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

Automatic Live Subtitling in Spain:
The Quality of Bilingual TV Newscasts in Galicia

María Rico-Vázquez

Universidade de Vigo - GALMA

QuaLiSub

QuaLiSub
The Quality of Live Subtitling

https://qualisub.webs.uvigo.es/

- Measure the quality of live subtitles in TV programs and events
  - Spain: Catalan, Basque, Galician and Spanish
  - UK, US, Canada: English
- Compare the quality of respoken vs. automatic subtitles (NER, WER)
- Develop a technological solution to automatize the NER model
- Issue recommendations for improving access

The study

- In collaboration with RTVE (the largest Spanish state-owned public media corporation)
  - *General Law of Audiovisual Communication*: TVE is required to provide subtitles for 90% of their TV programs → automatic subtitling since 2019
- Comprehensive quality assessment of machine-generated subtitling
- Galician territorial newscasts
- April-August 2021 (16 weeks)

Materials

- 275 minutes of audiovisual material

  Fifty-five 5-min samples

  - Accuracy
  - Speed
  - Delay
- Materials provided by TVE
- Original dialogues: Galician + Spanish
  - ≈5 sec for the ASR to detect change of language

The analysis: accuracy (I)

- WER (Word Error Rate)
  - Accuracy rate
  - WER rate

N = total number of spoken words

Errors = deletion (D), substitution (S), insertion (I)

$$\text{Accuracy rate} = \frac{N - \text{Errors}}{N} \times 100 = \%$$

$$\text{Accuracy rate} = \frac{N - D - S - I}{N} \times 100 = 16\%$$

The analysis: accuracy (II)

- Alphanumeric code
  - Speed up qualitative analysis
  - Facilitate quantitative analysis

TYPE OF ERROR	CODE
Incorrect punctuation*	**PU**
Incorrect word	**PI**
Repeated word	**PR**
Missing word	**PP**
Added word	**PA**
Incorrect number identification	**IN**
Change of language	**CL**
Change of speaker	**-**
Song	**CA**
Background conversation	**C2**

The analysis: accuracy (III)

- Alphanumeric code
  - Speed up qualitative analysis
  - Facilitate quantitative analysis

TYPE OF ERROR	CODE
Incorrect punctuation*	**PU**
Incorrect word	**PI**
Repeated word	**PR**
Missing word	**PP**
Added word	**PA**
Incorrect number identification	**IN**

*Punctuation errors

- Included in the accuracy rate
- Not included in the WER rate

$$\frac{PI + PR + PP + PA + IN}{\text{Total Words}}$$

The analysis: accuracy (IV)

Universidade Vigo GALMA

TABLA DE ANÁLISIS

N.º	Subtítulo	Transcripción	Error
1	meta cinco.	vinte ~~meta~~ e cinco~~.~~	IN (vinte) PP (e) PU (.MAY/ min)
2	Menos helenas de entre cinco y Océanos de ida deban ser hacinados	nenos ~~Menos~~ e nenas ~~helenas~~ de entre cinco e ~~y~~ doce anos ~~Océanos~~ de idade van ~~ida deban~~ ser vacinados ~~hacinados~~	PI (nenos) PI (e) PI (nenas) PI (e) PU (MAY/min) IN (doce) PI (anos) PI (idade) PI (van) PI (vacinados)

The analysis: speed

- Sample speed (average speed of all subtitles)
- Minimum speed (lowest speed observed in a subtitle)
- Maximum speed (highest speed observed in a subtitle)

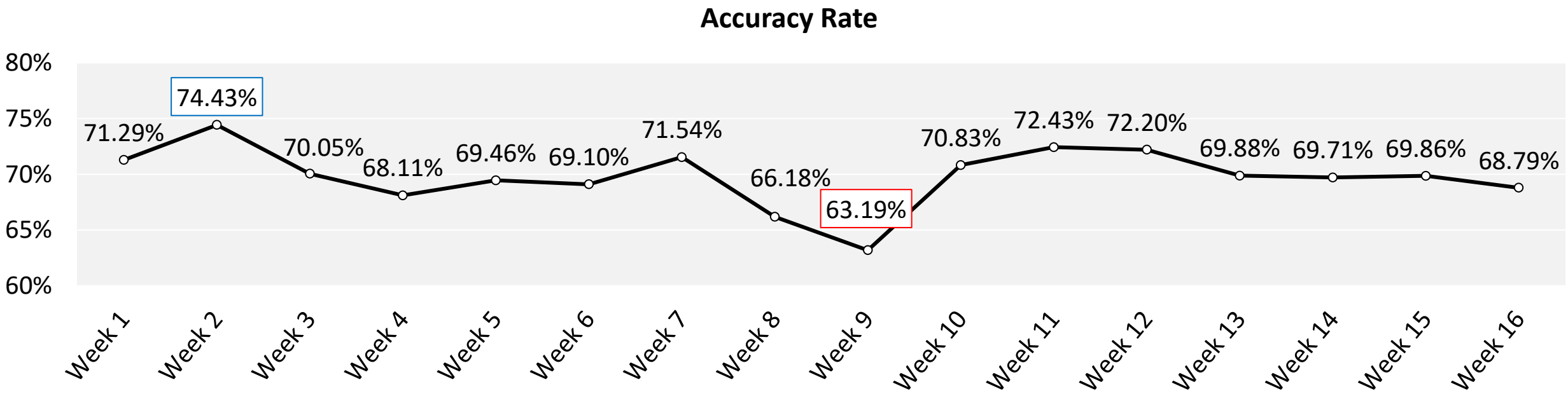
The analysis: delay

- Delay calculated in accordance with the **Annex C** of the **UNE-153010-2012 Standard**

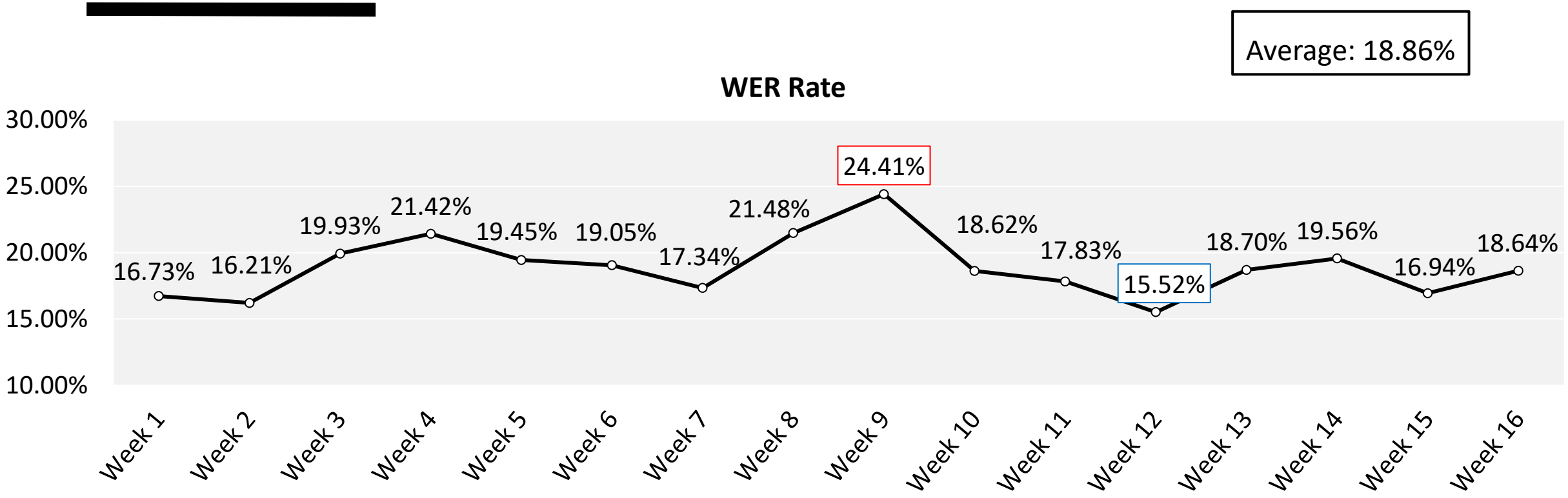
To measure the delay of the subtitles of a specific program, it is convenient to analyze a random and representative sample of all the subtitles of said program. (...) The sample is divided into two groups of different sizes. The first group of subtitles will be used to measure the delay in the appearance of the first word of the subtitles ***(start delay)****, while the second group will serve to measure the delay in the appearance of the last word* ***(end delay)****. The ratio between the two groups of subtitles in the sample is 2 to 1. Out of every 3 subtitles chosen in the sample,* ***2 will be used to measure the start delay and 1 will be used to measure the end delay.*** *(...) The final delay result is obtained by taking an* ***average*** *between the two partial delays.* (AENOR, 2012:28)

Results: accuracy (I)

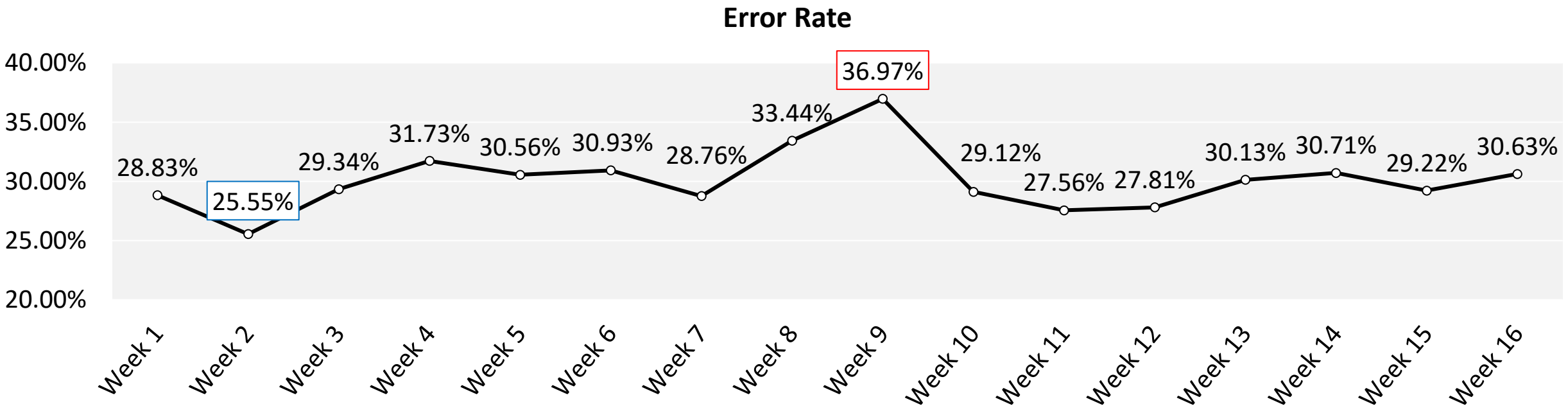
Average: 69.88%



Results: accuracy (II)



Results: accuracy (III)

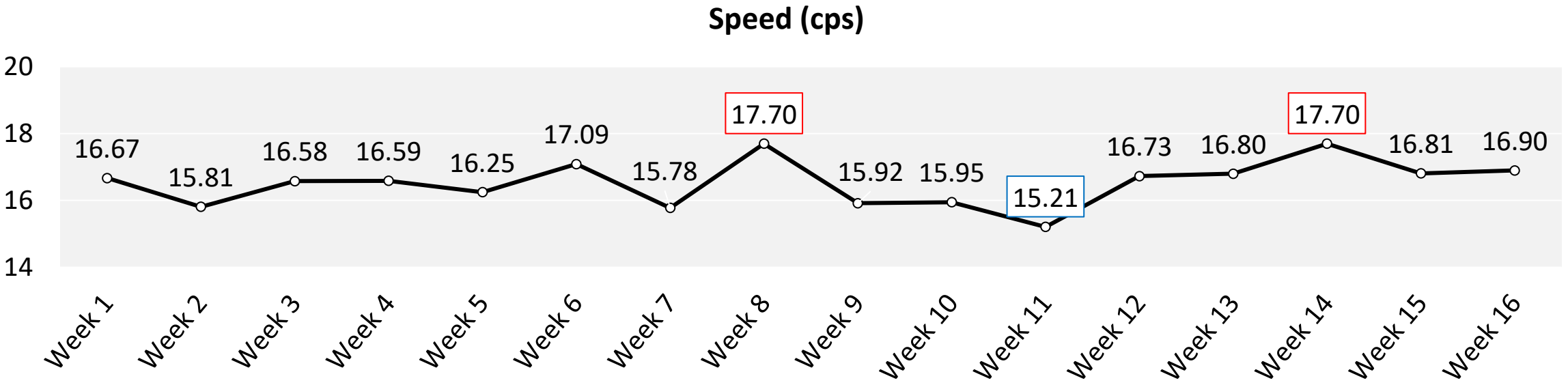


Results: accuracy (IV)

- Errors:
  - Numbers
  - COVID argot (Janssen, AstraZeneca, Pfizer, PCR…)
  - Foreign references (James Joyce, Baudelaire, Reverdie, Nishikawa, Jack Sparrow…)
  - Proper names
  - Acronyms, initialisms, abbreviations

Results: speed (I)

Average: 16.54 cps



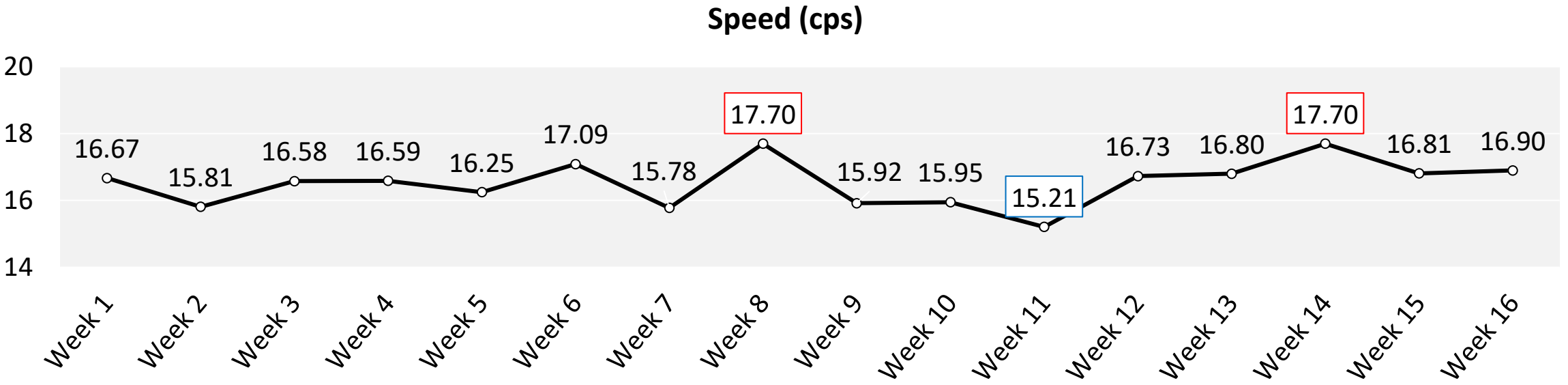
Results: speed (II)

>15 cps: 57%

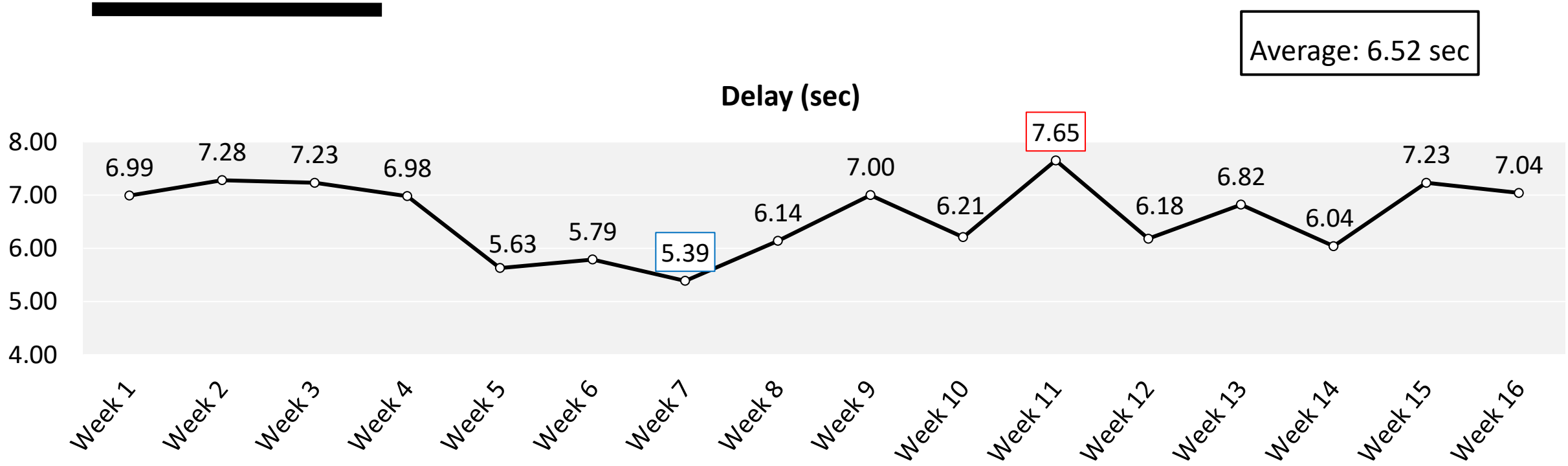
>18 cps: 33%

>20 cps: 19%

>25 cps: 3%



Results: delay



Follow-up reports (I)

- Examine the software's performance and gradually enhance its functioning based on the data collected

↓

Reports

(weekly, monthly)

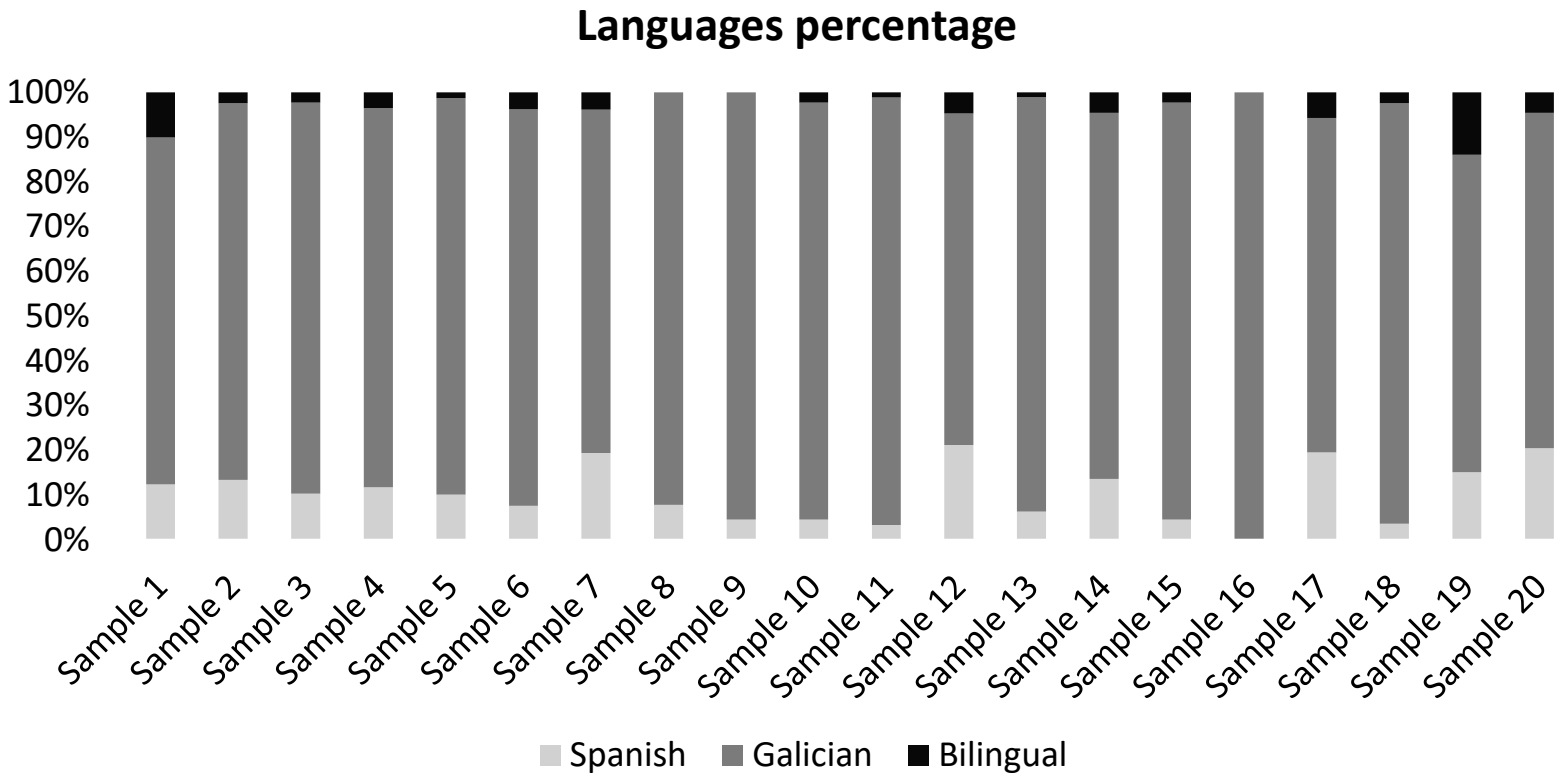
Follow-up reports (II)

+ RECURRING ERRORS

- Misrecognition of words
- Word omission
- Introduction of terms inexistent in the original audio
- Misrecognition of place names and proper names
- Misrecognition of numbers
- No capitalization after a period
- Incorrect capitalization
- Incorrect use of periods and commas
- Absence of periods and commas
- Incorrect question marks
- Incorrect accentuation
- Absence of speaker ID

More results: WER vs. NER (I)

- 20 samples: WER + NER
- High percentage of Galician



More results: WER vs. NER (II)

- 20 samples: WER + NER
- High percentage of Galician
- Preliminary results
- Very low accuracy
  - No sample reaches the minimum NER quality threshold (98%)

Sample	Accuracy WER (%)	Accuracy NER (%)	WER rate (%) (w/o punctuation)	NER rate (%) (w/o punctuation)
Sample 1	71.74	**93.17**	18.79	**94.89**
Sample 2	77.29	95.45	11.65	97.12
Sample 3	64.23	93.76	20.86	96.15
Sample 4	71.93	95.22	15.63	96.79
Sample 5	69.56	95.27	19.63	96.90
Sample 6	74.86	94.29	17.64	96.11
Sample 7	71.43	95.05	17.46	96.83
Sample 8	78.11	95.81	13.38	97.05
Sample 9	78.21	**95.92**	12.97	97.04
Sample 10	64.87	94.72	23.37	96.06
Sample 11	**81.95**	95.53	**10.01**	**97.29**
Sample 12	68.83	94.53	22.22	96.10
Sample 13	75.64	95.21	15.69	96.21
Sample 14	**58.97**	93.38	**28.36**	95.32
Sample 15	69.47	93.38	21.15	95.37
Sample 16	70.85	93.45	19.87	95.40
Sample 17	70.83	93.31	19.03	95.81
Sample 18	60.15	94.69	27.16	96.43
Sample 19	69.24	94.23	19.87	96.80
Sample 20	77.37	95.46	13.03	97.21
AVERAGE	**71.28**	**94.59**	**18.39**	**96.34**

Some conclusions (I)

- Over 16 weeks, some improvement of the ASR was observed, albeit not regularly nor consistently
  - Specific terms (COVID pandemic) still generated errors in the last weeks of the study
  - Not progressive or linear decrease in the error rate
- Encouraging evidence regarding subtitling speed and delay
  - Isolated samples with favorable results
  - Delay still high for automatic subtitles
  - Limiting the speed rate of all subtitles to 15 cps = increase in latency
  - Explore alternative combinations of speed and delay limit values for automatic live subtitles (antenna delay)

Some conclusions (II)

- The recognizer performs quite well within its capabilities
  - Highly controlled conditions are needed for good results
  - Quality of the subtitles for the main presenter vs. any other speaker (accent, noise, diction; masks)
- ASR may improve the original
  - Correct words, omit shuttering = clearer and cleaner subtitles
- No speaker identification
  - Problematic for some viewers
  - Advisable to incorporate in order to increase accessibility

Some conclusions (III)

- Coexistence of Galician and Spanish
  - First attempt at bilingual software
  - Change of language (Galician>Spanish>Galician): problematic for the ASR
  - Galician: minoritized language, highly influenced by Spanish
  - Galician speakers incorporate non-idiomatic words and Castilianisms, which “confuse” the recognizer
    - Incorporate the most frequent ones into the Galician dictionary used by the SR?
    - Dilemma: errors to be corrected / speech features to be preserved in the subtitles?

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

Thank you!

María Rico-Vázquez

Universidade de Vigo - GALMA