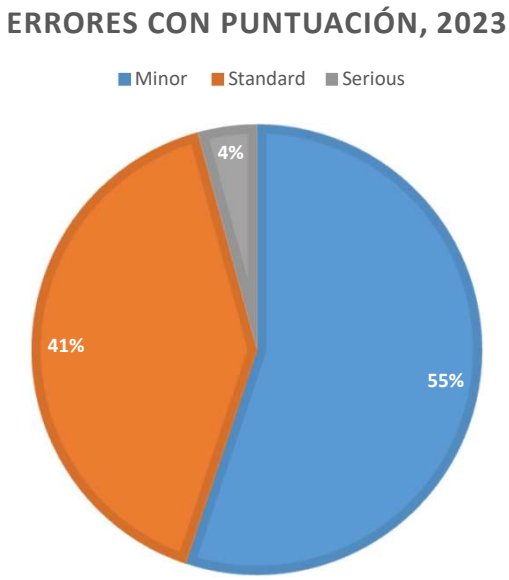


Average: 94.63% and 96.09%

Average: 94.95% and 96.15%

Basque, errors with punctuation

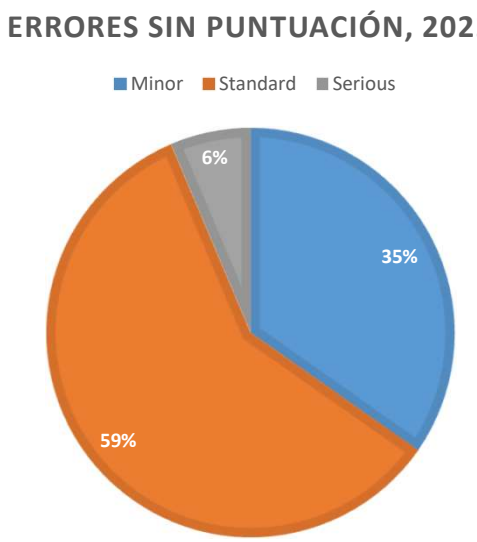


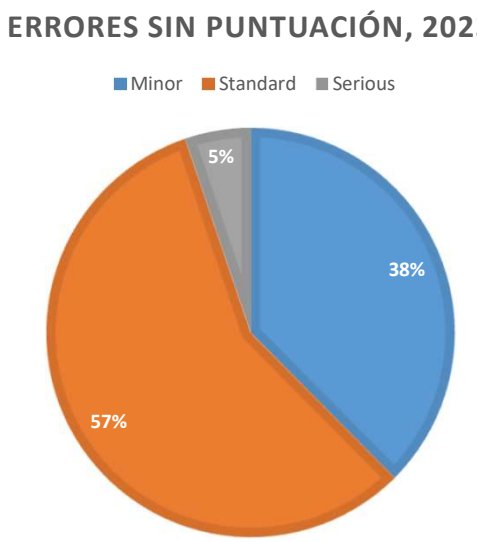


Total errors: 1967
Errors/min: 20.28

Total errors: 1975
Errors/min: 19.55

Basque, errors without punctuation

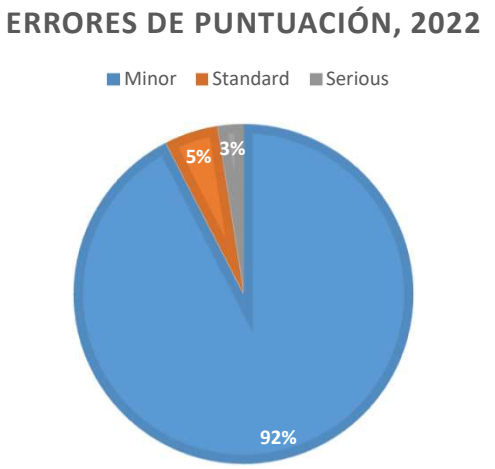


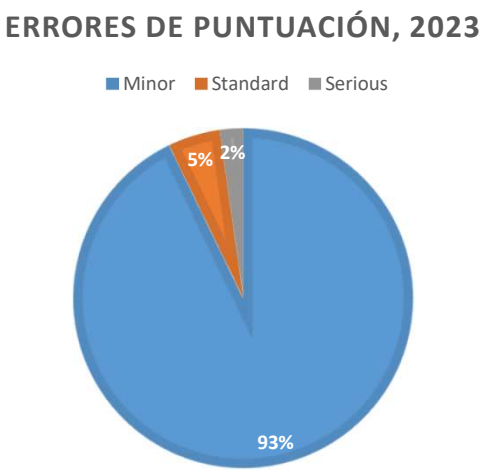


Total: 1241

Total: 1347

Basque, punctuation errors

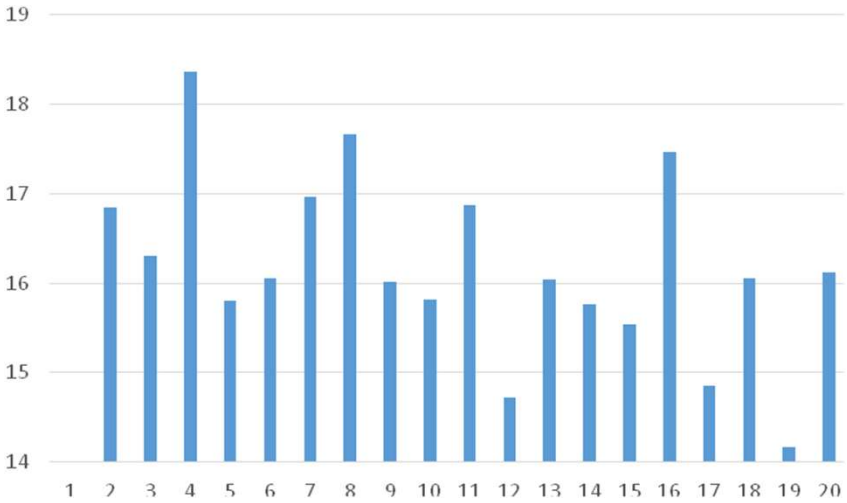




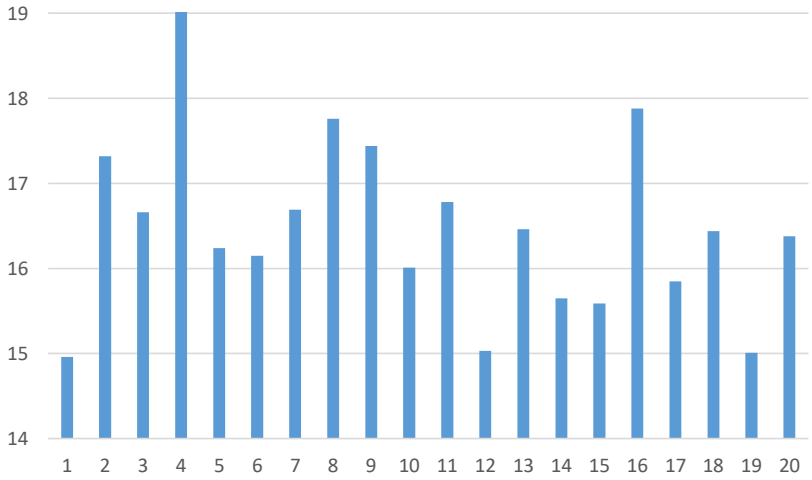
Total: 726, 36.91%

Total: 628, 31.80%

Basque, CPS



Average: 16.18 CPS



Average: 16.5 CPS

Basque, factors that influence AR

- Batua (standardized language) or the use of euskalkiak (dialects)
- Registry
- Diction of speakers
- Prefabricated discourse or spontaneity
- Number of speakers
- Speakers being anchors, reporters or informants
- Background noise and language interference in the audio
- Proper nouns

Basque, future research

- Analyse Idazle, Speechmatics and Wishper
- Analyse its implementation on real live subtitling
- Analyse it as a tool for recorded broadcast (actual current use)

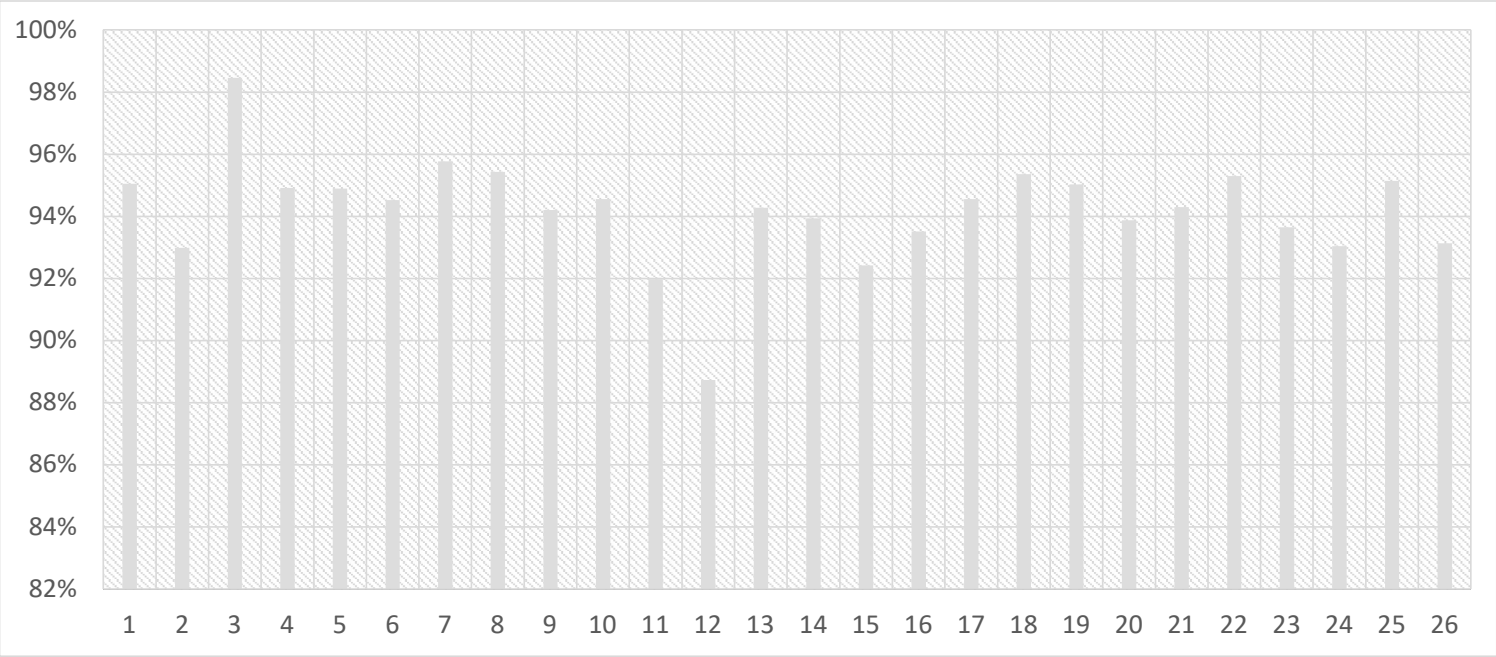
Catalan, introduction

- Catalan: three linguistic standards
  - Co-official status in Catalonia (since 1979), Region of Valencia (since 1982) and Balearic Islands (since 1983)
- RTVE regional news bulletins (*L'informatiu – Comunitat Valenciana*)
  - Bilingual speech
    - Catalan standard for Region of Valencia used by presenters, some reporters and interviewees
    - Spanish used by some reporters and interviewees
  - Broadcast with live automatic subtitles since February 2021 (without post-editing)

Catalan, methodology

- 26 samples (5 minutes each)
- Broadcast in La1 from April to July, 2021
- Automatic subtitles by Aicox with tecnology from Etiqmedia
- Analysis with NER model:
  - 130 minutes
  - 2116 subtitles

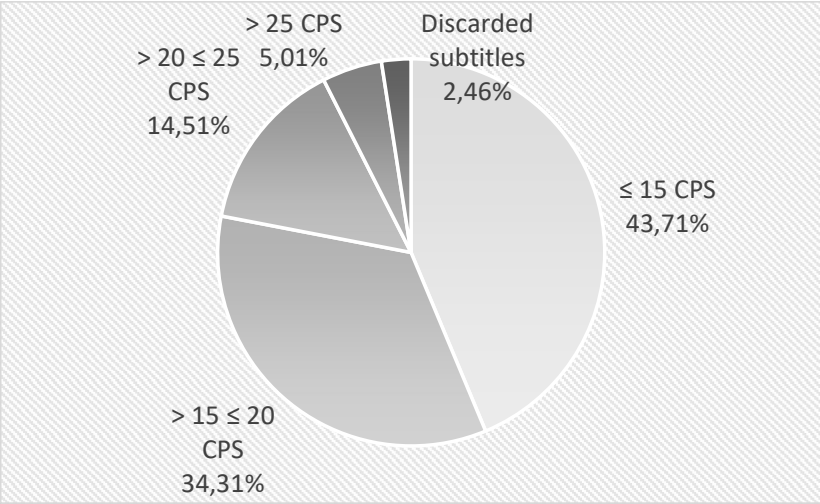
Bilingual subtitles, accuracy rate



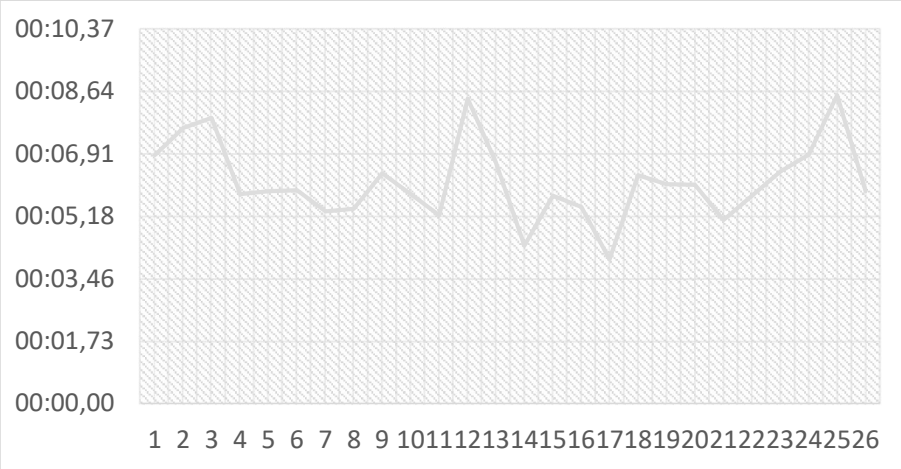
Average: 94.19%

Bilingual subtitles, CPS and delay

CPS



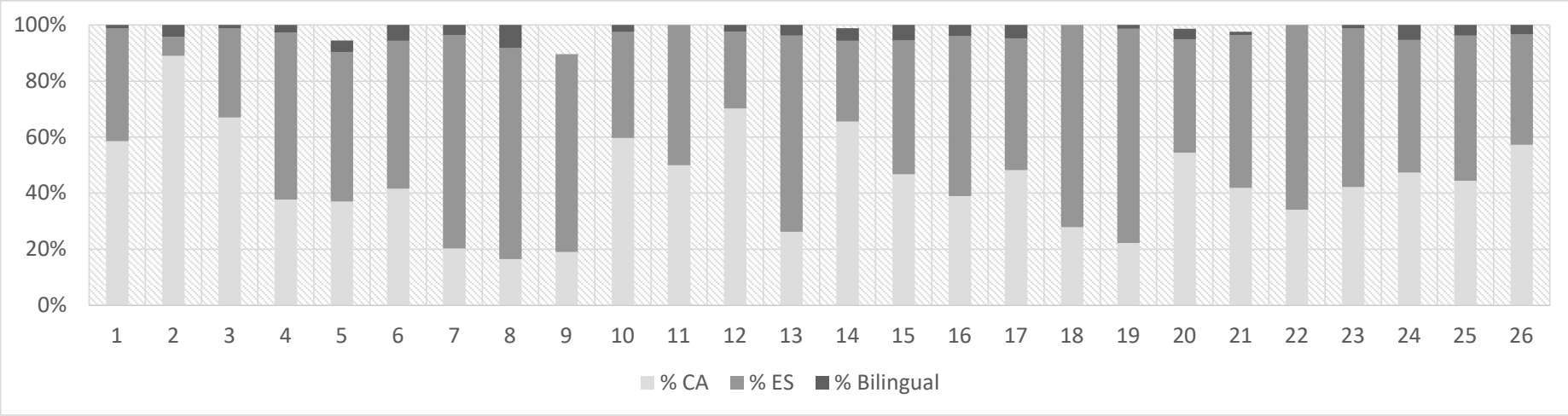
Delay



Average: 17.68 CPS

Average: 00:06,14

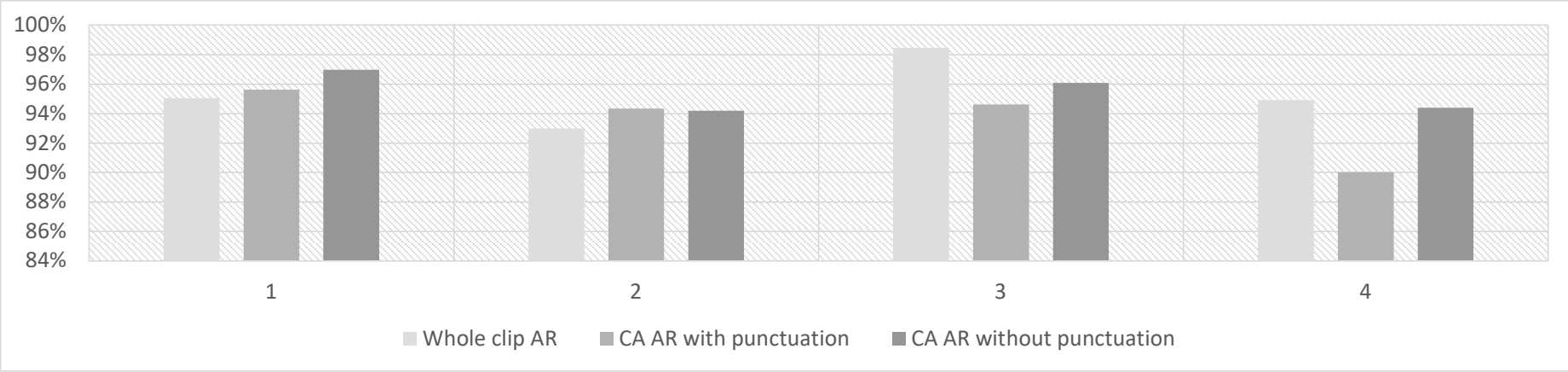
Catalan, language distribution



Average:

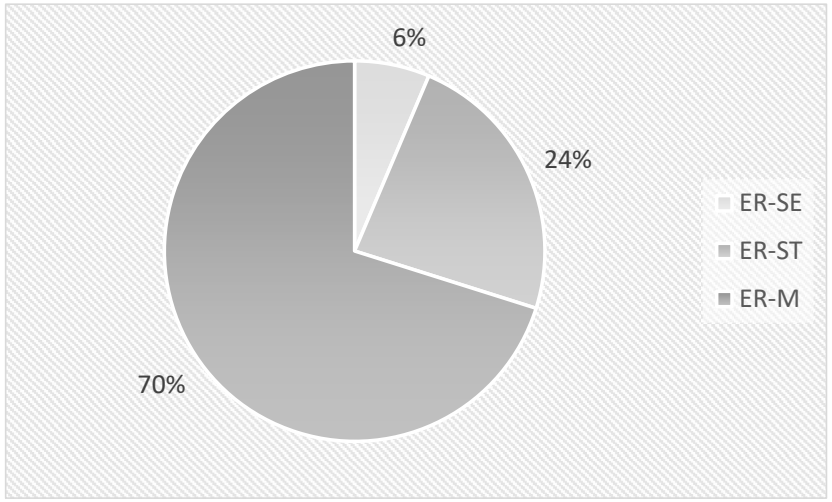
Bilingual: 3.02%
Spanish: 51.32%
Catalan: 45%

Catalan, only Catalan subtitles

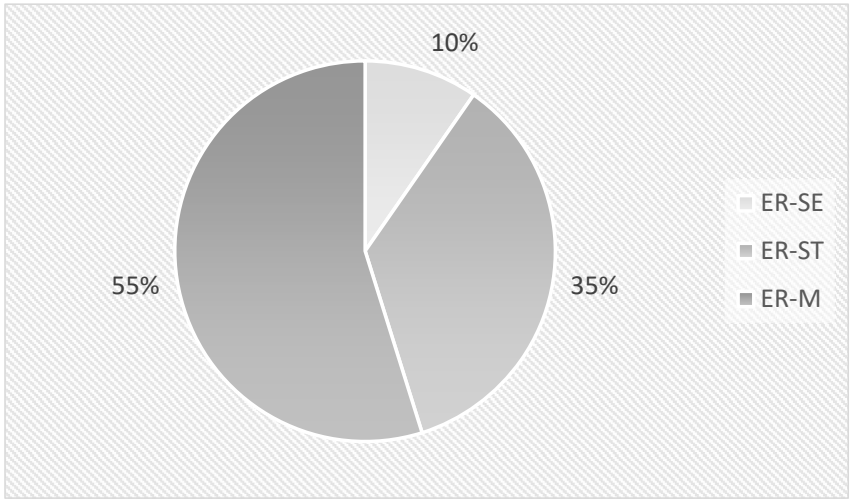


Only Catalan, types of errors

Errors with punctuation



Errors without punctuation



Total errors: 345
Errors/min: 17.25

Catalan, factors that influence AR

- Non-standard language, barbarisms and mixture of standards
- Registry
- Diction of speakers (the use of masks was still compulsory or recommended in 2021)
- Prefabricated discourse or spontaneity
- Number of speakers
- Speakers being anchors, reporters or informants
- Background noise and language interference in the audio
- Proper nouns

Future research (Catalan)

- Calculate AR for ES, CA and bilingual subtitles (26 clips)
- Calculate AR excluding subtitles within a window of 5 seconds after language change
- Compare subtitles broadcast in RTVE with subtitles generated by MLLP platform (Universitat Politècnica de València)
- Qualitative analysis: What affects lower AR in Catalan language?

8th Live Subtitling and Accessibility Symposium
Barcelona, 19 April 2023

Universidad del País Vasco
Euskal Herriko Unibertsitatea

UNIVERSITAT JAUME I

Thank you!

Ana Tamayo ana.tamayo@ehu.eus

Irene de Higes Andino dehiges@uji.es

UNIVERSITY OF LEEDS
Universidad de Vigo
SWISS TXT
TRANS MEDIA CATALONIA