The Interpreter and Translator Trainer

Are interpreters better respeakers?

--Manuscript Draft--

Full Title:	Are interpreters better respeakers?
Manuscript Number:	RITT-2016-0072R1
Article Type:	Original Article
Keywords:	respeaking; interpreting; live subtitling; NER; audiovisual translation
Abstract:	In this study, we examined whether interpreters and interpreting trainees are better predisposed to respeaking than people with no interpreting skills. We tested 57 participants (22 interpreters, 23 translators and 12 controls) while respeaking 5-minute videos with two parameters: speech rate (fast/slow) and number of speakers (one/many). Having measured the quality of the respeaking performance using two independent methods: the NER model and rating, we found that interpreters consistently achieved higher scores than the other two groups. The findings are discussed in the context of transfer of skills, expert performance and respeaking training.

Author Bio

Agnieszka Szarkowska is currently Research Fellow at the Centre for Translation Studies, University College London (2016-2018). She is now working on the SURE project: Exploring Subtitle Reading Process with Eye Tracking Technology. Since 2007, she has also been Assistant Professor in the Institute of Applied Linguistics, University of Warsaw. She is the founder and head of the Audiovisual Translation Lab (AVT Lab, www.avt.ils.uw.edu.pl) and specializes in audiovisual translation, especially subtitling for the deaf and hard of hearing and audio description. She is a member of European Association for Studies in Screen Translation (ESIST), European Society for Translation Studies (EST) and an honorary member of the Polish Audiovisual Translators Association (STAW).

Krzysztof Krejtz (PhD, social and cognitive psychologist). Assistant Professor at Department of Psychology, University of Social Science and Humanities and the Head of Interactive Technologies Laboratory at National Information Processing Institute, Warsaw, Poland. He is the founder and leader of Eye Tracking Research Center at University of Social Sciences and Humanities. His research interests include visual attention, eye-tracking methodology, Human-Computer Interaction, psychological and social aspects of internet. Author of publications in the field of eye tracking methodology, statistics and applications in the context of new media and education as well as social psychology of internet. He is a member of Association of Computing Machinery (ACM) and Polish Social Psychology Association.

Łukasz Stanisław Dutka is an interpreter and audiovisual translator. As a practitioner of subtitling and a pioneer of respeaking in Poland, he currently works at the Institute of Applied Linguistics at the University Warsaw in “Respeaking - process, competences, quality” research project. He is also involved in training interpreters and respeakers. He regularly cooperates with theatres providing surtitles. He works on a PhD on respeaking competences and quality in live subtitling. A member of Audiovisual Translation Lab, Polish Association of Audiovisual Translators (STAW) and European Society for Translation Studies (EST).

Olga Pilipczuk graduated from the Faculty of Applied Linguistics and Faculty of Economics at Warsaw University. Her main professional interests are statistical programming, econometrics and statistics. She is currently employed in the Respeaking project.

Table 1. Clips in the intralingual respeaking task by number of speakers and speech rate

	Slow speech rate	**Fast speech rate**
One speaker	Speech	News
Many speakers	Entertainment show	Political chat show

Table 2. Characteristics of videos used in the study

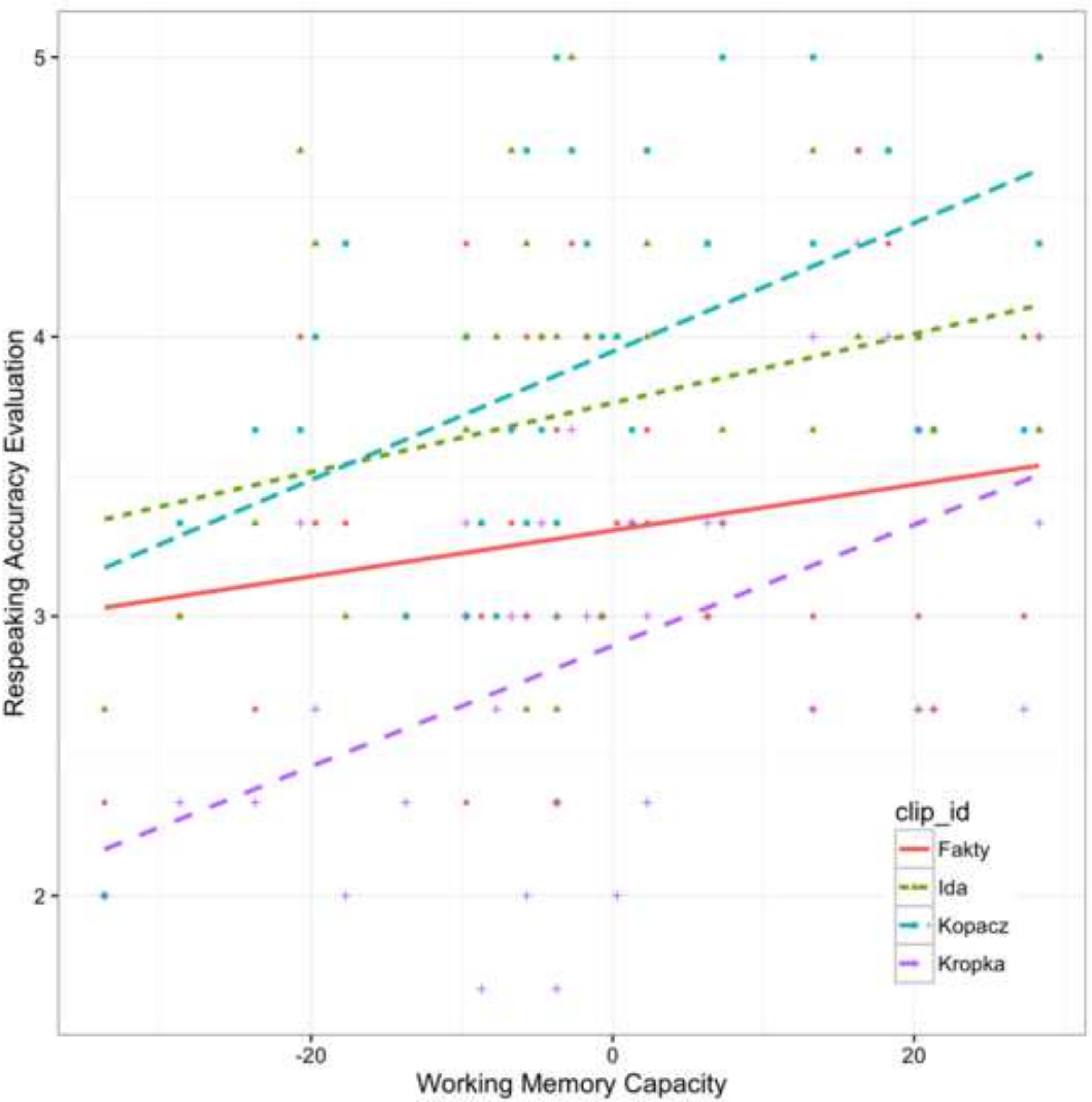
	Duration	Number of words	Speech rate	Number of sentences
Intralingual				
Speech	4:47	520	108 wpm	40
News	6:10	935	152 wpm	99
Entertainment show	5:20	707	133 wpm	74
Political chat show	5:35	916	165 wpm	107
Interlingual				
Speech	4:55	458	94 wpm	35

Table 3. Translation and interpreting experience of the participants

Experience	**Interpreters and interpreting trainees**	**Translators and translation trainees**
none[*]	1 (9.09%)[**]	1 (4.35%)
1–2 years	11 (50.00%)	8 (34.78%)
3–4 years	4 (18.18%)	9 (39.13%)
4 + years	6 (27.27%)	5 (21.74%)

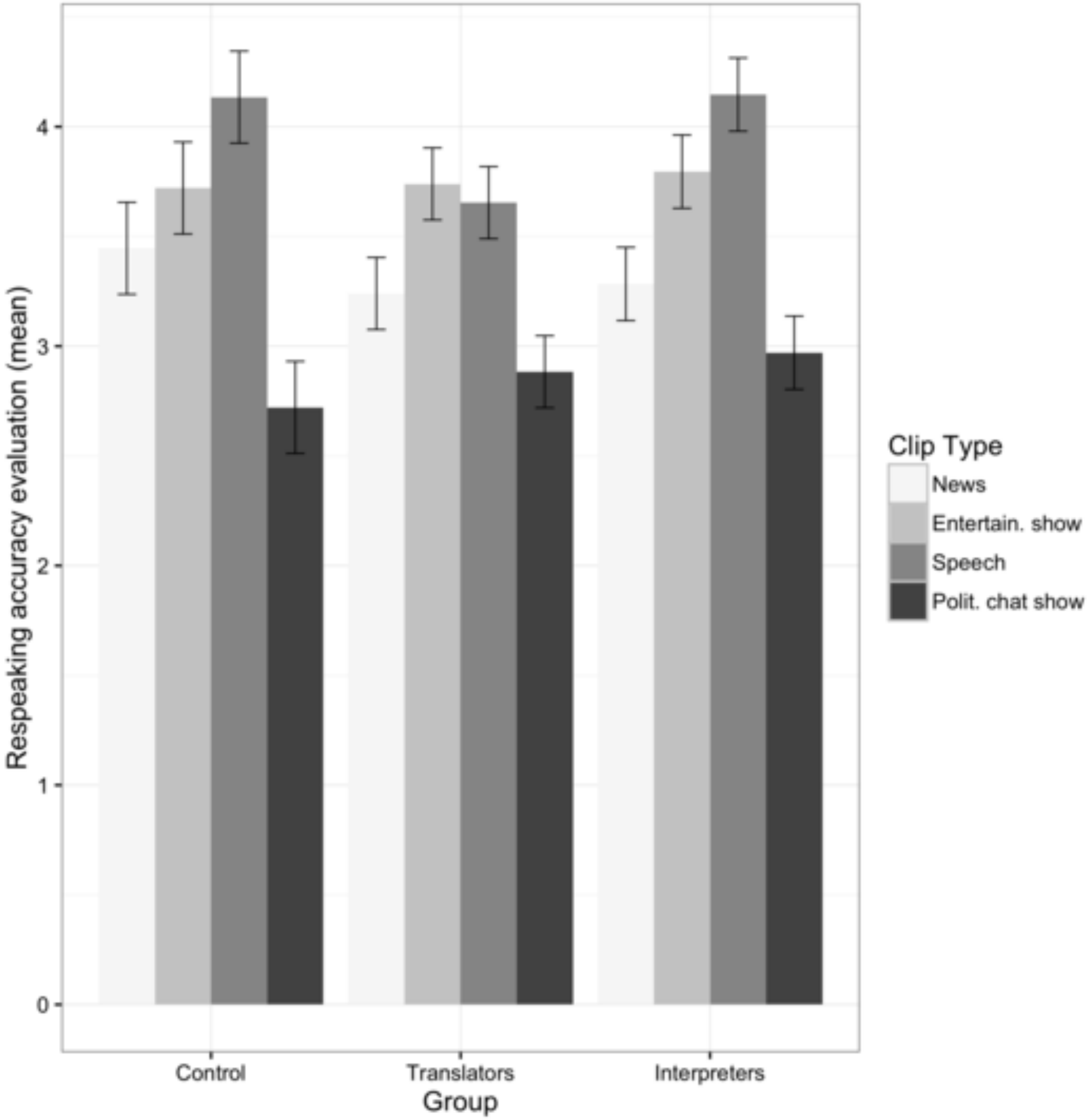
Note: [*] *participants with no experience were interpreting/translation students*

[**] *number of participants (percentage of the whole sample)*



Respeaking Accuracy Evaluation
5
4
3
2
-20
0
20
Working Memory Capacity
clip_id
Fakty
Ida
Kopacz
Kropka

Figure2


Respeaking accuracy evaluation (mean)
0
1
2
3
4
Control
Translators
Interpreters
Group
Clip Type
News
Entertain. show
Speech
Polit. chat show

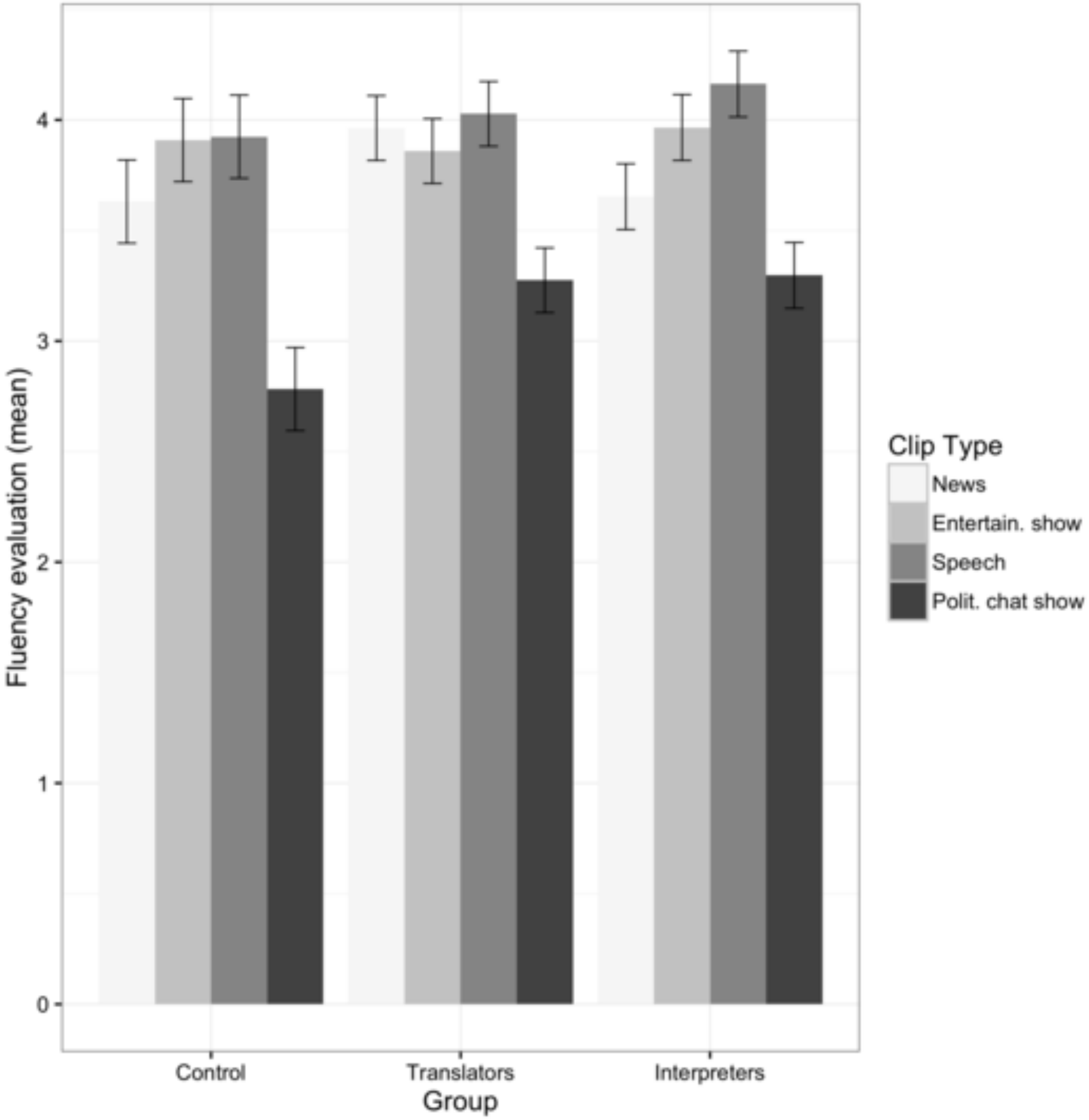

Fluency evaluation (mean)
0
1
2
3
4
Control
Translators
Interpreters
Group
Clip Type
News
Entertain. show
Speech
Polit. chat show

Figure4

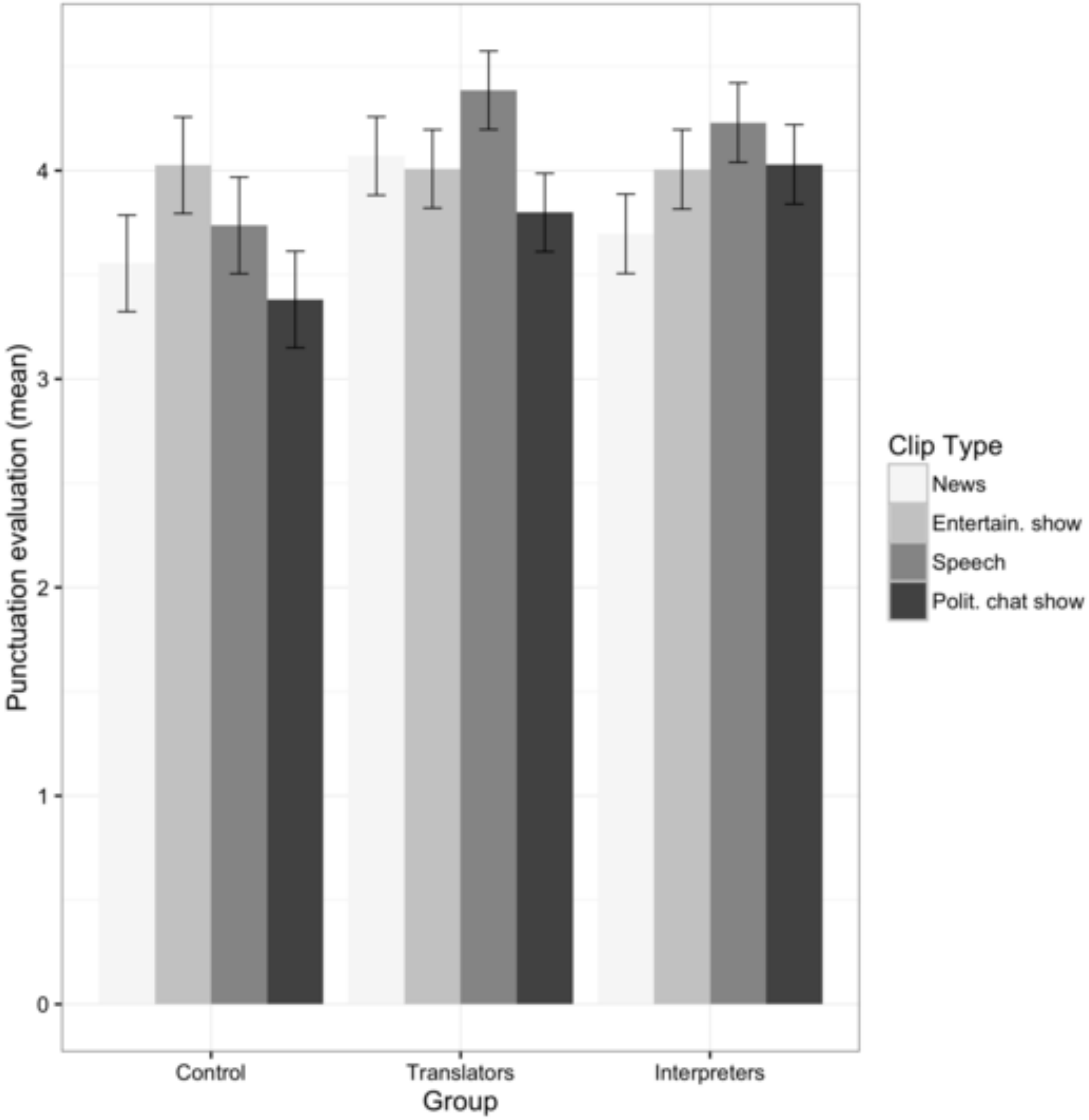
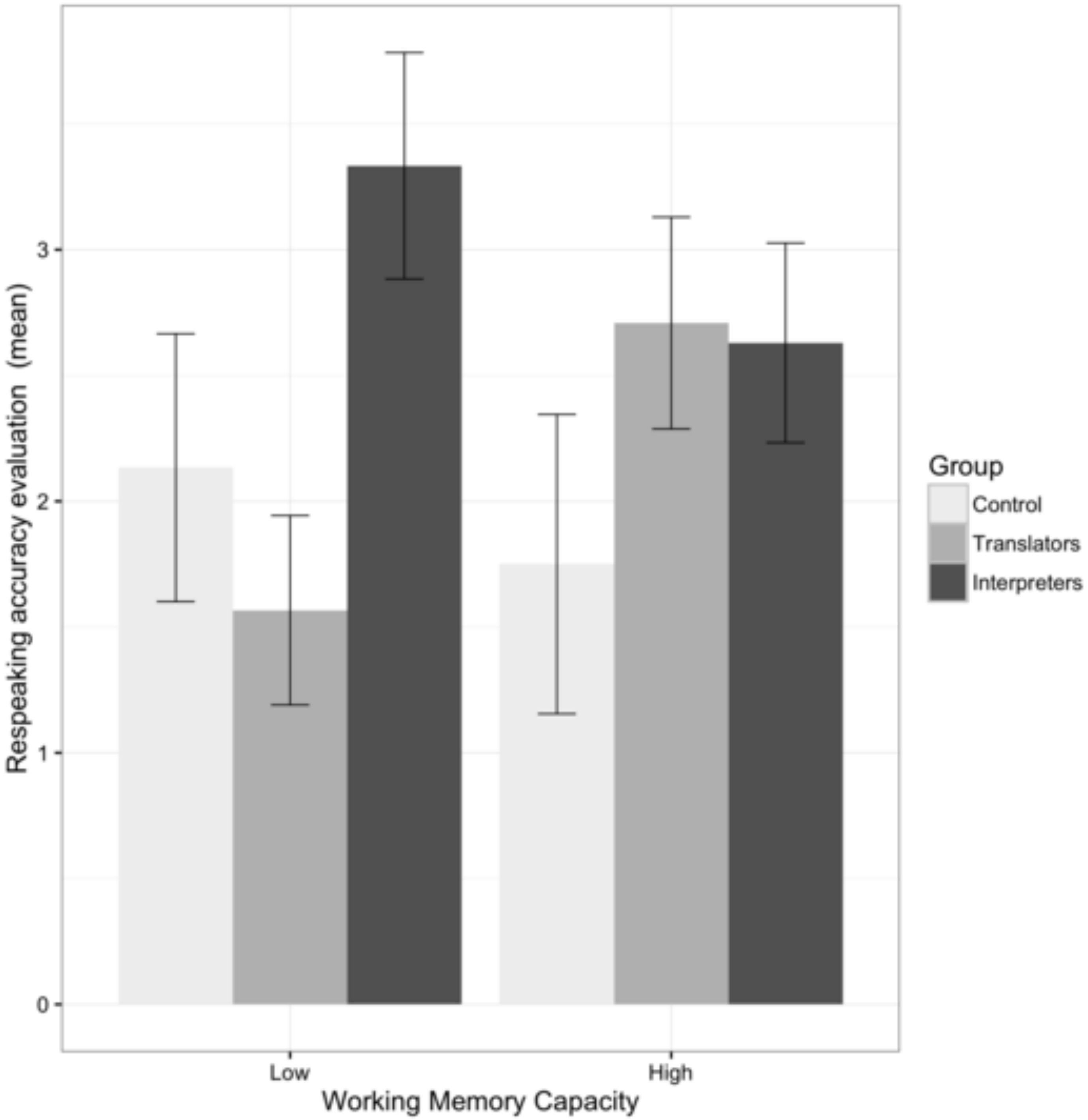

Punctuation evaluation (mean)
0
1
2
3
4
Control
Translators
Interpreters
Group
Clip Type
News
Entertain. show
Speech
Polit. chat show

Figure5


Respeaking accuracy evaluation (mean)
0
1
2
3
Low
High
Working Memory Capacity
Group
Control
Translators
Interpreters

Are interpreters better respeakers?

Abstract

In this study, we examined whether interpreters and interpreting trainees are better predisposed to respeaking than people with no interpreting skills. We tested 57 participants (22 interpreters, 23 translators and 12 controls) while respeaking 5-minute videos with two parameters: speech rate (fast/slow) and number of speakers (one/many). Having measured the quality of the respeaking performance using two independent methods: the NER model and rating, we found that interpreters consistently achieved higher scores than the other two groups. The findings are discussed in the context of transfer of skills, expert performance and respeaking training.

Keywords: respeaking, interpreting, live subtitling, rating, NER, audiovisual translation

1. Introduction

With the advent of modern technologies and their widespread use in audiovisual translation (AVT), new professions are emerging on the AVT market, like that of a respeaker. A respeaker "listens to the original sound of a live programme or event and respeaks it, including punctuation marks and some specific features for the deaf and hard of hearing audience, to a speech recognition software, which turns the recognized utterances into subtitles displayed on the screen with the shortest possible delay" (Romero Fresco 2012, 93). Respeaking is now a major method of producing real-time subtitling to live TV broadcasts using speech recognition technology (Lambourne 2006; Marsh 2006).

Used since 2001, respeaking is a relative newcomer to the field of AVT. However, given its dynamic growth in the AVT industry, the need has arisen for higher education institutions to include respeaking in their study programmes, curricula and training courses (see Remael and van der Veer 2006; Romero-Fresco 2012). As aptly put by Hurtado Albir (2015, 257), "due to changes in the translation profession and constant academic and professional mobility, new curriculum designs that meet society's demands and offer score for international harmonization are required." As new respeakers are entering the market, respeaker trainers – both in the academia and in the industry – are now facing a number of new questions (see Arumí Ribas and Romero-Fresco 2008; Romero-Fresco 2012), including who to recruit, how to train newcomers to the profession, whether expert skills from other linguistic areas like interpreting can be transferred to respeaking, and finally who is better predisposed to becoming a respeaker? Studies on expert performance have shown that contrary to common belief, it is not innate abilities but rather skills acquired through supervised practice and experience that are key to achieving top performance in various areas of expertise (Ericsson and Charness 1994). Owing to a number of overlapping skills and competences between interpreters and respeakers, in this study we hypothesised that some interpreting skills may successfully be transferred to respeaking. If confirmed, this finding may improve our understanding of respeaker competences and also have important implications for interpreter and respeaker training.

Respeaking shares a number of characteristics with interpreting (Eugeni 2008, Marsh 2004). Similarly to simultaneous interpreting (see Jones 2002), respeaking involves rendering a live spoken linguistic input into another form while simultaneously following the speaker's words and monitoring one's own linguistic output. Romero-Fresco argues that "the two activities share clear time constraints, namely real-time production, little or no margin for correction or improvement" (2011, 46). Both interpreters and respeakers need to prepare for the job beforehand, for instance by conducting terminology searches and building up glossaries, etc. Romero-Fresco (2011) also stresses similarities in working conditions: in a booth with another colleague, taking turns every 30 minutes or so, speaking to a microphone and listening to the incoming input through a headset. In contrast to interpreting, respeaking is usually done intralingually. Interlingual respeaking, that is between two languages, does exist (de Korte 2006; den Boer 2001; Robert & Remael 2017; Romero-Fresco & Pöchhacker 2017), but at present it constitutes only a small, though admittedly growing, share of respeaking practices.

Unlike in the case of interpreters, respeakers' intonation needs to be rather flat and monotonous, which makes it easier for the speech recognition software. This is why some respeakers with interpreting experience "may have to 'unlearn' certain interpreting skills, such as the ability to speak in a pleasant tone, which is not particularly effective when dictating to speech recognition programmes" (Romero Fresco 2012, 100). Apart from repeating and/or rephrasing the original text, respeakers also need to insert the necessary punctuation marks by voicing them (Romero-Fresco 2011). Whereas the interpreters' output is instantly consumed by end users listening to the interpreted text through earphones, the respeakers' output is further transformed into written text by speech recognition software and is later displayed as subtitles on viewers' television screens. In some countries, like the UK, respeakers are also required to monitor and correct the errors in the output of the speech recognition software, which adds even more to the pool of the competences they are required to have (Romero-Fresco 2011). In other countries, like France, proof-reading is done by another person.

The link between interpreting and respeaking skills has been previously noticed both by the academia and the AVT industry. When discussing respeaker competences, den Boer (2001, 172) argued that "the ideal person for the job would be someone who is a qualified interpreter and a professional subtitler". According to Romero Fresco (2012, 98), a US-based subtitling provider, *Mark Hall Associates*, appreciated memory skills in respeakers and therefore eagerly recruited sign language interpreters, whereas SWISS TXT, a language service provider for the public TV in Switzerland, found that "trained interpreters perform particularly well as respeakers, not least thanks to their capacity to make instant decisions regarding language usage and their ability to multitask”. In spite of this theoretical interest in the overlap of respeaking and interpreting, to the best of our knowledge, no empirical studies have been conducted so far on respeaker competences and the transfer of skills from interpreting, which is the gap this study intends to fill.

2. Measuring quality in respeaking

While many countries are still struggling to achieve higher levels of subtitling quantity, others, like the UK, are now focussing their attention more on quality. The UK regulator, Ofcom, regularly monitors the quality of live subtitling in both public service broadcasters and private ones; it even set a quality threshold of 98% accuracy rate with the NER model for live subtitling in TV broadcasts (Ofcom 2015). The most challenging quality aspects of live subtitling through respeaking are accuracy, delay and presentation rates (Ofcom 2015). In this study we focus on accuracy by quantifying respeakers’ performance using two methods: the NER model and rating.

2.1 The NER model

The most commonly used method of measuring accuracy in live subtitling through respeaking is based on the NER model (see Martínez Pérez 2015; Ofcom 2014; Romero-Fresco 2009; Romero Fresco 2013; Romero Fresco and Martínez Pérez 2015; Romero Fresco 2016). Drawing on its predecessors, the WER (‘word error rate’) and CRIM (Centre de Recherche Informatique de Montréal) methods (see Boulianne et al. 2009; Romero Fresco 2011), the NER model is based on the following formula (Romero-Fresco and Martínez Pérez 2015):

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

N is the number of words in the respoken text, including different commands such as punctuation; E is the number of edition errors, and R is the number of recognition errors. Edition errors are usually a result of incorrect decisions taken by the respeaker, such as omission or introducing wrong information. Recognition errors are usually caused by misrecognition, by the software or by respeaker's mispronunciation. There are three types of recognition errors: insertions, deletions, or substitutions. Correct editions are instances where the respeaker edited the text without any loss of information, for instance in the case of removing unnecessary false starts or repetitions. Both edition and recognition errors may be classified as serious, standard or minor, and be allocated a 1, 0.5 and 0.25 score, respectively (Romero Fresco and Martínez Pérez 2015).

In order to facilitate the quality measurement of respoken subtitles, NERstar tool was designed by Juan Martínez Pérez on behalf of SWISS TXT, and developed in close collaboration with VerbaVoice and Pablo Romero-Fresco. The tool requires a human operator to manually go through the respoken subtitle file (.srt) and compare it against a transcription of the original text. The operator marks edition and recognition errors as well as punctuation errors, allocating them a score (0.25, 0.5 or 1). The software then automatically calculates the final accuracy rate as well as a number of other parameters, including the proportion of edition and recognition errors. We used the NERstar tool to calculate the accuracy rates of subtitles respoken intralingually by the participants of our study; the interlingual respeaking task could not be measured with the NER model or the NERstar tool, as they are inadequate for translation between languages.

We also used NERstar to calculate the degree of reduction between the original text and its respoken version. Reduction, which usually takes a form of omission or paraphrase, is a typical feature of live subtitling through respeaking. Respeakers are rarely able to provide verbatim subtitles, be it due to fast speech rates, overlapping speech, a production delay between the original broadcast and live subtitles or limited cognitive resources of respeakers (for more on reduction in respeaking see Luyckx et al. (2010) and Sandrelli (2013)). In our study, we examined reduction rates between the three participant groups with a view to finding whether participants from the interpreting group will be able to follow the original text more closely and have lower reduction rates, possibly thanks to their trained working memory.

2.2 Rating

Apart from the NER model, we also used another method to assess the quality of respeaking done by our participants: rating. Three independent raters, unaware of study hypotheses, were asked to rate each respoken clip. They were provided with the original video and its transcription as well as an audio recording of the respeaking and its transcription, produced manually by a human transcriber. After a training session with detailed guidelines, they were asked to rate the quality of the respoken text on a Likert-type 1-5 scale, where 1 was the lowest possible score and 5 – the highest.

The rates rated the respeaking quality on six parameters:

1. Overall quality
2. Fluency of delivery – where the lowest score (1) meant a lot of hesitations, false starts, repetitions, unfinished sentences, and significant variance in the speed of delivery, while the highest score (5) was meant to be used in the case of fluent delivery with (almost) no false starts, repetitions, sentences were complete, constant speed of delivery, clear division into respeaking units
3. Pronunciation – where the lowest score (1) was applied when most words were not pronounced clearly, were joined together (sounds from the previous words carried to the next), most word endings were not pronounced clearly, the loudness varied significantly, while the highest score (5) was applied to cases with good enunciation,

where words could be heard clearly and were distinguishable from one another, words endings were fully pronounced, loudness was roughly the same throughout the recording

4. Punctuation – where the lowest score (1) was given to cases where no or almost no punctuation marks were used, while the highest score (5) was given to cases where all or almost all punctuation marks were present, sentences were finished with a full stop
5. Spoken to written language - where the lowest score (1) was allocated to cases where all the unnecessary features of spoken language of the original dialogue were retained, while the highest score (5) was allocated to cases where all the unnecessary features of spoken language were edited out
6. Accuracy of content – where the lowest score (1) was assigned to cases where most ideas from the original video were incorrectly or were omitted, with significant misrepresentations of semantic content, most proper names and numbers were transferred incorrectly or omitted, while the highest score (5) was assigned to cases all or almost all ideas from the original were rendered correctly and fully, with no significant omissions or misrepresentation of semantic content, and proper names and numbers correctly transferred.

2.3 Hypotheses

Given the skills overlapping with respeaking they had already acquired in interpreting training and practice, we expected participants from the interpreting group to achieve highest accuracy rates in the NER model and highest rating scores, particularly in the interlingual task. We also hypothesised that interpreters would have lower reduction rates, as they would be able to retain more information in their working memory (see Baddeley and Hitch 1974; Timarová, Čeňková and Meylaerts 2015), again thanks to their expertise in interpreting. We also expected interpreters to be rated higher in terms of pronunciation, as this is one of the areas they are trained in (Giles 2009; Pöchhacker 2004). Finally, we thought the control group, i.e. people with no interpreting or translation experience, would achieve lowest results due to the lack of any linguistic training, which may be reflected in such rating parameters as punctuation.

3. Method

3.1 Materials

Five videos representing different TV genres approx. 5 minutes long each were used in the study. Four of them were in Polish (intralingual respeaking task) and one was in English (interlingual respeaking task). All clips were self-contained and no previous knowledge was required to understand them.

As shown in Table 1, the clips in the intralingual respeaking task were matched in terms of speech rate (slow and fast) and the number of speakers (one or many).

Insert Table 1.

Table 2 presents more detailed characteristics of both tasks and clips. The slow one-speaker video was a New Year's Address delivered by former Polish Prime Minister Ewa Kopacz in 2015. It was pre-scripted and read out from a prompter. The video was not visually complex, as it only featured a shot with the speaker in her office.

The fast one-speaker[1] video was a fragment of the evening news programme *Fakty* related to a conflict between the Polish government and miners. Apart from the news anchor, the visuals in this programme also contained some infographics and split screen. News programmes are usually characterised by fast speech rates and condensed meaningful content, which makes it "particularly challenging to edit without losing important information" (Ofcom 2014, 27).

The slow video with many speakers was a fragment of an entertainment show *Fakty po faktach* with a host journalist discussing a recent Oscar award with an actress and a film critic. The dialogue in this video was not pre-scripted and it involved some overlapping speech.

The fast video with many speakers was a political chat show *Kropka nad i* with a host journalist and her two guests: politicians from two opposite ends of the political scene. The video contained a lot of overlapping speech between the three speakers. Chat shows are among the most difficult genres to respeak due to their high speech rates and overlapping speech (Ofcom 2014).

Finally, the interlingual respeaking task was a fragment of the speech delivered by President Barrack Obama on the Freedom Day anniversary in June 2014 on the occasion of Poland's celebrating 25 years since the fall of communism.

Insert Table 2.

3.2 Participants

A total of 57 participants (50 women, 7 men) were tested in the study. Participants were recruited at the Institute of Applied Linguistics at the University of Warsaw, where the study took place, the Faculty of English at Adam Mickiewicz University in Poznań, as well as through social media (AVT Lab and RespeakingProject Facebook pages) and personal contacts. They were all volunteers who went through a two-day respeaking training conducted by experienced respeaker trainers from Switzerland, UK and Italy.

The mean age of the participants was 27.48 (*SD*=5.71). Based on their self-reported linguistic and professional background, including experience in translation/interpreting (see Table 3), they were divided into three groups: interpreters and interpreting trainees (*N*=22), translators and translation trainees (*N*=23), and control group with people with no translation/interpreting background (*N*=12).

Insert Table. 3.

3.3 Procedure

Participants were tested individually in a research lab. Prior to commencing the tests, written informed consent was obtained from each participant.

During the respeaking test, participants' eye movements were monitored using the screen recording functionality of SMI Red eye tracker at 120 Hz and their brain activity was recorded using Emotiv EEG at 128 Hz[2]. FAB Subtitler Live was used to display the respoken subtitles on the screen. Participants' output was audio recorded and later transcribed manually for further analysis. Each participant worked on their own voice profile in the speech recognition software for Polish manufactured by Newton Technologies. The profiles were created during the respeaking workshops.

Each respeaking test began with an explanation of the procedure. Participants went through a short mock respeaking task to familiarise themselves with the procedure; the data for this task were not recorded. The proper respeaking test consisted of four randomised intralingual videos, followed by the interlingual task. After each task, the participant had to answer a few questions related to their self-reported cognitive load (assessing the difficulty of the task, its pace as well as their mental effort and concentration, see Szarkowska et al. 2016). The test ended with a short semi-structured interview.

3.4 Design

The present study was 3 x 4 mixed-design experiment with one between-subject fixed factor (group type: control vs. interpreters vs. translators), and the clip type as a within-subjects factor. Additionally, we also controlled for subjects' working memory capacity (WMC). As the indicator of WMC, we used the traditional score of reading span task (Conway, Kane, Bunting, Hambrick, Wilhelm & Engle, 2005). The main dependent variables were: respeaking accuracy as measured by the NER value, reduction rate and a series of rating parameters: overall quality, fluency of delivery, pronunciation, punctuation, spoken to written language, and accuracy of content.

4. Results

Below we present results of respeaking performance of the three participant groups: first, as measured using the NER model, and second, the rating method. As the NER model and the NERstar tool were developed for intralingual respeaking only and cannot be used between two different languages, we do not report the results of the interlingual respeaking task using this quality measurement method. The interlingual task is only reported in the rating results.

4.1 NER value

The mixed-design Analysis of Covariance (ANCOVA) was conducted on NER value as the dependent variable. The sample group was treated as a between-subject factor and the clip type as a within-subject factor. The score of the reading span task was an indicator of working memory capacity (WMC) and was treated as a covariant in the analysis. The ANCOVA was followed by pairwise comparisons with Tukey correction whenever necessary.

The results of the analysis show two statistically significant main effects: of the group type, $F(2,32) = 3.46$, $p < 0.05$, *eta-squared* = 0.12, and of WMC, $F(1,32) = 15.27$, $p < 0.001$, *eta-squared* = 0.22. The highest NER value was achieved by the interpreting group in all clips ($M = 94.36$, $SE = 0.52$), compared to the translators ($M = 92.65$, $SE = 0.50$) and to control group ($M = 93.47$, $SE = 0.58$). Pairwise comparisons showed that interpreters differed significantly from the translators ($p < 0.05$), but not from the control group. The difference between the translators and the control group was not statistically significant.

The main effect of WMC can be interpreted with the simple correlation coefficient, *r* = 0.41, *t*(142) = 5.51, *p* < 0.001, which showed that the higher the working memory capacity, the higher the NER value achieved by the participants.

4.2 Edition Errors

An analogous ANCOVA analysis was conducted for the number of edition errors as a dependent variable. The analysis revealed a main effect of the clip type, *F*(2.49, 79.67) = 69.61, *p* < 0.001, *eta-squared* = 0.57, which was statistically significant and strong in terms of explained variance. The following post-hoc tests with Tukey correction showed that the political chat show (i.e. the fast clip with many speakers) (*M* = 69.23, *SE* = 2.52) differed significantly (*p* < 0.001) from news (*M* = 35.43, *SE* = 2.52), entertainment chat show (*M* = 28.83, *SE* = 2.52), and speech (*M* = 24.90, *SE* = 2.52). All other comparisons were not statistically significant.

4.3 Reduction rate

The reduction rate was analysed with an analogous ANCOVA described above. The analysis revealed two statistically significant main effects of the group type, *F*(2,32) = 6.35, *p* < 0.01, *eta-squared* = 0.25, and clip type, *F*(83.43, 27.59) = 325.84, *p* < 0.001, *eta-squared* = 0.64. Pairwise comparisons with Tukey correction for the main effect of group type showed that the smallest percentage of words removed from the original text was found in the interpreting group (*M* = 27.87, *SE* = 2.56), which was significantly (*p* < 0.01) different from those in the translators group (*M* = 39.79, *SE* = 2.51), but not significantly different from the control group (*M* = 31.48, *SE* = 2.90). The difference between the control group and the translators was also not significant (*p* > 0.1).

The post-hoc tests for the main effect of the clip type showed that the highest reduction rate was for the fast political chat show with many speakers (*M* = 50.40, *SE* = 1.65) and the lowest for the slow one-speaker speech (*M* = 12.87, *SE* = 1.65). Medium values were found in the news

(*M* = 37.56, *SE* = 1.65) and the slow entertainment clip with many speakers (*M* = 31.34, *SE* = 1.65). All comparisons between clips used in the study on reduction rate were statistically significant (*p* < 0.001).

The analysis also showed no significant effect of working memory capacity nor any interaction term.

4.3 Rating score

Before running statistical analyses, the raters agreement was calculated for each rating scale. The overall mean agreement was relatively high (*M* = 83.63%), as well as for fluency (*M* = 91.58%), pronunciation (*M* = 87.36%), punctuation (*M* = 79.66%), and accuracy of content (*M* = 91.22%). Only the level of inter-rater agreement on evaluation of spoken-to-written transfer (*M* = 67.38%) yields more careful interpretation of the following results on this variable.

A series of two-way ANCOVA analyses with the group type as a between-subject factor and clip type as a within-subject factor and WMC as a covariant was performed to test the rating scores. The dependent variables in the following analyses were as follows: overall rating, rating of fluency, pronunciation, punctuation, spoken-to-written transfer, and accuracy of content. The results of these analyses are presented below.

The ANCOVA on **overall rating of respeaking quality** as a dependent variable showed a significant main effect of clip type, *F*(2.95, 100.33) = 44.80, *p* < 0.001, *eta-squared* = 0.32, and of WMC, *F*(1,34) = 30.36, *p* < 0.001, *eta-squared* = 0.362. The overall mean rating score was the lowest for the fast political chat show with many speakers (*M* = 2.95, *SE* = 0.09) and it was the highest for the slow one-speaker speech (*M* = 3.92, *SE* = 0.09). These two clips differed significantly between each other (*p* < 0.001). They also differed significantly (*p* < 0.001) from the slow many-speaker entertainment show (*M* = 3.68, *SE* = 0.09) and the news (*M* = 3.44, *SE* = 0.09). However, pairwise comparisons showed that only the entertainment show and the slow speech were marginally significantly different from each other (*p* = 0.052). The significant main effect of working memory capacity means that the higher WMC score, the higher the overall rating, *r* = 0.51, *t*(150) = 7.32, *p* < 0.001.

The analysis on **respeaking accuracy** revealed several interesting statistically significant effects. First, we found the main effect of clip type, *F*(2.86, 97.07) = 48.73, *p* < 0.001, *eta-squared* = 0.31. The best accuracy score was achieved for the slow one-speaker speech (*M* = 3.97, *SE* = 0.10), which was significantly (*p* < 0.001) higher than for the news (*M* = 3.23, *SE* = 0.10), and fast political chat show with many speakers (*M* = 2.86, *SE* = 0.10) but not different from the slow entertainment show (*M* = 3.75, *SE* = 0.10). The accuracy score for the fast political chat show with many speakers was significantly (*p* < 0.001) lower in comparisons with any other clip used in the study.

The analysis also showed a significant main effect of WMC, *F*(1, 34) = 9.07, *p* < 0.01, *eta-squared* = 0.15. The following correlation analysis showed, in line with the expectations, that the higher WMC, the higher accuracy of respeaking, *r* = 0.32, t(150) = 4.17, *p* < 0.001. Interestingly, this relation was moderated by the clip type. The interaction effect of WMC and clip type was found, *F*(2.86, 97.07) = 2.78, *p* < 0.05, *eta-squared* = 0.03. The influence of WMC on respeaking accuracy differed between clips used in the study. The strongest relation was for the fast political chat show with many speakers, *r* = 0.51, t(36) = 3.53, *p* < 0.01, and the lowest (not statistically significant) for the news, *r* = 0.21, *t*(36) = 1.29, *p* > 0.1. For the slow speech, the relation was significant, *r* = 0.49, *t*(36) = 3.40, *p* < 0.01, and for the slow entertainment show it was only marginally significant, *r* = 0.28, *t*(36) = 1.78, *p* = 0.08, see Figure 1.

Insert Fig. 1 here

Fig. 1. The interaction effect of clip type and working memory capacity on respeaking accuracy measured by rating

The analysis also revealed a marginally significant interaction effect of the clip type and group type, *F*(97.07 0.17) = 1.92, *p* = 0.08, *eta-squared* = 0.03. The means and standard errors are presented in Figure 2.

Insert Fig. 2 here

Fig. 2. The interaction effect of the clip type and group type on respeaking accuracy evaluation by rating. The error bars represent +/- 1 standard error for the estimated means

The ANCOVA analysis on **fluency of delivery** also revealed a significant main effect of clip type, *F*(2.61, 88.73) = 37.44, *p* < 0.001, *eta-squared* = 0.28 and the main effect of WMC, *F*(1, 34) = 14.05, *p* < 0.001, *eta-squared* = 0.21. Fluency was highest for the speech (*M* = 4.04, *SE* = 0.09), which was significantly (*p* < 0.02) higher than for the news (*M* = 3.74, *SE* = 0.09) and the fast political chat show with many speakers (*M* = 3.12, *SE* = 0.09, p < 0.001), but not different compared to the slow entertainment show with many speakers (*M* = 3.91, *SE* = 0.09). The fluency evaluation for the fast political chat show with many speakers reached significantly (*p* < 0.001) lower scores than the evaluations for other clips used in the study.

The analyses also showed a marginally significant interaction effect of the experimental group type and clip type, *F*(5.22, 88.73) = 1.97, *p* = 0.09, *eta-squared* = 0.04. Pairwise comparisons with Tukey correction revealed that in all experimental groups the fluency of respeaking significantly (*p* < 0.001) differed between the fast political chat show with many speakers and all other clips used in the study. However, only interpreters achieved significantly higher fluency scores while respeaking the speech (*M* = 4.16, *SE* = 0.15) compared to the news (*M* = 3.65, *SE* = 0.15). The means and standard errors for the interaction term are presented in Figure 3.

Insert Fig. 3 here

Fig. 3. The interaction effect of group type and clip type on fluency of respeaking evaluation by rating. The error bars represent +/- 1 standard error for the estimated means.

The ANCOVA analysis on spoken-to-written transfer rating revealed only a main effect of clip type, *F*(2.52, 85.78) = 10.00, *p* < 0.001, *eta-squared* = 0.10. The highest score was achieved when respeaking the news (*M* = 3.59, *SE* = 0.10), which was significantly (*p* < 0.05) higher than for the speech (*M* = 3.07, *SE* = 0.10) and for the fast political chat show (*M* = 3.31, *SE* = 0.10), but not significantly different than for the slow entertainment show (*M* = 3.53, *SE* = 0.10). The analysis of the pronunciation parameter again showed a significant effect of clip type, *F*(2.87, 97.43) = 3.28, *p* < 0.05, *eta-squared* = 0.03, and a main effect of WMC, *F*(1, 34) = 13.99, *p* < 0.001, *eta-squared* = 0.22. The following correlation test showed that the higher WMC, the better the pronunciation parameter, *r* = 0.45, *t*(150) = 6.18, *p* < 0.001. Pairwise comparisons of clips on the pronunciation parameter showed that the only significant difference (p < 0.02) was between the slow entertainment show (*M* = 4.24, *SE* = 0.09) and the news (*M* = 3.98, *SE* = 0.09).

Finally, the analysis of the punctuation rating score revealed three significant effects. The main effect of clip type, *F*(2.65, 90.19) = 8.36, *p* < 0.001, *eta-squared* = 0.05. The news had a significantly lower score (*M* = 3.77, *SE* = 0.11) than the slow entertainment show (*M* = 4.01, *SE* = 0.11) and the speech clip (*M* = 4.11, *SE* = 0.11). The speech was significantly better in terms of punctuation than the fast political chat show with many speakers (*M* = 3.74, *SE* = 0.11). The main effect of WMC, *F*(1, 34) = 10.37, *p* < 0.01, *eta-squared* = 0.19 means that the higher WMC, the better punctuation score, *r* = 0.42, *t*(150) = 5.65, *p* < 0.001.

The analysis also showed a significant interaction effect of the clip type and group type, *F*(5.31, 90.19) = 3.45, *p* < 0.01, *eta-squared* = 0.04. The results of post-hoc tests showed that in the control group a significant difference was found between the fast political chat show (*M* = 3.38, *SE* = 0.23) and the slow entertainment show (*M* = 4.03, *SE* = 0.23); in the translators group: the speech (*M* = 4.38, *SE* = 0.19) had a significantly higher punctuation score than the fast political chat show with many speakers (*M* = 3.80, *SE* = 0.19), and in the interpreters group the speech (*M* = 4.23, *SE* = 0.19) again was significantly better in terms of punctuation compared to the news (*M* = 3.70, *SE* = 0.19), see Figure 4.

Insert Fig. 4 here

Fig. 4. The interaction effect of clip type and group type on the punctuation evaluation score by rating. The error bars represent +/- 1 standard error for the estimated means.

Interlingual respeaking

Finally, we hypothesised that interpreters are going to perform better and produce more accurate respeaking especially when performing the interlingual task. In order to test this prediction, we conducted two 3 x 2 Analysis of Variance with the between-subject design. The first factor was the experimental group (interpreters vs. translators vs. control) and the second factor was the working memory capacity score divided into two categories with the median split (low WMC vs. high WMC). For the purpose of these analyses, only the data from the interlingual task were selected.

The analysis on overall rating of respeaking quality revealed only a marginally significant main effect of the group, *F*(2,37) = 2.65, *p* = 0.08, *eta-squared* = 0.13. The following not-adjusted pairwise comparisons showed that interpreters (*M* = 3.15, *SE* = 0.33) performed significantly ($p < 0.05$) better than control group (*M* = 2.00, *SE* = 0.44) and marginally significantly (*p* =

0.09) better than translators (*M* = 2.36, *SE* = 0.31). The difference between translators and control group was not statistically significant (*p* > 0.1).

The ANOVA on respeaking accuracy revealed two marginally significant effects: the main effect of the group, *F*(2,37) = 2.97, *p* = 0.06, *eta-squared* = 0.14 and the interaction effect of the group and working memory capacity, *F*(2,37) = 2.78, *p* = 0.08, *eta-squared* = 0.13 (see Figure 5). The following pairwise comparisons for the main effect of the group with no *p*-value adjustment showed that interpreters achieved significantly (*p* < 0.05) higher accuracy rating scores (*M* = 2.98, *SE* = 0.30) than translators (*M* = 2.14, *SE* = 0.28) and control group (*M* = 1.94, *SE* = 0.40). Translators did not differ significantly from the control group (*p* > 0.1).

The interaction effect suggests that the difference in respeaking accuracy between the groups was moderated by the participants' working memory capacity. The following pairwise comparisons with no p-value adjustment revealed that when low on WMC, interpreters achieved significantly (*p* < 0.01) higher accuracy ratings (*M* = 3.33, *SE* = 0.45) than translators (*M* = 1.57, *SE* = 0.38) and marginally higher results than the control group (*M* = 2.13, *SE* = 0.53). The difference between the translators and control group participants low on WMC was not statistically significant (*p* > 0.1). Interestingly, high working memory capacity compensates for the interpreters skills/experience such that the differences between the groups of participants high on WMC were not statistically significant. Moreover, the results showed that translators with high working memory capacity achieved marginally (*p* = 0.05) higher accuracy scores (*M* = 2.71, *SE* = 0.42) than translators who were low on working memory capacity. The pattern of mean of the interaction effect is presented on Figure 5.

Insert Fig. 5 here

Fig. 5. The interaction effect of the group and working memory capacity on the accuracy evaluation score by rating. The error bars represent +/- 1 standard error for the estimated means.

5. Discussion

In this study, we wanted to find whether people with interpreting expertise have some advantage over translators and people with no interpreting skills when performing a series of respeaking tasks. We tested professional interpreters and interpreting trainees, comparing their performance with that of translators and translation trainees as well as the control group. The performance was measured using two independent methods: the NER model and rating.

In both methods it was the interpreting group that achieved the highest scores. We think this result may be regarded as evidence of a transfer of skills from interpreting to respeaking. Interpreting skills relevant to respeaking include previous training and experience in dealing with simultaneous input and monitoring one's own output, working memory trained in following spoken input and dividing it into comprehensible chunks, divided attention, knowing how to cope with stress, etc. (see Pöchhacker 2004; Gile 2009) as well as the "ability to develop emergency strategies when source text is not understood" (Romero Fresco 2011, 52), being able to identify and select relevant information, condense the original if necessary, keep up with fast speech rates, and deduce the meaning from a wider context. All these may have helped participants from the interpreting group in achieving top performance, discernible both in the NER value and rating scores.

The NER analysis also showed that the interpreters were not only the best in achieving highest respeaking quality, but they also had the lowest reduction rates. Being familiar with the 'salami technique' (Jones 1998; Romero Fresco 2011), interpreters may have skilfully chunked the incoming input, which helped them cope with the source text with a lesser need for reduction. This result may also be attributed to the skills that the interpreters developed during interpreting training and further strengthened in their professional experience, such as anticipation, concurrent management of two speech channels (i.e. listening and speaking at the same time) and managing their voice as a professional tool, which allowed them to re-express more of the text by speaking faster for longer periods of time while still articulating clearly.

Back in 1952, Herbert stated that “A good interpreter must be a trained public speaker” (cited after Pöchhacker 2004, 125). Indeed, during interpreting training much emphasis is placed on the fluency of delivery, voice quality, faithfulness and completeness of utterances (Pöchhacker 2004; Buhler 1986). Both interpreters and respeakers need to “express thoughts clearly and concisely” (Romero-Fresco 2011, 53), have a good diction and a rich vocabulary as well as avoid hesitations, false starts and repetitions. It is therefore not surprising that the highest scores in the pronunciation and fluency of delivery parameters in rating were achieved by the interpreting group. The largest difference between the control group and the other two groups, i.e. people who were trained in linguistics, was in punctuation in the rating scores. This may point to a higher linguistic awareness of punctuation rules, as translators and interpreters work directly with language on a daily basis.

As for the differences between the clips, in line with our predictions, the fast clip with many speakers turned out to be the most difficult to respeak given its high speech rates and a great deal of overlapping speech. This clip was found to have the highest edition errors and reduction rates as well as the lowest rating scores. We believe that the fast speech rates and overlapping speech caused a strain on respeakers’ cognitive load and working memory, leaving little or no time to think of inserting appropriate punctuation marks (hence the lowest rating score in the punctuation parameter) and delivering a fluent and accurate respoken text (as shown by the rating results). Highest scores were attained in slow clips, particularly the slow one-speaker speech, which also had the fewest edition errors.

One of the most important findings of this study is a strong link between respeaking quality and working memory capacity (Baddeley and Hitch 1974). We found that higher WMC correlated positively with higher respeaking accuracy as measured by both the NER model and the rating.

This correlation was particularly strong in the case of the most difficult clip in the study, i.e. the fast video with many speakers (the political chat show). WMC also positively affected the rating of the overall quality, accuracy, fluency of delivery and punctuation parameters. This means that people with high WMC could deliver better quality respeaking by being more accurate and fluent as well as better able to control punctuation.

Low capacity of the working memory may result in cognitive overload, which in turn tends to translate into poor performance (Paas et al. 2003), as was indeed the case in our study. The load on working memory can be reduced by developing automated schemas in long-term memory (Kalyuga et al. 1999). Such schemas are often used to explain differences between experts and novices (see Kalyuga et al. 2003); in this study, experts are understood as people with interpreting experience and novices as people without such experience. Given that working memory capacity was found to be highest in the interpreting group, we believe that interpreters may have developed schemas in their long-term memory (Kirschner et al. 2011), enabling them to surpass the limitations of the working memory (see Shlesinger 2000). This might have helped them to better manage their ear-voice span (the delay between hearing the audio and respeaking it) and it could have resulted in the lower cognitive load they experienced during respeaking (Szarkowska et al. 2016), which in turn translated into better performance.

When it comes to the interlingual respeaking task, we found that the interpreting group scored highest in the ratings, although the differences were not as pronounced as we had expected them to be. Interestingly, similarly to the intralingual tasks, we observed a strong link between WMC and respeaking quality. People who have a high WMC performed consistently better as respeakers, regardless of whether they are interpreters or not. However, among those with lower WMC, it was the interpreting group who achieved higher scores. This shows that although working memory has an undeniable impact on the quality of respeaking, other factors and skills from the interpreting world positively affect respeaking performance.

Our study also showed parallels between more objective method of assessing respeaking accuracy (NER) and a more subjective method, using different human subjects as raters. We believe this adds significance to the NER model, confirming its usefulness as a major method of assessing accuracy in intralingual respeaking.

Limitations of the study

This study is not free from limitations, including a small selection of audiovisual material and a limited number of participants. Other studies could possibly replicate our research on other languages, participants and audiovisual materials.

Conclusion

To the best of our knowledge, this was the first empirical study on the competences of interpreters in relation to respeaking. We hope to have shown that being an interpreter may be advantageous in becoming a respeaker, particularly when it comes to better retention of the original input thanks to trained working memory, endurance in dealing with multiple sources of information and communicating the complex spoken content fluently and accurately to the target audience.

This finding has some considerable implications and practical applications. One of them is the inclusion of respeaking in the training of interpreters. This could add to their pool of skills, making them more versatile and flexible on the quickly changing translation market. Another consequence of establishing closer links between respeaking and interpreting is to support respeaker trainers with training tasks which are already well-established and proven in the field of interpreting, particularly WMC techniques. Merging interpreting and respeaking training may also help in preparing trainees to become interlingual respeakers – a profession which is now niche, but may develop in the near future. In this way, higher education institutions may provide competent workforce for the quickly changing translation, interpreting and media accessibility market.

Acknowledgements

This work was supported by XXX.

References

Arumí Ribas, M., and Romero Fresco, P. 2008. “A Practical Proposal for the Training of Respeakers.” *The Journal of Specialised Translation, 10*, 106–127.

Baddeley, A. D. and Hitch, G. 1974. “Working memory.” In G. H. Bower (ed.) *The Psychology of Learning and Motivation: Advances in research and theory*, 47–89. New York: Academic Press.

Boulianne, G., J.-F. Beaumont, Boisvert, M., Brousseau, J., Cardinal, P., Chapdelaine, C., Comeau, M., Ouellet, P., Osterrath, F., and Dumuchel, P. 2009. “Shadow Speaking for Real-Time Closed-captioning of TV Broadcasts in French.” In A. Matamala and P. Orero (Eds.) *Listening to Subtitles. Subtitles for the Deaf and Hard of Hearing*, 191–208). Bern: Peter Lang.

Bühler, H. 1986. “Linguistic (Semantic) and Extra-linguistic (Pragmatic) Criteria for the Evaluation of Conference Interpretation and Interpreters.” *Multilingua* 5(4), 231–5.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., and Engle, R. W. (2005). “Working memory span tasks: A methodological review and user's guide.” *Psychonomic Bulletin and Review*, 12, 769 - 786.

de Korte, T. 2006. “Live inter-lingual subtitling in the Netherlands. Historical background and current practice.” *Intralinea*. http://www.intralinea.org/specials/article/Live_inter-lingual_subtitling_in_the_Netherlands den Boer, C. 2001. “Live interlingual subtitling.” In Y. Gambier and H. Gottlieb (Eds.) *(Multi)media translation: concepts, practices and research,* 167–172. Amsterdam/Philadelphia, John Benjamins.

Ericsson, K. A. and Charness, N. 1994. “Expert performance. Its structure and acquisition.” *American Psychologist,* 49(8), 725–747.

Eugeni, C. 2008. “A Sociolinguistic Approach to Real-time Subtitling: Respeaking vs. Shadowing and Simultaneous Interpreting.” *English in International Deaf Communication* 72, 357–382.

Gile, D. 2009. *Basic Concepts and Models for Interpreter and Translator Training*. Amsterdam/New York: John Benjamins.

Hurtado Albir, A. 2010. “Competence.” In Y. Gambier and L. vad Doorslaer (Eds.) *Handbook of Translation Studies,* 55–59. Amsterdam: John Benjamins.

Jones, R. 2002. *Conference interpreting explained*. Manchester: St. Jerome Publishing.

Kalyuga, S., Chandler, P. and Sweller, J. 1999. “Managing split attention and redundancy in multimedia instruction.” *Applied Cognitive Psychology,* 13. 351–371.

Kalyuga, S., Ayres, P., Chandler, P. and Sweller, J. 2003. “The expertise reversal effect.” *Educational Psychologist* 38(1). 23–31.

Kirschner, F., L. Kesterand G. Corbalan. 2011. “Cognitive load theory and multimedia learning, task characteristics, and learner engagement: The current state of the art.” *Computers in Human Behavior,* 27(1). 1–4.

Lambourne, A. 2006. “Subtitle respeaking. A new skill for a new age.” *Intralinea*. http://www.intralinea.org/specials/article/Subtitle_respeaking

Luyckx, B., Delbeke, T., Van Waes, L., Leijten, M., and Remael, A. 2010. “Live subtitling with speech recognition. Causes and consequences of text reduction.” *IDEAS Working Paper Series from RePEc,* IDEAS Working Paper Series from RePEc, 2010.

Marsh, A. 2004. *Simultaneous Interpreting and Respeaking: A Comparison*. MA thesis. University of Westminster.

Marsh, A. 2006. “Respeaking for the BBC.” *Intralinea* http://www.intralinea.org/specials/article/Respeaking_for_the_BBC

Martínez Pérez, J., and Lopes, O. 2013. “Tool for Assessing Accuracy Rate of Live Subtitles.” Paper presented at the *4th International Symposium on Live Subtitling*, Barcelona, March.

Martínez Pérez, J. 2015. “New Approaches to Improve the Quality of Live Subtitling on TV.” Paper presented at t*he Respeaking, Live Subtitling and Accessibility Conference,* Rome, June 12.

Neubert, A. 1994. “Competence in translation: a complex skill, how to study and how to teach it.” In M. Snell-Hornby, F. Pöchhacker and K. Kaindl (Eds.) *Translation Studies. An Interdiscipline,* 411–420. Amsterdam: John Benjamins.

Ofcom. 2015. *Measuring live subtitling quality. Results from the fourth sampling exercise.*

Paas, F., Renkl, A. and Sweller, J. 2003. “Cognitive load theory and instructional design: recent developments.” *Educational Psychologist* 38(1). 1–4.

Pöchhacker, F. 2004. *Introducing Interpreting Studies*. London: Routledge.

Remael, A. and van der Veer, B. 2006. “Real-Time Subtitling in Flanders: Needs and Teaching.” *Intralinea*, Special Issue: Respeaking, 2006.

Robert, I. and Remael, A. (2017) “Assessing quality in live interlingual subtitling: a new challenge”. *Linguistica Antverpiensia New Series: Themes in Translation Studies 14.*

Romero Fresco, P. 2009. “Quality in Respeaking: The Reception of Respoken Subtitles.” Paper presented at the *Media for All 3* conference, Antwerp, October 22–24.

Romero-Fresco, Pablo. 2011. *Subtitling through Speech Recognition: Respeaking*. Manchester: St Jerome.

Romero-Fresco, P. 2012. “Respeaking in Translator Training Curricula. Present and Future Prospects.” *The Interpreter and Translator Trainer, 6(1), 91–112.*

Romero-Fresco, P. 2013. “Quality and NER in the UK.” Paper presented at the *4th International Symposium on Live Subtitling* in Barcelona, March.

Romero-Fresco, P., and Martínez Pérez, J. 2015. “Accuracy rate in live subtitling: the NER Model.” In J. Díaz Cintas and R. Baños Pinero (Eds.) *Audiovisual translation in a global context. Mapping an Ever-changing Landscape,* 28–50. London: Palgrave Macmillan.

Romero-Fresco, P. 2016. “Accessing communication: The quality of live subtitles in the UK.” *Language & Communication 49,* 56-69.

Romero-Fresco, P., and Pöchhacker, F. (2017) Quality assessment in interlingual live subtitling: The NTR model. *Linguistica Antverpiensia New Series: Themes in Translation Studies 14*.

Sandrelli, A. 2013. “Reduction strategies and accuracy rate in live subtitling of weather forecasts: a case study.” Paper presented at the *4th International Symposium on Live Subtitling* in Barcelona, March.

Shlesinger, M. 2000. “Strategic allocation of working memory and other attentional resources in simultaneous interpreting.” PhD diss., Bar Ilan University.

Szarkowska, A., Krejtz, K., Dutka, Ł., and Pilipczuk, O. 2016. „Cognitive load in intralingual and interlingual respeaking – a preliminary study.” *Poznan Studies in Contemporary Linguistics,* 52:2, 209–233.

Timarová, Š., Čeňková, I., and Meylaerts, R. 2015. “Simultaneous interpreting and working memory capacity.” In A. Ferreira and J. W. Schwieter (Eds.) *Psycholinguistic and Cognitive Inquiries into Translation and Interpreting,* 101–126. Amsterdam: John Benjamins.

[1] The term “one-speaker” refers to the number of people talking at the same time. Therefore, while there were many speakers in the news programme, none of them spoke simultaneously with others, so there was no overlapping speech.

[2] Results of these are reported elsewhere.

Are interpreters better respeakers?

Agnieszka Szarkowska[1], Krzysztof Krejtz[2], Łukasz Dutka[1], Olga Pilipczuk[1]

[1] Institute of Applied Linguistics, University of Warsaw, Warsaw, Poland

[2] SWPS University of Social Sciences and Humanities, Warsaw, Poland

Corresponding author:

a.szarkowska@uw.edu.pl

Abstract

In this study, we examined whether interpreters and interpreting trainees are better predisposed to respeaking than people with no interpreting skills. We tested 57 participants (22 interpreters, 23 translators and 12 controls) while respeaking 5-minute videos with two parameters: speech rate (fast/slow) and number of speakers (one/many). Having measured the quality of the respeaking performance using two independent methods: the NER model and rating, we found that interpreters consistently achieved higher scores than the other two groups. The findings are discussed in the context of transfer of skills, expert performance and respeaking training.

Keywords: respeaking, interpreting, live subtitling, rating, NER, audiovisual translation

1. Introduction

With the advent of modern technologies and their widespread use in audiovisual translation (AVT), new professions are emerging on the AVT market, like that of a respeaker.

A respeaker “listens to the original sound of a live programme or event and respeaks it, including punctuation marks and some specific features for the deaf and hard of hearing audience, to a speech recognition software, which turns the recognized utterances into subtitles displayed on the screen with the shortest possible delay” (Romero Fresco 2012, 93). Respeaking is now a major method of producing real-time subtitling to live TV broadcasts using speech recognition technology (Lambourne 2006; Marsh 2006).

Used since 2001, respeaking is a relative newcomer to the field of AVT. However, given its dynamic growth in the AVT industry, the need has arisen for higher education institutions to include respeaking in their study programmes, curricula and training courses (see Remael and van der Veer 2006; Romero-Fresco 2012). As aptly put by Hurtado Albir (2015, 257), “due to changes in the translation profession and constant academic and professional mobility, new curriculum designs that meet society’s demands and offer score for international harmonization are required.” As new respeakers are entering the market, respeaker trainers – both in the academia and in the industry – are now facing a number of new questions (see Arumí Ribas and Romero-Fresco 2008; Romero-Fresco 2012), including who to recruit, how to train newcomers to the profession, whether expert skills from other linguistic areas like interpreting can be transferred to respeaking, and finally who is better predisposed to becoming a respeaker? Studies on expert performance have shown that contrary to common belief, it is not innate abilities but rather skills acquired through supervised practice and experience that are key to achieving top performance in various areas of expertise (Ericsson and Charness 1994). Owing to a number of overlapping skills and competences between interpreters and respeakers, in this study we hypothesised that some interpreting skills may successfully be transferred to respeaking. If confirmed, this finding may improve our understanding of respeaker competences and also have important implications for interpreter and respeaker training.

Respeaking shares a number of characteristics with interpreting (Eugeni 2008, Marsh 2004). Similarly to simultaneous interpreting (see Jones 2002), respeaking involves rendering a live spoken linguistic input into another form while simultaneously following the speaker's words and monitoring one's own linguistic output. Romero-Fresco argues that "the two activities share clear time constraints, namely real-time production, little or no margin for correction or improvement" (2011, 46). Both interpreters and respeakers need to prepare for the job beforehand, for instance by conducting terminology searches and building up glossaries, etc. Romero-Fresco (2011) also stresses similarities in working conditions: in a booth with another colleague, taking turns every 30 minutes or so, speaking to a microphone and listening to the incoming input through a headset. In contrast to interpreting, respeaking is usually done intralingually. Interlingual respeaking, that is between two languages, does exist (de Korte 2006; den Boer 2001; Robert & Remael 2017; Romero-Fresco & Pöchhacker 2017), but at present it constitutes only a small, though admittedly growing, share of respeaking practices. Unlike in the case of interpreters, respeakers' intonation needs to be rather flat and monotonous, which makes it easier for the speech recognition software. This is why some respeakers with interpreting experience "may have to 'unlearn' certain interpreting skills, such as the ability to speak in a pleasant tone, which is not particularly effective when dictating to speech recognition programmes" (Romero Fresco 2012, 100). Apart from repeating and/or rephrasing the original text, respeakers also need to insert the necessary punctuation marks by voicing them (Romero-Fresco 2011). Whereas the interpreters' output is instantly consumed by end users listening to the interpreted text through earphones, the respeakers' output is further transformed into written text by speech recognition software and is later displayed as subtitles on viewers' television screens. In some countries, like the UK, respeakers are also required to monitor and correct the errors in the output of the speech recognition software, which adds even more to the pool of the competences they are required to have (Romero-Fresco 2011). In other countries, like France, proof-reading is done by another person.

The link between interpreting and respeaking skills has been previously noticed both by the academia and the AVT industry. When discussing respeaker competences, den Boer (2001, 172) argued that "the ideal person for the job would be someone who is a qualified interpreter and a professional subtitler". According to Romero Fresco (2012, 98), a US-based subtitling provider, *Mark Hall Associates*, appreciated memory skills in respeakers and therefore eagerly recruited sign language interpreters, whereas SWISS TXT, a language service provider for the public TV in Switzerland, found that "trained interpreters perform particularly well as respeakers, not least thanks to their capacity to make instant decisions regarding language usage and their ability to multitask". In spite of this theoretical interest in the overlap of respeaking and interpreting, to the best of our knowledge, no empirical studies have been conducted so far on respeaker competences and the transfer of skills from interpreting, which is the gap this study intends to fill.

2. Measuring quality in respeaking

While many countries are still struggling to achieve higher levels of subtitling quantity, others, like the UK, are now focussing their attention more on quality. The UK regulator, Ofcom, regularly monitors the quality of live subtitling in both public service broadcasters and private ones; it even set a quality threshold of 98% accuracy rate with the NER model for live subtitling in TV broadcasts (Ofcom 2015). The most challenging quality aspects of live subtitling through respeaking are accuracy, delay and presentation rates (Ofcom 2015). In this study we focus on accuracy by quantifying respeakers' performance using two methods: the NER model and rating.

2.1 The NER model

The most commonly used method of measuring accuracy in live subtitling through respeaking is based on the NER model (see Martínez Pérez 2015; Ofcom 2014; Romero-Fresco 2009; Romero Fresco 2013; Romero Fresco and Martínez Pérez 2015; Romero Fresco 2016). Drawing on its predecessors, the WER ('word error rate') and CRIM (Centre de Recherche Informatique de Montréal) methods (see Boulianne et al. 2009; Romero Fresco 2011), the NER model is based on the following formula (Romero-Fresco and Martínez Pérez 2015):

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

N is the number of words in the respoken text, including different commands such as punctuation; E is the number of edition errors, and R is the number of recognition errors. Edition errors are usually a result of incorrect decisions taken by the respeaker, such as omission or introducing wrong information. Recognition errors are usually caused by misrecognition, by the software or by respeaker's mispronunciation. There are three types of recognition errors: insertions, deletions, or substitutions. Correct editions are instances where the respeaker edited the text without any loss of information, for instance in the case of removing unnecessary false starts or repetitions. Both edition and recognition errors may be classified as serious, standard or minor, and be allocated a 1, 0.5 and 0.25 score, respectively (Romero Fresco and Martínez Pérez 2015).

In order to facilitate the quality measurement of respoken subtitles, NERstar tool was designed by Juan Martínez Pérez on behalf of SWISS TXT, and developed in close collaboration with VerbaVoice and Pablo Romero-Fresco. The tool requires a human operator to manually go through the respoken subtitle file (.srt) and compare it against a transcription of the original text. The operator marks edition and recognition errors as well as punctuation errors, allocating them a score (0.25, 0.5 or 1). The software then automatically calculates the final accuracy rate as well as a number of other parameters, including the proportion of edition and recognition errors. We used the NERstar tool to calculate the accuracy rates of subtitles respoken intralingually by the participants of our study; the interlingual respeaking task could not be measured with the NER model or the NERstar tool, as they are inadequate for translation between languages.

We also used NERstar to calculate the degree of reduction between the original text and its respoken version. Reduction, which usually takes a form of omission or paraphrase, is a typical feature of live subtitling through respeaking. Respeakers are rarely able to provide verbatim subtitles, be it due to fast speech rates, overlapping speech, a production delay between the original broadcast and live subtitles or limited cognitive resources of respeakers (for more on reduction in respeaking see Luyckx et al. (2010) and Sandrelli (2013)). In our study, we examined reduction rates between the three participant groups with a view to finding whether participants from the interpreting group will be able to follow the original text more closely and have lower reduction rates, possibly thanks to their trained working memory.

2.2 Rating

Apart from the NER model, we also used another method to assess the quality of respeaking done by our participants: rating. Three independent raters, unaware of study hypotheses, were asked to rate each respoken clip. They were provided with the original video and its transcription as well as an audio recording of the respeaking and its transcription, produced manually by a human transcriber. After a training session with detailed guidelines, they were asked to rate the quality of the respoken text on a Likert-type 1-5 scale, where 1 was the lowest possible score and 5 – the highest.

The rates rated the respeaking quality on six parameters:

1. Overall quality
2. Fluency of delivery – where the lowest score (1) meant a lot of hesitations, false starts, repetitions, unfinished sentences, and significant variance in the speed of delivery, while the highest score (5) was meant to be used in the case of fluent delivery with (almost) no false starts, repetitions, sentences were complete, constant speed of delivery, clear division into respeaking units
3. Pronunciation – where the lowest score (1) was applied when most words were not pronounced clearly, were joined together (sounds from the previous words carried to the next), most word endings were not pronounced clearly, the loudness varied significantly, while the highest score (5) was applied to cases with good enunciation, where words could be heard clearly and were distinguishable from one another, words endings were fully pronounced, loudness was roughly the same throughout the recording
4. Punctuation – where the lowest score (1) was given to cases where no or almost no punctuation marks were used, while the highest score (5) was given to cases where all or almost all punctuation marks were present, sentences were finished with a full stop
5. Spoken to written language - where the lowest score (1) was allocated to cases where all the unnecessary features of spoken language of the original dialogue were retained, while the highest score (5) was allocated to cases where all the unnecessary features of spoken language were edited out
6. Accuracy of content – where the lowest score (1) was assigned to cases where most ideas from the original video were incorrectly or were omitted, with significant misrepresentations of semantic content, most proper names and numbers were transferred incorrectly or omitted, while the highest score (5) was assigned to cases all or almost all ideas from the original were rendered correctly and fully, with no

significant omissions or misrepresentation of semantic content, and proper names and numbers correctly transferred.

2.3 Hypotheses

Given the skills overlapping with respeaking they had already acquired in interpreting training and practice, we expected participants from the interpreting group to achieve highest accuracy rates in the NER model and highest rating scores, particularly in the interlingual task. We also hypothesised that interpreters would have lower reduction rates, as they would be able to retain more information in their working memory (see Baddeley and Hitch 1974; Timarová, Čeňková and Meylaerts 2015), again thanks to their expertise in interpreting. We also expected interpreters to be rated higher in terms of pronunciation, as this is one of the areas they are trained in (Giles 2009; Pöchhacker 2004). Finally, we thought the control group, i.e. people with no interpreting or translation experience, would achieve lowest results due to the lack of any linguistic training, which may be reflected in such rating parameters as punctuation.

3. Method

3.1 Materials

Five videos representing different TV genres approx. 5 minutes long each were used in the study. Four of them were in Polish (intralingual respeaking task) and one was in English (interlingual respeaking task). All clips were self-contained and no previous knowledge was required to understand them.

As shown in Table 1, the clips in the intralingual respeaking task were matched in terms of speech rate (slow and fast) and the number of speakers (one or many).

Insert Table 1.

Table 2 presents more detailed characteristics of both tasks and clips. The slow one-speaker video was a New Year's Address delivered by former Polish Prime Minister Ewa Kopacz in 2015. It was pre-scripted and read out from a prompter. The video was not visually complex, as it only featured a shot with the speaker in her office.

The fast one-speaker[1] video was a fragment of the evening news programme *Fakty* related to a conflict between the Polish government and miners. Apart from the news anchor, the visuals in this programme also contained some infographics and split screen. News programmes are usually characterised by fast speech rates and condensed meaningful content, which makes it "particularly challenging to edit without losing important information" (Ofcom 2014, 27).

The slow video with many speakers was a fragment of an entertainment show *Fakty po faktach* with a host journalist discussing a recent Oscar award with an actress and a film critic. The dialogue in this video was not pre-scripted and it involved some overlapping speech.

The fast video with many speakers was a political chat show *Kropka nad i* with a host journalist and her two guests: politicians from two opposite ends of the political scene. The video contained a lot of overlapping speech between the three speakers. Chat shows are among the most difficult genres to respeak due to their high speech rates and overlapping speech (Ofcom 2014).

Finally, the interlingual respeaking task was a fragment of the speech delivered by President Barrack Obama on the Freedom Day anniversary in June 2014 on the occasion of Poland's celebrating 25 years since the fall of communism.

Insert Table 2.

3.2 Participants

A total of 57 participants (50 women, 7 men) were tested in the study. Participants were recruited at the Institute of Applied Linguistics at the University of Warsaw, where the study took place, the Faculty of English at Adam Mickiewicz University in Poznań, as well as through social media (AVT Lab and RespeakingProject Facebook pages) and personal contacts. They were all volunteers who went through a two-day respeaking training conducted by experienced respeaker trainers from Switzerland, UK and Italy.

The mean age of the participants was 27.48 (*SD*=5.71). Based on their self-reported linguistic and professional background, including experience in translation/interpreting (see Table 3), they were divided into three groups: interpreters and interpreting trainees (*N*=22), translators and translation trainees (*N*=23), and control group with people with no translation/interpreting background (*N*=12).

Insert Table. 3.

3.3 Procedure

Participants were tested individually in a research lab. Prior to commencing the tests, written informed consent was obtained from each participant.

During the respeaking test, participants' eye movements were monitored using the screen recording functionality of SMI Red eye tracker at 120 Hz and their brain activity was recorded using Emotiv EEG at 128 Hz[2]. FAB Subtitler Live was used to display the respoken subtitles on the screen. Participants' output was audio recorded and later transcribed manually for further analysis. Each participant worked on their own voice profile in the speech recognition software for Polish manufactured by Newton Technologies. The profiles were created during the respeaking workshops.

Each respeaking test began with an explanation of the procedure. Participants went through a short mock respeaking task to familiarise themselves with the procedure; the data for this task were not recorded. The proper respeaking test consisted of four randomised intralingual videos, followed by the interlingual task. After each task, the participant had to answer a few questions related to their self-reported cognitive load (assessing the difficulty of the task, its pace as well as their mental effort and concentration, see Szarkowska et al. 2016). The test ended with a short semi-structured interview.

3.4 Design

The present study was 3 x 4 mixed-design experiment with one between-subject fixed factor (group type: control vs. interpreters vs. translators), and the clip type as a within-subjects factor. Additionally, we also controlled for subjects' working memory capacity (WMC). As the indicator of WMC, we used the traditional score of reading span task (Conway, Kane, Bunting, Hambrick, Wilhelm & Engle, 2005). The main dependent variables were: respeaking accuracy as measured by the NER value, reduction rate and a series of rating parameters: overall quality, fluency of delivery, pronunciation, punctuation, spoken to written language, and accuracy of content.

4. Results

Below we present results of respeaking performance of the three participant groups: first, as measured using the NER model, and second, the rating method. As the NER model and the NERstar tool were developed for intralingual respeaking only and cannot be used between two different languages, we do not report the results of the interlingual respeaking task using this quality measurement method. The interlingual task is only reported in the rating results.

4.1 NER value

The mixed-design Analysis of Covariance (ANCOVA) was conducted on NER value as the dependent variable. The sample group was treated as a between-subject factor and the clip type as a within-subject factor. The score of the reading span task was an indicator of working memory capacity (WMC) and was treated as a covariant in the analysis. The ANCOVA was followed by pairwise comparisons with Tukey correction whenever necessary.

The results of the analysis show two statistically significant main effects: of the group type, $F(2,32) = 3.46$, $p < 0.05$, *eta-squared* = 0.12, and of WMC, $F(1,32) = 15.27$, $p < 0.001$, *eta-squared* = 0.22. The highest NER value was achieved by the interpreting group in all clips ($M = 94.36$, $SE = 0.52$), compared to the translators ($M = 92.65$, $SE = 0.50$) and to control group ($M = 93.47$, $SE = 0.58$). Pairwise comparisons showed that interpreters differed significantly from the translators ($p < 0.05$), but not from the control group. The difference between the translators and the control group was not statistically significant.

The main effect of WMC can be interpreted with the simple correlation coefficient, $r = 0.41$, $t(142) = 5.51$, $p < 0.001$, which showed that the higher the working memory capacity, the higher the NER value achieved by the participants.

4.2 Edition Errors

An analogous ANCOVA analysis was conducted for the number of edition errors as a dependent variable. The analysis revealed a main effect of the clip type, $F(2.49, 79.67) = 69.61$, $p < 0.001$, *eta-squared* = 0.57, which was statistically significant and strong in terms of explained variance. The following post-hoc tests with Tukey correction showed that the political chat show (i.e. the fast clip with many speakers) ($M = 69.23$, $SE = 2.52$) differed significantly ($p < 0.001$) from news ($M = 35.43$, $SE = 2.52$), entertainment chat show ($M = 28.83$, $SE = 2.52$), and speech ($M = 24.90$, $SE = 2.52$). All other comparisons were not statistically significant.

4.3 Reduction rate

The reduction rate was analysed with an analogous ANCOVA described above. The analysis revealed two statistically significant main effects of the group type, $F(2,32) = 6.35$, $p < 0.01$, *eta-squared* = 0.25, and clip type, $F(83.43, 27.59) = 325.84$, $p < 0.001$, *eta-squared* = 0.64.

Pairwise comparisons with Tukey correction for the main effect of group type showed that the smallest percentage of words removed from the original text was found in the interpreting group (*M* = 27.87, *SE* = 2.56), which was significantly ($p < 0.01$) different from those in the translators group (*M* = 39.79, *SE* = 2.51), but not significantly different from the control group (*M* = 31.48, *SE* = 2.90). The difference between the control group and the translators was also not significant ($p > 0.1$).

The post-hoc tests for the main effect of the clip type showed that the highest reduction rate was for the fast political chat show with many speakers (*M* = 50.40, *SE* = 1.65) and the lowest for the slow one-speaker speech (*M* = 12.87, *SE* = 1.65). Medium values were found in the news (*M* = 37.56, *SE* = 1.65) and the slow entertainment clip with many speakers (*M* = 31.34, *SE* = 1.65). All comparisons between clips used in the study on reduction rate were statistically significant ($p < 0.001$).

The analysis also showed no significant effect of working memory capacity nor any interaction term.

4.3 Rating score

Before running statistical analyses, the raters agreement was calculated for each rating scale. The overall mean agreement was relatively high (*M* = 83.63%), as well as for fluency (*M* = 91.58%), pronunciation (*M* = 87.36%), punctuation (*M* = 79.66%), and accuracy of content (*M* = 91.22%). Only the level of inter-rater agreement on evaluation of spoken-to-written transfer (*M* = 67.38%) yields more careful interpretation of the following results on this variable.

A series of two-way ANCOVA analyses with the group type as a between-subject factor and clip type as a within-subject factor and WMC as a covariant was performed to test the rating scores. The dependent variables in the following analyses were as follows: overall rating, rating of fluency, pronunciation, punctuation, spoken-to-written transfer, and accuracy of content. The results of these analyses are presented below.

The ANCOVA on **overall rating of respeaking quality** as a dependent variable showed a significant main effect of clip type, $F(2.95, 100.33) = 44.80$, $p < 0.001$, *eta-squared* = 0.32, and of WMC, $F(1,34) = 30.36$, $p < 0.001$, *eta-squared* = 0.362. The overall mean rating score was the lowest for the fast political chat show with many speakers ($M = 2.95$, $SE = 0.09$) and it was the highest for the slow one-speaker speech ($M = 3.92$, $SE = 0.09$). These two clips differed significantly between each other ($p < 0.001$). They also differed significantly ($p < 0.001$) from the slow many-speaker entertainment show ($M = 3.68$, $SE = 0.09$) and the news ($M = 3.44$, $SE = 0.09$). However, pairwise comparisons showed that only the entertainment show and the slow speech were marginally significantly different from each other ($p = 0.052$). The significant main effect of working memory capacity means that the higher WMC score, the higher the overall rating, $r = 0.51$, $t(150) = 7.32$, $p < 0.001$.

The analysis on **respeaking accuracy** revealed several interesting statistically significant effects. First, we found the main effect of clip type, $F(2.86, 97.07) = 48.73$, $p < 0.001$, *eta-squared* = 0.31. The best accuracy score was achieved for the slow one-speaker speech ($M = 3.97$, $SE = 0.10$), which was significantly ($p < 0.001$) higher than for the news ($M = 3.23$, $SE = 0.10$), and fast political chat show with many speakers ($M = 2.86$, $SE = 0.10$) but not different from the slow entertainment show ($M = 3.75$, $SE = 0.10$). The accuracy score for the fast political chat show with many speakers was significantly ($p < 0.001$) lower in comparisons with any other clip used in the study.

The analysis also showed a significant main effect of WMC, $F(1, 34) = 9.07$, $p < 0.01$, *eta-squared* = 0.15. The following correlation analysis showed, in line with the expectations, that the higher WMC, the higher accuracy of respeaking, $r = 0.32$, t(150) = 4.17, $p < 0.001$. Interestingly, this relation was moderated by the clip type. The interaction effect of WMC and clip type was found, $F(2.86, 97.07) = 2.78$, $p < 0.05$, *eta-squared* = 0.03. The influence of WMC on respeaking accuracy differed between clips used in the study. The strongest relation was for the fast political chat show with many speakers, *r* = 0.51, t(36) = 3.53, *p* < 0.01, and the lowest (not statistically significant) for the news, *r* = 0.21, *t*(36) = 1.29, *p* > 0.1. For the slow speech, the relation was significant, *r* = 0.49, *t*(36) = 3.40, *p* < 0.01, and for the slow entertainment show it was only marginally significant, *r* = 0.28, *t*(36) = 1.78, *p* = 0.08, see Figure 1.

Insert Fig. 1 here

Fig. 1. The interaction effect of clip type and working memory capacity on respeaking accuracy measured by rating

The analysis also revealed a marginally significant interaction effect of the clip type and group type, *F*(97.07 0.17) = 1.92, *p* = 0.08, *eta-squared* = 0.03. The means and standard errors are presented in Figure 2.

Insert Fig. 2 here

Fig. 2. The interaction effect of the clip type and group type on respeaking accuracy evaluation by rating. The error bars represent +/- 1 standard error for the estimated means

The ANCOVA analysis on **fluency of delivery** also revealed a significant main effect of clip type, *F*(2.61, 88.73) = 37.44, *p* < 0.001, *eta-squared* = 0.28 and the main effect of WMC, *F*(1, 34) = 14.05, *p* < 0.001, *eta-squared* = 0.21. Fluency was highest for the speech (*M* = 4.04, *SE* = 0.09), which was significantly (*p* < 0.02) higher than for the news (*M* = 3.74, *SE* = 0.09) and the fast political chat show with many speakers (*M* = 3.12, *SE* = 0.09, p < 0.001), but not different compared to the slow entertainment show with many speakers (*M* = 3.91, *SE* = 0.09). The fluency evaluation for the fast political chat show with many speakers reached significantly (*p* < 0.001) lower scores than the evaluations for other clips used in the study.

The analyses also showed a marginally significant interaction effect of the experimental group type and clip type, $F(5.22, 88.73) = 1.97$, $p = 0.09$, *eta-squared* = 0.04. Pairwise comparisons with Tukey correction revealed that in all experimental groups the fluency of respeaking significantly ($p < 0.001$) differed between the fast political chat show with many speakers and all other clips used in the study. However, only interpreters achieved significantly higher fluency scores while respeaking the speech ($M = 4.16$, $SE = 0.15$) compared to the news ($M = 3.65$, $SE = 0.15$). The means and standard errors for the interaction term are presented in Figure 3.

Insert Fig. 3 here

Fig. 3. The interaction effect of group type and clip type on fluency of respeaking evaluation by rating. The error bars represent +/- 1 standard error for the estimated means.

The ANCOVA analysis on spoken-to-written transfer rating revealed only a main effect of clip type, $F(2.52, 85.78) = 10.00$, $p < 0.001$, *eta-squared* = 0.10. The highest score was achieved when respeaking the news ($M = 3.59$, $SE = 0.10$), which was significantly ($p < 0.05$) higher than for the speech ($M = 3.07$, $SE = 0.10$) and for the fast political chat show ($M = 3.31$, $SE = 0.10$), but not significantly different than for the slow entertainment show ($M = 3.53$, $SE = 0.10$). The analysis of the pronunciation parameter again showed a significant effect of clip type, $F(2.87, 97.43) = 3.28$, $p < 0.05$, *eta-squared* = 0.03, and a main effect of WMC, $F(1, 34) = 13.99$, $p < 0.001$, *eta-squared* = 0.22. The following correlation test showed that the higher WMC, the better the pronunciation parameter, $r = 0.45$, $t(150) = 6.18$, $p < 0.001$. Pairwise comparisons of clips on the pronunciation parameter showed that the only significant difference (p < 0.02) was between the slow entertainment show ($M = 4.24$, $SE = 0.09$) and the news ($M = 3.98$, $SE = 0.09$).

Finally, the analysis of the punctuation rating score revealed three significant effects. The main effect of clip type, *F*(2.65, 90.19) = 8.36, *p* < 0.001, *eta-squared* = 0.05. The news had a significantly lower score (*M* = 3.77, *SE* = 0.11) than the slow entertainment show (*M* = 4.01, *SE* = 0.11) and the speech clip (*M* = 4.11, *SE* = 0.11). The speech was significantly better in terms of punctuation than the fast political chat show with many speakers (*M* = 3.74, *SE* = 0.11). The main effect of WMC, *F*(1, 34) = 10.37, *p* < 0.01, *eta-squared* = 0.19 means that the higher WMC, the better punctuation score, *r* = 0.42, *t*(150) = 5.65, *p* < 0.001.

The analysis also showed a significant interaction effect of the clip type and group type, *F*(5.31, 90.19) = 3.45, *p* < 0.01, *eta-squared* = 0.04. The results of post-hoc tests showed that in the control group a significant difference was found between the fast political chat show (*M* = 3.38, *SE* = 0.23) and the slow entertainment show (*M* = 4.03, *SE* = 0.23); in the translators group: the speech (*M* = 4.38, *SE* = 0.19) had a significantly higher punctuation score than the fast political chat show with many speakers (*M* = 3.80, *SE* = 0.19), and in the interpreters group the speech (*M* = 4.23, *SE* = 0.19) again was significantly better in terms of punctuation compared to the news (*M* = 3.70, *SE* = 0.19), see Figure 4.

Insert Fig. 4 here

Fig. 4. The interaction effect of clip type and group type on the punctuation evaluation score by rating. The error bars represent +/- 1 standard error for the estimated means.

Interlingual respeaking

Finally, we hypothesised that interpreters are going to perform better and produce more accurate respeaking especially when performing the interlingual task. In order to test this prediction, we conducted two 3 x 2 Analysis of Variance with the between-subject design. The first factor was the experimental group (interpreters vs. translators vs. control) and the second factor was the working memory capacity score divided into two categories with the median split (low WMC vs. high WMC). For the purpose of these analyses, only the data from the interlingual task were selected.

The analysis on overall rating of respeaking quality revealed only a marginally significant main effect of the group, *F*(2,37) = 2.65, *p* = 0.08, *eta-squared* = 0.13. The following not-adjusted pairwise comparisons showed that interpreters (*M* = 3.15, *SE* = 0.33) performed significantly (*p* < 0.05) better than control group (*M* = 2.00, *SE* = 0.44) and marginally significantly (*p* = 0.09) better than translators (*M* = 2.36, *SE* = 0.31). The difference between translators and control group was not statistically significant (*p* > 0.1).

The ANOVA on respeaking accuracy revealed two marginally significant effects: the main effect of the group, *F*(2,37) = 2.97, *p* = 0.06, *eta-squared* = 0.14 and the interaction effect of the group and working memory capacity, *F*(2,37) = 2.78, *p* = 0.08, *eta-squared* = 0.13 (see Figure 5). The following pairwise comparisons for the main effect of the group with no *p*-value adjustment showed that interpreters achieved significantly (*p* < 0.05) higher accuracy rating scores (*M* = 2.98, *SE* = 0.30) than translators (*M* = 2.14, *SE* = 0.28) and control group (*M* = 1.94, *SE* = 0.40). Translators did not differ significantly from the control group (*p* > 0.1).

The interaction effect suggests that the difference in respeaking accuracy between the groups was moderated by the participants' working memory capacity. The following pairwise comparisons with no p-value adjustment revealed that when low on WMC, interpreters achieved significantly (*p* < 0.01) higher accuracy ratings (*M* = 3.33, *SE* = 0.45) than translators (*M* = 1.57, *SE* = 0.38) and marginally higher results than the control group (*M* = 2.13, *SE* = 0.53). The difference between the translators and control group participants low on WMC was not statistically significant (*p* > 0.1). Interestingly, high working memory capacity compensates for the interpreters skills/experience such that the differences between the groups of participants high on WMC were not statistically significant. Moreover, the results showed that translators with high working memory capacity achieved marginally (*p* = 0.05) higher accuracy scores (*M* = 2.71, *SE* = 0.42) than translators who were low on working memory capacity. The pattern of mean of the interaction effect is presented on Figure 5.

Insert Fig. 5 here

Fig. 5. The interaction effect of the group and working memory capacity on the accuracy evaluation score by rating. The error bars represent +/- 1 standard error for the estimated means.

5. Discussion

In this study, we wanted to find whether people with interpreting expertise have some advantage over translators and people with no interpreting skills when performing a series of respeaking tasks. We tested professional interpreters and interpreting trainees, comparing their performance with that of translators and translation trainees as well as the control group. The performance was measured using two independent methods: the NER model and rating.

In both methods it was the interpreting group that achieved the highest scores. We think this result may be regarded as evidence of a transfer of skills from interpreting to respeaking. Interpreting skills relevant to respeaking include previous training and experience in dealing with simultaneous input and monitoring one's own output, working memory trained in following spoken input and dividing it into comprehensible chunks, divided attention, knowing how to cope with stress, etc. (see Pöchhacker 2004; Gile 2009) as well as the "ability to develop emergency strategies when source text is not understood" (Romero Fresco 2011, 52), being able to identify and select relevant information, condense the original if necessary, keep up with fast speech rates, and deduce the meaning from a wider context. All these may have helped participants from the interpreting group in achieving top performance, discernible both in the NER value and rating scores.

The NER analysis also showed that the interpreters were not only the best in achieving highest respeaking quality, but they also had the lowest reduction rates. Being familiar with the ‘salami technique’ (Jones 1998; Romero Fresco 2011), interpreters may have skilfully chunked the incoming input, which helped them cope with the source text with a lesser need for reduction. This result may also be attributed to the skills that the interpreters developed during interpreting training and further strengthened in their professional experience, such as anticipation, concurrent management of two speech channels (i.e. listening and speaking at the same time) and managing their voice as a professional tool, which allowed them to re-express more of the text by speaking faster for longer periods of time while still articulating clearly.

Back in 1952, Herbert stated that “A good interpreter must be a trained public speaker” (cited after Pöchhacker 2004, 125). Indeed, during interpreting training much emphasis is placed on the fluency of delivery, voice quality, faithfulness and completeness of utterances (Pöchhacker 2004; Buhler 1986). Both interpreters and respeakers need to “express thoughts clearly and concisely” (Romero-Fresco 2011, 53), have a good diction and a rich vocabulary as well as avoid hesitations, false starts and repetitions. It is therefore not surprising that the highest scores in the pronunciation and fluency of delivery parameters in rating were achieved by the interpreting group. The largest difference between the control group and the other two groups, i.e. people who were trained in linguistics, was in punctuation in the rating scores. This may point to a higher linguistic awareness of punctuation rules, as translators and interpreters work directly with language on a daily basis.

As for the differences between the clips, in line with our predictions, the fast clip with many speakers turned out to be the most difficult to respeak given its high speech rates and a great deal of overlapping speech. This clip was found to have the highest edition errors and reduction rates as well as the lowest rating scores. We believe that the fast speech rates and overlapping speech caused a strain on respeakers' cognitive load and working memory, leaving little or no time to think of inserting appropriate punctuation marks (hence the lowest rating score in the punctuation parameter) and delivering a fluent and accurate respoken text (as shown by the rating results). Highest scores were attained in slow clips, particularly the slow one-speaker speech, which also had the fewest edition errors.

One of the most important findings of this study is a strong link between respeaking quality and working memory capacity (Baddeley and Hitch 1974). We found that higher WMC correlated positively with higher respeaking accuracy as measured by both the NER model and the rating. This correlation was particularly strong in the case of the most difficult clip in the study, i.e. the fast video with many speakers (the political chat show). WMC also positively affected the rating of the overall quality, accuracy, fluency of delivery and punctuation parameters. This means that people with high WMC could deliver better quality respeaking by being more accurate and fluent as well as better able to control punctuation.

Low capacity of the working memory may result in cognitive overload, which in turn tends to translate into poor performance (Paas et al. 2003), as was indeed the case in our study. The load on working memory can be reduced by developing automated schemas in long-term memory (Kalyuga et al. 1999). Such schemas are often used to explain differences between experts and novices (see Kalyuga et al. 2003); in this study, experts are understood as people with interpreting experience and novices as people without such experience. Given that working memory capacity was found to be highest in the interpreting group, we believe that interpreters may have developed schemas in their long-term memory (Kirschner et al. 2011), enabling them to surpass the limitations of the working memory (see Shlesinger 2000). This might have helped them to better manage their ear-voice span (the delay between hearing the audio and respeaking it) and it could have resulted in the lower cognitive load they experienced during respeaking (Szarkowska et al. 2016), which in turn translated into better performance.

When it comes to the interlingual respeaking task, we found that the interpreting group scored highest in the ratings, although the differences were not as pronounced as we had expected them to be. Interestingly, similarly to the intralingual tasks, we observed a strong link between WMC and respeaking quality. People who have a high WMC performed consistently better as respeakers, regardless of whether they are interpreters or not. However, among those with lower WMC, it was the interpreting group who achieved higher scores. This shows that although working memory has an undeniable impact on the quality of respeaking, other factors and skills from the interpreting world positively affect respeaking performance.

Our study also showed parallels between more objective method of assessing respeaking accuracy (NER) and a more subjective method, using different human subjects as raters. We believe this adds significance to the NER model, confirming its usefulness as a major method of assessing accuracy in intralingual respeaking.

Limitations of the study

This study is not free from limitations, including a small selection of audiovisual material and a limited number of participants. Other studies could possibly replicate our research on other languages, participants and audiovisual materials.

Conclusion

To the best of our knowledge, this was the first empirical study on the competences of interpreters in relation to respeaking. We hope to have shown that being an interpreter may be advantageous in becoming a respeaker, particularly when it comes to better retention of the original input thanks to trained working memory, endurance in dealing with multiple sources of information and communicating the complex spoken content fluently and accurately to the target audience.

This finding has some considerable implications and practical applications. One of them is the inclusion of respeaking in the training of interpreters. This could add to their pool of skills, making them more versatile and flexible on the quickly changing translation market. Another consequence of establishing closer links between respeaking and interpreting is to support respeaker trainers with training tasks which are already well-established and proven in the field of interpreting, particularly WMC techniques. Merging interpreting and respeaking training may also help in preparing trainees to become interlingual respeakers – a profession which is now niche, but may develop in the near future. In this way, higher education institutions may provide competent workforce for the quickly changing translation, interpreting and media accessibility market.

Acknowledgements

This work was supported by the National Science Centre Poland under Grant DEC-2013/11/B/HS2/02762.

References

Arumí Ribas, M., and Romero Fresco, P. 2008. "A Practical Proposal for the Training of Respeakers." *The Journal of Specialised Translation, 10*, 106–127.

Baddeley, A. D. and Hitch, G. 1974. "Working memory." In G. H. Bower (ed.) *The Psychology of Learning and Motivation: Advances in research and theory,* 47–89. New York: Academic Press.

Boulianne, G., J.-F. Beaumont, Boisvert, M., Brousseau, J., Cardinal, P., Chapdelaine, C., Comeau, M., Ouellet, P., Osterrath, F., and Dumuchel, P. 2009. "Shadow Speaking for

Real-Time Closed-captioning of TV Broadcasts in French." In A. Matamala and P. Orero (Eds.) *Listening to Subtitles. Subtitles for the Deaf and Hard of Hearing*, 191–208). Bern: Peter Lang.

Bühler, H. 1986. "Linguistic (Semantic) and Extra-linguistic (Pragmatic) Criteria for the Evaluation of Conference Interpretation and Interpreters." *Multilingua* 5(4), 231–5.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., and Engle, R. W. (2005). "Working memory span tasks: A methodological review and user's guide." *Psychonomic Bulletin and Review*, 12, 769 - 786.

de Korte, T. 2006. "Live inter-lingual subtitling in the Netherlands. Historical background and current practice." *Intralinea*. http://www.intralinea.org/specials/article/Live_inter-lingual_subtitling_in_the_Netherlands

den Boer, C. 2001. "Live interlingual subtitling." In Y. Gambier and H. Gottlieb (Eds.) *(Multi)media translation: concepts, practices and research,* 167–172. Amsterdam/Philadelphia, John Benjamins.

Ericsson, K. A. and Charness, N. 1994. "Expert performance. Its structure and acquisition." *American Psychologist,* 49(8), 725–747.

Eugeni, C. 2008. "A Sociolinguistic Approach to Real-time Subtitling: Respeaking vs. Shadowing and Simultaneous Interpreting." *English in International Deaf Communication* 72, 357–382.

Gile, D. 2009. *Basic Concepts and Models for Interpreter and Translator Training*. Amsterdam/New York: John Benjamins.

Hurtado Albir, A. 2010. "Competence." In Y. Gambier and L. vad Doorslaer (Eds.) *Handbook of Translation Studies*, 55–59. Amsterdam: John Benjamins.

Jones, R. 2002. *Conference interpreting explained*. Manchester: St. Jerome Publishing.

Kalyuga, S., Chandler, P. and Sweller, J. 1999. “Managing split attention and redundancy in multimedia instruction.” *Applied Cognitive Psychology*, 13. 351–371.

Kalyuga, S., Ayres, P., Chandler, P. and Sweller, J. 2003. “The expertise reversal effect.” *Educational Psychologist* 38(1). 23–31.

Kirschner, F., L. Kesterand G. Corbalan. 2011. “Cognitive load theory and multimedia learning, task characteristics, and learner engagement: The current state of the art.” *Computers in Human Behavior*, 27(1). 1–4.

Lambourne, A. 2006. “Subtitle respeaking. A new skill for a new age.” *Intralinea*. http://www.intralinea.org/specials/article/Subtitle_respeaking

Luyckx, B., Delbeke, T., Van Waes, L., Leijten, M., and Remael, A. 2010. “Live subtitling with speech recognition. Causes and consequences of text reduction.” *IDEAS Working Paper Series from RePEc*, IDEAS Working Paper Series from RePEc, 2010.

Marsh, A. 2004. *Simultaneous Interpreting and Respeaking: A Comparison*. MA thesis. University of Westminster.

Marsh, A. 2006. “Respeaking for the BBC.” *Intralinea* http://www.intralinea.org/specials/article/Respeaking_for_the_BBC

Martínez Pérez, J., and Lopes, O. 2013. “Tool for Assessing Accuracy Rate of Live Subtitles.” Paper presented at the *4th International Symposium on Live Subtitling*, Barcelona, March.

Martínez Pérez, J. 2015. “New Approaches to Improve the Quality of Live Subtitling on TV.” Paper presented at t*he Respeaking, Live Subtitling and Accessibility Conference*, Rome, June 12.

Neubert, A. 1994. “Competence in translation: a complex skill, how to study and how to teach it.” In M. Snell-Hornby, F. Pöchhacker and K. Kaindl (Eds.) *Translation Studies. An Interdiscipline*, 411–420. Amsterdam: John Benjamins.

Ofcom. 2015. *Measuring live subtitling quality. Results from the fourth sampling exercise.*

Paas, F., Renkl, A. and Sweller, J. 2003. “Cognitive load theory and instructional design: recent developments.” *Educational Psychologist* 38(1). 1–4.

Pöchhacker, F. 2004. *Introducing Interpreting Studies*. London: Routledge.

Remael, A. and van der Veer, B. 2006. “Real-Time Subtitling in Flanders: Needs and Teaching.” *Intralinea,* Special Issue: Respeaking, 2006.

Robert, I. and Remael, A. (2017) “Assessing quality in live interlingual subtitling: a new challenge”. *Linguistica Antverpiensia New Series: Themes in Translation Studies 14.*

Romero Fresco, P. 2009. “Quality in Respeaking: The Reception of Respoken Subtitles.” Paper presented at the *Media for All 3* conference, Antwerp, October 22–24.

Romero-Fresco, Pablo. 2011. *Subtitling through Speech Recognition: Respeaking*. Manchester: St Jerome.

Romero-Fresco, P. 2012. “Respeaking in Translator Training Curricula. Present and Future Prospects.” *The Interpreter and Translator Trainer, 6(1), 91–112.*

Romero-Fresco, P. 2013. “Quality and NER in the UK.” Paper presented at the *4th International Symposium on Live Subtitling* in Barcelona, March.

Romero-Fresco, P., and Martínez Pérez, J. 2015. “Accuracy rate in live subtitling: the NER Model.” In J. Díaz Cintas and R. Baños Pinero (Eds.) *Audiovisual translation in a global context. Mapping an Ever-changing Landscape,* 28–50. London: Palgrave Macmillan.

Romero-Fresco, P. 2016. “Accessing communication: The quality of live subtitles in the UK.” *Language & Communication 49,* 56-69.

Romero-Fresco, P., and Pöchhacker, F. (2017) Quality assessment in interlingual live subtitling: The NTR model. *Linguistica Antverpiensia New Series: Themes in Translation Studies 14.*

Sandrelli, A. 2013. “Reduction strategies and accuracy rate in live subtitling of weather forecasts: a case study.” Paper presented at the *4th International Symposium on Live Subtitling* in Barcelona, March.

Shlesinger, M. 2000. “Strategic allocation of working memory and other attentional resources in simultaneous interpreting.” PhD diss., Bar Ilan University.

Szarkowska, A., Krejtz, K., Dutka, Ł., and Pilipczuk, O. 2016. „Cognitive load in intralingual and interlingual respeaking – a preliminary study.” *Poznan Studies in Contemporary Linguistics*, 52:2, 209–233.

Timarová, Š., Čeňková, I., and Meylaerts, R. 2015. “Simultaneous interpreting and working memory capacity.” In A. Ferreira and J. W. Schwieter (Eds.) *Psycholinguistic and Cognitive Inquiries into Translation and Interpreting*, 101–126. Amsterdam: John Benjamins.

[1] The term “one-speaker” refers to the number of people talking at the same time. Therefore, while there were many speakers in the news programme, none of them spoke simultaneously with others, so there was no overlapping speech.

[2] Results of these are reported elsewhere.