ILSA Interlingual Live Subtitling for Access

Co-funded by the Erasmus+ Programme of the European Union

IO2 REPORT

Development of speech-recognition software and competence analysis for ILS

Grant Agreement: 2017-1-ES01-KA203-037948
Funding Scheme: Erasmus+, KA2: Cooperation for innovation and the exchange of good practices
KA203: Strategic Partnerships for Higher Education
Project Duration: 01/09/2017 – 31/08/2020 (36 months)
Project Coordinator: Universidade de Vigo (UVIGO)
Partners: Universiteit Antwerpen (UANTWERPEN)
Uniwersytet Warszawski (UW)
Universitat Wien (UNIVIE)
De Vlaamse Radio en Televisieomroeporganisatie Nv (VRT)
Intro Pr Monika Szczygielska (INTRO)

Document title:
IO2 Report: O2 – Development of speech-recognition software and competence analysis for ILS

Authors:
Pablo Romero-Fresco (UVIGO)
Isabelle Robert (UANTWERP)
Franz Pochhacker (UNIVIE)
Lukasz Dutka (UW)
Monika Szcygielska (INTRO PR)
Marlies Decuyper (VRT)

Version: 1

Abstract:
This document reports on the completion of Intellectual Output 2 of the ILSA (Interlingual Live Subtitling for Access) project (2017-1-ES01-KA203-037948), devoted to identifying the professional skills from subtitling and interpreting required to perform interlingual live subtitling (ILS) and to the development of a new speech recognition software in Galician. This IO includes the largest experiment conducted so far on ILS, with a pilot study and three 4-week experiments analysing the performance of interpreters and subtitlers in this new discipline, along with targeted focus groups. The results have already been presented at nine international conferences, as well as in Multiplier Event 3 in Vienna, and have been accepted for publication at The Interpreter and Translator Trainer, a leading peer-reviewed journal on translation and interpreting. The results of this IO have informed the skills map developed in IO3 and the interlingual live subtitling (ILS) course planned in IO4.

Funding Scheme: Erasmus+, KA2: Cooperation for innovation and the exchange of good practices
KA203: Strategic Partnerships for Higher Education

Dissemination Level:

P	Public	X
C	Confidential, only for members of the consortium and the Commission Services	

Copyright and disclaimer:
This document is proprietary of the ILSA consortium members, and no copying or distributing, in any form of by any means, is allowed without the written agreement of the owner of the property rights. This project has been funded with support from the European Commission. This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

The following 4 sections include information (as presented in several international conferences) about the experiments conducted in the ILSA project to identify, for the first time, the professional skills from subtitling and interpreting required to perform interlingual live subtitling (ILS).

1. Aims

Overall aim for ILSA: To design, develop, test and validate the first training course for ILS and provide a protocol for this discipline for TV, the classroom and parliament.

Aim of the main experiment:

To train and test participants in an ILS course to answer the following questions:

- Is ILS feasible?
- Who is better suited?
- What are the main challenges?

2. Description of the trials

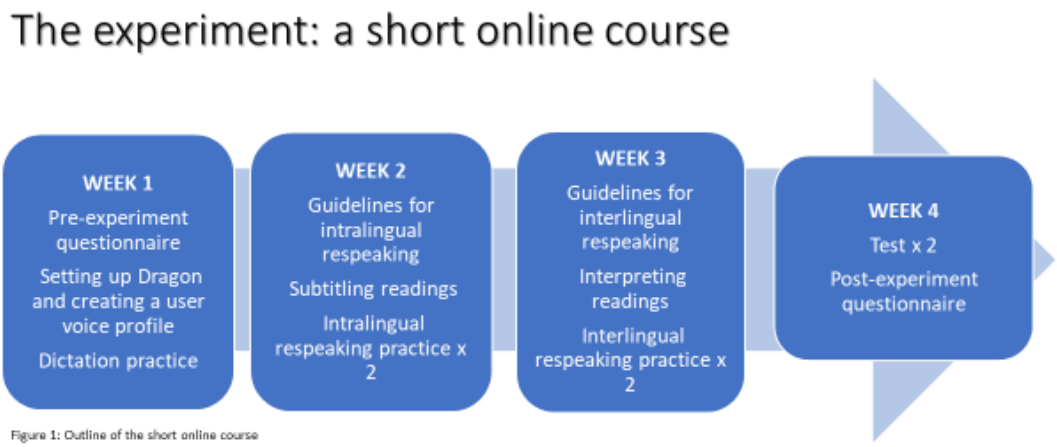


Figure 1: Outline of the short online course

Breakdown of professional profiles

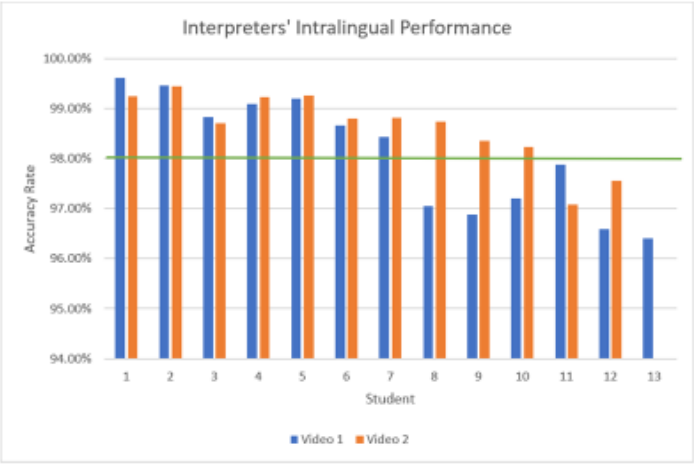
46 participants with the following profiles:

- 22% of students had a **clear-cut subtitling** profile.
- 28% of students had a **clear-cut interpreting** profile.
- The remaining students had a mixed background of subtitling and interpreting (46%), or no experience whatsoever of subtitling and interpreting (4%).
- Some students (12%) had previous experience of intralingual respeaking.

Videos

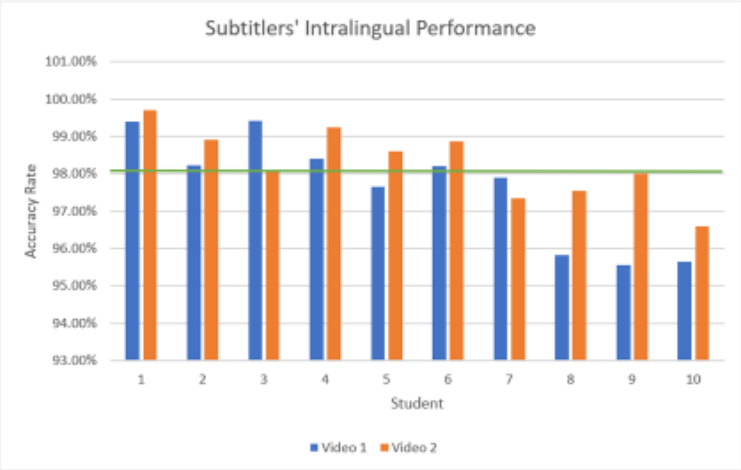
Week of Course	Intra/interlingual	Title	Genre	Duration	wpm
2	Intralingual	La vida en el arrecife	Documentary	00:05:48	76
2	Intralingual	DELE	Online Class	00:05:16	110
3	Interlingual	Beer	Talking Head	00:05:00	145
3	Interlingual	Médicos sin Fronteras	Interview	00:05:00	125
4	Interlingual	Emma Watson	Speech	00:05:21	107
4	Interlingual	Gardening	Talking Head	00:05:00	159

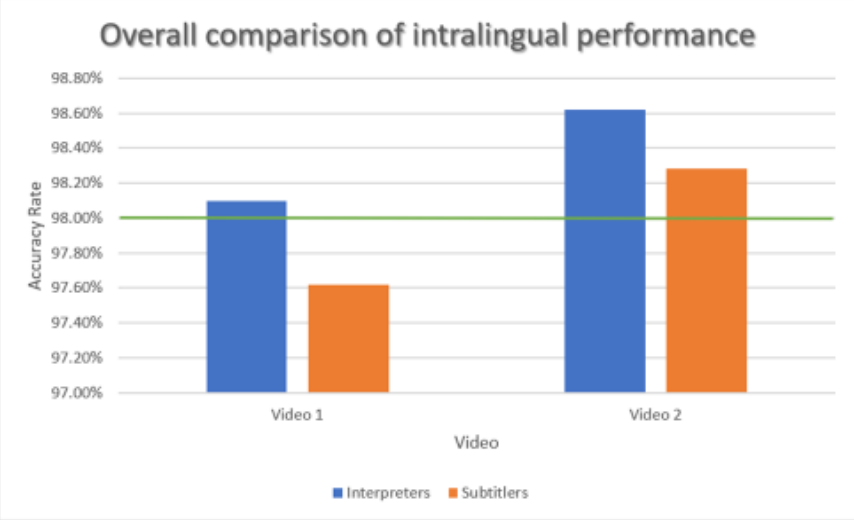
3. Results



- Average accuracy rate 98.10% in video 1 and 98.62% in video 2 – 98.36% overall.
- 100% of 'good performers' and 16% of 'poor performers' reached 98% in video 1.
- 100% of 'good performers' and 50% of 'poor performers' reached 98% in video 2.
- Edition and recognition errors are balanced.

- Average accuracy rate is 97.62% in video 1 and 98.28% in video 2 – 97.95% overall.
- 50% of subtitlers reached 98% in video 1 and 70% reached 98% in video 2.
- There are some very low accuracy rates of around 95%, which we did not see with the interpreters.





ILSA Interlingual Live Subtitling for Access

Co-funded by the Erasmus+ Programme of the European Union

Interpreters' and subtitlers'

interlingual respeaking performance

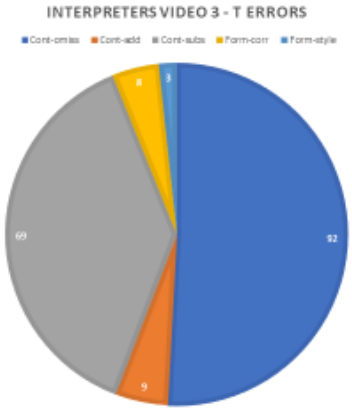
Video 3 – Interlingual 'Beer – talking head'

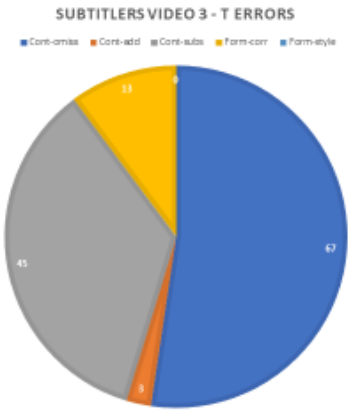
Interpreters - Video 3			
Student	T errors	R errors	Accuracy
1	7	7	99.17%
2	7	14	99.04%
3	11	13	98.47%
4	8	17	98.68%
5	14	13	98.80%
6	18	6	98.53%
7	23	29	97.32%
8	11	9	98.19%
9	16	7	98.33%
10	15	19	97.96%
11	19	10	97.08%
12	13	31	96.75%
13	19	17	97.09%
Averages	14	14.7	98.10%

Subtitlers - Video 3			
Student	T errors	R errors	Accuracy
1	10	14	98.65%
2	7	25	98.16%
3	3	30	98.06%
4	22	8	96.74%
5	14	4	98.47%
6	22	20	97.08%
7	11	36	97.05%
8	19	23	95.81%
9	12	22	97.85%
10	9	42	96.71%
Averages	12.9	22.4	97.45%

- 62% of interpreters and 40% of subtitlers reached the threshold of 98%.
- Subtitlers made on average 1.1 fewer T errors, which is a small difference so perhaps not statistically significant.
- Subtitlers' R errors are much higher suggesting they struggle with dictation.

Translation errors





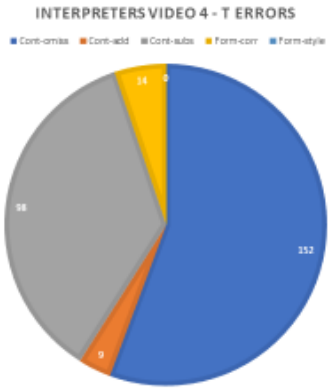
Video 4 – Interlingual 'MSF interview'

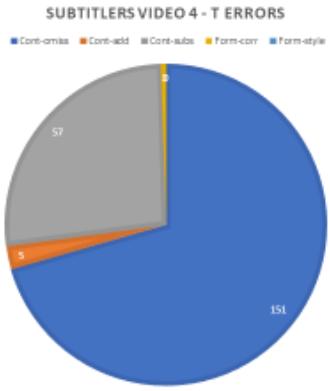
Interpreters - Video 4			
Student	T errors	R errors	Accuracy
1	21	0	98.48%
2	18	6	98.25%
3	8	8	99.19%
4	21	10	97.72%
5	17	10	97.89%
6	18	6	98.06%
7	27	17	97.17%
8	23	2	97.67%
9	22	12	97.28%
10	18	18	97.25%
11	18	5	96.61%
12	31	12	96.08%
13	31	9	93.57%
Averages	21	8.8	97.32%

Subtitlers - Video 4			
Student	T errors	R errors	Accuracy
1	13	9	98.51%
2	16	11	97.76%
3	8	17	98.75%
4	28	3	95.95%
5	28	9	96.57%
6	23	4	96.78%
7	23	18	96.90%
8	29	3	97.11%
9	25	11	96.33%
10	23	31	94.58%
Averages	21.6	11.6	96.92%

- 31% of interpreters and 20% of subtitlers reached the 98% threshold.
- Both groups had a high number of translation errors, suggesting students struggled with the specialised terminology.
- Four interpreters and two subtitlers managed to exceed 98% and three others reached at least 97.70%, suggesting that even difficult texts are feasible with little training.

Translation errors





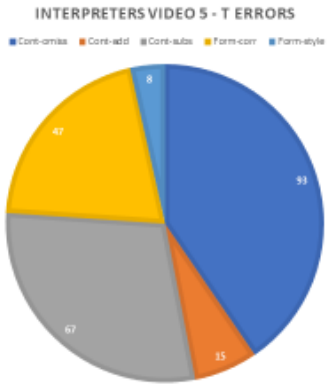
Video 5 – Interlingual Test 'Emma Watson'

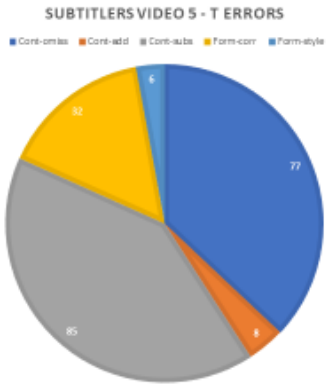
Interpreters - Video 5			
Student	T errors	R errors	Accuracy
1	11	10	98.67%
2	17	12	98.56%
3	10	13	98.64%
4	8	13	98.70%
5	17	15	97.93%
6	18	9	98.65%
7	13	25	97.88%
8	19	13	98.05%
9	24	16	97.47%
10	21	22	95.41%
11	32	23	96.40%
12	24	16	96.20%
13	16	21	96.89%
Averages	17.6	16	97.65%

Subtitlers - Video 5			
Student	T errors	R errors	Accuracy
1	9	15	98.92%
2	12	10	98.58%
3	4	32	98.09%
4	26	6	97.34%
5	28	8	96.65%
6	37	16	95.81%
7	12	29	97.46%
8	33	5	96.69%
9	30	13	95.47%
10	17	33	97.16%
Averages	20.8	16.7	97.21%

- 46% of interpreters and 30% of subtitlers reached 98%.
- 5 out of 6 'good performing' interpreters and 1 out of 7 'poor performers' reached the 98% threshold.
- There is a larger difference between good and poor performing interpreters than between interpreters and subtitlers.
- Both groups scored very similar in terms of R errors. Subtitlers had more T errors than interpreters.

Translation errors





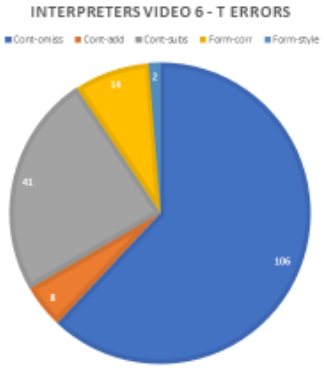
Video 6 – Interlingual Test 'Gardening'

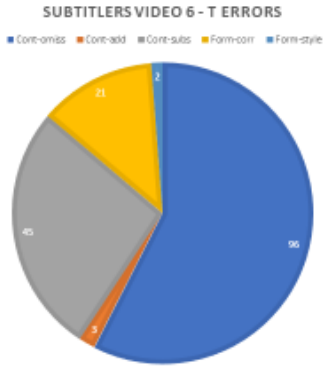
Interpreters - Video 6			
Student	T errors	R errors	Accuracy
1	11	8	98.81%
2	8	12	99.24%
3	5	15	98.80%
4	13	21	98.58%
5	14	13	98.46%
6	19	7	98.26%
7	14	28	97.66%
8	22	13	97.06%
9	13	14	98.06%
10	11	31	97.18%
12	22	8	97.83%
13	19	21	96.70%
Averages	14.25	16	98.05%

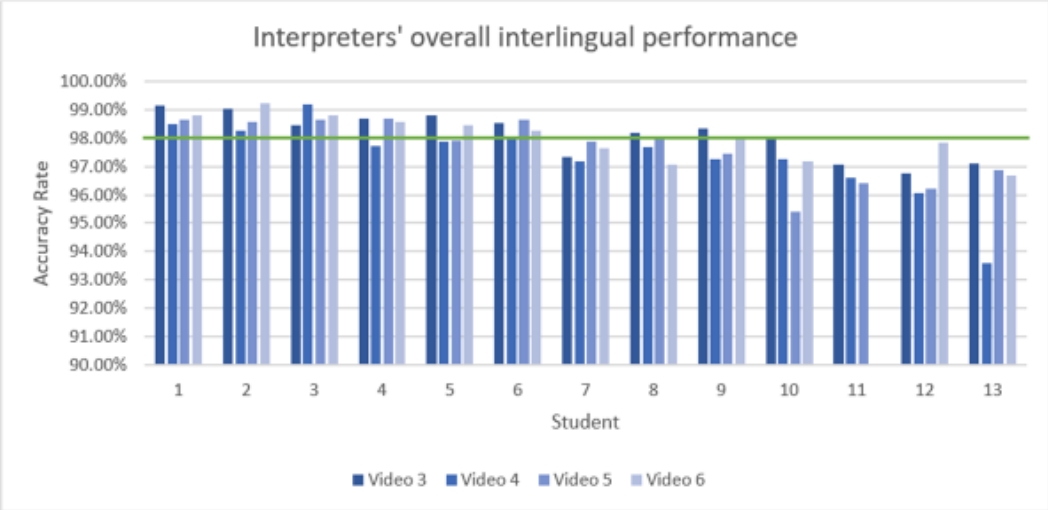
Subtitlers - Video 6			
Student	T errors	R errors	Accuracy
1	11	8	99.16%
2	16	13	98.31%
3	7	44	97.32%
4	15	4	98.46%
5	26	6	97.36%
6	18	15	98.23%
7	12	45	97.28%
8	27	8	96.88%
9	22	25	96.10%
10	13	44	97%
Averages	16.7	21.2	97.61%

- 53% of interpreters and 40% of subtitlers reached 98%.
- All 6 'good performing' interpreters and 1 'poor performer' reached 98%.
- Interpreters found this the second easiest interlingual video and subtitlers found it the easiest video to respeak.
- Some students struggled with recognition reaching up to 31 errors for interpreters and 45 errors for subtitlers.

Translation errors







Interpreters' | Overall performance

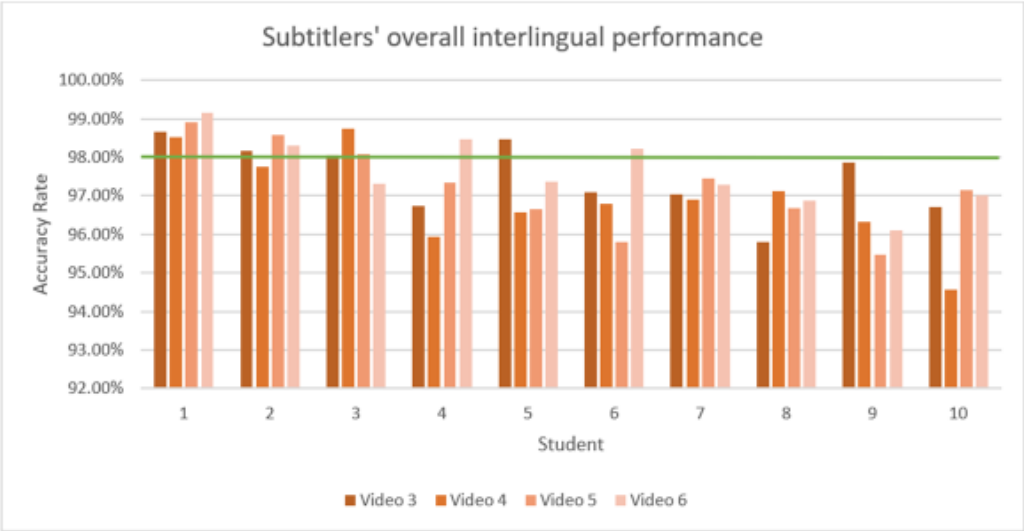
Video 1	Video 2	Video 3	Video 4	Video 5	Video 6
99.62%	99.25%	99.17%	98.48%	98.67%	98.81%
99.47%	99.45%	99.04%	98.25%	98.56%	99.24%
98.84%	98.71%	98.47%	99.19%	98.64%	98.80%
99.09%	99.24%	98.68%	97.72%	98.70%	98.58%
99.20%	99.27%	98.80%	97.89%	97.93%	98.46%
98.67%	98.80%	98.53%	98.06%	98.65%	98.26%
98.44%	98.82%	97.32%	97.17%	97.88%	97.66%
97.05%	98.75%	98.19%	97.67%	98.05%	97.06%
96.88%	98.36%	98.33%	97.28%	97.47%	98.06%
97.21%	98.23%	97.96%	97.25%	95.41%	97.18%
97.88%	97.08%	97.08%	96.61%	96.40%	N/A
96.59%	97.56%	96.75%	96.08%	96.20%	97.83%
96.41%	N/A	97.09%	93.57%	96.89%	96.70%
Averages					
98.10%	98.62%	98.10%	97.32%	97.65%	98.05%

The interpreters produced 76 respoken texts of which the following met or exceeded the 98% threshold:

- 17/25 (68%) intralingual texts
- 25/51 (49%) interlingual texts
- 13/25 (52%) interlingual tests

Video 3 had fewer average translation errors at 14 per text.

Video 4 had fewer recognition errors, with 8.8 errors per text. This was the most difficult video to translate live due to specialised terminology. Students may have decided to focus on dictation to control their errors.



Subtitlers' | Overall performance

Video 1	Video 2	Video 3	Video 4	Video 5	Video 6
99.40%	99.71%	98.65%	98.51%	98.92%	99.16%
98.22%	98.90%	98.16%	97.76%	98.58%	98.31%
99.41%	98.06%	98.06%	98.75%	98.09%	97.32%
98.41%	99.25%	96.74%	95.95%	97.34%	98.46%
97.64%	98.59%	98.47%	96.57%	96.65%	97.36%
98.21%	98.86%	97.08%	96.78%	95.81%	98.23%
97.89%	97.34%	97.05%	96.90%	97.46%	97.28%
95.81%	97.55%	95.81%	97.11%	96.69%	96.88%
95.56%	98%	97.85%	96.33%	95.47%	96.10%
95.65%	96.60%	96.71%	94.58%	97.16%	97%
Averages					
97.62%	98.28%	97.45%	96.92%	97.21%	97.61%

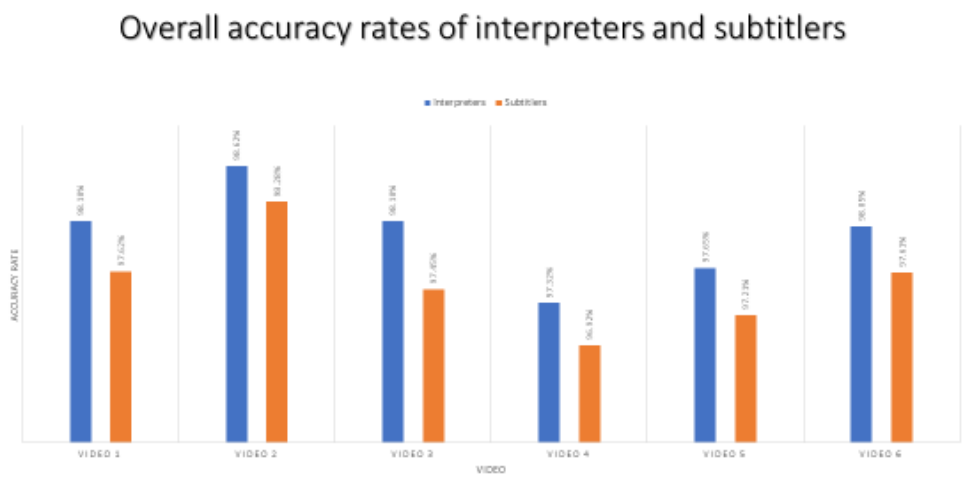
The subtitlers produced 60 respoken texts of which the following met or exceeded the 98% threshold:

- 12/20 (60%) intralingual texts
- 13/40 (32%) interlingual texts
- 7/20 (35%) interlingual tests

This suggests subtitlers cannot cope with the element of live translation, which is required for interlingual respeaking.

Some subtitlers can be good interlingual respeakers, but perhaps when they are taken as a group there is no guarantee they can be good respeakers.

4.



4. Conclusions

- For the interlingual tests, 50% of interpreters reached the 98% threshold, while only 20% of subtitlers reached 98%.
- Interpreters make consistently fewer R errors than subtitlers with an average of 4.1 fewer errors per text.
- Differences in T errors are much closer for both groups, with interpreters making on average 1.8 fewer errors than subtitlers.
- Omission and substitution errors pose more problems than any other:
  - Interpreters had an average of 8.5 omission errors per text, subtitlers had 9.8 omissions per text.
  - Interpreters had an average of 5.2 substitution errors per text, subtitlers had an average of 5.8.
- In terms of error severity, both groups maintained a similar pattern of making more minor, then major then critical errors.
- There are a few differences: subtitlers made more errors in general; interpreters made more critical content substitutions than major substitutions.
- Good interpreters and good subtitlers all start off well and reach 98% in intralingual respeaking, suggesting they quickly master the multitasking element of respeaking.
- Good interpreters and good subtitlers have not experienced technical issues and therefore have better recognition.
- Good interpreters and good subtitlers manage to make smaller omissions.

- ILS seems feasible (97.6%, 4/10)
- Interpreters perform better than subtitlers
- There is a greater difference between 'good' and 'poor' performing interpreters than there is between interpreters and subtitlers.
- Interpreter ≠ good performer / Subtitler ≠ poor performer
- Translation and recognition are equally important and challenging
- Good performers have around 50% fewer translation and recognition errors than bad performers, including consistently less serious errors.
- Bad performers struggle to keep up and as a result omit too many full sentences, mistranslate the source text and dictate less clearly.
- Subtitlers seem to struggle trying to keep up with the text, as a result they have more omissions, more mistranslations and more recognition errors.

5. Development of speech recognition software in Galician

As explained in the abstract above and in the application of the ILSA project, another aim of this IO, in line with the need to develop access at a regional level, was to contribute to the development of a new speech recognition software in Galician. The collaboration between UVIGO, the research group GTM (Engineering Faculty, UVigo), the Galician Parliament and the public Galician broadcaster TVG has resulted in the first speech recognition software in Galician, which can enable the provision of live subtitling for news programmes watched in Galician by 1.5M people daily.

The following slides, presented by Maria Rico, Laura Docío and Carmen García Mateo (UVigo) at the international conference Languages and the Media, provide an account of the work carried out to develop the software.

INTRODUCTION

Audiovisual Material

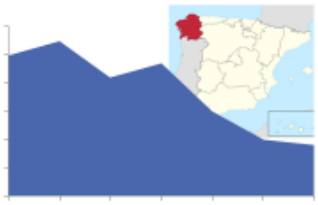
Live Subtitling

Speech Recognition

Automatization

GALICIA. GALICIAN LANGUAGE

- Few speakers (youth)
- Limited presence in mass media
- Vehicular language



Audiovisual Media Accessibility in Galicia

GALMA Galician Observatory for Media Accessibility

Multimedia Technology Group (GTM)

Speech Recognition software for live subtitling in Galicia

SR SOFTWARE. Functioning and Potential

EARLY STAGE

↓

NO punctuation

NO capitalization

NO subtitle format

Language Models

Software training

SR SOFTWARE. Materials and Assessment

21 samples

TELEVISIÓN DE GALICIA

7 samples: news

7 samples: weather

7 samples: sports

NER model

$$Accuracy = \frac{N - E - R}{N} \times 100$$

N = Number of words, E = Editing errors, R = Recognition errors

EXCELLENT	>99.5%
VERY GOOD	99%-99.5%
GOOD	98.5%-99%
ACCEPTABLE	98%-98.5%
POOR	<98%

SR SOFTWARE. (Initial) Results and Discussion

	Errors				Punctuation Errors Excluded				Punctuation + Capitalization Errors Excluded			
				Total				Total				Total
Minor	1733	2163	1970	5866	753	981	1049	2783	486	599	487	1572
Standard	175	122	328	625	171	115	321	607	171	114	320	605
Serious	2	5	6	13	2	2	6	10	2	2	6	10
Total	1910	2290	2304	6504	926	1098	1376	3400	659	715	813	2187
Accuracy Rate (AR)	95.24%	94.23%	94.22%	**94.59%**	97.49%	97.10%	96.25%	**96.95%**	98.10%	98.00%	97.47%	**97.86%**

- Most errors would not affect viewers' comprehension
- Unclear speech creates errors that affect comprehension
- Few errors that omit information/misinform
- Average AR below quality threshold (NER)

- Punctuation errors represent a large percentage of the total
- Minor errors are significantly reduced
- Lower reduction in more severe errors
- Average AR improves considerably

- Capital letter in proper names, place names, institutions, players or teams...
- Minimal changes in more severe errors
- Average AR almost reaches the quality threshold (NER)

SR SOFTWARE. Strengths and Weaknesses

- Most errors are **Minor**: they should not affect viewers' comprehension
  - However, a large number of **Minor** errors in a sentence may hinder reading and understanding
- The amount of **Serious** errors is low: meaning of original ideas hardly changes in written text
  - However, quite a few errors (**Standard**) disrupt the flow/meaning of the original text and might cause surprise
- Quite **sensitive software**: most words are recognized
  - However, not only words but **other sounds are recognized**, turning them into misplaced words.

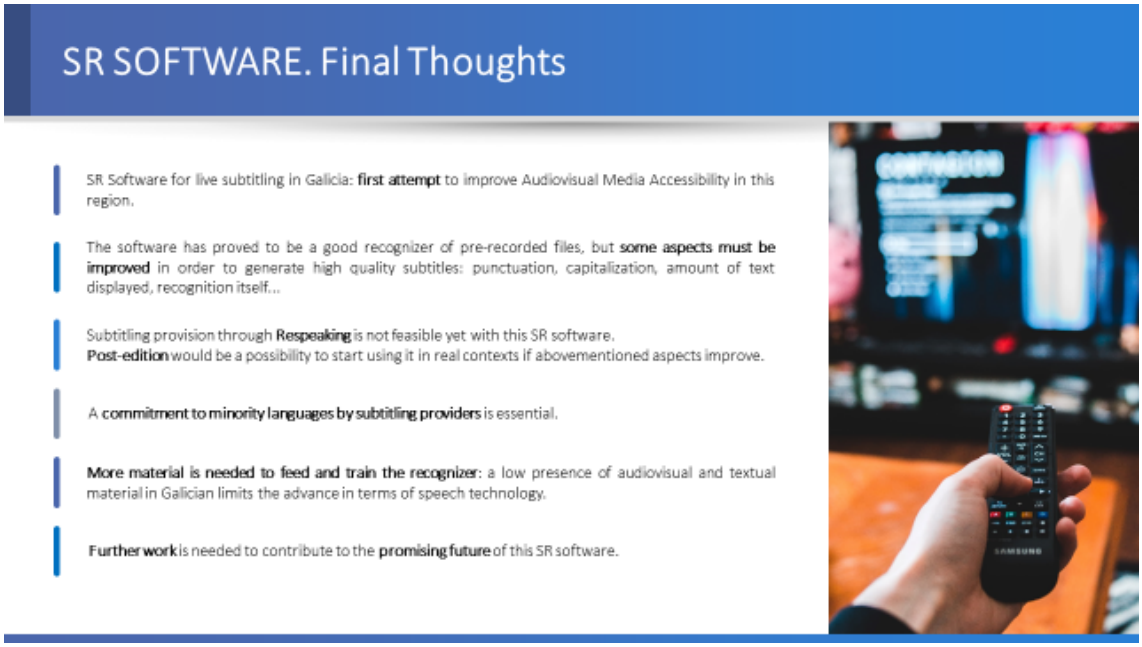
Recognition must be further improved

- Software **not** ready to be used for subtitling provision through **respeaking**
  - Needs to be activated for live use
  - Need to recognize punctuation marks and capital letters

How far are we?

Possibility of using the software for subtitling provision with **post-editing**

- Punctuation and Capitalization must me implemented
- Amount of time to correct errors: ≈10 sec.
  - Total Errors: 52 min. post-editing (per 10-min. sample)
  - Without P. Errors: 27 min. post editing (per 10-min. sample)
  - Without P. + C. Errors: 17 min. post-editing (per 10-min. sample)



As can be seen in the information provided, in order to provide live subtitles in Galician, the software developed as part of this project must be used for post-edition rather than respeaking. In other words, once the recognition has been made, a human operator must correct the errors and then broadcast the subtitles.

Indeed, whereas it is estimated that it normally takes 6 hours to subtitle a 1-hour programme, with the software as it is, this time could be reduced to 5 hours and, should automatic punctuation and capitals be included and refined, this time could be further reduced to 1,5 hours.

It thus looks like the software developed here can therefore not only facilitate the first ever provision of live subtitles for the public Galician broadcaster (TVG) and the Galician Parliament but can also become an extremely useful tool to expedite the subtitling of pre-recorded material too, such as TV programmes, films, etc.

6. References

Ofcom. (2015). Measuring live subtitling quality. Results from the fourth sampling exercise. Retrieved from https://www.ofcom.org.uk/__data/assets/pdf_file/0011/41114/qos_4th_ report.pdf

Robert, I. S., Remael, A., & Bastin, G. L. (2016). Quality Control in the Subtitling Industry: An Exploratory Survey Study. *Meta*, 61(3), 578-605.

Romero-Fresco, P. (2009). More haste less speed: Edited versus verbatim respoken subtitles. *Vigo International Journal of Applied Linguistics*, 6, 109-133.

Romero-Fresco, P. (2011). *Subtitling through speech recognition: respeaking*. Manchester: St. Jerome Publishing.

Romero-Fresco, P. (2012). Respeaking in Translator Training Curricula. *The Interpreter and Translator Trainer*, 6(1), 91-112. doi:10.1080/13556509.2012.10798831

Romero-Fresco, P., & Martínez, J. (2015). Accuracy Rate in Live Subtitling: The NER Model. In J. Díaz-Cintas & R. Baños Piñero (Eds.), *Audiovisual Translation in a Global Context. Mapping an Ever-changing Landscape* (pp. 28-50). London: Palgrave.

Romero-Fresco, P. (2016). Accessing communication: The quality of live subtitles in the UK. *Language & Communication*, 49, 56-69. doi:10.1016/j.langcom.2016.06.001

Romero-Fresco, P., & Pöchhacker, F. (2017). Quality assessment in interlingual live subtitling: The NTR Model. *Linguistica Antverpiensia New Series-Themes in Translation Studies*, 16, 149-167.

Romero-Fresco, P. (2018). Respeaking. In L. Pérez-González (Ed.), *The Routledge Handbook of Audiovisual Translation* (pp. 96-113).

Szarkowska, A., Dutka, Ł., Krejtz, K., & Pilipczuk, O. (2017). Respeaking crisis points. An exploratory study into critical moments in the respeaking process. In M. Deckert (Ed.), *Audiovisual Translation – Research and Use* (pp. 179-201). Bern: Peter Lang.

Szarkowska, A., Krejtz, K., Dutka, Ł., & Pilipczuk, O. (2018). Are interpreters better respeakers? *The Interpreter and Translator Trainer*, 12(2), 207-226. doi:10.1080/175039 9x.2018.1465679

Szczygielska, M., & Dutka, Ł. (2019). Historia napisów na żywo tworzonych metodą respeakingu w Polsce [The history of live subtitling through respeaking in Poland]. In K. Hejwowski, K. Dębska, & D. Urbanek (Eds.), *Tłumaczenie wczoraj, dziś i jutro* (pp. 129-164). Warsaw: Institute of Applied Linguistics University of Warsaw.

Waes, L., Leijten, M., & Remael, A. (2013). Live subtitling with speech recognition. Causes and consequences of text reduction. *Across Languages and Cultures*, 14(1), 15-46.tt

Disclaimer

Publication and preparation of this report was supported by ILSA (Interlingual Live Subtitling for Access), financed by the European Union under the Erasmus+ Programme, Erasmus+, KA2: Cooperation for innovation and the exchange of good practices, KA203: Strategic Partnerships for Higher Education, Project number: 2017-1-ES01-KA203-037948.

The information and views set out in this report are those of the authors and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

How to quote this document

You can quote this report as follows: ILSA (2018). Report on IO2: Development of speech-recognition software and competence analysis for ILS from http://www.ilsaproject.eu/project/