Does it help to see the speaker's lip movements?
An investigation of cognitive load and mental effort in simultaneous interpreting

Anne Catherine Gieshoff
Zurich University for Applied Sciences (ZHAW)

Simultaneous interpreting combines auditory and visual information. Within a multitude of visual inputs that interpreters receive, the one from the speaker seems to be particularly important (Bühler 1985; Seubert 2019). One reason might be that lip movements enhance speech perception and might thus reduce cognitive load in simultaneous interpreting and hence, induce lower mental effort. This effect may be even more pronounced when noise is added to the source speech. This study was conducted to investigate cognitive load and mental effort during simultaneous interpreting (a) with and without the ability to see speaker's lip movements, and (b) with and without interfering noise. A group of listeners was included to control for task-related effects. Mental effort and cognitive load were measured using pupillometry, interpreting accuracy measures, and subjective reports. The facilitation hypothesis for lip movements was not confirmed. However, the pupillometric data suggests that lip movements may increase arousal.

Keywords: simultaneous interpreting; cognitive load; arousal; pupillometry; visual input.

1. Introduction

Simultaneous interpreting is a multi-sensory task. It involves auditory information as well as a large amount of visual input that needs to be processed at the same time as the auditory input (Bühler 1985; Anderson 1994; Setton 1999; Rennert 2008; Seeber 2017; Seubert 2019; Stachowiak-Szymczak 2019). Within the multiple types of visual information, interpreters seem to particularly appreciate the ability to see the speaker (Bühler 1985; Seubert 2019). This preference may be explained by a facilitation effect of lip movements. Models of interpreting (Setton 1999; Seeber 2017) as well as psycholinguistic research suggest that the speaker's lip movements may enhance speech perception and thereby facilitate simultaneous interpreting (Bernstein et al 2004; Thomas & Jordan 2004; Brancazio et al 2006; von Kriegstein et al 2008). In this study, I examined the effects of the speaker's lip movements on interpreters' performance. Background noise was an additional variable, and a group of listeners worked as a control group. A combination of methods, including pupillometry, assessment of interpreting accuracy and subjective reports were used to facilitate studying the effects of lip movements from different angles.

This section provides an overview in two parts of the theoretical background for this study. The first part examines the impact of lip movements. The second part deals with measuring cognitive load in simultaneous interpreting.

1.1 *Lip movements as visual input in simultaneous interpreting*

In a series of surveys and interviews, Bühler (1985) found that 95% of professional conference interpreters emphasized the importance of having a view of the speaker. She states that it might "provide additional information about the speaker and help[s] to understand his words" (51). Seubert's (2019) eye-tracking study seems to corroborate the importance of the speaker as source of visual input: professional interpreters indeed did tend to fixate on the speaker most of the time.[1]

Bühler's speculation that a view of the speaker might aid comprehension matches the predictions of Seeber's *Cognitive Load Model for Visual Input in Simultaneous Interpreting* (Seeber 2017). *Cognitive*

[1] However, her data also shows a huge variability between participants and presentation slides.

This is an Accepted Manuscript of an article published by John Benjamins in *Translation, Cognition & Behavior* on 19 May 2021. The Version of Record is available online: https://doi.org/10.1075/tcb.00049.gie

load refers to processing demands resulting from subtasks in interpreting—storage, perceptual auditory processing, cognitive verbal processing and verbal response processing—and to the degree to which these subtasks interfere with each other.[2] According to Seeber (2017), the cognitive load of visual input depends on its complementarity to the auditory input. He distinguishes in particular between *redundant* and *complementary* visual input. In redundant visual input, such as a written speech manuscript, the visual information duplicates the auditory input. Processing redundant information requires focusing the full amount of attentional resources on both the auditory and the visual channel and may thus considerably increase overall cognitive load. In contrast, complementary visual input, such as the speaker's lip movements and gestures, provides additional information that can help to disambiguate the auditory input. The amount of attentional resources required to process complementary information may thus be more or less neutral with regard to cognitive load (Seeber 2017). The idea of information complementarity is largely consistent with psycholinguistic models of speech perception, such as the *Fuzzy Logical Model of Speech perception* (Massaro & Cohen 1999) and the *Neighborhood Activation Model* (Tye-Murray et al 2007), according to which visual information can boost lexical candidates that fit the visual properties.

Empirical research using short monologues (Jordan & Sergeant 2000) or single words (Thomas & Jordan 2004; Brancazio et al 2006; Peelle & Sommers 2015) as stimuli suggest not only a neutral—i.e., no effect—but even a facilitation effect for lip movements. This facilitation effect seems particularly strong in adverse listening conditions; for instance, when speech stimuli are masked by noise. Lower response accuracy for single word recognition in noise suggests that noise negatively impacts speech perception in general (Sumby & Pollack 1954; Bernstein et al 2004); especially, in the L2 (Lecumberri et al 2010; Tabri et al 2010). Gerver (1974) made similar observations in shadowing and simultaneous interpreting: the numbers of errors, omissions and self-corrections was higher at lower signal-to-noise ratios. In these adverse conditions, listeners seem to benefit more from lip movements. The lower the signal-to-noise ratio, the higher the percentage of fixations on the speaker's lips (Vatikiotis-Bateson et al 1998) and the more the lip movements contribute to response accuracy (Sumby & Pollack 1954; Bernstein et al 2004). Taken together, these studies suggest that (a) lip movements enhance speech perception; (b) noise negatively affects speech perception and interpreting; and (c) the facilitation effect of lip movements is stronger at low signal-to-noise ratios.

Surprisingly, the facilitation effect for lip movements in speech perception was not corroborated in simultaneous interpreting. In three experiments using a simulated talker face, Jesse et al (2000) observed a clear advantage of lip movements to recognise single words in noise, but they did not find this effect in simultaneous interpreting. The interpreting performance of participants did not improve or decline with lip movements in the language pairs English-German and English-Spanish. There are, however, several limitations to their conclusion: first, the simulated talker face might have other effects than a human speaker. Second, the participants were untrained bilinguals and—as the authors themselves admitted—their performance was rather poor. Therefore, the complexity of the task might have masked the effects of lip movements. Studies on simultaneous interpreting are very different from psycholinguistic studies in that the former combine several subtasks whereas psycholinguistic studies only measure the recognition score for single words or even phonemes. Even though participants' renditions were apparently analysed with great care, the effects of lip movements might have been too weak compared to other effects, like task difficulty. Third, Jesse et al (2000) did not manipulate source speech intelligibility in the interpreting task by overlaying noise on the source speech, as they did in the word recognition task. Sumby & Pollack (1954), Jesse et al (2000) and Bernstein et al (2004) suggest that lip movements may only have a measurable effect when the speech signal is noisy. Thus, the facilitation effect for lip movements might only be noticeable when the source speech is covered by noise.

Four methodological improvements may contribute to shedding more light on the effects of lip movements in simultaneous interpreting and to countering the limitations of Jesse et al's (2000) study. First, the study should choose participants with training in interpreting, since they are expected to be less sensitive to the difficulty of the task. Second, results from experimental informants should be compared with those from a control group of listeners who are not asked to perform any task. This might make it possible to observe the effects of lip movements without interfering effects of interpreting subtasks. Third, adding noise to the source speech may help to establish more precisely

[2] These processing demands are assumed to require a certain amount of attentional resources to be successfully processed (Seeber 2011). A distinction will be made later between the notions of *cognitive load* and *cognitive effort* (Gile 2009). References to research by other authors, however, respect their original usage.

when lip movements are an advantage. Fourth, Jesse et al (2000) only used task performance measurements to determine the effect of lip movements. However, interpreting performance might not have been sufficiently sensitive. Physiological measures and self-reports may contribute to gaining more insights into the effects of lip movements on cognitive load and mental effort in interpreting. The next section addresses measurements of cognitive load and mental effort in interpreting.

1.2 *Tapping into cognitive load and mental effort*

According to Seeber (2015), cognitive load in simultaneous interpreting can be empirically studied in three ways:[3] (1) by asking participants about their experience (self-reports); (2) by measuring their interpreting performance; and (3) by taking psychophysiological measurements, which involve different parameters of the physiological stress response. The study reported on here combines all three ways in a multi-method approach (see Halverson 2017), to obtain a comprehensive picture of the effects of lip movements on cognitive load in interpreting.

The first way, participants' self-reports, provides insights into participants' own perception of the task. Self-reports with rating scales and retrospective comments seem—often, but not always—to match interpreting performance (Moser-Mercer et al 1998; Ivars & Calatayud 2001; Korpal 2016; Wu 2019). In particular, studies on visual input (Rennert 2008) and remote interpreting (Roziner & Shlesinger 2010) report a mismatch between self-reports and measures of interpreting quality: interpreters' experience and performance satisfaction were in contrast to more objective measures of interpreters' renditions. Still, self-reports seemed interesting for this study to check whether the task difficulty as perceived by the participants would match task manipulations.

Measuring task performance requires operationalising interpreting quality. This is not an easy endeavour, as interpreting quality is a multidimensional concept with aspects like sense consistency, completeness, logical cohesion, correct use of terminology, grammar, fluency and intonation (Lee 2008; Zwischenberger 2010). So far, the main methods for measuring and describing interpreting quality include ratings (Anderson 1994; Jesse et al 2000; Roziner & Shlesinger 2010; Rosendo & Galván 2019), qualitative analyses (Rennert 2008), error/omission analyses (Gerver 1974, 2002; Moser-Mercer et al 1998) and analysis of accuracy and completeness based on discourse linguistics (Dillinger 1990; Tommola & Lindholm 1995; Hild 2015). As interpreters rank sense consistency and completeness among the most important criteria (Zwischenberger 2010), the present study operationalises interpreting quality as "interpreting accuracy" and articulates interpreting accuracy as the number of segments that were (a) rendered and (b) consistent with the corresponding segment in the source speech.

As for the third way to study cognitive load in simultaneous interpreting, Seeber (2015) mentions pupillometry in particular, which has a high temporal resolution. Psychological research suggests that pupil size increases with task difficulty (Kahneman 1973; Granholm et al 1996; Engelhardt et al 2010; Krejtz et al 2018) and this is usually interpreted as an indicator for cognitive load (review of pupillometric studies in van der Wel & van Steenbergen 2018). This effect seems to hold true for simultaneous interpreting as well. The first study with pupillometry, by Hyönä, Tommola & Alaja (1995), found significant differences in mean pupil size during listening, shadowing and interpreting. Pupil sizes were largest during simultaneous interpreting (mean difference between shadowing and interpreting about 0.5 mm) and smallest during listening (Hyönä et al 1995). Seeber & Kerzel (2012) contrasted verb-initial and verb-final sentences during simultaneous interpreting and identified statistically significant differences between both conditions: mean pupil dilation was about 0.05 mm larger in verb-final sentences compared to verb-initial sentences (Seeber & Kerzel 2012). Pupillometry was also used in the present study as a physiological indicator of cognitive load.

Besides task difficulty (Hyönä et al 1995; Seeber & Kerzel 2012; van der Wel & van Steenbergen 2018), pupils respond to a range of rather undifferentiated and confounding factors, like light (Brown & Page 1939), wakefulness (Lowenstein, Feinberg & Loewenfeld 1963), pain (Chapman et al. 1999) and emotional state (Oliva & Anikin 2018). In view of this large range of factors, Kahneman (1973) attributed pupillary responses to a state of general arousal, where cognitive load can be one factor of arousal. Physiological arousal occurs when the body responds to a stressor—like cognitively demanding task—but can also depend on circadian rhythms (Fisk et al 2018) or be affected by

[3] Seeber (2015) suggests a fourth, analytical approach, by which he essentially means developing a model of cognitive load.

stimulating substances like coffee (Yoon & Danesh-Meyer 2019). According to Kahneman (1973), it is therefore more accurate to interpret pupil dilations as a physiological arousal that reflects mental effort, rather than cognitive load.

Even though Kahneman (1973) assumes that mental effort usually corresponds to cognitive load or task demands, this is not necessarily always the case: A person might be tired, distracted or overwhelmed by the task, independently of task demands. According to van der Wel & van Steenbergen's (2018) meta-review on pupillometric studies, mental effort seems so far to be a meaningful approach for pupillary responses in cognitive tasks. As suggested by Ehrensberger-Dow et al (2020), I will distinguish between both concepts. I will use *cognitive load* to refer to task demands as perceived by the informants (task difficulty ratings), and *mental effort* to refer to the reaction to task demands as observed (pupillometry, interpreting accuracy). Based on the literature reviewed above, the relationship between the measurements and the concept are assumed to be such that higher measures correspond to higher cognitive load (Moser-Mercer et al 1998; Ivars & Calatayud 2001; Korpal 2016; Wu 2019) or mental effort (Hyönä et al., 1995; Seeber & Kerzel, 2012) respectively. The case is assumed be inverse for interpreting accuracy: higher task demands require higher mental effort, but typically lead to a drop in performance (see Hild 2015; Moser-Mercer et al 1998; Gerver 1974, 2002).

2. Materials and methods[4]

The experiment included two variables—(1) visual presentation (lip movements vs no lip movements); (2) auditory presentation (noise vs no noise)—within two tasks: interpreting, for the experimental group, and listening, for the control group (Table 1). This set-up aimed to find answers to the following *research questions* (RQ):

1. Do lip movements decrease cognitive load and induce lower mental effort in simultaneous interpreting and listening?
2. Does noise increase cognitive load and induce higher mental effort in interpreting and listening?
3. Is the effect of lip movements stronger when the source speech is overlaid with noise?
4. Are cognitive load and mental effort higher in interpreting than in listening?

If lip movements decrease cognitive load and induce lower mental effort (RQ1), the facilitation effect of lip movements should translate into lower task difficulty ratings, smaller pupil sizes, and (for interpreting) better interpreting accuracy. If noise increases cognitive load and induces higher mental effort in interpreting and listening (RQ2), interpreting accuracy should be lower, while task difficulty ratings and pupil sizes should be larger. If the effect of lip movements is indeed stronger when the source speech is overlaid with noise, then task difficulty ratings, interpreting accuracy, and pupil sizes should show an interaction of noise and lip movements. RQ4 was not central to this study but may still be visible in the results. As reported by Hyönä, Tommola & Alaja (1995), pupil sizes may expected to be larger in interpreting than in listening.

INSERT TAB 1 ABOUT HERE

Interpreting		Listening	
lip movements, noise	no lip movements, noise	lip movements, noise	no lip movements, noise

[4] The study was part of a larger research project (Gieshoff 2018). This paper discusses the results of self-reports, interpreting accuracy and pupillometry. Further analyses are included in Gieshoff (2018, forthcoming).

lip movements, no noise	no lip movements, no noise	lip movements, no noise	no lip movements, no noise

Table 1. Overview of experimental conditions regarding source speech.

2.1 *Participants*

Thirty-one participants, 17 translation trainees and 14 conference interpreting trainees in their last year at the University of Mainz, gave their informed consent to participate in the study after having received information about the experimental procedure and the data to be collected. In accordance with the declaration of Helsinki, they were further informed that they could withdraw from the experiment at any moment without providing a reason. They received 10 € or ECTS points in exchange for their participation. All participants confirmed that they felt good and were in good health. Self-reports confirmed no caffeine or drug intake two hours prior to the experiment. All participants spoke German as their A-language (L1) and English as their B- or C-language (L2 or L3). At that point in time, the interpreting trainees had completed at least three full semesters of simultaneous and consecutive interpreting training from their B-and C-language into their A-language. Translation trainees had at least the same experience in translating. While the interpreting trainees interpreted the four speeches ("talkers"), the translation trainees ("listeners") actively listened to the speeches without any coinciding secondary task. Three interpreting trainees were excluded from the analysis of the pupillometric data due to the large amount of track losses (>50%).

2.2 *Experimental stimuli*

The procedure consisted of two parts: a pre-test and the actual main experiment. The material for the pre-test consisted of 64 English nouns selected from among the 5000 most frequent lemmatized words of American English (Davies 2008). All words were concrete and monosyllabic (word length in letters, *M*=4.63, *SD*=0.2). They were spoken by an American native speaker and recorded as sound files. All words were grouped into four lists of 16 words each. The experiment was programmed in Psychopy (Peirce 2007; Peirce et al 2019), a programming tool designed to set up experiments. In Psychopy, the volume of a sound object can be set on a scale from 0.0 (silent) to 1.0 where 1.0 is the maximum volume of the soundcard. For this reason, the volume is not given in dB but based on Psychopy's volume scale. Each list was then overlaid with noise at four different signal-to-noise ratios ranging from level 0.1 to level 0.4 (10% to 40% of the maximal sound level of the sound card). The signal-to-noise ratio that was applied to each list was randomized between participants.

The stimuli for the main study consisted of four English speeches from the EU Speech Repository (European Commission 2009a, 2009b, 2012a, 2012b) on different topics: air travel, the Greek economic crisis, work conditions, and demographic change. They were shortened to approximately 590 words (*M*= 588, *SD*=5.23) and modified in order to obtain four speeches as comparable as possible. The structure of the first paragraph was the same across all speeches and served as a "warm-up" for the talkers. It contained the usual introductory expressions and mentioned the topic of the text twice. The speeches did not contain any numbers or proper names, which are potential local problem triggers. Filler sentences—i.e., evident information with no effect on text coherence—were introduced. All underlying logical relations within the text were made explicit by using conjunctions. All sentences were in the active mode and had a maximum of one subordinate clause (words per sentences, *M*=12.5, *SD*=2.2). The number of functional (grammatical) words and type/token ratio were used as indexes for information density. In every text, functional words made up approximately 40% of all words (ratio of functional words, *M*=0.4, *SD*=0.03; type/token ratio, *M*=0.48, *SD*=0.05). Finally, beyond this quantitative evaluation of information density, several measures were applied to reduce the information density, to allow talkers to catch up if they missed the message. Essential messages were repeated in different words.

All four speeches were read aloud by the same American native speaker and videotaped, in order to ensure the same conditions for each participant. The speech rate was kept constant, at 140 words per minute, within and between texts. These videos were used to create the factorial design *visual presentation × auditory presentation,* so that each speech was presented in one of the four resulting

conditions: lip movements + no noise; lip movements + noise; no lip movements + no noise; and no lip movements + noise. The condition with visible lip movements ("dynamic condition") showed the speaker's whole face as a video stream, so that the lip movements were visible. The speaker held his head as still as possible, so that head movements were reduced to a minimal extent. The video was displayed at a resolution of 1920 × 1080, a sampling rate of 16384 kbits/s and 29.97 frames per second. In the condition without lip movements ("still condition"), the sound stream remained the same, but the video stream was replaced by a freeze frame of the speaker's face in the same resolution. Hence, the two conditions differed only with regard to whether the lips were moving or not. This design should also help to avoid distortions of the pupillary data due to illumination effects. Sound was played in stereo at a volume level of 0.1. In the noise condition, white noise was added to the audio stream; in the no noise condition the source speech was played without noise. In order to reduce the potential of speech-related effects, I created two counterbalancing subgroups (A and B) for both experimental and control informants and switched the speeches in the dynamic condition and the still condition (Figure 1). The order of presentation was randomized.

INSERT FIG 1 ABOUT HERE

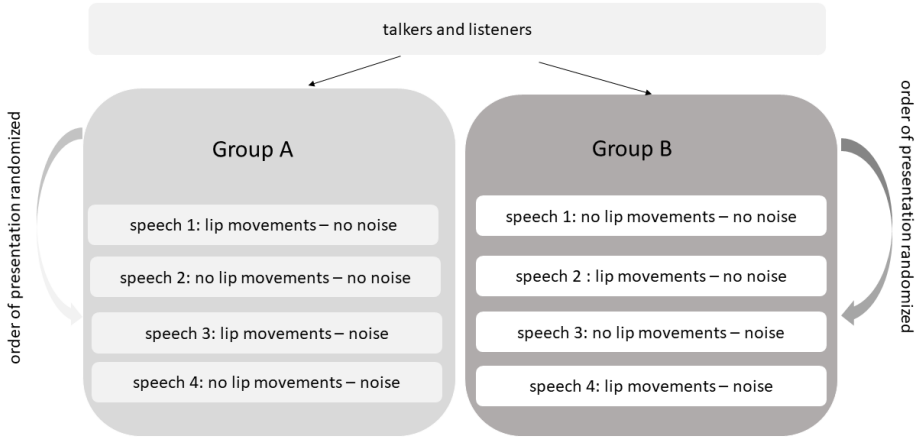

Figure 1. Talkers and listeners were randomly assigned to two groups in order to counterbalance speech related effects.

2.3 *Apparatus*

Pupillary data was recorded with a Tobii TX300 eye tracker with a sampling frequency of 120 Hz. The eye tracker was placed on a desk, underneath a 23"-computer screen with a resolution of 1920 × 1080 pixels. Participants sat in front of the eye tracker at a distance of approximately 64 cm from the screen, the distance at which the eye tracker achieves the highest accuracy, according to the manufacturer (Tobii 2010, 41). The chair was moved until the participants' eyes appeared in the centre of the calibration screen. As all participants were Caucasian, the bright pupil eye-tracking method was used. No chin or forehead rest was used because (1) participants were required to speak during the experiment and a chin rest would compromise articulatory movements; and (2) previous experiences with forehead rests at the laboratory of the Johannes Gutenberg-Universität Mainz were unsatisfactory. Participants felt impaired by the forehead rest and no benefit for data accuracy or precision was reported. According to the company, the Tobii TX300 corrects the raw pupil size during recording for changes in gaze direction or distance to the eye tracker (Tobii support, personal communication, 20 July 2016). A post-hoc analysis revealed a very weak correlation of pupil sizes with the gaze coordinates—correlation of the right pupil with the x-coordinates: $r(3971500)=-0.019$, $p<0.01$; correlation with the y-coordinates: $r(3971500)=0.027$, $p<0.01$—suggesting that the internal algorithms of the eye-tracker system effectively corrected pupil sizes for the gaze coordinates.

Auditory and visual stimuli were presented using Psychopy. The videos were displayed in full screen mode. The refreshment rate of the computer screen was 60 Hz. The sound card used during the experiment had a sampling rate of 192kHz. Recordings were collected using a microphone mounted on the headphones. The auditory stimuli were played through headphones and the overall sound pressure level was kept constant during the whole experiment.

2.4 *Procedure*

Participants were tested individually at the laboratory of the Faculty of Translation Studies, Linguistics and Cultural Studies of the University of Mainz. Sessions took place between 11 am and 5 pm in order to avoid circadian artefacts in the pupillary data. The lightning conditions remained constant in the testing room (no natural light) over all sessions. The entire procedure lasted about 40 minutes per participant; the pre-test lasted about 10 minutes and the main experiment lasted about 30 minutes. The aim of the pre-test was to identify participants' individual signal-to-noise ratios where they could correctly recognize 75% of random, but frequent and concrete English words (12 words) that were presented at four different signal-to-noise ratios. The resulting signal-to-noise ratio was then applied to the source texts in the condition with noise.

After the pre-test, the experiment, with four trials, started. Each trial included one of the speeches so that every listener listened to and every talker interpreted each one of the speeches. Participants were first instructed to fixate on a white cross on a black background (that served as a baseline epoch in order to obtain the pre-trial pupil size) and a cross centred on a picture of the speaker for ten seconds each. The picture of the speaker, instead of a white screen, was chosen in order to avoid light adaption of the pupil during the speech. Participants started the source speech by pressing a key.

The source speech was either presented as a video with visible lip movements (dynamic condition) or as an audio track with a freeze frame of the speaker's face (still condition). In both conditions, the speaker's face was visible, but the speaker's lips only moved in the dynamic condition. In the noise condition, white noise at the predefined volume was played simultaneously with the audio track. Talkers simultaneously interpreted the source text from English into German; listeners listened to the source text in English. After the interpreting or listening task, participants were asked to rate the video and audio quality, the speech difficulty and the speech rate on a four-point scale. Rating video and audio quality should ensure that the experimental conditions were clearly distinct.

3. Analysis and results

3.1. *Subjective reports*

A paired Wilcoxon signed rank test revealed significant differences between the dynamic and the still conditions ($V=2817$, $N=121$, $p<0.001$). A similar result was found for the distribution of the ratings for the auditory presentations: in a paired Wilcoxon signed rank test, the distribution of the ratings differed significantly between the noise and the no noise conditions ($V=4206$, $N=124$, $p<0.001$), suggesting that the experimental conditions were very clearly distinguishable.

Regarding speech difficulty and speech rate ratings, I conducted an ordinal regression for each rating, with the respective rating as response variable and random intercepts ($SD=1.05$) for participants. In both cases, fixed effects included *auditory presentation*, *visual presentation*, *task*, the individual *signal-to-noise ratio,* and *group* (A or B). P-values of the effects were estimated comparing models with and without the predictor in question, with likelihood ratio tests of cumulative link models. P-values of each level were estimated using a Wald test. All analyses were carried out with the R version 3.3.2 (R Core Team 2016) with the package *Ordinal* (Christensen 2015).

For text difficulty, the predictor variable *auditory presentation* (reference level, noise condition) was highly significant (estimate= –0.912, SE = 0.243, Z= –3.783 $p=0.0002$). The probability of a one-unit decrease of the rating (for example, *good-okay; okay-difficult*) dropped by 17.9% in the condition without noise compared to the condition with noise. There was also a tendency for a *task effect* (reference level, listeners, estimate= –0.7951, SE=0.433, $p<0.066$). The probability for a one-unit decrease of the rating *(very easy-easy, easy-okay,* or *okay-difficult)* dropped by 21.3% for the listeners compared to the talkers. The predictor variable *speech* did not improve the model when the model already contained the predictor variable *auditory presentation* ($X^2(7,2)=1.3548$, $p=0.507$). Table 2

shows the results for each predictor obtained by comparing the model with and without the predictor in question.

INSERT TAB 2 ABOUT HERE
---

predictor variable	likelihood ratio	DF	p-value
auditory presentation	15.553	1	< 0.001
task	3.201	1	0.069
visual presentation	1.034	1	0.390
group	5.957	3	0.178
signal-to-noise ratio	1.240	1	0.265

Table 2. Likelihood ratio, degrees of freedom and p-value obtained by model comparison for the response variable speech difficulty rating.

For speech rate, none of the tested predictor variables reached significance. The results for each predictor obtained by model comparison are summarized in Table 3.

INSERT TAB 3 ABOUT HERE
---

predictor variable	likelihood ratio	DF	p-value
auditory presentation	1.900	1	0.168
task	0.841	1	0.359
visual presentation	0.070	1	0.791
group	0.122	3	0.989
signal-to-noise ratio	0.026	1	0.871
speech	3.645	3	0.303

Table 3. Likelihood ratio, degrees of freedom and p-value obtained by model comparison for the response variable speech rate rating.

3.2 *Interpreting accuracy*

As in Jesse et al (2000), interpreting accuracy was assessed on the basis of the recordings from the talkers and defined as whether the rendition of a given segment was consistent with the corresponding segment in the source speech or not—i.e., if the translation of a given segment was either entirely missing or inconsistent with the corresponding segment of the source speech. This approach was refined with the use of categories. In order to account for the eventuality that talkers—especially, under more difficult conditions—may omit redundant or repetitive information or restructure segments, I divided all source speeches into small segments and grouped speech segments into five categories: core segments, secondary information, repetitions, context information, and fillers. *Core segments* were the "main theme" of the speech. They generally included the verb and its complements (subjects, objects, adverbial complements); i.e., all segments that are essential to understanding the sentence or the speaker's intention. *Secondary information* included information like adjectives, intensity particles or adverbs and adverbial structures such as information on time and place that could be omitted without affecting sentence structure or text coherence. *Repetitions* corresponded to segments that had already appeared in the speech, even if formulated differently. Discourse markers and—in a broader sense—logical sentence connectors or phrases that give information about how the sentence connects to the previous one were labelled *co-text information* (examples: I will give you an example; first, ... second, ...; because; I point out, etc.). Finally, *fillers* covered all segments with no

information at all. These sentences were deliberately introduced in the experimental material in order to allow participants to catch up if necessary.

Four recordings per participant were analysed, a total of 55 recordings. One recording was missing for technical reasons. On average, 39.75% of all interpreting segments were omitted (MD=39.41, range 16.07–76.88%). Omitted segments were treated as mistranslations, so participants rendered 55.33% of all segments correctly (MD=54.73%, range: 39.04–76.88%). Core segments were correctly rendered in 61.24% of the cases, followed by context information (52.45%), repetitions (51.17%), filler sentences (49.29%) and secondary information (45.82%). Core segments also had the lowest proportion of missing segments (32.75%), followed by repetitions (43.71%), context information (45.31%), filler sentences (48.57%) and secondary information (50.77%).

A logistic mixed model was constructed to analyse translation accuracy in each speech. Missing segments were treated as mistranslations, as they could result from cognitive overload. The effect on interpreting accuracy was captured with participant-by-trial random intercepts (SD=0.526) and random intercepts for participant (SD=0.331). Fixed effects covered *visual presentation* (reference level, condition with lip movements), *auditory presentation* (reference level, no noise condition), *signal-to-noise ratio* (reference level, condition with noise level=0.1) and *segment category* (reference level, core information). P-values for each fixed effect were approximated with maximum likelihood by adding fixed effects one by one and comparing the model with and without the effect in question. Results of model comparisons are reported in Table 4. All analyses were carried out using the package *lme4* (Bates et al 2015). Plotting was done using the R-package *ggplot2* (Wickham 2009).

There was a statistically significant effect for *auditory presentation,* noise–no noise (estimate= –0.738, SE=0.148, z= –4.994, p<0.001). According to the model, the probability for correctly rendering a speech segment decreases across all segment categories by approximately 17.26% when noise is added to the speech. There was a statistically significant effect for all segment categories. Compared to core segments, context information was about 5.74% less probable to be correctly rendered (estimate= –0.261, SE=0.099, z= –2.620, p<0.009). For filler segments, the probability decreased by about 14.86% (estimate= –0.642, SE=0.128, Z= –5.020, p<0.001). The response accuracy for repetitions decreased by 8.62% (estimate= –0.384, SE=0.075, z= –5.108, p<0.001) and by 15.56% for secondary information (estimate= –0.670, SE=0.052, z= –12.886, p<0.001) compared to core information segments. Further predictors like *visual presentation* or *signal-to-noise ratio* were not statistically significant (Table 4). A visual-by-auditory presentation interaction did not reach significance ($X^2(10,2)=3.887$, p=0.143).

INSERT TAB 4 ABOUT HERE
---

predictor variable	likelihood ratio	DF	p-value
auditory presentation	18.941	4,1	<0.001
visual presentation	2.059	4,1	0.151
auditory presentation * visual presentation	3.887	10,2	0.143
signal-to-noise ratio	2.179	4,1	0.140
segment category	182.657	7,4	<0.001

Table 4. Likelihood ratio, degrees of freedom and p-value obtained by model comparison for the response variable translation accuracy.

3.3 *Pupillometry*

Prior to the analysis, I removed blink artefacts and invalid data points. *Blink artefacts* are defined as sudden drops in pupil size (a difference to the preceding data point that exceeds the lower or the upper interquartile range of the mean of all differences by more than 1.5 times). *Invalid data points* were identified based on the validity codes provided by the eye-tracker. Gaps of up to 500 ms (the approximate duration of a blink) were replaced by linear interpolation. Of the data, 2% were considered a blink artefact or an invalid observation and were replaced by linear interpolation. Subsequently, pupil sizes were standardized separately for each participant and each trial based on the baseline epoch.

I used a growth curve analysis (Mirman 2014) on the standardized pupil sizes to analyse the time course of pupil dilation during the source speeches. For the time course, time windows of 500 ms were chosen because the model failed to converge with smaller time windows. Visual inspection revealed a polynomial slope of the data (compared to the normalized data depicted as point ranges in Figure 1). I thus chose to add a quadratic term to fit the model.

The random effect structure covered trial-by-participant random slopes on all time terms (trial-by-participant random slope on the intercept: *SD*=0.078, trial-by-participant random slope on the linear term: *SD*=5.59, trial-by-participant random slope on the quadratic term: *SD*=3.632). Fixed effects were *task* (listening–interpreting), *auditory presentation* (noise–no noise), *visual presentation* (dynamic–still). In order to exclude confounders, I added the participants' text difficulty and speech rate ratings, the source speech and the applied signal-to-noise ratio. Fixed effects were added one by one and compared to the reference level (still condition with noise). P-values were approached using maximum likelihood.

The model revealed a significant effect for the linear (*b*= –6.004, *SE*=1.008, *t*= –5.956, *p* <0.01) and the quadratic time term (*b*=3.417, *SE*=0.359, *t*=9.513, *p*<0.01), indicating a rapid decrease of the pupil size at the beginning of the speech that flattens towards the end of the speech. Furthermore, there was a significant effect of *task* on the intercept (*b*= –0.165, *SE*=0.017, *t*= –8.762, *p* <0.01) and the linear term (*b* = –2.639, *SE*= 1.091, *t*= –2.418, *p*<0.05), reflecting overall smaller pupil sizes during listening than during interpreting, and a faster decline of pupil size during listening than during interpreting. *Visual presentation* was significant on the intercept indicating overall smaller pupil sizes during the still condition than during the dynamic condition (*b*=0.039, *SE*=0.016, *t*=2.461, *p<0.*05). The effect of *visual presentation* on the linear time term was marginally significant (*b=1.970, SE*=1.069, *t*=1.843 *p*=0.067). *Auditory presentation*, *text difficulty ratings*, *speech rate ratings*, *signal-to-noise ratio*, *speech* or *trial* did not improve the model. The results of the model comparisons for each predictor variables are summarized in the appendix. In Figure 2, the model fit (ribbons) is plotted against the observed data (point ranges). The left facet shows the model fit for the talkers; the right facet, the model fit for the listeners. The shades of grey code the visual presentation (dynamic and still).

INSERT FIG 2 ABOUT HERE

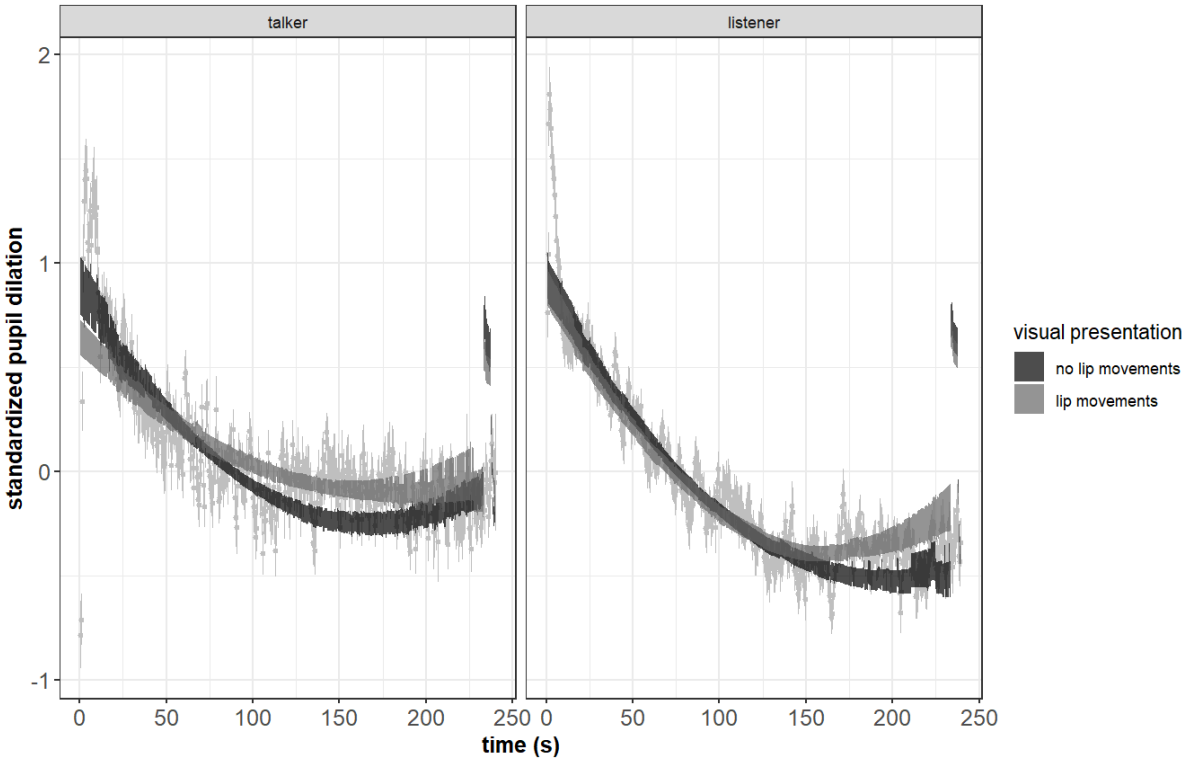

Figure 2. Model fit for a linear mixed model with visual presentation and task as fixed effects. The colour coding corresponds to the visual presentation (lip movements vs no lip movements). Fitted

values for talkers are plotted in the left facet; fitted values for listeners are plotted in the right facet. The point ranges in the background correspond to the normalized data.

4. Discussion and conclusions

Translation trainees ('listeners') listened to, and interpreting trainees ('talkers') interpreted four different speeches and rated their difficulty and the speech rate. Speeches were matched for linguistic complexity and presented in four experimental conditions (lip movements+no noise, lip movements+noise, no lip movements+no noise, no lip movements+noise). Self-reports indicated that the experimental conditions were indeed clearly distinguishable, and the speeches comparable, regarding speech rate and speech complexity. This suggests that speech-related variables did not influence the results. Mental effort was measured with pupillometry and interpreting accuracy, cognitive load was measured with self-reports.

Regarding the research questions, results can be summarized as follows: first, the results do not lend support to the facilitation hypothesis for lip movements. No main effect of visual presentation was found in self-reports (neither in interpreting nor in listening) or interpreting accuracy. These results confirm those by Jesse et al (2000), who did not find any effect of lip movements on interpreting performance. There was, however, an effect of visual presentation on pupil sizes. Pupils were larger with lip movements than without, for talkers and listeners alike. This result will be specifically addressed in the next section. Second, in line with Gerver (1974), noise significantly decreased interpreting accuracy. Interpretations were less consistent with the source speech when noise was overlaid. Similarly, both talkers and listeners perceived the speech as being more difficult in the noise condition. This suggests that noise overlaid on the source speech increases cognitive load in listening and interpreting and induces higher mental effort in interpreting. Third, no interaction of auditory and visual presentation was observed in self-reports, interpreting accuracy or pupillometry. The hypothesis that the facilitation effect of lip movements is stronger when the signal-to-noise ratio is low, was thus not confirmed. The fourth research question was answered positively: Talkers perceived speeches as being more difficult than listeners. In line with Hyönä, Tommola & Alaja (1995), pupil sizes were larger in talkers than in listeners. Taken together, these findings indicate higher cognitive load and mental effort in interpreting than in listening.

Still, the results of the three cognitive load and mental effort indicators barely overlap: Auditory presentation affected self-reports and interpreting accuracy, but not pupillometry. Visual presentation affected pupillometry, but not self-reports or interpreting accuracy. This mismatch between pupillary data, on the one hand, and performance and self-reports, on the other, raises two questions: (1) what is the effect of lip movements on interpreting? and (2) what do pupil sizes in this study tell us exactly?

4.1 *What is the effect of lip movements on interpreting and what do pupil sizes tell us?*

Even though pupil sizes were larger in the condition with lip movements, the results of this study do not seem to support the idea that lip movements induce higher mental effort or cognitive load as neither performance nor self-reports were affected (positively or negatively) by lip movements. Moreover, noise did not have an effect on pupil sizes even though noise significantly impacted interpreting accuracy and self-reports. Both observations suggest that, in this study, pupil sizes did not reflect mental effort, but rather a general state of arousal (Kahneman 1973). According to this interpretation, the effect on pupil sizes was most probably not caused by lip movements but more generally by the video of the speaker's face compared to the frozen image of his face. Indeed, moving pictures seem to elicit a higher arousal than still pictures (Courtney 2010).

Several further observations support this idea: In line with research by Hyönä, Tommola & Alaja (1995), pupil sizes were larger at the beginning of the speech and flattened towards the end, possibly indicating variations in a state of "alertness" (Hyönä et al 1995, p. 602) that may also be interpreted as arousal. The difference between interpreting and listening in pupil dilation can equally be understood in terms of arousal: interpreting requires constant responses whereas listening does not. The simple fact of responding to a stimulus might increase arousal. Of course, these speculations do not rule out that pupil sizes can indicate mental effort but, in this study, arousal seems to be a better explanation for the results.

4.2 *Limitations*

The first limitation concerns the object of study. Some may object that lip movements are irrelevant for the profession because the booths are often far away from the ST speaker, so that the interpreter can barely see the ST speaker's lip movements. Even though the facilitating effect of lip movements seems to be surprisingly stable up to a distance of 10 meters (Jordan & Sergeant 2000), this distance only covers small conference rooms. However, the purpose of this study is not to claim that lip movements are indispensable for interpreters, but rather to offer an approach to a systematic investigation of one type of visual input in interpreting studies that may or may not spark further studies.

The second limitation concerns the method. Pupillometry is typically used in relatively simple cognitive tasks, like pitch discrimination (Beatty 1982) and digit span (Granholm et al 1996; Kahneman 1973; Wong & Epps 2016) and word recognition (Kramer et al 1997; Kuchinsky et al 2013; Zekveld et al 2014) and other simple linguistic tasks (Engelhardt et al 2010). To the best of my knowledge, only two studies using pupillometry in simultaneous interpreting have been reported (Hyönä, Tommola & Alaja 1995 and Seeber & Kerzel 2012). Further studies using pupillometry in interpreting would provide more insights into whether pupillary responses are reliable in these conditions and whether arousal is a suitable construct to interpret them.

The study was conducted with interpreting trainees and not with professional conference interpreters. The patterns in information processing might change with experience, and professional interpreters might react very differently to audiovisual speech input or background noise. Several studies established an advantage in terms of accuracy for professional interpreters compared to interpreting trainees, which may suggest that skills in speech analysis increase with experience (Liu et al 2004; Díaz-Galaz et al 2015). Similar effects may be expected for integrating visual input.

4.3 *Conclusion*

Conference interpreters regard visual input and, in particular, the speaker as indispensable for rendering the source speech. The ability to see the speaker has been suggested to help to understand and anticipate what the speaker is saying (Bühler 1985). In particular, lip movements may contribute to disambiguate the auditory signal (Seeber 2017). A study to investigate the impact of lip movements on cognitive load and mental effort in simultaneous interpreting was conducted with 14 interpreting trainees, plus 17 translation trainees as control group. The results did not reveal any particularly positive or negative effect of lip movements on cognitive load or mental effort, even when noise was added to the source speech, but seem instead to suggest that lip movements increase arousal in interpreting.

Acknowledgments

The findings reported here were part of my PhD-project. I gratefully thank Professors Silvia Hansen-Schirra, Jukka Hyönä, Dörte Andres and Michaela Albl-Mikasa for their valuable advice and support in conducting and documenting the experiment.

References

Anderson, Linda. 1994. "Simultaneous Interpretation: Contextual and Translation Aspects." In *Benjamins Translation Library*: *Vol. 3*. *Bridging the Gap: Empirical Research in Simultaneous Interpretation.* Edited by S. Lambert and B. Moser-Mercer, 101–248. Amsterdam: John Benjamins.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1). https://doi.org/10.18637/jss.v067.i01

Beatty, Jack. 1982. "Phasic Not Tonic Pupillary Responses Vary With Auditory Vigilance Performance." *Psychophysiology*, *19*(2): 167–172. https://doi.org/10.1111/j.1469-8986.1982.tb02540.x

Bernstein, Lynne. E., Edward T. Auer, Sumik Takayanagi. 2004. "Auditory Speech Detection in Noise Enhanced by Lipreading." *Speech Communication*, *44*(1–4): 5–18. https://doi.org/10.1016/j.specom.2004.10.011

Brancazio, Lawrence, Catherine T Best and Carol A. Fowler. 2006. "Visual Influences on Perception of Speech and Nonspeech Vocal-Tract Events." *Language and Speech*, *49*(1): 21–53. https://doi.org/10.1177/00238309060490010301

Brown, Robert and Howard. E. Page. 1939. "Pupil Dilatation and Dark Adaptation." *Journal of Experimental Psychology*, *25*(4): 347–360. https://doi.org/10.1037/h0060296

Bühler, Hildegund. 1985. "Conference Interpreting: A Multichannel Communication Phenomenon." *Meta: Journal des traducteurs*, *30*(1) : 49-54. https://doi.org/10.7202/002176ar

Chapman, C. Richard, Shunichi Oka, David H. Bradshaw, Robert C. Jacobson and Gary W. Donaldson. 1999. « Phasic Pupil Dilation Response to Noxious Stimulation in Normal Volunteers: Relationship to Brain Evoked Potentials and Pain Report." *Psychophysiology*, *36*(1): 44–52. https://doi.org/10.1017/S0048577299970373

Christensen, Rune H. B. 2015. Ordinal - Regression Models for Ordinal Data. Version 2015.1-21. http://www.cran.r-project.org/package=ordinal (accessed 01 December 2015)

Courtney, Christopher G., Michael E. Dawson, Anne M. Schell, Arvind Iyer, Thomas D. Parsons. 2010. "Better than the Real Thing: Eliciting fear with Moving and Static Computer-generated Stimuli." *International Journal of Psychophysiology*, *78*(2): 107–114. https://doi.org/10.1016/j.ijpsycho.2010.06.028

Davies, M. 2008. Word frequency data. Retrieved from The Corpus of Contemporary American English (COCA): https://www.english-corpora.org/coca/ (accessed 19 March 2021).

Díaz-Galaz, Stephanie, Presentacion Padilla, and M. Teresa Bajo. 2015. "The Role of Advance Preparation in Simultaneous Interpreting: A Comparison of Professional Interpreters and Interpreting Students". *Interpreting 17* (1): 1–25. https://doi.org/10.1075/intp.17.1.01dia.

Dillinger, Mike. 1990. "Comprehension during Interpreting: What Do Interpreters Know that Bilinguals Don't?" *The Interpreters' Newsletter*, *3*: 41–58. Accessed March 03 17, 2021. http://hdl.handle.net/10077/2154

Ehrensberger-Dow, Maureen, Michaela Albl-Mikasa, Katrin Andermatt, Andrea Hunziker Heeb and Caroline Lehr. 2020. "Cognitive Load in Processing ELF: Translators, Interpreters, and Other Multilinguals". *Journal of English as a Lingua Franca 9*(2):217-238. https://doi.org/10.1515/jelf-2020-2039

Engelhardt, Paul E., Fernanda Ferreira and Elana G. Patsenko. 2010. „Pupillometry Reveals Processing Load during Spoken Language Comprehension." *Quarterly Journal of Experimental Psychology*, *63*(4): 639–645. https://doi.org/10.1080/17470210903469864

European Commission. 2009a. "United Airlines rewards fittest people." Retrieved from *Speech Repository*: https://webgate.ec.europa.eu/sr/speech/united-airlines-rewards-fittestpeople (accessed 4 February 2021)

European Commission. 2009b. "Disenchantment at work." Retrieved from *Speech Repository*: https://webgate.ec.europa.eu/sr/speech/disenchantment-work (accessed 4 February 2021)

European Commission. 2012a. "Demographic shift in Europe." Retrieved from *Speech Repository*: https://webgate.ec.europa.eu/sr/speech/demographic-shift-europe (accessed 4 February 2021).

European Commission. 2012b. "Greece in the doldrums." Retrieved from Speech Repository: https://webgate.ec.europa.eu/sr/speech/greece-doldrums (accessed 4 February 2021).

Fisk, Angus S., Shu K. E. Tam, L. A. Brown, Vladyslav V. Vyazovskiy, David M. Bannerman and Stuart N. Peirson. 2018. "Light and Cognition: Roles for Circadian Rhythms, Sleep, and Arousal." *Frontiers in Neurology*, *9:* 56. https://doi.org/10.3389/fneur.2018.00056

Gerver, David. 2002. "The Effects of Source Language Presentation Rate on the Performance of Simultaneous Conference Interpreters." *The Interpreting Studies*. *Reader.* Edited by F. Pöchhacker and M. Shesinger, 52–66. London: Routledge.

Gerver, David. 1974. "The Effects of Noise on the Performance of Simultaneous Interpreters: Accuracy of Performance." *Acta Psychologica*, *38*(3): 159–167. https://doi.org/10.1016/0001-6918(74)90031-6

Gieshoff, Anne Catherine. Forthcoming. "The impact of visible lip movements on silent pauses in simultaneous interpreting." *Interpreting*.

Gieshoff, Anne Catherine. 2018. "The Impact of Visual Input on Cognitive Load in Simultaneous Interpreting". Mainz: Johannes Gutenberg Universität.

Gile, Daniel. 2009. *Basic Concepts and Models for Interpreter and Translator Training: Revised edition* (2nd ed.). https://doi.org/10.1075/btl.8

Granholm, Eric, Robert F. Asarnow, Andrew J. Sarkin and Karen L. 1996. “Pupillary Responses Index Cognitive Resource Limitations.” *Psychophysiology*, *33*: 457–461. https://doi.org/10.1111/j.1469-8986.1996.tb01071.x

Halverson, Sandra. 2017. “Multimethod approaches.” *The Handbook of Translation and Cognition.* Edited by J. Schwieter and A. Ferreira. 195–212. Hoboken: Wiley Blackwell. https://doi.org/10.1002/9781119241485.ch11

Hild, Adelina. 2015. “Discourse Comprehension in Simultaneous Interpreting: The Role of Expertise and Redundancy.” *Psycholinguistic and Cognitive Inquiries into Translation and Interpreting.* Edited by A. Ferreira and J. W. Schwieter. 67–100. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.115.04hil

Hyönä, Jukka, Jorma Tommola, Anna-Maria Alaja. 1995. “Pupil Dilation as a Measure of Processing Load in Simultaneous Interpretation and Other Language Tasks.” *The Quarterly Journal of Experimental Psychology Section A*, *48*(3): 598–612. https://doi.org/10.1080/14640749508401407

Ivars, Amparo J. Daniel P. Calatayud. 2001. “‘I Failed because I Got Nervous’. Anxiety and Performance in Interpreter Trainees: An Empirical Study.” *The Interpreter’s Newsletter*, *11*: 105–120. Accessed March 17, 2021. http://hdl.handle.net/10077/2452.

Jesse, Alexandra, Nick Vrignaud, Michael M. Cohen and Dominic W. Massaro. 2000. “The Processing of Information from Multiple Sources in Simultaneous Interpreting.” *Interpreting*, *5*(2): 95–115. https://doi.org/10.1075/intp.5.2.04jes

Jordan, Timothy R. and Paul Sergeant. 2000. “Effects of Distance on Visual and Audiovisual Speech Recognition.” *Language and Speech*, *43*(1): 107–124. https://doi.org/10.1177/00238309000430010401

Kahneman, Daniel. 1973. *Attention and Effort*. Englewood Cliffs, N.J: Prentice-Hall.

Korpal, Paweł. 2016. “Interpreting as a Stressful Activity: Physiological Measures of Stress in Simultaneous Interpreting.” *Poznan Studies in Contemporary Linguistics*, *52*(2):297-316. https://doi.org/10.1515/psicl-2016-0011

Kramer, Sophia E., Theo S. Kapteyn, Josst M. Festen, and Dirk J. Kuik. 1997. “Assessing Aspects of Auditory Handicap by Means of Pupil Dilation.” *Audiology*, *36*: 155–164. https://doi.org/10.3109/00206099709071969

Krejtz, Krzysztof, Andrew T. Duchowski, Anna Niedzielska, Cezary Biele, and Izabela Krejtz. 2018. “Eye Tracking Cognitive Load Using Pupil Diameter and Microsaccades with Fixed Gaze”. *PLOS ONE 13* (9): e0203629. https://doi.org/10.1371/journal.pone.0203629.

Kuchinsky, Stefanie E., Jayne B. Ahlstrom, Kenneth I. Vaden, Stephanie L Cute, Larry E. Humes, Judy R. Dubno and Mark A. Eckert. 2013. “Pupil Size Varies with Word Listening and Response Selection Difficulty in Older Adults with Hearing Loss: Pupil Size in Older Adults.” *Psychophysiology*, *50*(1): 23–34. https://doi.org/10.1111/j.1469-8986.2012.01477.x

Lecumberri, Maria L. G., Martin Cooke and Anne Cutler. 2010. “Non-native Speech Perception in Adverse Conditions: A Review.” *Speech Communication*, *52*(1): 864–886. https://doi.org/doi:10.1016/j.specom.2010.08.014

Lee, Jieun. 2008. “Rating Scales for Interpreting Performance Assessment.” *The Interpreter and Translator Trainer*, *2*(2): 165–184. https://doi.org/10.1080/1750399X.2008.10798772

Liu, Minhua, Diane L. Schallert, and Patrick J. Carroll. 2004. ‘Working Memory and Expertise in Simultaneous Interpreting’. Interpreting 6 (1): 19–42. https://doi.org/10.1075/intp.6.1.04liu.

Lowenstein, Otto, Richard Feinberg and Irene Loewenfeld (1963). “Pupillary Movements During Acute and Chronic Fatigue A New Test for the Objective Evaluation of Tiredness.” *Investigative Ophthalmology, 2(2)*: 138–158.

Massaro, Dominic W. And Michael M. Cohen. 1999. “Speech Perception in Perceivers with Hearing Loss: Synergy of Multiple Modalities.” *Journal of Speech, Language, and Hearing Research*, *42*(1), 21–41. https://doi.org/10.1044/jslhr.4201.21

Mirman, Daniel. 2014. *Growth Curve Analysis and Visualization Using R*. Boca Raton: CRC Press.

Moser-Mercer, Barbara, Alexandra Künzli and Marina Korac. 1998. “Prolonged Turns in Interpreting: Effects on Quality, Physiological and Psychological Stress (Pilot study).” *Interpreting*, *3*(1): 47–64. https://doi.org/10.1075/intp.3.1.03mos

Oliva, Manuel and Andrey Anikin. 2018. “Pupil Dilation Reflects the Time Course of Emotion Recognition in Human Vocalizations.” *Scientific Reports*, *8*, 4871. https://doi.org/10.1038/s41598-018-23265-x

Peelle, Jonathan E. and Mitchell S. Sommers. 2015. “Prediction and Constraint in Audiovisual Speech Perception.” *Cortex*, *68*: 169–181. https://doi.org/10.1016/j.cortex.2015.03.006
Peirce, Jonathan, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Eric Kastman and Jonas Kristoffer Lindeløv. 2019. “PsychoPy2: Experiments in Behavior made Easy.” *Behavior Research Methods*, *51*(1): 195–203. https://doi.org/10.3758/s13428-018-01193-y
Peirce, Jonathan W. 2007. “PsychoPy—Psychophysics Software in Python.“ *Journal of Neuroscience Methods*, *162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017
R Core Team. 2016. “R: A language and environment for statistical computing.” Version 3.2.4. Retrieved from *R Foundation for Statistical Computing*: https://www.R-project.org (accessed 1 March 2016)
Rennert, Sylvie. 2008. “Visual Input in Simultaneous Interpreting.” *Meta: journal des traducteurs*, *53*(1): 204–217. https://doi.org/10.7202/017983ar
Rosendo, Lucía R. and María C. Galván. 2019. “Coping with Speed: An Experimental Study on Expert and Novice Interpreter Performance in the Simultaneous Interpreting of Scientific Discourse.” *Babel*, *65*(1) : 1–25. https://doi.org/10.1075/babel.00081.rui
Roziner, Ilan & Miriam Shlesinger. 2010. “Much Ado about Something Remote: Stress and Performance in Remote Interpreting.” *Interpreting*, *12*(2): 214–247. https://doi.org/10.1075/intp.12.2.05roz
Seeber, Kilian G. 2017. “Multimodal Processing in Simultaneous Interpreting.” *The Handbook of Translation and Cognition.* Edited by J. Schwieter and A. Ferreira. 461–475. Hoboken: Wiley Blackwell. https://doi.org/10.1002/9781119241485.ch25
Seeber, Kilian G. 2015. “Cognitive Load in Simultaneous Interpreting: Measures and Methods.” *Interdisciplinarity in translation and interpreting process research.* Edited by M. Ehrensberger-Dow, S. Göpferich & S. O’Brien (Eds.). 18–33. https://doi.org/10.1075/bct.72.03see
Seeber, Kilian G. and Dirk Kerzel. 2012. “Cognitive Load in Simultaneous Interpreting: Model Meets Data.” *International Journal of Bilingualism*, *16*(2): 228–242. https://doi.org/10.1177/1367006911402982
Seeber, Kilian G. 2011. “Cognitive Load in Simultaneous Interpreting.” *Interpreting*, *13*(2): 176–204.
Setton, Robin (1999). *Simultaneous Interpretation: A Cognitive-Pragmatic Analysis*. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.28
Seubert, Sabine. 2019. *Visuelle Informationen beim Simultandolmetschen: Eine Eyetracking-Studie*. Berlin: Frank & Timme GmbH.
Stachowiak-Szymczak, Katarzyna. 2019. *Eye Movements and Gestures in Simultaneous and Consecutive Interpreting*. https://doi.org/10.1007/978-3-030-19443-7
Sumby, William H. and Irwin Pollack. 1954. “Visual Contribution to Speech Intelligibility in Noise.” *Journal of the Acoustical Society of America*, (26): 212–215. https://doi.org/10.1121/1.1907309
Tabri, Dollen, Kim M. S. A. Chacra & Tim Pring. 2010. “Speech Perception in Noise by Monolingual, Bilingual and Trilingual Listeners.” *International Journal of Language & Communication Disorders*, 46(4):411-422. https://doi.org/10.3109/13682822.2010.519372
Thomas, Sharon M. Timothy R. Jordan. 2004. “Contributions of Oral and Extraoral Facial Movement to Visual and Audiovisual Speech Perception.” *Journal of Experimental Psychology: Human Perception and Performance*, *30*(5): 873–888. https://doi.org/10.1037/0096-1523.30.5.873
Tobii support. 2016, July 20. *Personal communication*.
Tobii Technology. 2010. *Tobii TX300 Eye Tracker. User Manual*. Accessed March 17, 2021. https://www.tobiipro.com/siteassets/tobii-pro/user-manuals/tobii-pro-tx300-eye-tracker-user-manual.pdf?v=2.0
Tommola, Jorma and Johan Lindholm. 1995. “Experimental Research on Interpreting: Which Dependent Variable?” *Topics in interpreting research*. Edited by J. Tommola. 121–133. Turku: University of Turku.
Tye-Murray, Nancy, Mitchell Sommers and Brent Spehar. 2007. “Auditory and Visual Lexical Neighborhoods in Audiovisual Speech Perception.” *Trends in Amplification*, *11*(4): 233–241. https://doi.org/10.1177/1084713807307409
van der Wel, Pauline and Henk van Steenbergen. 2018. “Pupil Dilation as an Index of Effort in Cognitive Control Tasks: A Review.” *Psychonomic Bulletin & Review*, *25*(6): 2005–2015. https://doi.org/10.3758/s13423-018-1432-y

Vatikiotis-Bateson, Eric, Inge-Marie Eigsti, Sumio Yano and Kevin G. Munhall. 1998. “Eye Movement of Perceivers during Audiovisual Speech Perception.” *Perception & Psychophysics*, *60*(6): 926–940. https://doi.org/10.3758/BF03211929

von Kriegstein, Katharina, Özgür Dogan, Martina Grüter, Anne-Lise Giraud, Christian A. Kell, Thomas Grüter, Andreas Kleinschmidt and Stefan A. Kiebel. 2008. “Simulation of Talking Faces in the Human Brain Improves Auditory Speech Recognition.” *Proceedings of the National Academy of Sciences*, *105*(18): 6747–6752. https://doi.org/10.1073/pnas.0710826105

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Wong, Hoe K. and Julien Epps. 2016. “Pupillary Transient Responses to Within-Task Cognitive Load Variation.” *Computer Methods and Programs in Biomedicine*, *137*: 47–63. https://doi.org/10.1016/j.cmpb.2016.08.017

Wu, Zhiwei. 2019. “Text Characteristics, Perceived Difficulty and Task Performance in Sight Translation: An Exploratory Study of University-level Students.” *Interpreting*, *21*(2): 196–219. https://doi.org/10.1075/intp.00027.wu

Yoon, Jinny J. and Helen V. Danesh-Meyer. 2019. “Caffeine and the Eye.” *Survey of Ophthalmology*, *64*(3): 334–344. https://doi.org/10.1016/j.survophthal.2018.10.005

Zekveld, Adriana A., Dirk J. Heslenfeld, Ingrid S. Johnsrude, Niek J. Versfeld and Sophia E. Kramer 2014. “The Eye as a Window to the Listening Brain: Neural Correlates of Pupil Size as a Measure of Cognitive Listening Load.” *NeuroImage*, *101*: 76–86. https://doi.org/10.1016/j.neuroimage.2014.06.069

Zwischenberger, Cornelia. 2010. “Quality Criteria in Simultaneous Interpreting: An International vs. A National View.” *The Interpreters’ Newsletter*, *15*: 127–142. Accessed 17 March 2021. http://hdl.handle.net/10077/4754

Appendix

	X^2	DF	p-value
linear term	56.76	9.1	>0.01
quadratic term	65.34	10.1	>0.01
task	48.50	11.1	>0.01
task*linear term	5.477	12.1	>0.05
task*quadratic term	1.213	13.1	0.271
visual presentation	3.913	13.1	>0.05
visual presentation*linear term	3.344	14.1	0.067
visual presentation*quadratic term	0.011	15.1	0.915
auditory presentation	0.007	14.1	0.93
auditory presentation*linear term	0.000	15.1	0.98
auditory presentation*quadratic term	0.005	16.1	0.95
text difficulty ratings	0.414	15.2	0.813
text difficulty ratings*linear term	0.215	17.2	0.899
text difficulty ratings*quadratic term	0.842	19.2	0.656
speech rate ratings	1.406	15.2	0.495
speech rate ratings*linear term	0.806	17.2	0.669
speech rate ratings*quadratic term	0.000	19.2	1.000
signal-to-noise ratio	0.010	14.1	0.919
signal-to-noise ratio*linear term	0.112	15.1	0.112
signal-to-noise ratio*quadratic term	0.356	16.1	0.356
speech	1.704	16.3	0.636
speech *linear term	0.283	17.3	0.868
speech*quadratic term	5.362	19.3	0.999
trial	0.198	17.3	0.978
trial*linear tern	7.143	19.3	0.067
trial*quadratic term	6.645	22.3	0.084

Table 2. Chi-square value, degrees of freedom and p-value obtained by model comparison for the response variable pupil dilation.