Ear-voice span and pauses in intra- and interlingual respeaking: an exploratory study into temporary aspects of the respeaking process

Journal:	*Applied Psycholinguistics*
Manuscript ID	APS-Aug-16-0079.R1
mstype:	Original Article
Specialty Area:	Spoken Language Comprehension, Speech Production, Speech Processing, Cognition

Together with technological advancements, new forms of linguistic mediation are being developed in the area of media accessibility and audiovisual translation (AVT). One such form is respeaking – a method of producing real-time subtitles to live television programmes using speech recognition (SR) software (Marsh, 2006; Romero-Fresco, 2011). Respeaking was first used in 2001 by two public service broadcasters: the BBC in the UK and VRT in Belgium (Lambourne, 2006). Since then, this form of AVT has grown tremendously and it is now a major method used to produce live subtitling on TV.

Respeaking plays an important role in delivering the original spoken text in the form of written subtitles to people who are deaf and hard of hearing as well as language learners and others who use subtitling to support their TV viewing (Eugeni, 2008a; Romero-Fresco, 2011). In daily life, many of us probably encountered subtitles to live programmes and wondered how it is possible for the spoken dialogue to be converted into subtitles within a matter of seconds. The respeaking process is a complex activity (see Fig. 1), in both technical and cognitive terms (Boulianne et al., 2009; Luyckx, Delbeke, Van Waes, Leijten, & Remael, 2010; Romero-Fresco, 2011; Romero Fresco, 2012). A respeaker needs to listen to what the original speaker is saying in a TV programme and to respeak it, i.e. repeat or rephrase the text, adding the necessary punctuation marks and important information for viewers who are deaf or hard of hearing, related to speaker identification and sounds. The words uttered by the respeaker are then turned into text using speech recognition software. This text is later displayed on viewers' screens as subtitles with a delay of several seconds (Ofcom, 2015). In some countries, respeakers are also required to correct the output of SR programme if they spot any errors. In other countries, the error correction is done by another person, known as editor or moderator. On the one hand, the error correction process improves the quality of respoken subtitles, but on the other it may increase a delay between the time when the original speaker said something in a TV programme and when the corresponding subtitles appeared.

Since delay is one of the most frequently voiced complaints by deaf and hard of hearing viewers (Mikul, 2014), every effort has to be taken to ensure that this delay be as short as possible.

Insert Fig. 1 here

Fig. 1. The respeaking process.

Given that respeaking is a relative newcomer on the audiovisual translation scene – both as an AVT practice and as an area of academic research, there are still a number of fundamental questions about respeaking that remain unanswered. They include, but are not limited to, the similarities and differences between respeaking and interpreting, the competences of respeakers, the methods of respeaker training aimed at achieving good quality of live subtitling, particularly when it comes to reducing the delay (Mikul, 2014).

In this paper, we address the previously unexplored temporal aspects of respeaking which affect the delay in live subtitles: ear-voice span (EVS) and pauses. We examine how the characteristics of to-be-respoken materials modulate temporal aspects of respeaking performance. Additionally, by examining the common ground that respeaking shares with interpreting, we hope to start a discussion on respeaker competences and to better understand the respeaking process. This, we believe, can in turn translate into concrete solutions in respeaking training, mainly aimed at optimising the delay in live subtitling. Because intralingual respeaking shares temporal constraints with interpreting and because interlingual respeaking shares the process of transferring message from one language into another with both interpreting and translation, we wanted to find out if interpreters and translators are

better predisposed to producing respoken subtitles with shortest delay possible than average bilinguals without any interpreting or translation experience.

Respeaking as a hybrid modality

Despite being a newcomer on the AVT scene, respeaking seems to have some 'elder siblings' in the translation/interpreting family. First of all, respeaking (especially its interlingual variety) can be likened to media interpreting on television, where the interpreter interprets live televised content for the TV audience (Kurz, 2002; Pignataro, 2011; Pöchhacker, 2010). Both respeakers and media interpreters are bound by strict time constraints since they are often instructed to keep as close to the original speaker as possible (Pignataro, 2011) and to translate at a "supersonic pace" (Bros-Brann, 1994, p. 26), which undoubtedly affects their ear-voice span. Similarly to media interpreting on TV, long EVS is not desirable in respeaking as it increases the delay in displaying subtitles on screen. Secondly, the media interpreter addresses two types of audiences at the same time: on-screen participants (speakers in the TV studio such as a talk show host and guests) and off-screen participants (TV viewers) (Sergio, 2013). In the same vein, respeakers deal with the televised dialogue on the one hand, and with their deaf and hard of hearing audience at home, on the other. In 2013, Sergio noted that it is possible for media interpreting and subtitling to coexist in the same TV programme: when the interpreter's words are subtitled by another person. We now know that these two roles can be performed by a single person: an interlingual respeaker. Sergio (2013) claims that media interpreting "requires additional skills and a new professional profile" (p. 2) and cites Bros-Brann (1993, p. 1) stating that it requires "an entirely new mind set compared to everyday practice of conference interpretation and to what all of us have learned and taught in various schools of interpretation". We believe the same applies to respeaking as it now calls for new research and training.

Respeaking also shares some ground with simultaneous film interpreting (Russo, 2005), found at film festivals and other live events. The common ground lies in a complex audiovisual situation, combining inputs from the visual, oral and written codes. Among the similarities shared by simultaneous film interpreting, subtitling and respeaking are also the need to synchronise the translation with the image and to reduce the original text, as well as the fact of being "caught in the relationship between the written and the oral" (Gambier, 2003, p. 178).

Finally, respeaking has often been likened to simultaneous conference interpreting (Eugeni, 2008b; Marsh, 2004; Romero-Fresco, 2011). Indeed, both interpreting and respeaking demand excellent multi-tasking and split attention skills, as they require listening to the original text, rendering it into the same or another language, simultaneously monitoring the output (Jones, 2002; Pöchhacker, 2004; Romero-Fresco, 2011). Unlike interpreters, however, respeakers also need to add the necessary punctuation marks, and condense the original text to fit the limited space of subtitles on the screen. One of the most important similarities in the process of respeaking and interpreting is what is known as the ear-voice span.

Ear-voice span

Similarly to respeaking, simultaneous interpreting involves concurrent management of two speech channels: listening to the source language as well as producing and monitoring the target language. In order to do that, interpreters may to a certain extent produce the target text during pauses in the source text (Barik, 1973; Goldman-Eisler, 1972; Goldman-Eisler, Dechert, & Raupach, 1980), but for most of the time – estimated at 70% by Chernov (1994) – they listen and speak at the same time. Thus, the time lag or delay between the moment the original utterance is spoken out and the moment when the interpreter produces his/her equivalent in the target language, known as the ear-voice span or décalage, is one of the most

important features of processing involved in simultaneous interpreting (Lee, 2002; Pöchhacker, 2004). According to Timarová, Dragsted, and Hansen (2011), EVS "provides insight into the temporal characteristics of simultaneity in interpreting, speed of translation and also into the cognitive load and cognitive processing" involved in simultaneous interpreting (p. 121) and it is a reliable and quantifiable measure of cognitive processing (Lee, 2002). It is also an important skill to be practiced in simultaneous interpreter training (Bartłomiejczyk, 2015).

A number of previous studies have attempted to establish a minimum or optimum EVS unit. EVS can be measured in units of time (e.g. seconds) and/or linguistic units, such as a number of words or the nature of syntactic phrases. Paneth (1957) measured EVS values and found they were between 2 and 4 seconds, a result which was later confirmed by Barik (1973) and Oléron and Nanpon (1965/2002). Lederer (1978) found that EVS fell between 3 and 6 seconds. Other scholars reported the average EVS amounting to 2 seconds (Christoffels & de Groot, 2004), 2.68 seconds (Defrancq, 2015) and 4.7 seconds (Timarová et al., 2011). Schweda-Nicholson (1987) found EVS to range from 5 to 10 words or several seconds, while Gerver (1969) argued that the average delay at average presentation rates is about 4-5 words. Goldman-Eisler (1972) suggested that EVS units depend more on the syntax than on the lexis and established the minimum EVS unit to be a predicative expression (noun phrase and a verb phrase). Adamowicz (1989) also measured EVS in terms of meaningful syntactic units and nominal/verbal phrases as frequent EVS units in English-Polish interpreting. Donato (2003) applied an extended typology of syntax-based EVS units and found that the most frequently applied EVS units consisted of a noun phrase and a verb phrase. Christoffels and de Groot (2004) noticed a certain consistency pertaining to the average EVS duration reported in various studies with varied methodologies. They claimed that the upper boundary of the EVS is related to memory capacity (longer EVS increases memory load) while the lower boundary

is related to the minimum meaningful unit (Lederer, 1981; Setton, 1999) needed for the interpreter to decode meaning and perform interpreting.

EVS has also been linked to other factors, such as source text delivery rate (Lee, 2002), working with or without text (Lamberger-Felber, 2001) and sentence length (Lee, 2002). Barik (1973) measured temporal characteristics of interpretation of four types of texts: spontaneous speech, semi-prepared material, prepared "oral" material (a written-to-be-read-out speech) and prepared "written" material (an article). He found that interpreted texts usually included speaking for a greater proportion of the time than the original texts and this proportionality was greater for scripted than unscripted texts, as the latter have higher information density. However, due to a low number of participants, Barik did not find a consistent pattern of results when it comes to text types. Timarová et al. (2011) found a significantly smaller EVS when interpreting figures as opposed to verbs or beginnings of sentences, which means that interpreters try to produce figures as quickly as possible in order not to burden the memory with such difficult non-contextual items with high informative content (Chmiel, 2015; Mazza, 2001). Adamowicz (1989) found smaller EVS in interpreting a spontaneous text as compared to prepared texts, while Díaz-Galaz, Padilla, and Bajo (2015) reported smaller EVS following advance preparation for the simultaneous interpreting task. It also seems that more experienced interpreters work with smaller EVS as Timarová, Čeňková, and Meylaerts (2015) found a negative correlation between median EVS and days of interpreting experience. Taken together, these studies suggest that EVS depends on a combination of global (such as language combination) and local (such as propositions in the text) factors.

A generally held view is that longer EVS is better than keeping very close to the speaker (Kade, 1967). Interpreting trainees are taught to prolong their EVS in order to avoid word-for-word interpreting and in order to have enough time to restructure and reformulate the message

and to express it naturally in the target language (Bartłomiejczyk, 2015; Chmiel, 2015; Gorszczyńska, 2015). Shorter EVS is advised when information density in the text increases (for instance due to enumerations or numerical data) because it lowers the memory load and leads to fewer omissions. In fact, empirical studies have shown that shorter EVS does not have to lead to poorer quality (Defrancq, 2015) or even may be associated with better quality and higher accuracy (Lee, 2002; Timarová et al., 2014). It seems that the interpreter has to strike a balance between the EVS that is not too short (so that meaning can be constructed and restructuring is possible) and not too long (so that there is no memory overload leading to lower accuracy of interpretation) (Christoffels, de Groot, & Kroll, 2006; Kade & Cartellieri, 1971). In a sense, EVS lengthening and shortening can be the interpreter's strategic choice and is listed as one of tactics to be used by interpreters to prevent the occurrence of problems (Gile, 2009).

As opposed to interpreting, where longer EVS is usually not problematic for the target audience, in respeaking "décalage is less desirable" (Romero-Fresco, 2011, p. 107) because it causes delay in the live subtitling production process and effectively increases the gap between the on-screen images and the subtitles accompanying them. Such delay, as mentioned above, is a cause of great distraction to deaf and hard of hearing viewers and makes a TV programme with live subtitles difficult to follow as the subtitles simply cannot catch up with the images (Mikul, 2014).

Pauses

Important as it is, EVS is not the only time-related factor which is crucial in respeaking. Another significant factor are pauses made by respeakers in their speech. In its traditional sense, pauses "serve to divide discourse into tone groups and organize it into information units" (Shlesinger, 1994, p. 229); some pauses, however, are a result of hesitations (Pignataro, 2011, p. 87). We can therefore see that some pauses are more desirable than others which are

unwanted and can be considered "linguistically detectable faults" (Goffman, 1981, p. 172) and "disfluencies" (Garnham, 1985, p. 206). The undesirable pauses may be an indication of increased cognitive effort on the part of the speaker, or, as Mead (2000) put it, "manifestations of the effort of reasoning and formulation which accompany linguistic production" (p. 91). Skilled professional speakers are able to control their linguistic output, minimising the unwanted pauses in production which may "betray moments of doubt or distraction" (Mead, 2000, p. 91); see also Goffman (1981)).

In interpreting, pauses are an important element of fluency, which is thought to be a distinguishing feature between a professional and trainee interpreter (Mead, 2000). In respeaking, pauses are used as an important element of interaction between the respeaker and the SR software. In order to work properly, SR requires a short pause (less than 1 second) to be made by respeakers before it will transfer speech into text (Jurafsky & Martin, 2008; Romero-Fresco, 2011). Dictating one word at a time is not a good option, as it would remove the language context from the SR process, which is responsible for its high word accuracy (Romero-Fresco, 2011). Therefore, respeakers need to stop dictating after a few words and make a pause in appropriate moments in order for the SR programme to turn the spoken words into text which can then be displayed as subtitles. It needs to be noted here that subtitling is a constrained medium in itself: there is a limit to the number of lines and to the number of characters per line that can be used so that people can both read the subtitles and follow the on-screen images. Given all the above, respeakers are advised to divide the text they respeak into short, self-contained meaningful chunks, known as 'respeaking units'. Respeaking units are, as defined by Romero-Fresco (2011, p. 108), "idea units that lend themselves to accurate recognition by the SR software (phrases as opposed to single words) and to comfortable reading for the viewers (around one line in a one-, two- or three-line subtitle)". An ideal respeaking unit is believed to consist of five to seven words (Romero-

Fresco, 2011). Longer stretches of text may cause problems with subtitle segmentation, extend over too many lines and as such be difficult to follow by viewers.

Pauses – both in respeaking and in interpreting – may result from pauses made by the original speaker and/or the speech production process conducted by the interpreter/respeaker. In regular speech production, pauses serve a number of functions: they enable breathing, may have a semantic or rhetorical function and "provide time to cope with difficulties which can arise at any point in the speech production cycle" (Tóth, 2013, p. 3). In simultaneous interpreting, silent pauses can additionally manifest problems with source text comprehension, lexical search for the translation equivalent, difficulties with expressing a given sense in the target (Bartłomiejczyk, 2006; Piccaluga, Nespoulous, & Harmegnies, 2005; Tóth, 2011). As Piccaluga et al. (2005, p. 151) state: "it would be reasonable to think that the proliferation and/or lengthening of these pauses is linked to difficulties of various kinds in performing the complex task of interpreting."

Cecot (2001) examined silent pauses produced by professional interpreters and concluded that the most frequent were segmentation pauses (defined as grammatical and serving a communicative function) followed by hesitation pauses (defined as non-grammatical without a communicative function) and rhetorical pauses. Cecot's comparison of pauses in the source and target texts led her to conclude that interpreters generally followed the speaker's pattern of segmentation and rhetorical pauses. However, these patterns also reflect additional linguistic processing and cognitive effort (Goldman-Eisler, 1972). This is why studies have shown that there are more pauses in the interpreted text as compared to the source text (Tissi, 2000). Tissi (2000) found that interpretations included fewer but longer silent pauses than source texts, which suggests that pausing in interpreting might to a large extent stem from linguistic processing that differs from language planning in regular speech production. Pauses might be strategic (to gain time before a correction) or stem from

problems in the simultaneous interpreting technique experienced by interpreting trainees (Tissi, 2000). In fact, trainees have been found to pause more than professional interpreters, suggesting more cognitive effort expended by trainees for restructuring and lexical retrieval (Tóth, 2013).

Pausing patterns are also influenced by other factors, such as the direction of interpretation, source text speed and sound quality. Interpreters were more fluent and paused less when working into their native language both in simultaneous (Piccaluga et al., 2005) and consecutive interpreting (Mead, 2000). Piccaluga et al. (2005) studied a simultaneous interpreting task with manipulated source text speed and sound quality (added noise interference). They analyzed both the number and the length of pauses. Their participants paused more when the source text was of lower quality and delivered faster.

Interpreting experience may also influence pausing patterns. In general, the more experienced the interpreter, the shorter the produced pauses. In a study by Tóth (2013), pauses made by trainees lasted for 250-750 ms on average and they were longer than the pauses made by professionals (those ranged between 200 and 500 ms). The findings by Piccaluga et al. (2005) suggest that with increased interpreting experience, the participants seem to produce shorter but more numerous pauses. On the other hand, less experienced participants produced longer pauses, manifesting processing difficulties and breakdowns and leading to speech flow disruptions. The mean length of pauses in this study was 768 ms across all participants.

Taken together, the studies on silent pauses in interpreting suggest that pauses do reflect linguistic processing difficulties and serve as an index of disfluency (Pöchhacker, 2004). Pauses tend to be longer when the source text is more difficult (delivered faster or with a lower sound quality) and when the interpreter is less experienced and works from the mother tongue into the foreign language.

To recap the discussion on EVS and pauses, respeakers need to minimise their EVS, following the original utterance quite closely, and on the other hand, they must make short pauses between self-contained units of meaning ('respeaking units') in order for the SR software to work efficiently. Respeakers must therefore find "a good speech-to-pause rhythm, given that no subtitle will be shown until a pause is made" (Romero-Fresco, 2011, p. 108).

The present study

The goal of the present study is to better understand the process of intra- and interlingual respeaking of different TV genres, in particular in terms of temporal characteristics like EVS and pauses. By conducting this study, which, to the best of our knowledge, is the first of its kind, we also set out to investigate whether previous interpreting and translation experience may be an important factor affecting temporal characteristics of respeaking.

The profession of a respeaker is in its nascent stage and there were no professional respeakers in Poland at the time when this study was conducted. Thus, in order to establish whether interpreters and translators would produce respoken subtitles with less delay than non-interpreting and non-translating bilingual controls, we exposed them to intensive two-day training in respeaking fundamentals. In the first experiment, we compared their performance in intra- and interlingual respeaking tasks. In the second experiment, we decided to probe deeper into text characteristics that could influence temporal features of respeaking in intralingual tasks only.

Method

Experiment 1

In the first experiment we wanted to see how EVS and pauses are modulated by the nature of the respeaking task: interlingual vs. intralingual. Additionally, we wanted to compare three groups of participants: interpreters (as those who have experience both with temporal constraints characteristic for interpreting and respeaking, and with interlingual processing),

translators (as those who have experience in copying with interlingual processing in translation) and bilingual non-interpreting controls (who have experience neither with temporal constraints nor with interlingual processing). Thus, we used a mixed factorial design with task (intralingual and interlingual respeaking) as a within-subject independent variable and group (interpreters, translators and bilingual controls) as a between-groups independent variable.

We predicted that the interlingual respeaking task will be more difficult than intralingual respeaking, which will be manifested by longer EVS and longer pauses in the interlingual condition. We also expected that interpreting and translation experience would modulate EVS and pause length in both tasks. We predicted that in the interlingual condition, interpreters would respeak with shorter EVS and would produce shorter pauses than translators and controls, because they are used to coping with demanding time constraints when interpreting and because they are regularly exposed to interlingual processing. We also expected that translators would outperform controls (i.e. manifest shorter EVS and pauses) in the interlingual, but not in the intralingual condition, because they are used to interlingual processing in their translation experience. Additionally, we predicted that translators and controls would have more pauses than interpreters as compared to the pauses in the original text.

Materials

In the interlingual condition we used a slow-paced fragment of a speech delivered by President Barrack Obama on the 25th Anniversary of Polish Freedom Day in June 2014 in Warsaw. In the intralingual condition we used a slow one-speaker speech, the New Year's address by Poland's Prime Minister Ewa Kopacz.

Table 1.

The clips were matched in terms of the number of speakers, genre and speech rate measured by words per minute. As shown in Table 1, although matched for duration and number of words, the clips differed in terms of the number of syllables and speech rate measured as the number of syllables per duration. This discrepancy stems from the fact that Polish words are on average longer than English words. Both clips represented pre-scripted speech.

Participants

In this study we tested a total of 57 participants (50 women, 7 men). They were volunteers recruited among professional interpreters and translators as well as graduates and final year students at the Institute of Applied Linguistics at the University of Warsaw, the Faculty of English at Adam Mickiewicz University in Poznań, University of Social Sciences and Humanities in Warsaw as well as through social media (AVT Lab and RespeakingProject Facebook pages).

Their mean age was 27.48 (SD=5.71), ranging from 21 to 51. Based on the self-reported experience in interpreting, participants were divided into three groups: 22 interpreters (with at least two years of exposure to interpreting either in a professional context or during an intensive academic programme), 23 translators (with at least two years of exposure to translation either in a professional context or during an intensive academic programme), and a control group including 12 participants with no previous experience in translation or interpreting.

None of the participants had any respeaking experience. Therefore, prior to the respeaking test, all participants underwent a two-day (16 hours in total) intensive training in respeaking. They were trained in the fundamentals of respeaking, including linguistic and technical skills, such as management of simultaneous resources, working memory, monitoring their respoken output, enunciation, punctuation as well as the creation of voice profiles in the

Newton Technologies speech recognition software and the FAB Subtitler Live subtitling programme.

The participants were instructed to respeak in "idea units" or "respeaking units" (Romero-Fresco, 2011, pp. 60, 108), that is to say whole phrases rather than individual words and to pause after the entire phrase has been uttered for the SR language and acoustic model to work effectively. An ideal respeaking unit is considered to be between five to seven words (Romero-Fresco, 2011); longer stretches of text may not be comfortable for viewers to read and may result in increasing the delay between the image and the accompanying subtitles.

Procedure

Participants were tested individually in a research lab. Before the test, informed consent was obtained from each participant. The respeaking test started with the experimenter familiarising the participant with the procedure and the equipment.

Data recording

The data analysed in this paper come from a larger project which also included tracking the participants' eye movements and brain activity (which are reported elsewhere). A laptop with a custom-built software (Prompter) was used to record the two synchronised audio channels (the original speech and the respeaker's output). To display the respoken subtitles on the screen, we used FAB Subtitler Live. Each participant worked on their own voice profile in the speech recognition software for the Polish language manufactured by Newton Technologies. The order of clips was randomised.

The test started with the equipment testing and calibration. Then, participants went through a short mock respeaking task to familiarise themselves with the procedure; the data for this task were not recorded. The respeaking test proper consisted of respeaking one clip in the intralingual condition and one clip in the interlingual condition. After each task, the participants had to answer a few questions related to their self-reported cognitive load.

Calculation of EVS and pauses

To calculate EVS and pauses, we used automatic time alignment (described below) to generate precise segmentation of audio with respect to the spoken words. Given the information where each word begins and ends in the audio, we computed the time difference between the utterances heard and respoken (for EVS) and the durations of gaps between words (for pauses). However, the automatic method could not be used to measure EVS in the interlingual task as it involved different languages – instead, manual alignment was used for selected words.

Time alignment procedure

In order to accurately appraise the statistics related with the timing of spoken events, an annotation known as time-segmentation is required. This can be achieved either manually, which is an arduous and labour-intensive process, or using automated tools that perform a task known as automatic time alignment. Automatic time alignment can be quite precise given the right input conditions. While the quality of time alignment is difficult to assess for any given input, it relies on an automatic speech recognition engine and thus suffers from a lot of the same problems. The minimum resolution for such systems is usually 10 ms and provided that the right word sequence is matched correctly, the difference between the automatic and reference alignment boundaries is generally below 50 ms (Chen, Liu, Harper, Maia, & Mcroy, 2004). For various pragmatic reasons, the boundary shift is not often measured in such systems and other metrics are used instead (Räsänen, Laine, & Altosaar, 2009).

Time alignment is a well-known and thoroughly studied problem in the field of automatic speech processing (Benesty, Sondhi, & Huang, 2007; Jelinek, 1997). It is usually solved using a variant of the Viterbi algorithm (Rabiner, 1989), but this approach does not always work with fairly long and noisy audio sequences, as the ones studied in this paper. Instead, a technique inspired by SailAlign (Katsamanis, Black, Georgiou, Goldstein, &

Narayanan, 2011) is employed, the main idea of which is to perform simple continuous speech recognition, which normally produces a segmentation of the words, but does not guarantee their correct ordering. This initial match of words is then aligned to the original word sequence using a text-to-text alignment based on the Levenshtein-like algorithm (Levenshtein, 1966). This procedure produces a list of matches for individual words: correct, insertions, deletions and substitutions. The correctly matched words are assumed to match with correct segments of the audio and all the incorrect matches are re-aligned recursively, until convergence.

In order to produce time segmentation, the following steps were performed. First, the audio was recorded in a series of sessions, using a special custom-made program that can record multiple streams of audio simultaneously. Each recording comprised two audio streams: 1) the sound of the original material being respoken, and 2) the audio of the respeaker's voice, recorded with a professional grade microphone, used normally for speech recognition purposes. The purpose of having both streams was to be able to synchronize the segmentation times later in the experiment. This was crucial to obtain accurate time delay between what the respeakers heard and what they respoke.

The time-alignment procedure requires accurate transcriptions in order to work. While this step could have been performed automatically, it was important to avoid any mistakes at such an early stage, thus this step was performed completely by hand. After human-made transcriptions in orthographic form were produced, they had to be converted to a phonetic script to perform the actual alignment. This is because the recorded speech exists only in the phonetic, i.e. spoken form and this is the actual information that is being aligned. For languages like Polish, grapheme-to-phoneme conversion is not very difficult, but it can produce many false positives in real-world data, which contains information like foreign

words, names or characters with no obvious phonetic representation (e.g. numbers, abbreviations) (Brocki, Marasek, & Koržinek, 2012).

Calculating EVS

The time-alignment described above produced a segmentation of the audio into individual word segments. This segmentation is usually generated in the form a TextGrid file used by Praat software (Boersma, 2002) for phonetic analysis.

To complete the analysis in the experiment, we needed to compare the sequence of segments from the reference audio (as heard by the respeaker) to the sequence of segments from the respoken audio (of individual respeakers). Since it is very likely that the respeaker will not repeat what they hear verbatim (Luyckx et al., 2010; Romero-Fresco, 2011), this comparison is not as trivial as simply matching each segment in sequence. The respeaker can omit, substitute and in some cases even insert words that were not in the reference material (such as punctuation). Therefore, we needed to treat this problem as another candidate for the Levenshtein algorithm (Levenshtein, 1966). Myers' difference algorithm (Myers, 1986) was used to find the optimal alignment of the sequences, such that the minimum number of editing operations (substitutions, deletions, insertions) needs to be performed to convert one sequence into the other. In this study, only the correctly matching segments were compared.

As the alignment for the interlingual condition could not be managed automatically, it was performed manually. Automatic alignment would require matching words between languages, which would have to rely on some form of machine translation. It was decided that such a strategy would introduce too much error into the experiment and the manual approach was more cost effective in this case. Every tenth word from the original speech was selected for alignment with the appropriate translation equivalent from the transcribed respeaking output. If that word happened to be a function word, a cognate or a proper name, the subsequent eligible word was selected. We thus arrived at manually aligned selected words

from the interlingual condition and all automatically aligned words from the intralingual condition. In order to make the two comparable, we also selected words from the intralingual text following the same criteria we used for the interlingual condition and analysed the data only for those words.

Calculating pauses

The pauses in respeaking follow automatically from the segmentation as presented above. Since the experiment involves a single person speaking in a quiet environment, anything that is not segmented as the words of that person can be assumed to be a pause. In other words, to compute the pauses, we simply take the difference of the beginning of one word and the end of the one preceding it. Because the participants were asked to insert punctuation marks by voicing words such as "comma" and "full stop", it was easy to exclude from our analysis any functional pauses stemming from syntactic boundaries. We excluded all pauses following the words denoting punctuation marks. We also excluded pauses shorter than 200 ms, which is typically done to eliminate from the analysis those moments of silence which are due to the articulatory characteristics of e.g. voiceless consonants or due to the latency of the recording equipment (Warren, 2013).

We also calculated a pause ratio in order to examine the potential influence of the pausing pattern of the original speaker on the pausing pattern of the respeaker. Pause ratio was calculated by dividing the sum of pauses in the original clip by the sum of pauses in the respoken output for each participant and each clip. If the value was 1, it meant that the total length of pauses in the original equalled the total length of pauses made by the respeaker (excluding the functional pauses after pronouncing the punctuation marks). If the ratio was smaller than 1, the respeaker made longer pauses than there were in the original clip. If the ratio was greater than 1, the total length of pauses made by the respeaker was smaller than the total length of pauses in the original clip.

Results

The initial study design was 3 (group: interpreters, translators, bilingual controls) by 2 (clip type: interlingual, intralingual). However, during data collection many bilingual participants from the control group found the interlingual respeaking task too difficult and they failed to complete the task. Hence, due to insufficient data in the control group, further analyses of bilingual participants could not be performed. Therefore, finally our study followed a 2 (group: interpreters, translators) x 2 (clip type) mixed factorial design. The dependent variables were: EVS, length of pauses and the ratio of the pauses in the original to the pauses in the respoken output.

EVS

After visual inspection of Q-Q plots and histograms of EVS, values longer than 6900ms were considered outliers and removed from the analysis (30.67% of data). In order to normalise the distribution, the remaining EVS data were then log-transformed and subsequently analysed with linear mixed effects (LME) models via the lme4 package (Bates, 2013) within R (Baayen, 2008; R Development Core Team, 2010). This type of analysis combines the traditional ANOVA F1 and F2 analyses by treating participants and items as random effects and does not necessitate the aggregation of data over items or participants as it analyses them at the trial level. The following model was fitted to the data (with sliding contrasts and random intercepts and slopes): EVS~group*clip_type + (1+clip_type | subject) + (1 | item). This analysis revealed that EVS in the intralingual condition (M=2381ms, SD=1193) was significantly shorter than in the interlingual condition (M=4164ms, SD=1678) (b = 0.7944; SE=0.2059; t =3.858; p<.001). For reasons of clarity, we report the observed means and SDs rather than the predicted ones. No other predictors or interactions reached the significance level.

Pauses

Similarly to EVS data, after visual inspection of Q-Q plots and histograms of pauses, those longer than 2000ms were considered outliers and removed from the analysis (4.7% of data). For the purposes of this analysis, all pauses were then log-transformed to normalise the distribution of the data. The following model was fitted to the data (with sliding contrasts and random intercepts and slopes): logPauses~group*clip_type + (1+clip_type | subject) + (1 | item). The only statistically significant result was that pauses in the intralingual condition (M=538ms, SD=374) turned out to be shorter than in the interlingual condition (M=853ms, SD=651) (b = 0.36; SE=0.05; t = 7.43; p<.001).

Even though there is a possibility that EVS and pauses might be two reflections of a similar processing (as pointed out by one of the reviewers), we found only a very weak correlation between EVS and pausing data (r=.117; n=38441, p<.001). Furthermore, when pauses were added as a covariate to a post-hoc LME model of EVS as a dependent variable, their influence on EVS turned out to be unreliable (p>.05).

Since the analysis of pauses in the respoken output was likely affected by the pauses naturally produced by the speakers in the to-be-respoken clips, we decided to look at pause ratio to check how much more pausing there was in the respoken output compared to the source clips themselves. These individual ratios were next log-transformed and initially fitted in the following LME model: pauseRatio~group*clip_type + (1 | subject) with random intercept and sliding contrasts. However, since this model failed to converge, a model with the same fixed predictors but only random intercepts was used to analyse pause ratio data. Apart from a significant difference between pause ratio in the intralingual (M=1.89, SD=0.61) and interlingual condition (M=1.25, SD=0.39) (b =-0.45; SE=0.06; t =-7.19; p<.001), the model revealed a marginally significant effect of group, where interpreters had a higher ratio, i.e. made shorter pauses relative to the original (M=1.82, SD=0.56) than translators (M=1.50, SD=0.63) (b =-0.25352; SE=0.13160; t =-1.927; p=.067).

Discussion

In general, we expected the main effect of task type, with interlingual condition generating more cognitive effort (longer EVS, longer pauses, lower pause ratio) than the intralingual condition. These predictions were corroborated by our findings. The interlingual clip was respoken with longer EVS and longer pauses than the comparable intralingual clip. The pause ratio was lower for the interlingual condition, suggesting that the participants added more pauses in their respoken output than there were in the original. This is in line with studies comparing pauses in source texts and their interpretations where interpreted speeches included more pauses than the original ones, reflecting additional linguistic processing and cognitive effort (Goldman-Eisler, 1958; Tissi, 2000). Our findings confirm that respeaking is more demanding interlingually than intralingually, as demonstrated by the temporal characteristics of the output.

The other set of predictions was related to the main effect of group. We expected interpreters to outperform the other groups. Unfortunately, we can only discuss the predictions regarding interpreters and translators here. We found no significant difference between the performance of interpreters and translators both in EVS and pause length data. We found a marginally significant difference in pause ratio, suggesting that interpreters paused less relatively to the original speakers than translators.

Contrary to our expectations, interpreters did not differ from translators despite their previous exposure to temporal constraints of interpreting. In spite of similarities in processing between interpreting and interlingual respeaking, respeaking may be additionally challenging to interpreters due to the need to provide punctuation and to speak in subtitle-length units. These novel challenges may offset any potential interpreter advantage in respeaking (interpreter advantage is understood here as the extensive experience with working under strict temporal constraints). Since respeaking is more similar to media interpreting than

regular conference interpreting, who were tested in our study, maybe such interpreter advantage could be seen if we compared media interpreters with translators.

Experiment 2

To find out if there are any differences in EVS and pauses between different TV genres (speech, news, entertainment show and political chat show) characterised by different speaking speeds, varying numbers of speakers and different levels of scriptedness of language, we further examined four videos respoken intralingually. We selected examples of both scripted and unscripted dialogue (Remael, 2008) to find out if the character of the dialogue had an impact on the temporal characteristics of respeaking. The main difference between scripted and unscripted dialogue lies in the fact that the former is pre-prepared in written form to be later delivered orally, and the latter is spontaneous and does not follow any pre-prepared script. In linguistic terms, unscripted speech tends to be less formal and less fluent owing to more frequent pauses, hesitations and repetitions; it also contains elements typical of spoken dialogue, certain discourse interpersonal markers (*I mean, you know*), vague language like hedges (*kind of*), references (*stuff*) or coordination tags (*or something*) (Adrian, 2013; Quaglio, 2009). From the subtitling perspective, unscripted dialogues tend to contain some irrelevant information and as such need to be "'cleaned up' grammatically and structurally" (Remael, 2008, p. 65) in order to make an utterance more explicit. This is important in live subtitling through respeaking, because subtitles need to be easy to read and broadcasters usually "want subtitles to be written in standard language, focusing on content rather than idiosyncrasies of the speaker" (Remael, 2008, p. 65).

We were also interested to see whether the number of speakers and, in consequence, overlapping speech, has any impact on EVS and pauses in respeaking. With this goal in mind, we selected videos with single speakers (speech, news) as well as those with multiple

speakers, whose utterances were sometimes overlapping (entertainment show and political chat show).

We had a mixed factorial design for the intralingual respeaking task with clip type as an independent within-subject variable with four levels: speech, news, entertainment chat show, political chat show, and group as a between-groups independent variable with two levels: interpreters and bilingual controls. We expected to obtain the main effect of group, that is we predicted that interpreters would provide respeaking with shorter EVS and shorter pauses than controls as they are used to linguistic processing constrained by similar time demands in professional interpreting assignments. We also expected the clip type to modulate EVS and pause length across participants. We predicted that the political chat show would generate the longest EVS, the longest pauses and the smallest pause ratio due to its high speech rate and the presence of multiple speakers and overlapping speech. This would be followed by news (due to information density and pre-scripted nature of the content), the entertainment chat show (due to its medium-paced speech and multiple speakers). Speech was expected to generate the shortest EVS, the shortest pause length and the highest pause ratio due to its slow speech rate and because it was delivered by one speaker only.

Materials

Participants were asked to intralingually respeak four clips. The clips represented different genres from Polish TV channels: (1) pre-scripted speech: the New Year's address by Poland's Prime Minister Ewa Kopacz, (2) pre-scripted news: an excerpt from *Fakty*, an evening news broadcast, reporting on miners' protests, (3) unscripted entertainment chat show (*Fakty po Faktach* with an actress and a movie critic discussing the movie *Ida* after it was awarded an Oscar for best foreign language film, and (4) unscripted political chat show with speakers from opposite ends of the political spectrum (an excerpt from *Kropka nad i*, with numerous cases of overlapping speech). In this analysis, we used data from clip 1 from Experiment 1

(speech). It turned out that the automatic analysis of EVS and pauses was impossible for clip 4 due to numerous occurrences of overlapping speech. Therefore, further characteristics of the clips (see Table 2) and results are given for the first three clips only.

Table 2.

Participants

The participants included two groups involved in Experiment 1, interpreters (N=22) and bilingual controls without any interpreting experience (N=12).

Procedure

The procedure was similar to the one used in Experiment 1. The experiment proper included intralingual respeaking of the experimental clips in a randomised order. The test ended with a short semi-structured interview.

Calculation of EVS and pauses

The automatic calculation of EVS and pauses was performed similarly to that in Experiment 1. We used the data calculated for every word in the three video clips. Pause ratio was calculated as in Experiment 1.

Results

The study followed a 2 (group: interpreters, bilingual controls) x 3 (clip type) mixed factorial design. The dependent variables were: EVS, length of pauses and the ratio of original pauses to pauses in respeaking. Table 3 presents mean results for each dependent variable and for each condition.

Table 3.

EVS

After visual inspection of Q-Q plots and histograms of EVS, values longer than 5000ms were considered outliers and removed from the analysis (4.14% of data). The remaining EVS data was analysed with linear mixed effects (LME) models via the lme4 package (Bates, 2013) within R (Baayen, 2008; R Development Core Team, 2010). This type of analysis combines the traditional ANOVA F1 and F2 analyses by treating participants and items as random effects and does not necessitate the aggregation of data over items or participants (analyses them at the trial level). The following model was fitted to the data (with sliding contrasts and random intercepts and slopes): EVS~group*clip_type + (1+clip_type | subject) + (1 | item). Since sliding contrasts allowed us to look only at comparisons of adjacent levels of factors, we refitted the same model but with a different order of the levels of the variable clip type. Below we report observed rather than predicted data. The following predictors came out as significant from these analyses: EVS in Chat Show (M=1829ms) was only marginally significantly shorter than in Speech (M=2139ms) (b = 259.06; SE=132.58; t = 1.95; p=.07). In turn, EVS in News (M=2526ms) was significantly longer than in Speech (b = 716.84; SE=110.44; t = 6.49; p<.001) and in Chat Show (b = 975.90; SE=47.430; t = 20.63; p<.001). This suggests more cognitive effort in respeaking fast scripted speech. Even though EVS values were not significantly different across the two groups (b = −35.03; SE=247.15; t = −0.14; p>.05), we found a significant interaction showing that the groups differed significantly in their EVS values between Speech and News (b = -717.84; SE=216.61; t = −3.31; p<.01) and between News and Chat Show (b=686.73; SE=85.85; t=7.99; p<.001). Controls worked with a shorter EVS than interpreters both when respeaking Speech and Chat Show. The trend reversed in the case of News, where interpreters outperformed controls by respeaking with a shorter EVS.

Pauses

Similarly to EVS data, after visual inspection of Q-Q plots and histograms of pauses, those longer than 2000ms were considered outliers and removed from the analysis (2.9% of data). For the purposes of this analysis, all pauses were then log-transformed to normalise the distribution of the data. Below we report observed and not predicted data. We also report real means instead of log-transformed means for clarity of presentations. However, the coefficients are presented as log-transformed and calculated in the respective model. The following model was initially fitted to the data (with sliding contrasts and random intercepts and slopes): logPauses~group*clip_type + (1+clip_type | subject) + (1 | item), however it failed to converge. A model with the same set of fixed predictors and only with random intercepts did converge and yielded the following results (since sliding contrasts allowed us to look only at comparisons of adjacent levels of factors, we refitted the same model but with a different order of the levels of the variable clip type). Pauses in Chat Show (M=589ms) were significantly longer than in Speech (M=495ms) (b = .13; SE=.03; t = 4.78; p<.001) and in News (M=502ms) (b =.17; SE=.04; t =4.62; p<.001). The difference in pauses between Speech and News was only numeric (b = .04; SE=.38; t = 1.14; p=.25). This suggests that respeaking unscripted speech produced by multiple speakers (Chat Show) is more challenging than respeaking single speakers, even when producing fast scripted text (News). Even though pauses were not significantly different across the two groups (interpreters: M=534ms; controls: M=552ms) (b = -.02; SE=-.05; t = -0.43; p>.05), we found a marginally significant interaction showing that the groups differed in how much they paused when respeaking Speech and News (b = .01; SE=.07; t = 1.79; p=.07). A similar interaction was found in pauses made by interpreters and controls in Chat Show and News (b = -.01; SE=.07; t = -1.66; p=.09). This again suggests a different behaviour of both groups depending on the clip.

Controls paused more than interpreters when respeaking Chat Show and Speech, while interpreters paused more when respeaking News.

Similarly to Experiment 1, we checked for correlations between EVS and pausing data in Experiment 2, and found no reliable results (r=-.003, p=.882). Furthermore, when pauses were added as a covariate to a post-hoc LME model of EVS, they again failed to explain more variance (p>.05).

As in the analyses of pausing in Experiment 1, we decided to look at pause ratio to test the relationship between pauses in the respoken output in the three types of clips themselves. To that end, we calculated a pause ratio for every participant for every clip. These individual ratios were next analysed with the following LME model: pauseRatio~group*clip_type + (1 | subject) with random intercept and sliding contrasts. These analyses revealed a significant effect of group where the pause ratio was significantly smaller for controls (M=1.50) than for interpreters (M=1.99) (b=0.54; SE=0.26; t =2.09; p<.05), suggesting that both groups paused less than the original speakers and that the total length of pauses relative to pauses in the to-be-respoken original was smaller for interpreters as compared to controls. This means that both interpreters and controls used pauses in the original to continue with their production of respeaking and interpreters did that to a greater extent. Also, the pause ratio was significantly shorter for Chat Show (M=1.65) than for Speech (M=2.19) (b =0.44; SE=0.13; t = 3.25; p<.01) and marginally shorter for Speech than for News (M=1.75) (b =-0.36; SE=0.19; t = -1.98; p=.06), suggesting that the respeakers did produce respoken subtitles while using pauses, especially in the slow-paced clip (Speech).

Discussion

By comparing interpreters and controls on an intralingual respeaking task, we expected that interpreters would produce respeaking with better temporal characteristics (shorter EVS and shorter pauses) than controls thanks to their interpreting experience. We found no reliable

differences on two out of the three measures. Interpreters did not outperform controls as regards EVS, contrary to Timarová et al. (2015), who found shorter EVS for more experienced interpreters. Interpreters did not have shorter pauses than controls, but they had a higher original-to-respoken output pause length ratio than controls. This suggests that, relative to pauses made by the speakers, interpreters introduced less additional pausing to the respoken output. They were more skilled at using the original pauses to produce output, which is in line with the performance typical of interpreters (Barik, 1973; Goldman-Eisler, 1972; Goldman-Eisler, Dechert, & De Gruyter, 1980). Four significant or marginally significant interactions show an interesting pattern. These interactions were driven mainly by the difference between News (fast and scripted) and the other two clips (slow scripted and medium-paced unscripted). As regards EVS, controls had shorter EVS than interpreters when respeaking Speech and Chat Show while interpreters outperformed controls when respeaking News. As regards pauses, the results were exactly the opposite. It is difficult to clearly explain these results on the basis of the temporal characteristics only. It seems that interpreters had a shorter delay and paused more than controls when respeaking News, the most difficult clip in the study. It was delivered fast, read out from the script and included various visuals and numbers. Interpreters may have decided to shorten the delay in order to grasp as much information from the fast-delivered text as possible, as it is generally advised in interpreter training to shorten EVS when working with high information density texts (Chmiel, 2015).

We were also interested to see if clip type would modulate temporal characteristics of respeaking and expected to see better performance on Speech (slow delivery, single speaker, but prescripted) than on Chat Show (medium-paced delivery, unscripted, but multiple speakers) and worse on News (fast delivery, scripted, single speaker). The results were mixed and our predictions were not confirmed fully with any of the dependent variables. As regards EVS, News proved to be the most difficult with the longest delay. It differed significantly

both from Speech and Chat Show. This is in line with Adamowicz (1989), who found shorter EVS for spontaneous speeches than pre-prepared ones. However, despite numerical differences, we found no reliable difference between Speech and Chat Show. It seems that fast and scripted text entails the longest EVS as compared with slower both scripted and unscripted clips.

The pause length data shows a different pattern of results. Here, Chat Show generated the longest pauses and differed from both News and Speech in this respect. Chat Show was the only clip with multiple speakers and in the post-test interview over 80% participants stated it was the multiple speaker programmes that were more problematic for them than single speaker clips. When asked about what was more difficult: fast speech rate and the number of speakers, 63% declared that it was the number of speakers that was making respeaking more demanding. This may explain why Chat Show proved the most demanding in the analysis of pause length data. The findings are at a variance with the results of Piccaluga et al. (2005), who found that faster texts entail longer pauses in interpreting. Chat Show was not the fastest clip used in the study. Together with the introspective data elicited from interviews, this suggests that the number of speakers might be a factor that influences pausing patterns more than the delivery rate of the original text.

The analysis of the pause ratio data showed that Speech generated the highest pause ratio and differed significantly both from News and Chat Show. This means that in this slowly delivered text, the participants introduced the shortest pauses relative to the pauses already present in the original. It seems that slow pace is beneficial to the respeakers as they have time to respeak the content without major disruptions.

General discussion

By conducting the two experiments reported in this paper, we wanted to examine temporal aspects of the respeaking process, in particular EVS and pause length in intra- and interlingual

respeaking of different TV genres as performed by three groups of participants: interpreters, translators and controls. We were interested in how the characteristics of the to-be-respoken materials would influence EVS and pausing patterns. We also wanted to see whether the EVS and pauses in respeaking were affected by interpreting and translation skills, in other words – whether interpreters and translators manifest any difference in respeaking tasks as compared to controls due to their experience.

EVS

In general, our findings show that the average EVS was 2100ms in intralingual respeaking and 4160ms in interlingual respeaking. Thus, the former is shorter while the latter is similar to most EVS values in previously reported studies on interpreting (Lederer, 1978; Oléron & Nanpon, 1965/2002; Schweda-Nicholson, 1987; Timarová et al., 2011). This suggests that interlingual respeaking can be likened to interpreting, while intralingual respeaking requires less cognitive effort than interpreting.

The findings confirmed our predictions regarding the effect of to-be-respoken material on EVS. Interlingual respeaking generated greater EVS than intralingual respeaking. Additionally, the longest EVS – among the three intralingual clips analysed – was found in the news programme, followed by the speech and the entertainment chat show. What the news programme and the speech have in common is that they were both pre-scripted, mostly delivered from a prompter, with high density of information and syntactic complexity. The entertainment chat show, on the other hand, consisted of unscripted dialogue and as such contained numerous elements of spoken language like repetitions, false starts, hesitations, unfinished sentences and reformulations. It also featured three speakers engaged in one conversation, taking turns and sometimes interrupting one another. Due to such dynamic turn-taking, some participants may have deliberately chosen to keep close to the speakers in order not to get lost with who is saying what. In this way, it was probably easier for them to have

one speaker per subtitle: longer EVS could mean that they would have to respeak the utterances of two different speakers in the same subtitle, which would also be more difficult for viewers to follow.

Because long delays are not desired in respeaking (Mikul, 2014), live respeaking of scripted content with fast delivery should be avoided whenever possible. Instead, subtitling should be done using a pre-prepared script – either as pre-recorded subtitling or as semi-live subtitling (where the subtitle chunks are pre-prepared, but not time-coded). Since this is not always possible (for instance, there is rarely prior access to an official speech read out by a politician), respeaker training should focus on such content at an advanced stage.

The EVS data did not confirm our predictions about the interpreters having shorter delays in both inter- and intralingual respeaking as compared to controls and about the translators having smaller EVS as compared to controls in the interlingual respeaking. We found that interlingual respeaking was generally too difficult for bilingual controls to perform but they were not outperformed by interpreters in the intralingual respeaking. It seems that temporal constraints and similarities between interpreting and respeaking are not sufficient to see any transfer of experience from one to the other. This might be due to the new challenges posed by respeaking, such as intra- rather than interlingual processing, voicing punctuation marks and providing output in subtitle-length chunks. Interpreters outperformed controls only when respeaking the news bulletin. They may have applied a general strategy usually successful when interpreting highly informative speeches delivered fast, i.e. shorten their EVS in order not to miss too much content. However, this explanation is highly speculative and requires further research. In general, EVS data provided no evidence for the greater predisposition of interpreters relative to non-interpreters to become respeakers.

Pauses

The length of pauses in respeaking may be taken as an indication of processing effort on the part of the respeaker (the longer the pause, the greater the processing effort) and the difficulty of the original text to be respoken (the more difficult the task, the longer the pauses). In this study we found that in the interlingual respeaking task, pause length was indeed longer than in the intralingual task. Additionally, the entertainment chat show (with unscripted text delivered by many speakers) generated longer pauses than the news bulletin and the speech. It seems that coping with numerous speakers proves more difficult than coping with fast single-speaker text, which was also confirmed by introspective data from the post-test interviews. More studies are needed to disentangle the effect of the speed of delivery and the number of speakers on pausing patterns in respeaking. Contrary to our expectations, pause length data did not reveal any differences between the experimental groups, suggesting that interpreters have no clear advantage over translators and controls when training to become respeakers.

The pause ratio data revealed that clip type did modulate the participants' performance. Again, in line with our predictions, interlingual respeaking entailed introducing more pauses relative to the respoken original as compared to intralingual respeaking. Among the intralingual clips, it was the slow-paced speech that generated the highest pause ratio, i.e. required the introduction of the least additional pausing in the respoken output. We found group differences on that measure: interpreters introduced the least additional pausing, which suggests that they used the pauses in the original text to continue their production, in line with pausing patterns observed in simultaneous interpreting.

Conclusions

Taken together, our findings on the temporal aspects of respeaking show that the characteristics of to-be-respoken material modulate the respeaking performance. Pre-scripted speech with fast delivery resulted in the highest EVS, which might entail certain recommendations for the respeaking practice. When working with such texts, in order to

reduce the delay, subtitles should preferably also be pre-prepared (pre-recorded or semi-live). When respeaking is done live on such materials, greater delays between the original speech and the subtitle should be expected. Our findings can also be taken as empirical evidence confirming previous intuitive conjectures according to which interlingual respeaking requires more cognitive effort than intralingual respeaking as it combines two complex tasks: respeaking and interpreting. Additionally, we found no reliable group differences in EVS and pause length data suggesting that previous exposure to lexical processing under time constraints in the form of interpreting does not seem to make interpreters better predisposed to become respeakers than translators or controls. Finally, following a suggestion by an anonymous reviewer, we examined the relation between pause length and EVS in respeaking and found very weak or no correlation between the two.

Admittedly, our study was not free from limitations. One of them was the lack of gender balance among the participants, which may have affected the results, as noted by Cecot (2001), who found pauses to be gender-specific. Additionally, our groups were not completely homogenous because – for technical reasons – they included both interpreters and interpreting trainees, translators and translation trainees. We had to exclude a lot of data due to great variability, which may be avoided in the future by involving more homogenous and numerous experimental groups. Finally, we chose authentic materials for our study, which meant limited control over numerous variables. Thus, we could not use the delivery rate, scriptedness and the number of speakers as independent variables. Instead, we opted for having a clip type as an independent variable, which made it impossible to disentangle the effects of separate clip characteristics on the temporal aspects of respeaking.

In general, this preliminary study has shown a number of interesting research avenues to be pursued in the future. Future research could look into replicating our results on other types of clips to disentangle the effect of delivery rate, scriptedness and the number of speakers on

the temporal characteristics of respeaking. It would also be interesting to further explore potential benefits of having interpreting skills in respeaking by involving more homogenous and separate groups of conference interpreters, media interpreters and interpreting trainees. This, in turn, may affect interpreting and audiovisual translation curricula by expanding them with aspects of respeaking, particularly optimum EVS.

All in all, empirical research on respeaking is still in its infancy and we hope to have opened new research avenues in this area. We also believe that future research can overcome the limitations of this study and successfully replicate our results, fine-tuning the analysis methods and contributing to further development of respeaking.

Tables and figures

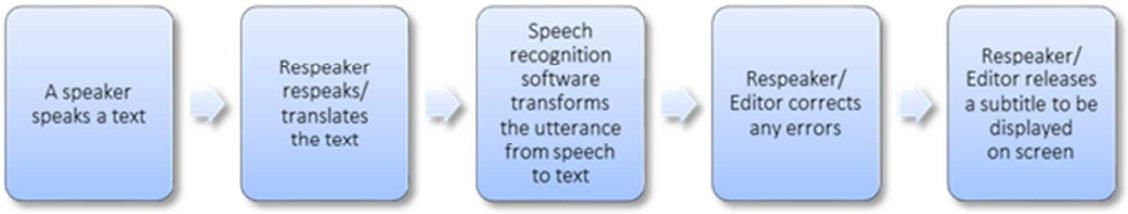


Fig. 1

References

Adamowicz, A. (1989). The role of anticipation in discourse: Text processing in simultaneous interpreting. *Polish Psychological Bulletin, 20*(2), 153-160.

Adrian, R. (2013). *Talking Television. Viewer identification of unscripted conversation and scripted television dialogue and their corresponding features*.

Baayen, R. H. (2008). *Analyzing linguistic data : a practical introduction to statistics using R*. Cambridge: Cambridge : Cambridge University Press.

Barik, H. (1973). Simultaneous interpretation: temporal and quantitative data. *Language and Speech, 16*(3), 237.

Bartłomiejczyk, M. (2006). Strategies of simultaneous interpreting and directionality. *Interpreting, 8*(2), 149-174. doi:10.1075/intp.8.2.03bar

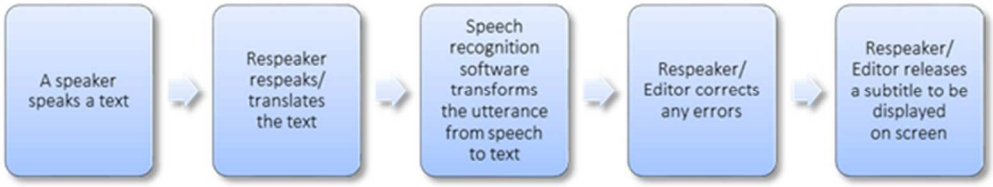
Bartłomiejczyk, M. (2015). Wprowadzenie do tłumaczenia symultanicznego. In A. Chmiel & P. Janikowski (Eds.), *Dydaktyka tłumaczenia ustnego* (pp. 207-226). Katowice: Stowarzyszenie Inicjatyw Wydawniczych.

Bates, D. (2013). *Linear mixed model implementation in lme4*. Madison.

Benesty, J., Sondhi, M. M., & Huang, Y. (2007). *Springer Handbook of Speech Processing*.
Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot international, 5*(9/10), 341-345.
Boulianne, G., Beaumont, J.-F., Boisvert, M., Brousseau, J., Cardinal, P., Chapdelaine, C., . . . Dumuchel, P. (2009). Shadow Speaking for Real-Time Closed-captioning of TV Broadcasts in French. In A. Matamala & P. Orero (Eds.), *Listening to Subtitles. Subtitles for the Deaf and Hard of Hearing* (pp. 191-208). Bern: Peter Lang.
Brocki, Ł., Marasek, K., & Koržinek, D. (2012). Multiple model text normalization for the Polish language. In L. Chen, S. Felfernig, J. Liu, & Z. W. Raś (Eds.), *Foundations of Intelligent Systems* (pp. 143-148). Berlin Heidelberg: Springer.
Bros-Brann, E. (1994). *Interpreting live on television: some examples taken from French television*. report. UIMP course The Interpreter as a Communicator. La Coruna.
Cecot, M. (2001). Pauses in simultaneous interpretation: A contrastive analysis of professional interpreters' performances. *The Interpreters' Newsletter, 11*, 63-85.
Chen, L., Liu, Y., Harper, M., Maia, E., & Mcroy, S. (2004). *Evaluating factors impacting the accuracy of forced alignments in a multimodal corpus.* Paper presented at the LREC, Lisbon.
Chernov, G. V. (1994). Message redundancy and message anticipation in simultaneous interpreting. In S. Lambert & B. Moser-Mercer (Eds.), *Bridging the gap: empirical research in simultaneous interpretation* (pp. 139-154). Amsterdam, Philadelphia: John Benjamins.
Chmiel, A. (2015). Przetwarzanie w tłumaczeniu symultanicznym. In A. Chmiel & P. Janikowski (Eds.), *Dydaktyka tłumaczenia ustnego* (pp. 227-247). Katowice: Stowarzyszenie Inicjatyw Wydawniczych.
Christoffels, I. K., & de Groot, A. M. B. (2004). Components of Simultaneous Interpreting: Comparing Interpreting with Shadowing and Paraphrasing. *Bilingualism: Language and Cognition, 7*(3), 227-240. doi:10.1017/S1366728904001609
Christoffels, I. K., de Groot, A. M. B., & Kroll, J. F. (2006). Memory and language skills in simultaneous interpreters: The role of expertise and language proficiency. *Journal of Memory and Language, 54*(3), 324-345. doi:10.1016/j.jml.2005.12.004
Defrancq, B. (2015). Corpus-based research into the presumed effects of short EVS. *Interpreting, 17*(1), 26-45. doi:10.1075/intp.17.1.02def
Díaz-Galaz, S., Padilla, P., & Bajo, M. T. (2015). The role of advance preparation in simultaneous interpreting: A comparison of professional interpreters and interpreting students. *Interpreting, 17*(1), 1-25. doi:10.1075/intp.17.1.01dia
Donato, V. (2003). Strategies adopted by student interpreters in SI: A comparison between the English-Italian and the German-Italian language-pairs. *Interpreters' Newsletter, 12*, 101-134.
Eugeni, C. (2008a). Respeaking the TV for the Deaf: For a Real Special Needs-Oriented Subtitling. *Studies in English Language and Literature, 21*, 37-47.
Eugeni, C. (2008b). A Sociolinguistic Approach to Real-time Subtitling: Respeaking vs. Shadowing and Simultaneous Interpreting. *English in International Deaf Communication, 72*, 357-382.
Gambier, Y. (2003). Introduction. *The Translator, 9*(2), 171-189. doi:10.1080/13556509.2003.10799152
Garnham, A. (1985). *Psycholinguistics: Central Topics*. London, New York: Routledge.
Gerver, D. (1969). Effects of grammaticalness, presentation rate, and message length on auditory short-term memory. *Quarterly Journal of Experimental Psychology, 21*(3), 203-208. doi:10.1080/14640746908400214
Gile, D. (2009). *Basic concepts and models for interpreter and translator training*. Amsterdam, Philadelphia: John Benjamins.
Goffman, E. (1981). *Forms of talk*. Oxford: Blackwell.
Goldman-Eisler, F. (1958). The predictability of words in context and the length of pauses in speech. *Language and Speech, 1*(3), 226.
Goldman-Eisler, F. (1972). Segmentation of input in simultaneous translation. *J Psycholinguist Res, 1*(2), 127-140. doi:10.1007/BF01068102
Goldman-Eisler, F., Dechert, H. W., & De Gruyter, D. (1980). *Temporal Variables in Speech : Studies in Honour of Frieda Goldman-Eisler*.

Goldman-Eisler, F., Dechert, H. W., & Raupach, M. (1980). *Temporal variables in speech: studies in honour of Frieda Goldman-Eisler*. The Hague: Mouton.
Gorszczyńska, P. (2015). Produkcja w tłumaczeniu symultanicznym. In A. Chmiel & P. Janikowski (Eds.), *Dydaktyka tłumaczenia ustnego* (pp. 248-288). Katowice: Stowarzyszenie Inicjatyw Wydawniczych.
Jelinek, F. (1997). *Statistical methods for speech recognition*. Cambridge, Mass: MIT Press.
Jones, R. (2002). *Conference interpreting explained* (2nd ed. ed.). Manchester: St. Jerome Publishing.
Jurafsky, D., & Martin, A. (2008). *Speech and Language Processing: an introduction to natural language processing, computational linguistics, and speech recognition*. New York: Prentice Hall.
Kade, O. (1967). Zu einigen Besonderheiten des Simultandolmetschens. *Fremdsprachen, 11*(1), 8-17.
Kade, O., & Cartellieri, C. (1971). Some Methodological Aspects of Simultaneous Interpreting. *Babel, 17*(2), 12-16.
Katsamanis, A., Black, M., Georgiou, P. G., Goldstein, L., & Narayanan, S. (2011). SailAlign: Robust long speech-text alignment. *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*.
Kurz, I. (2002). Physiological stress responses during media and conference interpreting. In G. Garzone & M. Viezzi (Eds.), *Interpreting in the 21st Century* (pp. 195-202). Amsterdam, Philadelphia: John Benjamins.
Lamberger-Felber, H. (2001). Text-oriented Research into Interpreting: Examples from a Case-study. *Hermes, 26*, 39-63.
Lambourne, A. (2006). Subtitle respeaking. A new skill for a new age. *Intralinea*.
Lederer, M. (1978). Simultaneous Interpretation: Units of Meaning and Other Features. In D. Gerver & H. W. Sinaiko (Eds.), *Language Interpretation and Communication* (pp. 323-332). New York: Plenum Press.
Lederer, M. (1981). *La traduction simultanée: expérience et théorie*. Paris: Minard.
Lee, T.-H. (2002). Ear Voice Span in English into Korean Simultaneous Interpretation. *Meta, XLVII*(4), 596-606.
Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady 10*(8), 707-710.
Luyckx, B., Delbeke, T., Van Waes, L., Leijten, M., & Remael, A. (2010). Live subtitling with speech recognition causes and consequences of text reduction. *IDEAS Working Paper Series from RePEc*.
Marsh, A. (2004). *Simultaneous Interpreting and Respeaking: A Comparison*. MA thesis. University of Westminster. London.
Marsh, A. (2006). Respeaking for the BB. *Intralinea*.
Mazza, C. (2001). Numbers in simultaneous interpretation. *The Interpreters' Newsletter, 11*, 87-104.
Mead, P. (2000). Control of pauses by trainee interpreters in their A and B languages. *The Interpreters' Newsletter, 10*, 89-102.
Mikul, C. (2014). *Caption Quality: Approaches to standards and measurement*. Retrieved from Sydney:
Myers, E. W. (1986). AnO (ND) difference algorithm and its variations. *Algorithmica, 1*(1-4), 251-266.
Ofcom. (2015). *Measuring live subtitling quality. Results from the fourth sampling exercise*. Retrieved from http://stakeholders.ofcom.org.uk/market-data-research/other/tv-research/live-subtitling/sampling_results_4/
Oléron, P., & Nanpon, H. (1965/2002). Research into simultaneous translation. In F. Pöchhacker & M. Shlesinger (Eds.), *The Interpreting Studies Reader* (pp. 43-50). London: Routledge.
Paneth, E. (1957). *An Investigation into conference interpreting (with special reference to the training of the interpreter).* (MA), University of London Institite of Education.
Piccaluga, M., Nespoulous, J.-L., & Harmegnies, B. (2005, 10-12 September 2005). *Disfluencies as a window on cognitive processing. An analysis of silent pauses in simultaneous interpreting.* Paper presented at the Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop, Aix-en-Provence, France

Pignataro, C. (2011). Skilled-based and knowledge-based strategies in television interpreting. *The Interpreters' Newsletter, 16*, 81-98.
Pöchhacker, F. (2004). *Introducing interpreting studies*. London: Routledge.
Pöchhacker, F. (2010). Media Interpreting. In Y. Gambier & L. van Doorslaer (Eds.), *Handbook of Translation Studies* (Vol. 1, pp. 224-226): John Benjamins.
Quaglio, P. (2009). *Television dialogue : the sitcom Friends vs. natural conversation*. Amsterdam, Philadelphia: John Benjamins.
R Development Core Team. (2010). A Language and Environment for Statistical Computing.
Rabiner, L. R. (1989). *A tutorial on hidden Markov models and selected applications in speech recognition.* Paper presented at the Proceedings of the IEEE.
Räsänen, O. J., Laine, U. K., & Altosaar, T. (2009). *An improved speech segmentation quality measure: the r-value.* Paper presented at the Interspeech.
Remael, A. (2008). Screenwriting, scripted and unscripted language: What do subtitlers need to know? . In J. Diaz-Cintas (Ed.), *The Didactics of Audiovisual Translation* (pp. 57-67). Amsterdam, Philadelphia: John Benjamins.
Romero-Fresco, P. (2011). *Subtitling through speech recognition: respeaking*. Manchester: St. Jerome Publishing.
Romero Fresco, P. (2012). Respeaking in Translator Training Curricula. Present and Future Prospects. *The Interpreter and Translator Trainer, 6*(1), 91-112.
Russo, M. (2005). Simultaneous film interpreting and users' feedback. *Interpreting, 7*(1), 1-26. doi:10.1075/intp.7.1.02rus
Schweda-Nicholson, N. (1987). Linguistic and extra-linguistic aspects of simultaneous interpretation. *Applied Linguistics, 8*, 194.
Sergio, F. S. (2013). Media Interpreting *Encyclopedia of Applied Linguistics*.
Setton, R. (1999). *Simultaneous interpretation: A cognitive-pragmatic analysis*. Amsterdam, Philadelphia: John Benjamins.
Shlesinger, M. (1994). Intonation in the production and perception of simultaneous interpretation. In S. Lambert & B. Moser-Mercer (Eds.), *Bridging the Gap: Empirical research in simultaneous interpretation* (pp. 225-236). Amsterdam, Philadelphia: John Benjamins.
Timarová, Š., Čeňková, I., & Meylaerts, R. (2015). Simultaneous interpreting and working memory capacity. In A. Ferreira & J. W. Schwieter (Eds.), *Psycholinguistic and cognitive inquiries into translation and interpreting* (pp. 101-126). Amsterdam: John Benjamins.
Timarová, Š., Čeňková, I., Meylaerts, R., Hertog, E., Szmalec, A., & Duyck, W. (2014). Simultaneous interpreting and working memory executive control. *Interpreting, 16*(2), 139-168. doi:10.1075/intp.16.2.01tim
Timarová, Š., Dragsted, B., & Hansen, I. G. (2011). Time lag in translation and interpreting: A methodological exploration. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Methods and strategies of process research: Integrative approaches in translation studies* (pp. 121-146). Amsterdam: John Benjamins.
Tissi, B. (2000). Silent pauses and disfluencies in simultaneous interpretation: A descriptive analysis. *The Interpreters' Newsletter, 10*, 103-127.
Tóth, A. (2011). Speech disfluencies in simultaneous interpreting: A mirror on cognitive processes. *SKASE Journal of Translation and Interpretation, 5*(2), 23-31.
Tóth, A. (2013). *The study of pauses and hesitations in conference interpreters' target language output*. Doctoral thesis. Eötvös Loránd University. Budapest.



Respeaking process

163x33mm (96 x 96 DPI)

Table 1. Characteristics of video clips used in Experiment 1.

	Interlingual respeaking clip	Intralingual respeaking clip
Duration (min)	4:55	4:47
Number of words	458	520
Speech rate (words per minute)	94	108
Number of sentences	35	40
Number of syllables	661	1214
Pause length (s)	109	66
Phonation time (s)	186	221
Speech rate 2 (number of syllables / duration) (s)	2.24	4.22
Articulation rate (number of syllables / phonation time) (s)	3.55	5.49
Pre-scripted	yes	yes
Number of speakers	one	one

Table 2. Characteristics of clips used in Experiment 2.

	Speech	News	Chat Show
Duration (min)	4:47	6:10	5:20
Number of words	520	935	707
Speech rate (words per minute)	108	152	131
Number of sentences	40	99	74
Number of syllables	1214	1890	1307
Pause length (s)	66	201	150
Phonation time (s)	221	169	170
Speech rate 2 (number of syllables / duration) (s)	4.22	5.10	4.08
Articulation rate (number of syllables / phonation time) (s)	5.49	11.18	7.69
Pre-scripted	yes	yes	no
Number of speakers	one	one	multiple

Table 3. Mean results for each dependent variable by condition (SDs in parentheses).

	Interpreters	Bilingual controls
EVS in ms		
Speech	2294 (1098)	1941 (916)
News	3442 (957)	3004 (838)
Chat Show	1857 (1145)	1744 (1134)
Pause length in ms		
Speech	485 (286)	505 (297)
News	508 (309)	477 (273)
Chat Show	575 (366)	620 (379)
Pause ratio		
Speech	2.57 (0.88)	1.7 (0.42)
News	1.76 (0.44)	1.73 (0.90)
Chat Show	1.83 (0.62)	1.19 (0.55)

Reply to Reviewers

Manuscript "Ear-voice span and pauses in intra- and interlingual respeaking: an exploratory study into temporary aspects of the respeaking process."

We would like to thank the Reviewers for their invaluable comments and suggestions. Following these comments, we decided to extensively rewrite the manuscript. We have added new data about interlingual respeaking and analysed it with comparable intralingual respeaking data (now in Experiment 1). We have also re-analyzed the data regarding intralingual respeaking (now in Experiment 2). We followed a rigorous data trimming procedure (exclusion of outliers, log transformations, exclusion of syntax-related functional pauses in respoken output, exclusion of short pauses below 200 ms). We analysed the data with linear mixed models rather than the analysis of variance, which made it possible to examine a potential link between our two dependent variables – EVS and pause length. We also introduced a new dependent variable: *pause ratio*, i.e. the length of pauses in the original clip to the length of pauses in the respoken output – in order to see the connection between pausing patterns in the experimental materials and the participants' behaviour. We also justified the selection of experimental groups in both experiments. Please find detailed replies to the Reviewers' comments below in blue.

Reviewer 1

We are very thankful for the Reviewer's comments, which have helped us improve the manuscript, especially with respect to the methodological concerns raised in the review. We hope that a thorough revision of the paper and the new statistical analyses which we performed to address the concerns make the interpretation of our results even more convincing now.

Did the subjects of the study receive some kind of instruction on the principles of respeaking? As the author(s) state(s), respeaking involves deliberate pausing by the respeaker in order to allow the software to correctly segment meaningful units. If subjects apply this principle, deliberate pauses are evidently going to pollute the data set. In that case, pausing is not related to cognitive effort, but by the training received. It is unclear to me at this stage what kind of instruction the subjects received on the process of respeaking itself.

We fully agree with the Reviewer as the instruction our participants received prior to data collection likely had an impact on the pauses made by the participants in the experiments. In order to make this issue clearer we have added the following information in the revised manuscript (Participants section):

None of the participants had any respeaking experience. Therefore, prior to the respeaking test, all participants underwent a two-day (16 hours in total) intensive training in respeaking. They were trained in the fundamentals of respeaking, including linguistic and technical skills, such as management of simultaneous resources, working memory, monitoring their respoken output, enunciation, punctuation as well as the creation of voice

profiles in the Newton Technologies speech recognition software and the FAB Subtitler Live subtitling programme.

The participants were instructed to respeak in "idea units" or "respeaking units" (Romero Fresco 2011, p. 60 and p. 108), that is to say whole phrases rather than individual words and to pause after the entire phrase has been uttered for the SR language and acoustic model to work effectively. An ideal respeaking unit is considered to be between five to seven words (Romero Fresco 2011); longer stretches of text may not be comfortable for viewers to read and may result in increasing the delay between the image and the accompanying subtitles.

Furthermore, in the new analysis reported in the revised version of the manuscript, we excluded all pauses following punctuation marks added by the participants when respeaking. This was done to eliminate deliberate pauses (which we refer to as 'functional pauses') from the analysis.

There is a considerable gender imbalance among the subjects. This cannot be changed, of course, but it should be mentioned, especially because part of the literature, including the literature referred to by the author(s), finds that pausing patterns are gender-specific (see Cecot 2001).

Indeed, our study was not balanced for gender. We added this information to the limitations discussed towards the end of the paper.

Some aspects of the transcription and alignment procedure need some clarification. For instance, the alignment of source (the video-recording) and target (respeaking) texts is done automatically using algorithms. Usually when using NLP, authors report on the accuracy of the NLP output, which is not the case here. It is hard to imagine that the author(s) didn't at least manually check part of the automatic alignments and determine the error margin. In addition, it is stated towards the end of the paper that the alignment of the interlingual respeaking task was done manually. It would be helpful to mention that earlier on.

The following text has been added in order to clarify the alignment procedure:

Automatic time alignment can be quite precise given the right input conditions. While the quality of time alignment is difficult to assess for any given input, it relies on an automatic speech recognition engine and thus suffers from a lot of the same problems. The minimum resolution for such systems is usually 10 ms and provided that the right word sequence is matched correctly, the difference between the automatic and reference alignment boundaries is generally below 50 ms (Chen et al., 2014). For various pragmatic reasons, the boundary shift is not often measured in such systems and other metrics are used instead (Räsänen at al., 2009).

Additionally, a sample of recordings was processed manually, revealing the metrics included in the table below. This shows that automatic and manual alignment results are similar. Due to space constraints, this table was not added to the manuscript.

Metric	Result
Number of boundaries in reference segmentation	210

Number of boundaries in studied segmentation	208
Number of hits	196
Hit rate (higher=>better)	93.33%
Over-segmentation rate (closer-zero=>better)	-0.95
Precision (higher=>better)	94.23%
Recall (higher=>better)	93.33%
F-measure (higher=>better)	93.78%
r1 (closer-zero=>better)	6.73
r2 (closer-zero=>better)	-4.04
R-value (higher=>better)	94.61%
Average WEBS	23 ms
Average WBBS	44 ms

In the revised version of the manuscript we described the manual alignment of the interlingual respeaking task in the description of Experiment 1, much earlier than in the first version.

The measurements and statistical analysis lack crucial information: (a) what pauses are taken into consideration? Usually, in interpreting studies and pausological studies, a minimum of 200 ms of absence of articulated sounds is required to be able to speak of a real pause. Given the reported averages of pause length, which seem rather low, I assume the author(s) used a lower minimum. That minimum should be made explicit.
Given that, it is likely that the distribution of pause lengths is extremely skewed, including many more ultrashort and short pauses than long pauses. If that is correct, the author(s) should discuss the usefulness of the ANOVA test.

We are very grateful to the Reviewer for pointing out these problems. Now, the entire data set has been re-analysed. Following Warren (2013), we excluded pauses that were shorter than 200ms from the new analyses. We also excluded functional pauses, which we took to be those related to inserting punctuation. Furthermore, we applied linear mixed effects (LME) models, instead of the analysis of variance to analyse our data. The following justification for the LME models has been added to the revised text:

This type of analysis combines the traditional ANOVA F1 and F2 analyses by treating participants and items as random effects and does not necessitate the aggregation of data over items or participants as it analyses them at the trial level.

We also log-transformed the data prior to fitting the models in order to work with more normally distributed data. The following information has been added to the text:

For the purposes of this analysis, all pauses were then log-transformed to normalise the distribution of the data.

As pauses necessarily increase EVS, the two dependent variables of this study are not mutually independent. So it might very well be that EVS variation can completely be accounted for in terms of pauses, in which case studying EVS is superfluous. Therefore, the author(s) should provide a method capable of determining what part of the EVS variation can be accounted for in terms of pause length.

We would like to thank the Reviewer for this helpful suggestion. The use of linear mixed models allowed us to include pauses as an additional predictor in modelling EVS data to see how much variance of the EVS is explained by pauses. This post-hoc analysis revealed no reliable influence of pauses. Additionally, we tested whether EVS and pauses are correlated on the assumption that if two measures are measuring the same construct, they should correlate (Kimberlin & Winterstein, 2008).

The following text has been added to report on the analyses in the revised manuscript.

Experiment 1
Even though there is a possibility that EVS and pauses might actually be two reflections of a similar processing (as pointed out by one of the reviewers), we found only a very weak correlation between EVS and pausing data (r=.117, p<.001). Furthermore, when pauses were added as a covariate to a post-hoc LME model of EVS, they failed to explain more variance in the data (p>.05).

Experiment 2
Similarly to Experiment 1, we checked for correlations between EVS and pausing data in Experiment 2, and found no reliable results (r=-.003, p=.882). Furthermore, when pauses were added as a covariate to a post-hoc LME model of EVS, they again failed to explain more variance (p>.05).

Since the analyses revealed very weak or no reliable correlations between pausing and EVS, and pauses did not reliably explain any additional variance, we decided to pursue the analyses of EVS and pausing independently.

The methodology obviously involves a lot more than producing the results that are reported (EEG, eyetracking, retrospective interviews). Could the author(s) briefly clarify what the purpose of all the additional tests was.

This part has been removed from the revised manuscript so as increase clarity and cohesion of the text. Only some of the information on retrospective interviews and what they revealed has remained as this was directly related to the difficulty of the respeaking task, which is an key issue in this paper.

It is surprising that the author(s) report(s) so little from the retrospective interviews. Was there nothing relevant to be collected from them on pauses and EVS?

The retrospective interviews turned out to be far less informative than we had expected. The only relevant findings have been used when discussing findings from Experiment 2:

Chat Show was the only clip with multiple speakers and in the post-test interview over 80% participants stated it was the multiple speaker programmes that were more problematic for them than single speaker clips. When asked about what was more difficult: fast speech rate and the number of speakers, 63% declared that it was the number of speakers that was making respeaking more demanding.

The author(s) report(s) clearly on their findings and discuss their results at length. There is some overlap between the results section and the discussion as the latter repeats the findings.

The results and the discussion part of the paper has been revised following the re-analysis of the data. The results section now presents the results while the discussion section includes our explanations of the findings.

I take issue with one conclusion the author(s) repeatedly draw(s) from the study, namely that speech rate has an effect on EVS (p. 20 & p. 21). This is not the case: the longest EVS occurs in respeaking the news programme, followed by the speech and the entertainment show respectively (p. 20). However, Table 2 shows that the speech rate of the entertainment show is higher than the speech rate of the speech.

This part is no longer included in the manuscript because we have extensively rewritten the results and the discussion part of the paper.

Throughout the paper: “pause length” is more elegant than “pause duration”.

This has been revised.

Title: “temporary aspects” should be “temporal aspects”

This has been corrected as suggested by the Reviewer.

p. 5: Donato (2003) is misrepresented: subjects did not interpret “the same text from English and German into Italian”; they interpreted an English and a German translation of the same Swedish source text.

The misleading information has been removed and the text has been reformulated to include only the most relevant information:

Donato (2003) applied an extended typology of syntax-based EVS units and found that the most frequently applied EVS units consisted of a noun phrase and a verb phrase.

p 5: the sentence explaining Adamowicz’ findings is confusing as it seems to contrast “spontaneous speech” and “planned utterance”, which are two different levels of analysis.

This has been changed into “spontaneous text” and “prepared texts”. These are the categories Adamowicz uses in line with the distinction introduced by Kopczyński between these two (Adamowicz 1989, p. 155).

p. 6: Defrancq (2015) did not find that short EVS is associated with higher quality and accuracy, but found that in most cases, short EVS does not lead to poorer quality, as claimed in the literature.

Thank you very much for pointing out this infelicity. The sentence now reads:

Empirical studies have shown that shorter EVS does not have to lead to poorer quality (Defrancq, 2015) or even may be associated with better quality and higher accuracy (Lee, 2002; Timarová et al., 2014).

p. 18, line 2: “M=706” probably “M=760” given the information provided just before. Incidentally, the author(s) might consider not to repeat the information.

This sentence does not appear in the manuscript as the results section has been extensively rewritten.

p. 18: “we wanted to find out about temporal aspects”: could this be more precisely formulated?

This has been reformulated:
By conducting the two experiments reported in this paper, we wanted to examine temporal aspects of the respeaking process, in particular EVS and pause length in intra- and interlingual respeaking of different TV genres as performed by three groups of participants: interpreters, translators and controls.

p. 20: “In general, it seems that there was no interpreter advantage”: I do not agree with the “advantage”. That is an evaluative claim about long or short EVS which is not supported by the research, nor by the literature.

Indeed, thank you for pointing that out. We have added the following explanation to avoid any misunderstanding:

These novel challenges may offset any potential interpreter advantage in respeaking (interpreter advantage is understood here as the extensive experience with working under strict temporal constraints).

p. 22: “the more experience the respeakers have in conference interpreting”: it seems that this is stated with more precision than previous statements which only refer to “interpreting”.

This sentence no longer appears in the manuscript due to extensive rewriting.

Reviewer 2

We thank the Reviewer for her (his) comments, which have greatly improved the manuscript. We hope that the thorough revision of the article, in which we incorporated the suggested changes, and the new statistical analyses convince the Reviewer of the publication value of the paper.

The introduction provides definitions of some of the main concepts dealt with in the article. However, the problem and the research questions are not stated early in the manuscript. It is advisable that these are clearly formulated in the first paragraphs of the introduction.

We have revised the introductory section to include the problem and the questions.

In this paper, we address the previously unexplored issue of temporal aspects of respeaking which affect the delay in live subtitles: ear-voice span (EVS) and pauses. We examine how the characteristics of to-be-respoken materials (including rate of delivery, scriptedness and the number of speakers) modulate temporal aspects of respeaking performance. Additionally, by examining the common ground respeaking shares with interpreting, we hope to start a discussion on respeaker competences and to better understand the respeaking process. This, we believe, can in turn translate into concrete solutions in respeaking training, mainly aimed at optimising the delay in live subtitling. Because intralingual respeaking shares temporal constraints with interpreting and because interlingual respeaking shares the process of transferring message from one language into another with both interpreting and translation, we wanted to find out if interpreters and translators are better predisposed to producing respoken subtitles with shortest delay possible than average bilinguals without any interpreting and translation experience.

The concepts of EVS and pauses are well defined and go accompanied by an extensive review of the literature, especially from the perspective of simultaneous interpreting process research. However, the concepts associated to the independent variables are not covered in equal depth (the respeaking tasks; assumptions about the role of previous experience, etc.) or even go unmentioned (description of text types; previous experience as interpreter OR translator; studies on TV interpretation of scripted or unscripted speech, etc.).

We have clarified our independent variables. They now include clip type and group. We have therefore reviewed EVS and pause studies involving text types as independent variables. The following text has been included in the manuscript:

EVS has also been linked to other factors, such as source text delivery rate (Lee, 2002), working with or without text (Lamberger-Felber, 2001) and sentence length (Lee, 2002). Barik (1973) measured temporal characteristics of interpretation of four types of texts: spontaneous speech, semi-prepared material, prepared "oral" material (a written-to-be-read-out speech) and prepared "written" material (an article). He found that interpreted texts usually included speaking for a greater proportion of the time than the original texts and this proportionality was greater for scripted than unscripted texts, as the latter have higher information density. However, due to a low number of participants, Barik did not find a consistent pattern of results when it comes to text types. Timarová et al. (2011) found a significantly smaller EVS when interpreting figures as opposed to verbs or beginnings of sentences, which means that interpreters try to produce figures as quickly as possible in

order not to burden the memory with such difficult non-contextual items with high informative content (Chmiel, 2015; Mazza, 2001). Adamowicz (1989) found smaller EVS in interpreting a spontaneous text as compared to prepared texts, while Díaz-Galaz et al. (2015) reported smaller EVS following advance preparation for the simultaneous interpreting task. It also seems that more experienced interpreters work with smaller EVS as Timarová et al. (2015) found a negative correlation between median EVS and days of interpreting experience. Taken together, these studies suggest that EVS depends on a combination of global (such as language combination) and local (such as propositions in the text) factors.

However, the literature review is short as such studies are scarce. As regards respeaking and assumptions about the role of previous interpreting/translation experience, there are no experimental studies including such factors to review since research on respeaking is still in its infancy.

The literature review is based on the assumption that intralingual and interlingual respeaking are comparable to shadowing and simultaneous interpreting. However, this relation is not clear, especially when considering the specific constraints of respeaking as live subtitling. A reference to sight translation or simultaneous interpreting with text on television contexts could provide a closer reference framework for this hybrid modality from an interpretation perspective.

We would like to thank the Reviewer for this comment. We have added a section relating respeaking to its 'elder sibilings' such as media interpreting on TV (under the heading: Respeaking as a hybrid modality). Also, we removed the comparison between respeaking and shadowing to increase clarity.

From an audiovisual translation perspective, references to live translation of films could provide a closer comparison to the task, than just "simultaneous interpreting" (see for example Russo, M. (2005). Simultaneous film interpreting and users' feedback. Interpreting, 7(1), 1-26.).

We have added information on the live translation of films and its resemblance to respeaking in the section Respeaking as a hybrid modality.

The literature review section could be reduced/condensed in order to appropriately address the relationship between respeaking and interpreting, with explicit reference to interpreting of television material or in television contexts, in order to describe the context and specific characteristics and difficulties of the task. Eg. Pöchhacker, F. (1997). " Clinton speaks German": A case study of live broadcast simultaneous interpreting. BENJAMINS TRANSLATION LIBRARY, 20, 207-216; Kurz, I. (1990). Overcoming language barriers in European television. Interpreting–Yesterday, Today, and Tomorrow, 168-175. Pochhacker, F. (2010). Media interpreting. Handbook of Translation Studies, 1, 224. Pignataro, C. (2011). Skilled-based and knowledge-based strategies in Television Interpreting. The Interpreters Newsletter 16, 81-98. SERGIO, F. S. (2013). Media Interpreting. Encyclopedia of Applied Linguistics. DOI: 10.1002/9781405198431.wbeal0757

A section on the relationship between respeaking and interpreting has been added, together with relevant literature. We would like to thank the Reviewer for suggesting the references to us.

Process research in interpreting has not focused on interpreting on television settings, but instead on interpreting of narrative, expository genres, typical of conference settings. This is a major caveat that has to be made before comparing the results of interpreting process research to the task of respeaking or live subtitling through respeaking.

We have added this point when discussing results of Experiment 1:

Since respeaking is more similar to media interpreting than regular conference interpreting, maybe such interpreter advantage could be seen if we compared media interpreters only with translators.

We have also mentioned it in the concluding part of the paper:

It would also be interesting to further explore potential benefits of having interpreting skills in respeaking by involving more homogenous and separate groups of conference interpreters, media interpreters and interpreting trainees.

There are other gaps in the literature review. For example, the assumption to test interpreters and translators is not addressed. If the hypothesis is the role of previous experience in interpreting, then why testing experienced and inexperienced translators? Moreover, in the results sections, the reader is informed that these two subgroups were collapsed into one single group. This is confusing and leaves the reader wondering about the research strategy.

We have revised the paper with the groups division made clearer in the following manner:

Based on the self-reported experience in interpreting, participants were divided into three groups: 22 interpreters (with at least two years of exposure to interpreting either in a professional context or during an intensive academic programme), 23 translators (with at least two years of exposure to translation either in a professional context or during an intensive academic programme), and a control group including 12 participants with no previous experience in translation and interpreting.

We performed additional statistical analyses (not reported in the paper) to additionally check the influence of the amount of the interpreting/translation experience on the participants' performance in the study. No reliable effects have been found (all *p*s>.05).

In the first experiment (where we compare intra- to interlingual respeaking), we analysed the data for interpreters, translators and bilingual controls. We assumed that both interpreters and translators would outperform controls due to their previous exposure to interlingual transfer in interpreting and translation. Additionally, interpreters would also outperform translators because they can better cope with temporal constraints in respeaking as they are similar to those in interpreting. In the second experiment, where we focused on intralingual respeaking of various texts only, we analysed the data only for interpreters and controls, assuming that the former would outperform the latter due to

their former exposure to temporal constraints in interpreting. We hope that this structuring of our experimental tasks and groups is now clearer and justified.

The effect of speech rate on EVS is not addressed in sufficient detail, considering that it is one of the independent variables.

We have clarified the independent variables. They now include group and clip type, rather than speech rate. Unfortunately, as we opted for authentic videos as experimental stimuli, speech rate was not matched well enough in the clips to serve as a dependent variable along with the number of speakers. Also, due to overlapping speech in one of the clips, its automatic analysis turned out to be impossible.

The design is not described in the Methods section. The authors may wish to provide the details of design early in the methods section, by describing the experimental design, dependent variables, independent variables and control of confounding variables.

We have added a description of the design to each of the Experiments in the following manner:

Experiment 1
The initial study design was 3 (group: interpreters, translators, bilingual controls) by 2 (clip type: interlingual, intralingual). However, during data collection many bilingual participants from the control group found the interlingual respeaking task too difficult and they failed to complete the task. Hence, due to insufficient data in the control group, further analyses of bilingual participants could not be performed. Therefore, finally our study followed a 2 (group: interpreters, bilingual controls) x 2 (clip type) mixed factorial design. The dependent variables were: EVS, length of pauses and the ratio of the pauses in the original to the pauses – length, pauses-ratio.

Experiment 2
We had a mixed factorial design for the intralingual respeaking task with clip type as an independent within-subject variable with four levels: speech, news, entertainment show, political chat show, and group as a between-groups independent variable with two levels: interpreters and bilingual controls.

Research questions and hypothesis are poorly stated (see below in Presentation of results).

We have revised the research questions and hypotheses by introducing detailed predictions for each experiment.

Experiment 1:
In the first experiment we wanted to see how pauses and EVS are modulated by the nature of the respeaking task: interlingual vs. intralingual. Additionally, we wanted to compare two groups of participants: interpreters (as those who have experience both with temporal constraints characteristic for interpreting and respeaking and with interlingual processing), and bilingual non-interpreting controls (who have experience neither with interlingual processing nor with temporal constraints). Thus, we used a mixed factorial design with task

(intralingual and interlingual respeaking) as a within-subject independent variable and group (interpreters and bilingual controls) as a between-groups independent variable.

We predicted that the interlingual respeaking task will be more difficult than intralingual respeaking, which will be manifested by longer EVS and longer pauses in the interlingual condition. We also expected that interpreting and translation experience would modulate EVS and pause length in both tasks. We predicted that in the interlingual condition, interpreters would respeak with shorter EVS and would produce shorter pauses than translators and controls, because they are used to coping with demanding time constraints when interpreting and because they are regularly exposed to interlingual processing. We also expected that translators would outperform controls (i.e. manifest shorter EVS and pauses) in the interlingual, but not in the intralingual condition, because they are used to interlingual processing in their translation experience.

Experiment 2:
We expected to obtain the main effect of group, that is we predicted that interpreters would provide respeaking with shorter EVS and shorter pauses than controls as they are used to linguistic processing constrained by similar time demands in professional interpreting assignments. We also expected the clip type to modulate EVS and pause length across participants. We predicted that the political chat show would generate the longest EVS, the longest pauses and the smallest pause ratio due to its high speech rate and the presence of multiple speakers and overlapping speech. This would be followed by news (due to information density and pre-scripted nature of the content), the chat show (due to its medium-paced speech and multiple speakers). Speech was expected to generate the shortest EVS, the shortest pause length and the highest pause ratio due to its slow speech rate and because it was delivered by one speaker only.

No information is provided as to the control of confounding variables such as, for example, previous experience of participants in real respeaking tasks. The number of participants is adequate and could be considered a relatively good sample for a population of translators and interpreters.

Since there were no professional TV respeakers in Poland at the time of conducting the study, no participant had any experience in that. This information has been added to the revised text:

The profession of a respeaker is in its nascent stage and there were no professional respeakers in Poland at the time when this study was conducted.

Speed of delivery. The authors should provide objective and more fine-grained data to compare a "slow" speech and a "fast" speech, such as total number of syllables; phonation time, articulation rate and speech rate. Since features such as pauses in the source speech are key in respeaking, more information about pause variation in the source speeches should be provided.

We have added a number of parameters describing the clips in the study in Table 1 and 2.

Table 1. Characteristics of video clips used in Experiment 1.

	Interlingual respeaking clip	Intralingual respeaking clip
Duration (min)	4:55	4:47
Number of words	458	520
Speech rate (words per minute)	94	108
Number of sentences	35	40
Number of syllables	661	1214
Pause length (s)	109	66
Phonation time (s)	186	221
Speech rate 2 (number of syllables / duration) (s)	2.24	4.22
Articulation rate (number of syllables / phonation time) (s)	3.55	5.49
Pre-scripted	yes	yes
Number of speakers	one	one

Table 2. Characteristics of clips used in Experiment 2.

	Speech	News	Chat Show
Duration (min)	4:47	6:10	5:20
Number of words	520	935	707
Speech rate (words per minute)	108	152	131
Number of sentences	40	99	74
Number of syllables	1214	1890	1307
Pause length (s)	66	201	150
Phonation time (s)	221	169	170
Speech rate 2 (number of syllables / duration) (s)	4.22	5.10	4.08
Articulation rate (number of syllables / phonation time) (s)	5.49	11.18	7.69
Pre-scripted	yes	yes	no
Number of speakers	one	one	multiple

As stated in one of the research questions, one independent variable is "clip difficulty", which is operationalized into two variables: speech rate and number of speakers. Then in the Methods section, the reader is informed that there is another variable: mode of delivery (scripted vs. unscripted), which has not been properly addressed in the literature review, and is not part of the research questions.

We have clarified the variables in the present version of the manuscript. The independent variables are now group and clip type. Because we used authentic materials, our clips turned out not to be matched well enough as regards speech rate to use the following orthogonal design 2 (speech rate: fast vs. slow) by 2 (number of speakers: one vs. many). Also, one of the clips had to be removed from analysis due to problems with automatic analysis of overlapping speech. Thus, we used clip type and included information about limitations of such a variable (impossibility to disentangle between the effects of speech rate and the number of speakers on the temporal aspects of respeaking performance).

Table 1. The information in this Table should be organized according to variable names. Eg. Speech rate; number of speakers, type of clip, type of show, and mode of delivery.

Table 2 is not referred to in the text.

Tables should have proper headings that describe the information they provide.

Tables and their references in the text have been revised according to the Reviewer's suggestions.

The procedure mentions that the experiment was conducted on an eye tracking device in which participants' eye movements were monitored and recorded. This is surprising since no mention had been made earlier in the manuscript of the study being an eye-tracking study. The reader then infers that the respeaking data was collected during an eye-tracking and EEG study. However, this could threaten the validity of the data collection process. The study reported is a study on language production in a very specific professional setting (respeaking on television for live subtitling). One may wonder about the effect of the whole eye-tracking and EEG configuration on the participants' pauses and EVS during the respeaking task. This issue should be addressed by the authors explicitly in the manuscript.

We fully understand Reviewer's concerns, however we feel that the fact that eye-tracking and EEG recordings were concurrently made had very little, if any, influence on the validity of data collection during the reported experiments. Any significant influence is highly unlikely due to the fact that eye-tracking was performed only remotely in a completely non-invasive manner – the tracking device (SMI RED 250) was attached just below the screen and looked more like an extended screen frame than an eye-tracking equipment. The EEG recordings were collected via Emotiv, which requires no cap, no gel nor any wiring. It is just a small 8-electrode plastic headset with a wireless connection placed on the back of the head. Emotiv is most often used by computer gamers for human-computer interaction.

The presentation of results is confusing and poorly organized. Authors should consider presenting the results according to each one of the research questions and hypothesis stated early in the manuscript. Descriptive statistics should be provided before inferential statistics. Descriptive statistics should be provided in a clear, unambiguous manner, either in text or table form (see APA guidelines).

The results and discussion have been thoroughly revised in line with the Reviewer's suggestions.

The first research question (average EVS for intralingual respeaking) is not answered explicitly in the text.

We described intra- vs. interlingual respeaking in Experiment 1. The following text has been added:

Results:
This analysis revealed that EVS in in the intralingual condition (M=2381ms) was significantly shorter than in the interlingual condition (M=4164ms) (b = 0.7944; SE=0.2059; t =3.858; p<.001).

Discussion:
In general, we expected the main effect of task type, with interlingual condition generating more cognitive effort (longer EVS, longer pauses, lower pause ratio) than the intralingual condition. These predictions were corroborated by our findings.

Temporal measures are provided in different units: seconds and milliseconds. Authors should use milliseconds throughout the paper.

This has been revised. We used seconds only to provide the characteristics of the experimental materials, in line with the tradition of presenting such data. However, throughout the results section, we use milliseconds.

The results should also stick to clearly defined independent variables. For example, one factor tested through ANOVA tests is "clip type". However, clip type is not defined as a variable in research question and hypothesis number 2, which clearly refers to testing the effect of "speech rate" and "number of speakers". Since speech rate and number of speakers are not included in the statistical analysis, question number 2 is not appropriately answered in the results and the hypothesis, as stated, was not tested.

We have reanalysed the data and written the results and discussion sections anew. We also clarified the description of our independent variables, which are now condition (intra- vs. interlingual respeaking) in Experiment 1and clip type and group in Experiment 2.

A similar issue affects research question and hypothesis number 3, about the effect of "previous translation OR interpreting experience" on EVS in respeaking. The analysis conducted did not test the hypothesis as stated, for which it should have had to compare both translators AND interpreters to control group. Instead, the analysis conducted tests

the hypothesis that respeakers with interpreting experience have different (shorter) EVS than respeakers with translation experience AND control groups.

The revised manuscript now includes modified research questions and predictions. However, whenever relevant, we corrected the text as suggested by the Reviewer. Thank you very much for pointing out this logical lapse on our part.

No statistical analysis is reported to test significant differences in pauses in intralingual vs. interlingual respeaking, which contradicts the title of the manuscript and leaves research question number 4 unanswered.

In the revised manuscript, we presented Experiment 1, which addresses this question directly by comparing intralingual and interlingual respeaking. The following result has been reported:
The only statistically significant result was that pauses in the intralingual condition (M=538ms) turned out to be longer than in the interlingual condition (M=853ms) ($b = 0.36$; $SE=0.05$; $t = 7.43$; $p<.001$).

Results provided for group effect on pause length (page 18, paragraph 2) are not related to any research question or hypothesis.

We have revised the research questions and hypotheses as presented above.

Authors should organize the discussion section along the research questions stated in the introduction.

The results and discussion sections have been to a large extent re-written and thoroughly revised in line with the Reviewer's comments.

The average EVS reported on page 19, paragraph 1, "EVS in intralingual respeaking was 2.26 seconds", this is the average EVS across different types of clips. In order to compare the results with previous findings, the authors may wish to choose the type of speech that is similar to the type of speech used in previous studies, e.g. the speech (as opposed to news programs, and entertainment shows, which were not studied in the studies mentioned).

On page 19, paragraph 1, the authors say "this result is in line with our both predictions: on the one hand it is longer than the average EVS in shadowing tasks found in previous studies". This comparison is not so straightforward, since this study did not compare shadowing with interpreting, but entirely different tasks in entirely different settings. Absolute values reported in one study cannot be taken at nominal value, but as a relative reference value only if studies are similar and subjects, data collection procedure, etc. are comparable.

We fully agree with this comment. We changed our predictions in the new version of the manuscript. The comparison with shadowing has been excluded. In the general discussion we only generally compare EVS values for respeaking with what is generally reported in the literature on EVS in interpreting.

On the same page, paragraph 2, the authors say "we may also partly explain our findings by looking at the visual characteristics of the experimental clips". The explanation that follows is a hypothesis and it should be supported by previous research on visual processing in interpreting. While it can be an interesting avenue for future research, it cannot be offered as an explanation, since it was not part of the design of the study.

The discussion section has been thoroughly rewritten and this explanation is no longer included in the manuscript.

Along the same lines, discussion on page 20, paragraph 3 states "a possible explanation could be their better trained working memory". Control of participants' working memory is not reported in the methods section. If participants' working memory was not measured, then it cannot be claimed that the participants have better trained working memory (better as compared to what?). The "interpreter advantage hypothesis" has been challenged in recent literature (see for example García 2014, The interpreter advantage hypothesis. Preliminary data patterns and empirically motivated questions. Translation and Interpreting Studies 9:2, 219-238), and assumptions made on it should be taken with caution.

The discussion section has been thoroughly revised to meet the Reviewer's requirements and this explanation does no longer appear in the manuscript.

The Discussion section should be properly organized and structured as to address the issues raised in the previous sections of the manuscript, specifically the research questions, in the broader context of the disciplines involved (interpreting studies and audiovisual translation).

The discussion section has been thoroughly revised to meet the Reviewer's requirements.

The authors say on page 25 "we also believe that future research can overcome the limitations of this study" but no limitations are stated or discussed.

We have described limitations of our study towards the end of the revised manuscript.

Title should be reworded to represent the results reported. On page 14, the reader is informed that EVS data was obtained only for the intralingual respeaking task and not for the interlingual respeaking task. This would warrant a modification of the manuscript's title in order to accurately describe the information reported. Likewise, no statistical analysis is reported to test significant differences in pauses in intralingual vs. interlingual respeaking, which contradicts the title of the manuscript.

We have included an additional analysis of intra- vs. interlingual respeaking in Experiment 1. We believe that the title now reflects the contents of the manuscript.

The abstract lacks any mention to the context of the research, the research questions, etc. It only refers to the method, participants, procedure and results. The claim "We found that mean EVS in respeaking was shorter than in interpreting and longer than in shadowing, reflecting cognitive effort" is misleading, since the study did not compare respeaking with interpreting and shadowing. The abstract states that "the findings are discussed in the

context of … applied research on audiovisual translation", but this discussion is not found on the text.

The abstract has been changed to reflect the results of the reanalysed data and new data on interlingual respeaking.

Figures provided are redundant and do not add any useful information. Figures lack proper headings, titles and labels for proper comprehension of the information provided.

We excluded figures presenting the results and replaced them with tables.

The following works are cited in text but not included in the reference sections:[list]

The Reference section has been revised to include all the works cited in the paper.

The manuscript does not contain mention of approval by an institutional review board for studies involving human subjects.

According to the regulations within the institution where the testing took place (which is in line with the state regulations for research involving human subjects) the type of research which is reported on in the present manuscript does not necessitate an approval by a review board. Only informed consent needs to be obtained from the participants before they can take part in the study. The following information has been added to the manuscript:

Before the test, informed consent was obtained from each participant.

The authors should check the journal guidelines and APA Style for consistency in in-text citations.

This has been revised.

Reviewer 3

We would like to thank the Reviewer for pointing out those aspects of our manuscript that needed correction. We hope that our revision of the manuscript made it free from all mistakes and ambiguities. Below is a list of our revisions and responses to the Reviewer's comments.

The methodology is relevant for the object of study, but:

Why did you put interpreters and interpreting trainees in the same group (same for translators and translation trainees)? I question whether the level of expertise plays an important role here, especially when comparing interpreters and interpreting trainees.

We changed the labels for participants groups to make it less confusing in the following manner:
Based on the self-reported experience in interpreting, participants were divided into three groups: 22 interpreters (with at least two years of exposure to interpreting either in a professional context or during an intensive academic programme), 23 translators (with at

least two years of exposure to translation either in a professional context or during an intensive academic programme), and a control group including 12 participants with no previous experience in translation and interpreting.

However, we are aware of the fact that our groups are not homogenous. Their makeup results from general constraints. It was difficult to obtain sufficient samples from homogenous populations (either professionals or trainees). The participants had to participate in a two-day training and then take part in lengthy experimental sessions. We have included information about this limitation in the manuscript:

Our groups were not completely homogenous because – for technical reasons – they included both interpreters and interpreting trainees, translators and translation trainees. We had to exclude a lot of data due to great variability, which may be avoided in the future by involving more homogenous and numerous experimental groups.

In page 18, it is explained that the differences between the participants from the interpreting group and the control group was statistically significant, even though p=0.08.

We have reanalysed the data and extensively rewritten the results and discussion section. This sentence does not appear in the manuscript anymore. However, we do point to marginally significant results in our study (with p values around .06 or .07). The level of significance is a generally held assumption in psycholinguistic research, but researchers do report marginally significant results even in journals with high impact factors (see for instance the article from our list of references by Christoffels, I. K., & De Groot, A. M. B. (2004). Components of simultaneous interpreting: Comparing interpreting with shadowing and paraphrasing. *Bilingualism: Language and Cognition*, 7(3), 227-240).

Page 6:

Suggestion: a more detailed description of intra-subject variability here.

We actually removed this section since other Reviewers suggested that the section on EVS is too lengthy. In fact, we felt that such a detailed discussion of EVS in interpreting (especially conference interpreting, which is not as directly comparable to respeaking as media interpreting, as pointed out by one of the Reviewers) was not directly relevant to the focus of the paper.

Please check punctuation in: Shorter EVS is advised when information density in the text increases, for instance when interpreting enumerations and numerical data because it lowers the memory load and leads to fewer omissions. Please reformulate.

This has been reformulated as follows:

Shorter EVS is advised when information density in the text increases (for instance due to enumerations or numerical data) because it lowers the memory load and leads to fewer omissions.

Page 7: This paragraph should be expanded to explain better those factors that are linked to EVS. For instance, when author(s) mention(s) the delivery rate, text type, and sentence

length they could briefly explain the results found by Barik, Lee, and Lamberger-Felber, which are only cited in parenthesis.

This information has been added as follows:

EVS has also been linked to other factors, such as source text delivery rate (Lee, 2002), working with or without text (Lamberger-Felber, 2001) and sentence length (Lee, 2002). Barik (1973) measured temporal characteristics of interpretation of four types of texts: spontaneous speech, semi-prepared material, prepared "oral" material (a written-to-be-read-out speech) and prepared "written" material (an article). He found that interpreted texts usually included speaking for a greater proportion of the time than the original texts and this proportionality was greater for scripted than unscripted texts, as the latter have higher information density. However, due to a low number of participants, Barik did not find a consistent pattern of results when it comes to text types. Timarová et al. (2011) found a significantly smaller EVS when interpreting figures as opposed to verbs or beginnings of sentences, which means that interpreters try to produce figures as quickly as possible in order not to burden the memory with such difficult non-contextual items with high informative content (Chmiel, 2015; Mazza, 2001). Adamowicz (1989) found smaller EVS in interpreting a spontaneous text as compared to prepared texts, while Díaz-Galaz et al. (2015) reported smaller EVS following advance preparation for the simultaneous interpreting task. It also seems that more experienced interpreters work with smaller EVS as Timarová et al. (2015) found a negative correlation between median EVS and days of interpreting experience. Taken together, these studies suggest that EVS depends on a combination of global (such as language combination) and local (such as propositions in the text) factors.

Page 10: About Piccaluga (2005): The mean duration of pauses was 768ms of the group as a whole. Please add this info.

This information has been added to the revised text as suggested by the Reviewer.

Method: Were the tasks carried out sequentially? Did they take a break? What is the role of fatigue here? Please explain.

We revised the method section and clarified that the order of the tasks was randomised in order to control for such confounding variables as fatigue.

Page 18: … the differences between the participants from the interpreting group and the control group was statistically significant (p=.08). How is p=.08 statistically significant?

We have reanalysed the data and extensively rewritten the results and discussion section. This sentence does not appear in the manuscript anymore. However, we do point to marginally significant results in our study (with p values around .06 or .07). The level of significance is a generally held assumption in psycholinguistic research, but researchers do report marginally significant results even in journals with high impact factors (see for instance the article from our list of references by Christoffels, I. K., & De Groot, A. M. B. (2004). Components of simultaneous interpreting: Comparing interpreting with shadowing and paraphrasing. *Bilingualism: Language and Cognition*, 7(3), 227-240).

Timarová, along the text.

This has been revised.

Abstract: First sentence, “interlignual” respeaking.

This has been revised.