The impact of audio-visual speech input on work-load in simultaneous interpreting

Inauguraldissertation zur Erlangung des Akademischen Grades eines Dr. phil.,

vorgelegt dem Fachbereich Translations-, Sprach- und Kulturwissenschaft (06) der Johannes Gutenberg-Universität Mainz

von

Anne Catherine Gieshoff

aus Siegburg

Universität Mainz

2018

Tag des Prüfungskolloquiums: 13. Juni 2018

Thanks to

the supervisors for their valuable advice,

The team of the Division of English linguistics and translation studies for their precious help with the experiments,

to the students participating at the experiment,

to my "speaker" who lent his voice to my experiments,

to my partner and friends for their critical thinking and support during all stages of this endeavor.

Table of Contents

Abstract 1

0 Introduction 5

1 Simultaneous interpreting – a multimodal task 6

1.1 Interpreting and translating 7

1.2 The importance of visual information in simultaneous interpreting 10

1.3 Simultaneous interpreting: a complex task 13

2 Cognitive processing of stimuli 17

2.1 Vision 17

2.1.1 The physiology of the eye 18

2.1.2 Visual processing beyond the retina 19

2.1.3 Pupillary reflexes 21

2.2 Audition 23

2.2.1 The physiology of the ear 24

2.2.2 Processing of auditory stimuli 25

2.2.3 Interactions of multiple auditory streams 27

2.3 Audio-visual stimuli 29

2.3.1 Multisensory binding: facilitated response for multimodal stimuli 29

2.3.2 The superior colliculus: detecting temporal and spatial congruence 33

2.3.3 The superior temporal sulcus: ensuring semantic congruence 36

2.3.4 Feedback-mechanisms and further integration sites 37

2.3.5 Processing of audio-visual speech 38
2.4 Multimodal input in simultaneous interpreting 42
2.4.1 Types of visual input in simultaneous interpreting 43
2.4.2 Studies on visual information in simultaneous interpreting ... 47
3 Cognitive load and mental effort 50
3.1 Models of working memory 51
3.2 Models of cognitive load 55
3.2.1 Kahnemann's capacity model of attention 55
3.2.2 Sweller's cognitive load theory 58
3.2.3 Wickens' model of task interference 62
3.2.4 Lavie's load theory of selective attention and cognitive control 64
3.2.5 Barrouillet's time-based resource sharing model 66
3.2.6 Cognitive load versus mental effort 68
3.3 Mental effort and cognitive load in simultaneous interpreting 70
3.3.1 Gile's effort model 70
3.3.2 Seeber's model of cognitive load 71
3.3.3 Visual information as a load factor in simultaneous interpreting 74
3.3.4 Noise as load factor in simultaneous interpreting 76
3.3.5 Effects of visual input and noise: predictions from psychological and interpreting research 79
3.4 Measuring work-load 81
3.4.1 Subjective ratings 82
3.4.2 Performance 85
3.4.3 Physiological measures 90

4 Method and results 98
4.1 Aims and general approach 98
4.2 Pilot study 102
4.2.1 Experimental material used in the pilot study 102
4.2.2 Participants of the pilot study 104
4.2.3 Apparatus used in the pilot study 105
4.2.4 Procedure of the pilot study 106
4.2.5 Data preparation and results of the pilot study 107
4.2.6 Discussion of the pilot study's results 111
4.3 Main study 117
4.3.1 Experimental material used in the main study 117
4.3.2 Participants of the main study 120
4.3.3 Apparatus used in the main study 121
4.3.4 Procedure of the main study 121
4.3.5 Data preparation and results of the pretest 127
4.3.6 Data preparation and results of the main part 152
4.4 General discussion and limitations 227
4.4.1 Effects of audio-visual speech 228
4.4.2 Effects of noise 234
4.4.3 Effects of task 238
4.4.4 The impact of audio-visual speech input on work-load in simultaneous interpreting 241
4.4.5 Limitations 245
5 Conclusion 247
5.1 Methodological implications 248
5.2 Practical implications 251

5.3 Future research 253
6 References 257
7 Appendix 286
7.1 Pretest stimuli 286
7.2 Speeches 290
7.2.1 Speech “Greece” 290
7.2.2 Speech “demographic change” 291
7.2.3 Speech “air travel” 293
7.2.4 Speech “work” 295
7.3 Text-related questions 297
7.3.1 Speech “Greece” 297
7.3.2 Speech “demographic change” 299
7.3.3 Speech “air travel” 301
7.3.4 Speech “work” 303
7.4 Tables 305
7.4.1 Results of the tukey comparison of response accuracy between participants during the pretest 305
7.4.2 Results of the analyses of variance for the noise pretest. .. 316
7.4.3 Model estimates for the speech duration estimations 318
7.4.4 Cognate translations: Cognate pairs with high Levenshtein ratio 319

Tables

Table 1: Classification scheme for visual input in simultaneous interpreting 45

Table 2: Experimental conditions. 104

Table 3: Participant's ratings 108

Table 4: Critical cases of cognate pairs. 120

Table 5: Speeches in each experimental condition 124

Table 6: Model estimates of the linear mixed model on response times. 133

Table 7: Video quality ratings: Counts for the different levels 154

Table 8: Video quality ratings: results of model comparison with likelihood ratio test 155

Table 9: Sound quality: counts of the different ratings 156

Table 10: Sound quality ratings: model estimates, standard error, z-values and p-values..... 158

Table 11: Results of model comparison using likelihood ratio tests 158

Table 12: Counts of each text difficulty rating 160

Table 13: Results of post-hoc comparison of text difficulty ratings between speeches..... 160

Table 14: Text difficulty ratings: likelihood ratio, degrees of freedom and p-values..... 162

Table 15: Speech rate ratings: counts..... 163

Table 16: Speech rate ratings: likelihood ratio, degrees of freedom and p-values..... 164

Table 17: Speech duration judgments: Degrees of freedom, likelihood ratio and p-value 170

Table 18: Speech duration judgments: Results of the Tukey HSD post-hoc test 172

Table 19: Text-related questions: degrees of freedom, likelihood ratio and p-value 177

Table 20: Example of segment categorization 181

Table 21: Translation accuracy: Number of segments 182

Table 22: Translation accuracy (missing translations as missing values): Degrees of freedom, likelihood ratio and p-value 184

Table 23: Translation accuracy (missing translations as wrong translations): Degrees of freedom, likelihood ratio and p-value 187

Table 24: Number of cognates in each text 191

Table 25: Number of cognates in each category 192

Table 26: Cognate translations: Degrees of freedom, likelihood ratio and p-values 195

Table 27: Number and percentage of silent pauses in each category 201

Table 28: Number of silent pauses of different length in each experimental condition. 202

Table 29: Silence duration: degrees of freedom, likelihood ratio and p-values 203

Table 30: Silence duration: model estimates 204

Table 31: Fundamental voice frequency: estimates, standard error, t-values and p-values 211

Table 32: Pupil dilation: chi-squared, degrees of freedom and p-value .. 221

Table 33: Results of the analyses of variance for the noise pretest. 318

Table 34: Model estimates (intercept for all participants in the noise and no noise condition), standard error and t-value for estimation of speech duration. 319

Table 35: Cognate pairs with a ratio Levenshtein distance over 0.167 .. 321

Figures

Figure 1: A capacity model of attention 56

Figure 2: Seeber's cognitive load model 73

Figure 3: Fixed effects of both experimental conditions on pupil sizes ... 110

Figure 4: Fixed effects on the fundamental voice frequency 111

Figure 5: Levenshtein distance between the orthographic form of the English stimulus and its German translation 119

Figure 6: Procedure for the noise pretest. 123

Figure 7: Procedure of the main experiment (first part). 125

Figure 8: Procedure of the main experiment (second part). 127

Figure 9: Boxplot of observed response times of each participant. 129

Figure 10: Boxplot of observed response times for each stimulus. 130

Figure 11: Boxplot of mean response times depending on cognate status. 131

Figure 12: Boxplot of observed mean response times for each noise level. 131

Figure 13: Boxplot of response duration for correct (green) and wrong (yellow) responses. 132

Figure 14: Effect plots for response time as a function of response accuracy and noise level. 134

Figure 15: Percentage of correct (yellow) and wrong (green) responses to each stimulus. 135

Figure 16: Response accuracy depending on noise level. 136

Figure 17: Response accuracy of cognates and non-cognates. 137

Figure 18: Effects plot for response accuracy. 139

Figure 19: Pupil dilation depending on noise level and cognate status. . 143

Figure 20: Pupil dilation depending on noise level and response accuracy. The 144

Figure 21: Fitted values against the observed standardized pupil dilation. 145

Figure 22: P-values for the predictor variables cognate status, response accuracy and noise level. 147

Figure 23: Time course of the effect of noise level in the first three seconds after stimulus onset. 148

Figure 24: Time course of the effect of correct response in the first three seconds after stimulus onset. 149

Figure 25: Time course of the effect of cognate status in the first three seconds after stimulus onset. 150

Figure 26: Fitted probabilities for video quality ratings 156

Figure 27: Fitted probabilities for sound quality ratings 159

Figure 28: Fitted probabilities for text difficulty ratings 162

Figure 29: Proportion of speech rate ratings 165

Figure 30: Speech duration estimations of each participant 169

Figure 31: Model fit for speech duration judgements. 171

Figure 32: Text-related questions: Proportions of correct (yellow) and wrong (green) answers 176

Figure 33: Fitted probabilities for correct answers to text-related questions. 178

Figure 34: Model fit for translation accuracy (with missing translations treated as missing value). 185

Figure 35: Model fit for translation accuracy (missing translations as wrong translations) 188

Figure 36: Proportions of cognate and non-cognate translations 194

Figure 37: Cognate translations: Observed (black dots) versus model fit (colored lines)..... 197

Figure 38: Boxplot of silence durations for each participant. 199

Figure 39: Silence duration in each condition..... 200

Figure 40: Silence duration: barplots of fitted values with error bars. 205

Figure 41: Fundamental voice frequency and trend line for the noise/no noise condition. 209

Figure 42: Fundamental voice frequency: Observed (black dots) versus fitted values (yellow and green line). 211

Figure 43: Fundamental voice frequency and trend line for each speech. 214

Figure 44: Fundamental voice frequency and trend line for each speech. 215

Figure 45: Pupil sizes during speeches..... 220

Figure 46: Model fit for a linear mixed model 222

Abbreviations

BOLD	Blood oxygenation level dependent
dB	Decibel
EDA	Electrodermal activity
EEG	Electroencephalogramm
ERP	Event-related potential
F0	Fundamental voice frequency
fMRI	Functional magnetic resonance imaging
hz	Hertz
IQ	Intelligence quotient
L1	First language (mother tongue)
L2	Second language (first foreign language)
LGN	Lateral geniculate nuclei
M	Mean
MD	Median
MEG	Magnetoencephalography
ms	Millisecond(s)
NASA-TXL	NASA task-load Index
PET	Positron-emission tomography
pSTS	Posterior superior temporal sulcus
SD	Standard deviation
SE	Standard error
STAI	State-trait anxiety inventory
SWAT	Subjective work-load assessment technique
TE	Anterior inferior temporal cortex

TEPR	Task-evoked pupillary responses
V1-V6	Visual corteci 1-6

Abstract

Conference interpreters face various sources of visual input during their work: the speaker's facial movements or gestures, reactions from the audience, presentations, charts, notes, etc. Far from perceiving this additional visual information as increasing their work load, interpreters insist on having access to visual information and in particular to the speaker. The question that arises is the following: What is the role of visual input in simultaneous interpreting? Beyond providing clues to the meaning of what the speaker is saying, does visual input reduce work-load in simultaneous interpreting? Previous research on visual input in simultaneous interpreting had failed to observe any facilitating effects of visual input on simultaneous interpreting possibly due to methodological issues. The aim of this study was to investigate the impact of visual input and in particular visible lip movements on simultaneous interpreting using a more systematic and controlled approach. Based on previous research, I hypothesized that audio-visual speech should have a facilitating effect on speech comprehension and therefore should lower work-load in simultaneous interpreting. As I could not be sure whether audio-visual speech really has an impact on work-load in simultaneous interpreting or not, I introduced a second variable known to affect work-load in simultaneous interpreting, namely white noise that is added to the source speech and makes it partially unintelligible. The experiment contrasted thus simultaneous interpreting from English to German in four conditions: audio-visual speech without background noise, audio-visual speech with background noise, auditory-only speech without background noise and auditory-only speech with background noise.

In total, three experiments were conducted: a pilot study, a pretest and a main study. The aim of the pilot study was to test the experimental design with a limited number of participants (*N*=6) and to discover potential flaws in the experimental design. The main study (*N*=31, 17 listeners and 14

Abstract

interpreters) was preceded by the pretest, a word recognition task, in order to determine the participants' individual signal-to-noise ratio where participants correctly identified 75% of the stimuli. Pupillary responses, response accuracy and response time during the pretest were statistically investigated. The signal-to-noise ratio that resulted from the pretest was then applied to the speeches during the main experiment. The main experiment followed roughly the same procedure as the pilot study: participants orally translated four speeches. After each speech, participants estimated the speech duration, rated the video and sound quality, the text difficulty and the speech rate and finally answered five text-related questions. A control group of listeners matched for their level of English was included in order to control for possible task effects. The effects of audio-visual speech and white noise on pupil sizes, duration estimations, general ratings, text-related questions and (for interpreters only) fundamental voice frequency, silent pauses, translation accuracy and cognate translations were analyzed.

On the whole, the findings did not support the hypothesis that audio-visual speech input lowers work-load in simultaneous interpreting. No effect for audio-visual speech input was found when analyzing duration estimations, general ratings, text-related questions, translation accuracy and cognate translations. The analyses of general ratings, text-related questions, translation accuracy, however, showed an effect for background noise added to the speech, suggesting that those indicators are in principle sensitive to increases in work-load. The only indicator that hinted towards a facilitating effect of audio-visual speech in simultaneous interpreting is silent pauses: long silent pauses were more frequent when no audio-visual speech input was provided. Their number increased even more when white noise deteriorated speech intelligibility. For fundamental voice frequency and pupil dilations even the opposite pattern was found. Fundamental voice frequency rose faster during simultaneous interpreting with audio-visual speech input. Pupil sizes decreased over the time-course of the speech. During simultaneous interpreting with audio-visual input,

Abstract

this decrease was less pronounced than in the auditory-only condition. Surprisingly, background noise did not affect pupil sizes. Comparing listeners and interpreters, an effect of task was found in participants' ratings, text-related questions and pupil dilations. Listeners rated speeches as being less difficult than interpreters and gave on the whole more correct answers to text-related questions than interpreters. Moreover, listening was associated with a larger decrease in pupil sizes than simultaneous interpreting.

The fact that audio-visual speech input had no or very little effect on performance or perceived difficulty in simultaneous interpreting as indicated by translation accuracy, cognate translations, answers to text-related questions or the participants' ratings of text difficulty may be explained by predictions. During simultaneous interpreting, interpreters draw on the available speech context and their world knowledge in order to anticipate the next speech segments. This is very different to word recognition where the target word cannot be predicted and where lip movements can provide crucial clues. Anticipation during simultaneous interpreting may explain why interpreters do not seem to benefit from the speaker's lip movements. This does, however, not necessarily mean that audio-visual speech induces higher work-load in simultaneous interpreting. Increases in pupil dilation or fundamental voice frequency are indeed associated with higher arousal and commonly interpreted as a work-load indicator, larger pupil dilations or higher voice frequency meaning higher work-load. In the present study, pupil sizes were larger during simultaneous interpreting with audio-visual speech, but not during simultaneous interpreting with background noise, a factor having been found to affect performance and participants' ratings. In the light of this pattern, the results may be interpreted in terms of general arousal without necessarily being linked to work-load. Interpreters may simply be aroused when the speaker's face is moving compared to a static face. One concept that might provide some explanation in this regard is the notion of *sense of*

Abstract

presence: Seeing the speaker's face may help the interpreter to immerse in the situation and thereby increase her task engagement.

0 Introduction

Translating something from one language to another the minute you hear it – for most people, it seems an impossible task. Yet, for conference interpreters it is their daily work. Sitting in a booth, they listen to the speech via headphones and immediately give a rendition of what is said in another language. It is easy to imagine the complexity of such a task. It is also easy to imagine that an interpreter will draw on every source of information that might help her[1] to cope with the task. For instance, lip movements may improve the intelligibility of the speech. Likewise, a pointing gesture may be important to understand what speaker refers to. A presentation slide may help the interpreter to anticipate what the speaker will be saying next. Lip movements, gestures or presentation slides may provide precious clues but at the same time may represent a burden for the interpreter who needs to process this additional information. Nevertheless, interpreters usually insist on having visual contact with the speaker. The question is now: How does all this additional and mainly visual information affect the interpreter during her work? Or more specifically: apart from providing clues to what is said, does it make the task of interpreting easier or more difficult?

The present study seeks to answer this question using different methods like pupil dilation and voice frequency, interpreting performance, cognate translations and silent pauses, and subjective ratings of task difficulty and recall of speech content in order to obtain insight on how visual input impacts the interpreter during her assignment. White noise was used to compare the effects of visual input against the effects of adverse listening conditions in simultaneous interpreting. Due to the wide range of

[1] In order to facilitate reading throughout the text, interpreters are treated as being female in contrast to speakers who are treated as being male although there are also men that exert the profession of conference interpreting and women who speak at a conference.

information sources during interpreting (see above) the study is limited to lip movements and its effects on simultaneous interpreting.

Before exploring the topic any further, the next sections describe the key notion of this study, simultaneous interpreting, and the importance of visual information in interpreting. Chapter 2 explains visual and auditory processing and the phenomenon of *multimodal binding*, e.g. how the brain puts together visual and auditory information. It provides first hints under what conditions visual input in simultaneous interpreting could facilitate simultaneous interpreting. As I used pupil dilation and white noise in my experiment, two sections will briefly describe the mechanisms underlying pupillary reactions and the interactions between multiply auditory streams as it is the case when a speech is overlaid with white noise. Chapter 3 then moves on with the notion of *work-load* and how it is defined in general and in particular in simultaneous interpreting. The last section of chapter 3 describes various methods to measure work-load and in particular those methods that were used for the present study. Chapter 4 describes the three experiments - pilot study, pre-test and main study - I conducted in order to investigate the impact of lip movements on work-load during simultaneous interpreting, and reports the results of the experiments. The results are discussed in the last section of chapter 4. Finally, chapter 5 reflects on the methodological and practical implications of the present study and gives an outlook on research that may give further insights on the impact of visual input during simultaneous interpreting.

1 Simultaneous interpreting – a multimodal task

Rendering a speech in another language the moment it is given: Most persons will certainly agree that simultaneous interpreting is a highly complex task that may seem infeasible and maybe somewhat mysterious to a non-initiated person. The following sections give a description of simultaneous interpreting and explain the importance of visual input during

simultaneous interpreting. The chapter will close with an overview over the cognitive processes that are assumed to take place during simultaneous interpreting.

1.1 Interpreting and translating

Roughly speaking, simultaneous interpreting requires the interpreter to orally translate a speech into another language while it is given. At the center of this preliminary description stands the notion of translation. But what does *translation* actually mean? In a first attempt, translation was defined as substituting text [2] in one language (source language) by equivalent text in a second language (target language) *equivalent* meaning here that both texts relate to the same situations (Malmkjær, 2011a, pp. 57-70; Catford, 1965). With the advent of the *skopos theory* (Reiß & Vermeer, 1991, see also Vermeer, 1992) it became clear, however, that the definition of *equivalence* as an overlap in semantic meaning was insufficient because it did not take into account the purpose of a text. Commercials, song texts or press releases have very different purposes that guide the translator in his choices: depending on the text type, the translator may choose very different words or syntactical structures in order to preserve the persuasive character of a commercial, the rhythm of a song text or the content of a press release (Malmkjær, 2011b, pp. 108-122). The *skopos* of a text covers all factors that might guide the translator in his choices, like the text type, the purpose of the text, the reader to whom it is intended, the author and the client's reasons to translate the text, the place or medium where the translation will appear, etc.. This understanding of translation also implies that the translator needs not only to be aware of linguistic differences, but also of cultural differences. He needs to understand not only the core message that the author wants to get across, but also the author's intentions, the user he

[2] The term *text* may refer here to written as well as to spoken text.

has in mind and the role of the client. In this understanding, translation reflects a holistic approach to make a message in one language accessible to members of another (linguistic, cultural, social, etc.) community in the same way as it could be understood by members of the first community.

Commonly, *translating* is used to refer to written translation as opposed to *interpreting* which is usually oral in nature. But the notion of *translation* described above holds also for interpreting or maybe even more for interpreting because interpreting takes place in a specific communicative situation that determines the *skopos*[3] (Pöchhacker, 1995). For this reason, the verb *to translate* will also be used to refer to interpreting throughout this paper although the more accurate term would be *to interpret*. There are different interpreting modes and the difference between *interpreting* and *translating* is not always as clear-cut as it seems as there are also "in-between modes" like sight translation where written text in the source language is orally rendered in the target language. Similarly, the term *interpreting* may also be used when rendering sign language into spoken language (Pöchhacker, Simultaneous Interpreting, 2011). There are also interpreting modes where the interpretation is delivered *after* the speaker has finished his speech (*consecutive interpreting*) (Pöchhacker, 2011, pp. 294-306). In the following, I will only focus on *simultaneous interpreting* as this is the only interpreting mode that is relevant for this paper.

Simultaneous interpreting has several characteristics that have major implications for the interpreter. First, as already mentioned, simultaneous interpreting is an oral mode of translation. The source text as well the target text is spoken. Spoken information is time-bound, e.g. the source

[3] The notion of *functional equivalence* may, however, depend more strongly on aspects like culture and knowledge about other cultures and languages in interpreting than in written translation and is challenged by the use of English as *lingua franca* (Pöchhacker, 1995).

text advances gradually and earlier parts of the source text get lost as the speaker moves on in his speech. The interpreter needs thus to keep up with the speaker and needs to memorize the parts of the source text he has not translated yet. Simultaneous interpreting is thus associated with time constraints and a high memory burden. Second, simultaneous interpreting is provided in real time while a written translation is usually provided once the source text is completed. As a consequence, the interpreter needs to listen to the speech while giving *at the same time* her translation of the speech. This simultaneity of listening and speaking explains the complexity of the task and motivated a number of reflections on the process of simultaneous interpreting (see for example (Gerver, 1975; Moser, 1978; Setton, 1999).

Third, simultaneous interpreting is conditioned by the situational context and (visual) communicative interactions ("hic et nunc", Pöchhacker, 1999, p. 327). Knowledge about the type of situation and the norms that apply or the interacting partners and their role in a given situation provides the interpreter with important clues to what is being said and allows her to build up a background knowledge that may be useful to disambiguate the auditory input. As simultaneous interpreting takes place in a real time communicative setting, interactions between the speaker, co-speakers (at a panel) and the listeners or between the speaker and his environment may influence the translation choices of the interpreter. For instance, the listeners may laugh at a joke. The speaker may look at another panelist to cede his turn or he may comment on difficulties with the technical set-up. The interpreter needs thus to take the whole setting into account in order to deliver a meaningful translation of what is said[4] (Pöchhacker, 2011;

[4] Apart from the situational context, the constellation of the interacting partners or nonverbal communication, simultaneous interpreting may be constrained by further factors, like the technical equipment (quality of sound or booths) or the text type and delivery speed of the speech (Pöchhacker, 1999).

Pöchhacker, 1999; Riccardi, 2002). A somewhat special case is remote interpreting. During remote interpreting, interpreters still work in the simultaneous mode but they are not in a booth located in the conference room, but in a booth at some other place. Usually, the booth is equipped with screens to show the speaker and/or further elements like the panelists, the conference room and the audience (see for example Roziner & Shlesinger, 2010; Mouzourakis, 2006). Nevertheless, the interpreter will only have limited access to contextual and visual information like the speaker's face, his gestures, the audience or the conference room and may thus depend more heavily on the auditory input which may cause higher stress levels (Riccardi, Marinuzzi, & Zecchin, Interpretation and stress, 1998). This leads us to the next section on visual information and its importance during simultaneous interpreting.

1.2 The importance of visual information in simultaneous interpreting

Human communication is in most cases not restricted to the verbal code[5]. In a face-to-face conversation, lip movements, gestures, eye gaze, facial expressions or other body movements add to what is said. It tells us how the speaker feels about what he is saying, if he is making a joke or if something suddenly caught his interest. The environment or further objects or persons that are visible during a conversation may also provide important information as the speaker interacts with them by pointing at an object or nodding at a person. It is only when the other sees to which person the speaker is nodding that he knows to what person the speaker is referring when he says: " This is my colleague."

These visual cues can also be vital for the interpreter in order to correctly interpret what the speaker is saying (see also Rennert, 2008; Viaggio,

[5] An example for human communication in a verbal-only mode is telephone communication.

1997; Pöchhacker, 1999; Riccardi, Marinuzzi, & Zecchin, 1998; Riccardi, 2002). As this implies that visual input carries some sort of information, I will use the term *visual information* to refer to the visual cues the interpreter receives information from and the term *visual input* to speak of visual stimuli in more general terms (without necessarily implying that it carries relevant information). A sentence like “Here you can see where our company is located” makes only sense to the interpreter when she sees the speaker’s pointing gesture to the map. Likewise, turning the head towards a co-panelist may indicate to the interpreter that the speaker cedes his turn to the next panelist. This may have practical implications for the interpreter, for instance the next person may speak in another langue and the interpreters in the booth may want to switch turns, too. It is thus not surprising that interpreters judge visual information to be crucial for their work. According to a survey by Bühler (1985) 98% of the participating interpreters find it important to see the speaker and in particular his facial expressions. An eye-tracking study conducted by Seubert demonstrated the high relevance of the speaker and of presentation slides during simultaneous interpreting (Seubert, 2017). The author found that interpreters nearly exclusively fixated the speaker or the presentation slide while ignoring other potential sources of visual information like the video transmission of the speaker or the chair of the conference. According to the author, these findings – although tentative due to the low number of participants - suggest that interpreters deliberately choose visual input that provides most information to them. In the same vein, the ISO-norm 2603 (1998) requires interpreting booths to have a window across the whole width of the booth in order to enable visual contact with the speaker and the conference room (for a more thorough discussion of types of visual input in simultaneous interpreting, see chapter 2.4.1).

The importance of visual information during simultaneous interpreting probably gained more attention with the advent of remote interpreting, which has been introduced by international organizations like the United Nations Organization or the European Union to reduce costs

(Mouzourakis, 2006). In remote interpreting, the interpreting booth is not located in the conference room but somewhere else: in another room in the same building or even in another city. Remote interpreting is interesting for the present study to the extent that visual information is only partially available. The visual information that appears on screen varies with each conference. Often, it includes the speaker or the face of the speaker and sometimes the panelists and the conference room (Mouzourakis, 2006). In 2005, the European Parliament conducted a comprehensive study comparing remote and on-site interpreting (Roziner & Shlesinger, 2010). Similar studies were carried out around 2000 by the United Nations (Mouzourakis, 2006), again by the European Parliament (Mouzourakis, 2006) and by the University of Geneva in cooperation with International Telecommunication Union and the Swiss company Swisscom (Moser-Mercer, 2005b). It turned out that interpreters generally reported higher stress levels when working remote than on-site (Roziner & Shlesinger, 2010; Mousavi & Low, 1995; Moser-Mercer, 2005a; Moser-Mercer, 2005b). Moser-Mercer found interpreters to tire faster in the remote condition as indicated by a faster decline of performance (2005b). However, no global difference in performance was found (Roziner & Shlesinger, 2010). Similar findings were reported for on-site interpreting with and without any visual input: while no global difference in performance was found between simultaneous interpreting with or without visual input, interpreters found the task easier when visual input was provided (Rennert, 2008; Anderson L. , 1994) (for a more thorough discussion of studies on visual input in simultaneous interpreting, see chapter 2.4.2.).

The fact that albeit lower self-reported stress levels in the on-site condition (with unlimited access to visual information) performance was not better compared to simultaneous interpreting with limited or no visual information at all raises the question of the benefit of visual information. There may be cases where the benefit of visual information is obvious, like a pointing gesture. If the interpreter does not see where the speaker is pointing to it

will be more difficult or even impossible for her to understand what the speaker is saying. But the fact that interpreters feel less stressed when they have visual contact with the speaker and the conference room suggests that the effect of visual information goes beyond some clues to the meaning of what is said. According to Moser-Mercer (2005b), interpreters felt the effects of fatigue faster in the remote condition than in the on-site condition (although parts of visual information were still preserved in the remote condition). This finding gives rise to the speculation that visual information may not only help to understand the meaning of what is said but even lower task-load during simultaneous interpreting, e.g. visual input may facilitate the task of orally translate a speech in real-time.

Before moving on to investigate this idea more closely, it may be helpful to first get a more precise understanding of what processes are involved in the complex task that is called simultaneous interpreting and how visual input has been taken into account so far. The following section gives a short overview over the processes that are supposed to take place during simultaneous interpreting.

1.3 Simultaneous interpreting: a complex task

Most researchers agree that simultaneous interpreting basically involves perception and analysis of the source text, storage of segments, verbal production of the target text and monitoring of the target text (for models on cognitive load involved in simultaneous interpreting, see chapter 3.3). One of the first attempts to describe the complex task which is simultaneous interpreting was made 1975 by Gerver. He divided the interpreting process into four main components: „input procedures" (Gerver, 1975, p. 125), storage, decoding and encoding and "output procedures" (Gerver, 1975, p. 125). Input procedures encompass the perception of the auditory input and its buffering until the segment is further processed. New input units may be ignored if the buffer is already full and cannot be emptied, e.g. its content cannot be processed

immediately. The input unit in the buffer is retrieved, decoded and encoded in the target language. Gerver distinguishes decoding of “the sounds, words, and sentences heard by the interpreter from his understanding of their [the sounds, words, and sentences] underlying meaning” (Gerver, 1975, p. 126). Output procedures include essentially monitoring and checking mechanisms that take place either before or after the interpreter pronounced his translation (Gerver, 1975).

A few years later, Moser (1978) proposed another process model inspired by psycholinguistic findings. Similarly to Gerver’s model (1975), her model can roughly be divided into four main processes: perception and decoding of auditory input, storage and interaction with long-term memory, encoding of output and monitoring that are closely intertwined. The focus of her model, however, lies on input decoding which she describes very precisely: sound wave patterns are recognized as an auditory signal and compared with acoustic features and phonological rules stored in long-term memory. As soon as a feature has been recognized as syllable, it is held in a perceptual auditory storage and combined with new syllables until a word has been recognized. As soon as syntactical and semantic rules indicate that a combination of words yield a meaningful unit, relevant concepts stored in long-term memory are activated. If the activation is not successful, either new elements need to be processed and added or the unit is discarded. The description of the output is rather short: If the activation of the unit is complete, a paraphrase in the target language is encoded using syntactical, semantic and phonological rules of the target language stored in long-term memory. If the paraphrase allows the interpreter to predict the new unit, she may discard it and concentrate on the next unit. Before pronouncing the target paraphrase, the interpreter compares it against the source language unit to see if it is correct. If this is not the case, she may start a new paraphrase. If her verification is positive, she has completed the translation of a unit and may proceed with all other incoming units until the speech has finished (Moser, 1978; Jacobs, 2013).

For the purpose of the present study the two models briefly outlined above have one important drawback: they do not make any statement about visual input in simultaneous interpreting. More promising in this respect are interpreting models that focus on the interpreter as communicator.

One example that sees the interpreter as actor in the communication process is found in the model developed by Stenzl (Stenzl, 1983). In her model, the interpreter is not simply putting a message from one language into another based solely on what she hears. Instead, the interpreter includes in her target text her background knowledge about the situation and visual information like gestures, presentation slides or other. She also takes into account the speaker's and listeners' sociocultural backgrounds and makes assumptions about the speaker's intention and the listeners' knowledge. The resulting target text is thus not limited to a transcription of the (auditory) source text, but goes beyond the source text and reflects the interpreter's knowledge or understanding of the speaker's intention, the setting and the interacting partners. Another example that explicitly focusses on visual information is the model of nonverbal communication in simultaneous and consecutive interpreting by Poyatos (1987). Poyatos states that a message is not only encoded by (verbal) language, but also by paralanguage (prosody, loudness, word stresses, speech rate, etc.), and nonverbal language or kinesics (gestures, facial expressions, etc.) (see also Poyatos, 1984; Poyatos, 1997). He describes different types of nonverbal communication, like gestures that give indications about space and time, *pictographs* that imitate the shape of something or someone, *deictics* or pointing gestures or movements and others that carry different kinds of information. According to him, nonverbal language and paralinguistic features may differ across cultures. Poyatos therefore suggests that the interpreter should be able to correctly understand nonverbal signs or paralinguistic features in order to render the message correctly into the target language (Poyatos, 1987; Poyatos, 1997). According to Poyatos (Poyatos, 1997, pp. 250-255), the speaker, the audience and the interpreter perceive paralinguistic and nonverbal signs

from each other[6]. This interpretation of interpreting is interesting because it sees the interpreter not only as a passive "channel" through which messages pass but as an active part of the communication process. As such, the interpreter perceives the speaker's gestures and is paralanguage, e.g. his prosody, the stress on important words, etc., but she also emits paralinguistic features that bear the risk to alter the message. Rackow (Rackow, 2013) picked up this idea: In her model, the audience receives verbal information only from the interpreter (typically, the audience listens to the interpretation through headphones), but it receives nonverbal signs from the speaker[7]. Poyatos' (1987; 1997) and Rackow's (2013) models suggest that nonverbal signs are important in order to understand the speaker's message but they do not tell precisely how nonverbal information intervenes in the interpreting process.

A process model that does acknowledge the importance of visual information in simultaneous interpreting and is more precise with regard to the processes that are affected by visual information was developed by Setton (1999). Apart from the auditory input (phonetics and prosody), he assumes other input sources like the environment (for example the conference room or the audience), the speaker's gestures and lip movements. All these input sources interact with the auditory input at different stages of the source text analysis. For example, lip movements help during word recognition. Gestures or visual information from the

---

[6] Poyatos (1987) notes, however, that speaker, interpreter and audience may not perceive paralinguistic or nonverbal features to the same extent. For instance, in simultaneous interpreting, listeners may perceive nonverbal signs from the speaker, but not from the interpreter who is visually absent. Likewise, listeners may not perceive paralinguistic features from the speaker as they listen to the interpreter.

[7] If the interpreter is placed in a booth, she might not be visible to the audience and may not transmit any nonverbal signs to the audience. The case, however, is different in settings like courtroom interpreting, medical interpreting or others where the interpreter is usually visible and present in the same room as the speaker and his interlocutor.

environment may allow the interpreter to correctly recognize the underlying meaning of the source text. Combined with the interpreter's world knowledge they contribute to build a discourse record in memory that is used as basis to produce the target text. In this way, Setton includes visual input as an important clue to the comprehension of the source text. Although Setton emphasizes the importance of visual input in simultaneous interpreting (Setton, 1999, pp. 71-74), his model does not allow predictions about how visual input acts upon simultaneous interpreting. It might therefore be helpful to first look at how visual and auditory stimuli are processed in more general terms and how processing is enhanced by multimodal stimuli.

2 Cognitive processing of stimuli

This chapter deals with the processing of visual, auditory and audio-visual stimuli in the brain. The first section focuses on visual perception and how visual information finds its way into our brain. As the present study uses pupillary responses as physiological indicator for stress, the mechanisms of pupillary reflexes are briefly explained. The second section gives a short introduction to auditory perception and touches upon the question how our brain deals with multiple auditory streams. The focus of this chapter lies on the third section. It outlines how the brain combines auditory and visual stimuli and reviews studies on audio-visual speech processing.

2.1 Vision

Visual perception is enabled by the transformation of light waves - or more precisely: the reflections of the light waves in our environment - into electrical impulses that are processed and integrated in various parts in the brain in order to interpret our environment. The following paragraph gives a short introduction to the principles of vision and pupillary reflexes.

2.1.1 The physiology of the eye

When we look at something, light penetrates through the pupil in the eye. The pupil is actually not an actively working part of the eye, but a little opening. As the pupil does not reflect any light, it is black. Its size is manipulated by the muscles lying underneath the colorful pigmented surface of the iris. Light first passes the cornea that covers and protects the iris and the pupil. Then, light is focused in a biconvex lens. By modulating the thickness of the lens, the eye can adapt the optical power and the visual acuity for objects that are either far away or very near-by. Finally, light is projected on the retina. The retina is composed of two thin layers of cells, the inner and the outer retina, capable to translate the light waves in electrical impulses.

There are two kinds of photoreceptor cells: cones and rods. While cones only react to specific light spectrums, rods are less specified and fire in response to the presence of light. Researchers have been able to identify three kinds of cones: red spectrum cones, blue spectrum cones and green spectrum cones. They are concentrated in the fovea centralis, where the eye obtains the highest resolution. Cones on the outer retina that are situated in the fovea signal each to a cone on the inner retina as to preserve the spatial resolution. For rods, the pattern is reversed: Rods are distributed mainly in the periphery of the visual field and to a lesser extent in the fovea. Each rod in the outer retina receives information from several other rods in the inner retina. As more rods respond to one light spot, the spatial resolution, e.g. visual acuity, is decreased, whilst the visual signal is amplified, which, in turn, allows to us see in poor light conditions (Levin, Nilsson, Ver Hoeve, & Wu, 2011, p. 495; Bhatnagar, 2013, pp. 276-278). The photoreceptor cells form groups according to their specialty (light wave length, light intensity) and location. Each group signals to one ganglion cell, e.g. each ganglion cell receives the signals of one cell group (Levin, Nilsson, Ver Hoeve, & Wu, 2011, pp. 443-458). The ganglion cells axons of each eye are bundled up in the optical nerve of each eye (Levin,

Nilsson, Ver Hoeve, & Wu, 2011, p. 550). Both optical nerves act like a highway that carries all visual information to several cerebral areas involved in integrating visual information. They cross each other at the chiasm opticum where the retinal information is divided according to their visual field and fed back to the optical nerves so that each nerve carries the binocular information on one half of the visual field. The most important projections are the lateral geniculate nuclei (LGN), the superior colliculi, the pretectae and the pulvinar (Bhatnagar, 2013, p. 276; Levin, Nilsson, Ver Hoeve, & Wu, 2011, p. 545).

2.1.2 Visual processing beyond the retina

Ninety percent of the ganglion cells signal to the LGN which is situated in the thalamus. It controls the information flow to the primary visual cortex (V1) and is essential for visual awareness. It has six layers that each receives signals from one eye. Its input comes not only from the retina, but also from other cortical sources, like the primary visual cortex, the superior colliculi in the midbrain, and the pretectae. In addition, non-visual information, like motor and other sensory input, arrives from the hypothalamus and the raphné nuclei. These very diverse input sources provide feedback that modulates the retinal input (Levin, Nilsson, Ver Hoeve, & Wu, pp. 574-579). Research suggests moreover that the LGN links eye movements and attention. It has been shown, for instance, that LGN activation increases when subjects attend to the stimuli and diminishes for unattended stimuli (Levin, Nilsson, Ver Hoeve, & Wu, p. 583; Bhatnagar, 2013, p. 283). Minor projections of the retina are the superior colliculi which are associated with eye movements and binding of different sensory modalities (see chapter 2.3), the pretectae that is linked to the Edinger-Westphal nucleus and controls the pupillary reflexes (see chapter 2.1.3), and the pulvinar, that is situated in the thalamus and, like the LGN, connects directly to the primary visual cortex (V1). Several studies indicate that the pulvinar nucleus may contribute to encode the saliency of visual information or may have a role in visual attention. In

addition, the pulvinar seems to be involved in hand-eye-coordination (Levin, Nilsson, Ver Hoeve, & Wu, 2011, pp. 545-546). The LGN transmits the visual information of its six layers to V1. It is only at this level, in V1, that the visual information of the two eyes is combined, although it seems that 3D vision only emerges in higher-processing areas. Each cell of V1 corresponds to a specific group of photoreceptor cells. This way V1 forms a detailed map of the visual field that allows for exact localization of an object in the visual field. More precisely, it is organized in columns that each “analyzes one portion of visual space” (Levin, Nilsson, Ver Hoeve, & Wu, p. 594). Each column consists of six main cell layers with different cell types that respond to specific orientations or directions and encode information about color, motion, temporal and spatial frequency that is chiefly passed on to the adjacent visual area V2 and higher visual areas (Levin, Nilsson, Ver Hoeve, & Wu, 2011, pp. 586-598; Bhatnagar, 2013, p. 284).

From V1 on, visual processing seems to split up in two different pathways: a dorsal stream and a ventral stream. The dorsal stream is associated with spatial encoding. It signals to three visual areas: V3, V5 and V6. V5 and V3 are involved in the perception of depth and visual motion processing and receive information about direction and orientation from V1 and V2. As such, lesions affecting V5 make it impossible for the patients to identify motion (Levin, Nilsson, Ver Hoeve, & Wu, pp. 605-607). V6, an area that is situated in the parietal occipital sulcus, displays a detailed visual field map, including the border regions of the visual field (Levin, Nilsson, Ver Hoeve, & Wu, p. 607). The ventral stream is involved in object and face recognition. An important mechanism of the ventral stream is the loss of retinotopic information, e.g. information about where a stimulus is located on the visual field. This mechanism is essential in order to perceive an object as a whole. The ventral stream starts in V4 and reaches ultimately the anterior inferior temporal cortex. Activation patterns in the anterior inferior temporal cortex encode rather complex features that are transmitted to higher processing parts of the brain that are involved in the

integration of multisensory information, as for example in visual speech processing (lip reading) (Levin, Nilsson, Ver Hoeve, & Wu, 2011, pp. 607-612; Ungerleider & Pessoa, 2008).

To date, we do not know exactly how our brain processes lip movements and extracts linguistic information from it. Most studies agree that visual speech processing leads to activation in the posterior superior temporal sulcus (pSTS) (Putzar, et al., 2010; Blank & von Kriegstein, 2013; Campbell, 2008; McGettigan, et al., 2012; Calvert, Hansen, Iversen, & Brammer, 2001). This region receives input from the TE and auditory areas and is generally associated with the integration of form and motion (Barraclough, 2005). Campbell (2008) therefore suggests that the pSTS is involved in processing the dynamic features of audio-visual speech, e.g. the combination of visual and auditory speech signals (Campbell, 2008). This seems even to work for consecutive auditory and visual speech stimuli. Blank and von Kriegstein (2013) reported increased activity in the pSTS if the visual speech stimulus did not correspond to the preceding auditory speech stimulus. Importantly, activation correlates positively with performance. Based on this finding, the authors theorized that pSTS increases visual speech recognition by signaling prediction errors (Blank & von Kriegstein, 2013). Other sites involved in lip reading are the primary auditory cortex, left inferior frontal gyrus and the supplementary motor areas (Blank & von Kriegstein, 2013; Putzar, et al., 2010). The activation of both, perceptual and motor areas, suggests that the processing and execution of lip movements involve the same neural systems (Blank & von Kriegstein, 2013).

2.1.3 Pupillary reflexes

Pupil movements regulate how much light enters in the eye and contributes to accommodate to near-by objects (Bhatnagar, 2013, pp. 284-285). Thanks to pupillary reflexes, we can see in darkness and get a high resolution image of objects at only a few centimeters distance. But pupil diameter is not only an indicator for illumination or distance of an object. It

also tells us about neuronal damages and about the state of alertness or arousal of a person. Pupil assessment has therefore many applications for physicians and researchers (Levin, Nilsson, Ver Hoeve, & Wu, 2011, pp. 502-503; Trepel, 2012, pp. 315-330; Schmidt, Lang, & Heckmann, 2010, p. 352).

Pupil movements are controlled by the ciliary muscle. The ciliary muscle lies underneath the iris and contains two muscles, the dilator and the sphincter. The sphincter borders the pupil and mediates a pupil constriction. The dilator lies in the midperiphery of the iris and causes the pupil to dilate. As the sphincter is stronger than the dilator, the sphincter needs to relax so that the dilator can pull the pupil open. But when does the sphincter relax? In the case of light adaptation, the key to this question is the Edinger-Westphal nucleus. It is part of the parasympathetic pathway of the central nervous system, which in turn is regulated by the hypothalamus (Levin, Nilsson, Ver Hoeve, & Wu, pp. 508-509). Basically, the parasympathetic pathway effects physical reactions that could be summarized as “rest and digest”: it stimulates the digestive tract, reduces the heart beat rate and constricts the pupil. As long as the Edinger-Westphal nucleus is active, for example while sleeping or under anesthesia, the pupils stay small. But as soon as the Edinger-Westphal nucleus in inhibited, the sphincter relaxes and the pupil dilates (Schmidt, Lang, & Heckmann, 2010, pp. 295, 409,411).

To date it is unclear if the same mechanism also underlies pupillary responses in the case of arousal and stressors like pain or fear that activate the sympathetic system. The sympathetic system prepares the body to respond to these stimuli in mediating “fight and flight” reactions: increase of the heart rate and the respiratory rate, inhibition of digestion and activation of smooth muscles like the iris dilator[8] (Schmidt, Lang, &

[8] It is, however, important to note that sympathetic and parasympathetic pathways are not antagonist to each other, but complete each other (Schmidt, Lang, & Heckmann, 2010,

Heckmann, pp. 405-409). Gilzenrath and colleagues (2010) argue that in case of sympathetic activation, pupil dilations are elicited by the locus coeruleus-norepinephrine system. In short, they propose that the norephinephrine system mediates different control states: phasic firing facilitates the processing of task-relevant stimuli and increases performance (exploitation mode). The phasic state is associated to phasic pupil dilations during task execution. Tonic activation, in contrast, favors the exploration mode, e.g. the processing of irrelevant stimuli and distractibility. It is associated with tonic (pre-test) pupil dilations (Gilzenrath, Niewenhuis, Jepma, & Cohen, 2010).

Even if the underlying neural principles are the same for each subject, pupillary responses can differ very much. This is due to changes in the iris during dilation and contraction that restrict the pupil diameter (Levin, Nilsson, Ver Hoeve, & Wu). For instance, age seems to be a crucial factor determining maximal pupil movements. Older subjects show smaller pupil dilation in response to darkness than younger subjects do (van Gerven, Paas, van Merrienbroer, & Schmidt, 2004).

2.2 Audition

The sounds we hear are actually oscillations in the air. The wave's length defines the frequency of the sound perceived as pitch. The wave's amplitude defines its intensity which we perceive as loudness. Simple noises consisting of only one frequency like that of a sinus tone, however, are seldom. Usually, sounds do not display one, but a multitude of different frequencies all overlapping one another. The human ear can perceive frequencies between 20 hz and 20 000 hz, but it does not perceive them equally well. Within the same sound pressure level, extreme frequencies seem softer than sounds in the middle range

---

pp. 405-409). In most stress situations, both parts of the central nervous system are activated (Trepel, 2012, p. 295).

(Schiffmann, 1996, pp. 333-336). The ear is particularly sensitive for frequencies between 200 hz and 2000 hz and as such perfectly adapted to the perception of human speech. Furthermore, it is well designed to localize sounds and contributes thereby to direct attention to stimuli even outside the visual field (Bhatnagar, 2013, pp. 220-223; Ward, The Student's Guide to Cognitive Neuroscience, 2nd edition, 2010, pp. 211-212).

2.2.1 The physiology of the ear

The ear is divided into three parts: the external ear, the middle ear and the inner ear. The external ear captures the sound waves and conducts them through the auditory meatus to the tympanic membrane. The tympanic membrane vibrates as soon as a sound wave hits it. This in turn activates the ossicular chain in the middle ear, the maellus, the incus and the stapes that are attached to the membrane. They transmit the wave patterns to the oval window. The ossicular chain has a leverage effect on the sound pressure level. Furthermore, the oval window is smaller in size than the tympanic membrane. Together, these two mechanisms amplify the pressure level of the sound waves. The amplification of the pressure level is necessary for processing auditory stimuli in the inner ear that lies behind the oval window. The inner ear encompasses the vestibular organ which helps us to maintain the balance, and the snail-shaped cochlear for hearing. The cochlear duct is filled with a liquid called endolymph that allows the oscillations to propagate. It is divided into three parts: the scala vestibular that directly connects to the oval window, the scala tympani that is particularly important to compensate excessive pressure, and the scala media with the organ of corti. The organ of corti contains hair cells that are placed beneath the tectorial membrane. As the sound wave moves through the cochlear it deforms the tectorial membrane and compresses the hair cells, thereby converting the waves into electrochemical signals. Each hair cell is tuned to a specific frequency. The hair cells are tonotopically distributed, e.g. hair cells for higher frequencies are situated

near the oval window, while hair cells for lower frequencies are located at the apex of the cochlear. As the basilar membrane becomes less and less rigid to its apex, it responds better to lower frequencies at its center. The tonotopic arrangement allows thus for a maximal response of each hair cell (Bhatnagar, 2013, pp. 220-223; Ward, 2010, pp. 211-212).

2.2.2 Processing of auditory stimuli

The auditory pathway relays the information from the cochlear to the auditory cortex and decodes information about timing, frequency or intensity. It encompasses several cerebral regions like the cochlear nucleus, the superior olivary nucleus, the inferior colliculus and the medial geniculate body. The cochlear nucleus receives information directly from the hair cells. Each neuron in the cochlear nucleus corresponds to one hair cell, with higher frequencies being represented in deeper layers and lower frequencies being represented at more superficial layers. The cochlear nucleus projects to the superior olivary nucleus in the pons. The superior olivary nucleus combines the information from both ears and uses differences in time and intensity to localize sounds and prompt necessary eye movements. It also feeds back to the cochlear nucleus and the outer hair cells. This feedback helps to adjust the tectorial membrane and to intensify certain frequencies while softening others, for example in order to suppress background noise. The auditory pathway reaches further to the inferior colliculus where higher level auditory processing takes place. The inferior colliculus is situated right beneath the superior colliculus which receives visual information. Its proximity to areas of visual spatial processing triggers head and body movements when unexpected sounds occur, and contributes to attentional hearing. The medial geniculate body receives the auditory input from the inferior colliculus and transmits it to the auditory cortex. It has furthermore numerous projections to very divers regions like the basal ganglia, the amygdala or the temporal and parietal cortici and is therefore supposed to contribute to attentional and emotional processing (Bhatnagar, 2013, pp. 228-231; Ward, 2010, pp. 210-213). In

the auditory cortex, the processing of auditory information gets more and more complex as the auditory information moves on through the different auditory areas. Although the tonotopic representation is largely preserved, it is not the only information that is processed. For example, some cells respond to changes in frequency or intensity rather than to specific frequencies (Bhatnagar, 2013, p. 231; Ward, 2010, pp. 214-216). Research suggests that complex sounds are processed along different streams: While some areas process mainly spatial and timing information, others are responsible for sound recognition (Ward, 2010, p. 218).

The auditory areas then project to a large network of areas involved in language processing. Central in language processing are Wernicke's and Broca's areas. Wernicke's area is situated largely in the superior temporal gyrus, while Broca's area is located in the inferior frontal gyrus. Patients suffering from lesions in Wernicke's areas typically show fluent speech production, but impaired speech comprehension. Therefore, Wernicke's area has been related to semantic processing. Patients suffering from lesions in Broca's areas show the reversed pattern: preserved speech comprehension, but strongly affected speech production. Clinical linguistics therefore associate Broca's area with syntactic processing. The division between Broca's and Wernicke's areas, however, may not be as clear-cut as aphasia symptoms suggest. Friederici and her research team propose in several publications (2002, 2012, 2013a, 2013b) that both areas are connected through two ventral and two dorsal pathways. The ventral pathways play a central role in mapping the auditory input onto meaning and in local syntactic processing. Single cell studies suggest that the left superior temporal gyrus contributes to phoneme and auditory word processing (Friederici & Singer, 2015). In a very elegant study combining eye-tracking and fMRI data, Bonhage and colleagues (2015) distinguished areas involved in lexical-semantic and syntactic predictions. They contrasted regular, meaningful sentences with sentences in which they had replaced the central words by non-words while maintaining the grammatical structure (jabberwocky sentences). In both conditions, they

had replaced the final target words by two boxes, one for “noun” and one for “verb” and asked the participants to judge whether the final word was a verb or a noun. Their results revealed increased activation in the inferior and middle frontal gyri, the left superior and temporal sulci during syntactic word category predictions in jabberwocky conditions and supramarginal gyrus and middle temporal gyrus during lexical-semantic predictions in regular sentences. These findings suggest that syntactic and lexical-semantic processing indeed follow different pathways through the brain (Bonhage, Mueller, Friederici, & Fiebach, 2015). The dorsal pathways are involved in syntactic processing and in transforming meaning into articulatory processes. The first pathway covers essentially the anterior superior temporal cortex and the posterior superior temporal gyrus (retrieving basic syntactic information and verb-argument processing), as well as the frontal operculum and the pars opercularis (syntactic processing). Integration of syntactic and semantic information takes place in the inferior frontal gyrus and the posterior temporal cortex. The second pathway projects to the premotor cortex, linking sentence meaning and articulation (Friederici, 2002; Friederici, 2009; Friederici, 2012; Friederici & Singer, 2015; Friederici & Gierhan, 2013).

2.2.3 Interactions of multiple auditory streams

The section above described how individual auditory stimuli are processed. But most of the time our environment is rather noisy with several sounds overlapping each other (traffic, telephone ringing, several people talking at the same time, etc.). As a consequence, parts of one auditory stream “disappear” or become temporarily inaudible because of the second, interfering auditory stream or noise. If we engage in a conversation we often do not even notice that we actually missed a part of the conservation because we anticipate what our interlocutor will be saying next (Schiffmann, 1996, p. 384). But even taking apart linguistic aspects like lexical or sentential context, our brain efficiently deals with multiple auditory streams and is able to segregate them from each other

by using different cues. This enables us to listen to an interesting conversation while ignoring irrelevant sounds. Important cues are different onsets of auditory streams, frequency, amplitude or binaural masking level difference. Binaural masking level difference means that two competing auditory streams interfere differently on each ear. A listener will perceive a sound emanating at his left louder on the left ear than on the right ear. A second sound emanating in front of the listener will be equally loud on both ears. When both sounds occur simultaneously, the first sound will mask the second one stronger on the left ear than on the right ear. This difference in masking contributes to the segregation of different auditory streams (Carlyon, 2004).

Interference can occur at a purely perceptual level, for example if we hear the traffic during the rush hour in the background while discussing the shopping list with our partner on the way to the super market (*energetic masking*). At some point, the traffic noise may mask parts of our partner's reply and make it more difficult to discriminate similar sounds because the sound waves emanating from the traffic cover (partially) those of the speech. Another form of interference is *informational masking*. Informational masking affects speech perception not at the perceptual level, but affects the selection of one lexical candidate among other candidates. This is the case when the listener experiences high cognitive load induced by a concurrent task (visual search task, recall, etc.) (Mattys, Brooks, & Cook, 2009). Mattys and Wiget (2011) found that energetic masking impedes lexical-semantic access and leads native speakers to rely more heavily on sublexical cues, such as prosodic or allophonic features or stress, whereas during a concurrent visual search task (informational masking), native speakers favor lexical-sematic cues. The shift towards lexical-semantic cues in informational masking is less strong in non-native speakers, probably due to weaker lexical knowledge (Mattys & Wiget, 2011; Mattys, Caroll, Li, & Chan, 2010; Lecumberri, Cooke, & Cutler, 2010).

2.3 Audio-visual stimuli

The mechanisms involved in processing of audio-visual stimuli have been a topic of extensive research in the field of multisensory integration or multimodal binding. Each of our senses encodes different features of the world. Our eyes process information about luminosity and colors, our ears relay information about sound frequency and intensity, our skin allows us to perceive pressure or temperature. The words *multimodal* or *multisensory* refer to the different senses and the way they encode the different features of a stimulus. In the following, both terms are used interchangeably. Multisensory integration means that two or more stimuli of different modalities are not perceived separately, but as one combined scene. When infants play with their rattle, shaking it and exploring it with their hands and tongue, they get different sensory information and learn to put them together as to perceive one brightly colored round object with a smooth surface that makes a sound when it is shaken.

Another way to think about multisensory integration is hand-eye coordination or head and eye movements triggered by (unexpected) sounds. The visual field of the human eye covers about 170°, but the fovea, where the human eye has the highest acuity, is limited to 3°. Whenever we hear something outside of our visual field, we need to turn our head or to move our eyes. This happens at a very early stage of auditory processing even before the meaning of the sound has been analyzed (Bhatnagar, 2013, p. 231). It shows that auditory and visual processing is linked very closely to each other. The following chapters explore the principles governing multisensory integration, its localization and neural functioning and its behavioral effects.

2.3.1 Multisensory binding: facilitated response for multimodal stimuli

None of our senses works flawlessly, and no signal is unambiguous. There is always some noise. In other words, our brain can interpret

information from our senses in different ways. Integrating information from other senses helps to limit the number of potential interpretations because information from the second sense completes information from the first one (Koelewijn, Thomas, Bronkhorst, & Theeuwes, 2010). To understand how multisensory integration works at a neuronal level, researchers most commonly use imaging or recording techniques such as fMRI, EEG, single unit recording with electrodes that are directly positioned in the brain, PET or MEG. They try to identify cells that do not only respond to one modality, but also to a second (or a third) one. One major finding was that the cells' joint response to a bimodal stimulus is stronger than would be predicted on the basis of the respective unimodal responses (*superadditivity).* These cells were called multisensory cells, in contrast to unisensory cells that exhibited a depressed response in response to bimodal stimuli (Meredith & Stein, 1986). Calvert and colleagues used fMRI to identify integration sites in the human brain. Their stimulus was a checkerboard that alternated with a resting condition (gray screen) every 30 seconds, and white noise bursts, that alternated with a silent resting condition every 39 seconds. The visual and the auditory signal were thus presented simultaneously or separately. When they analyzed the BOLD-signals they had obtained during the three presentation modes (audio-visual, visual-only, auditory-only) they discovered a super-additive response in various regions; the highest response was found in the superior colliculi which will be discussed in more detail in chapter 2.3.2. The neuronal activity in the superior colliculi in response to the audio-visual stimulus was stronger than the sum of the response to the auditory only and the visual only stimulus (Calvert, Hansen, Iversen, & Brammer, 2001; see also Klemen & Chambers, 2012).

At a behavioral level, multisensory integration very soon caught psychologists' interest. As soon as 1954, Broadbent discovered that participants were able to recall different digits presented simultaneously on both ears (see Broadbent, 1956). In 1956, he tested whether this effect was present in other sensory conditions. He divided the participants in two

groups. In both groups, he tested recall of six numbers in a visual-only and an auditory-only condition. In group 1, he additionally presented two different digits simultaneously to one ear and the eyes, one pair each second. In group 2, he presented three digits successively to the ear and simultaneously three (other) digits to the eyes. He observed an advantage of audio-visual presentation on recall in group 2, but not in group 1, and concluded:

> It is probable from the results of these experiments that temporary storage of information in one sensory channel is possible with a number of different senses, and certainly it is not peculiar to the binaural situation. It may, however, be replaced by alternation of attention or confusion of the two channels if the sensory cues associated with the two channels are not very distinctive. (Broadbent, 1956, p. 151)

Since these first studies, researchers have studied multisensory integration using different stimuli and methods. They observed that multisensory integration has a *facilitating effect*: it speeds up the detection of target stimuli[9], lowers the detection threshold and facilitates recall. Moreover, input from a second modality can increase response accuracy. In a simple reaction time task, Molholm and her colleagues (2006) found shorter reaction times in response to audio-visual stimuli as compared to auditory or visual stimuli alone. Similar findings were reported by Teder-Salejärvi and his colleagues (2002), by Miller (1982) for letter search tasks, and by Giard and Peronnet (1999) in a decision task where participants were to decide to which object a visual, auditory or audio-

[9] An alternative view holds that target detection is sped up by preparation rather than multisensory integration. As auditory stimuli are processed faster than visual ones (see also (Meredith, Nemitz, & Stein, 1987), the auditory stimuli could act as a warning cue and prepare visual detection (Los & Van der Burg, 2013). This could explain why the facilitation effect occurs even if both stimuli are not semantically congruent. But even if this hypothesis is true and the facilitation effect is not due to multimodal binding, but preparation, it does not change the fact, that reaction times to audio-visual stimuli are shorter compared to auditory or visual stimuli.

visual stimulus belonged to. In addition, participants made fewer errors identifying the target stimuli in an audio-visual condition compared to an auditory or visual condition in all four studies.

However, there might be some factors that influence accuracy and recall. Lewandowski and Kobus (1989) presented simultaneously high or low tones and high or low lighted pixels to their subjects and the following conditions: 1) tones only with white noise, 2) pixels only with (visual) white noise, 3) tones only with auditory and visual noise, 4) pixels only with auditory and visual noise, 5) tones and pixels with auditory and visual noise. First of all, they found a lower detection threshold and faster reaction times in the bimodal condition. The effect on accuracy depended on whether both stimuli (auditory and visual) corresponded to each other (high tone and high-lighted pixel) or not. In trials where both stimuli were congruent accuracy increased significantly. Conversely, in cases where they did not correspond to each other (high tone and low lighted pixel) accuracy decreased. In a second study, Lewandowski and Kobus (Lewandowski & Kobus, 1993) asked the participants to recall words and to classify them according to their category. Words were presented in a visual-only, an auditory-only or in an audio-visual mode and in an audio-visual condition with two different words. In contrast to the first study, Lewandowski observed faster reaction times for the visual-only condition, but better recall in the congruent audio-visual condition. He concluded from these results that the facilitation effect of multisensory input on accuracy only holds for redundant information in two sensory modalities. More recent EEG-studies confirm the advantage for congruent versus incongruent stimuli. Baart, Stekelenburg and Vroomen (2014) observed a negative ERP-wave peaking at 200-500 ms after stimulus onset that was larger for incongruent than congruent sounds (Baart, Stekelenburg, & Vroomen, 2014).

The studies described above demonstrate that audio-visual presentation has a facilitation effect on behavioral measures like reaction times, error

rate or detection threshold and enhances the neural response compared to a unimodal presentation mode, but only if certain conditions are complied with. For example, stimuli have been presented simultaneously and at the same location. Apart from the spatial and temporal congruence, a third principle has emerged: semantic congruence: This means that multisensory integration occurs only if both stimuli convey the same meaning. The following sections look more closely at those three conditions, beginning with temporal and spatial congruence.

2.3.2 The superior colliculus: detecting temporal and spatial congruence

Multisensory integration does not occur randomly. The first condition that is essential for multisensory integration is that both stimuli share the same localization. If we see a dog in front of us, but hear something barking to our left, we would think that there is a second dog to our left, while the first dog in front of us is may not bark at all. The second condition necessary for multisensory integration of two stimuli is that both stimuli occur at the same time. If we see a barking dog, but hear something barking one minute later, we would assume that there is a second dog and that we did not hear the first one, maybe because there was too much surrounding noise. Both conditions go hand in hand as they are linked to the same cerebral structure: the superior colliculus.

As described in chapter 2.1.2, the superior colliculus has six layers. Its superficial layers receive its input largely from the retina. The intermediate and deeper levels are primarily associated with motor control, like eye or head movements (Groh & Werner-Reiss, 2002). This connection to motor control is what makes us turn our head or move our eyes when we hear a sound outside our visual field. It is therefore not wrong to think of the superior colliculus as a structure that triggers auditory reflexes. More recent studies have further revealed connections between the superficial and the deeper layers of the superior colliculus that even allow auditory input to modulate the firing activity of visual cells (Ghose, Maier, Nidiffer, &

Wallace, 2014), as well as cortical projections from the anterior ectosylvian sulcus that modulate multisensory integration (Alvarado, Stanford, Vaughan, & Stein, 2007). But more importantly, the deeper layers of the superior colliculus contain multisensory neurons. In contrast to the upper layers of the superior colliculus that react only to visual input coming from the retina, these neurons are not modality-specific. They fire each time they receive visual, auditory or tactile input. Their firing rate even increases whenever they receive input from different modalities at the same time. Another category of multisensory neurons inhibits the response when the multimodal input is asynchronous. This enables temporal coding of the sensory-specific input (Groh & Werner-Reiss, 2002).

Single cell studies in cats have shown that the critical time window for multisensory integration is about 100 ms. In cases where the asynchrony between two stimuli of different modalities exceeds the critical time window multisensory cells show a depressed response (Meredith, Nemitz, & Stein, 1987). Response depression for asynchronous visual and auditory stimuli was also obtained in humans (Noesselt, et al., 2007). Maier and colleagues (2011) have investigated the asynchrony detection in audio-visual speech. They presented a speaker uttering short sentences on video and asked their participants to judge whether speech and lip movements were synchronous or not. Speech asynchrony was manipulated between 0 ms and 333 ms. Both, the auditory stream or the visual stream, could be leading. Most synchronous judgements were made for a discrepancy between 0 ms and 200 ms. On the whole, participants tended to tolerate larger asynchronies when the video track was leading. This is in line with research by Stevenson, Zemtov and Wallace (2012) and not surprising as auditory signals are processed faster than visual ones (Meredith & Stein, 1986).

Multisensory neurons have receptive fields that do respond to stimuli of different modalities that occur either at the same location or at different,

but strongly related locations. How does the cell know that two stimuli share the same location? For visual stimuli, the location can be deduced from the activation site on the retina and the gaze direction. For auditory stimuli, inter-aural timing differences help to identify the location of the sound source. In some neurons, all receptive fields react to the same location; in other cells, the receptive fields do react to locations that are somehow linked, but not necessarily the same. For example, tactile sensations on the face correspond to visual input in the center of the visual field, while tactile stimuli on the back are related to peripheral areas of the visual field. If two stimuli that encode different modalities but the same location reach the corresponding multisensory cell, this cell fires faster than in response to a unimodal stimulus alone. If two stimuli of different modalities do not share their location, they do not activate the same cells (Groh & Werner-Reiss, 2002).

Neurophysiological evidence for the importance of spatial congruence came from an experiment by Meredith and Stein (1986). They stimulated multisensory cells with noise bursts, light spots and vibrations in anaesthetized cats while measuring the firing rate of multisensory neurons and observed an enhanced response for multisensory stimuli at the center of the receptive field and within the same time frame, and a depressed response when one of the stimuli was outside the receptive field, but still within the same time frame (Meredith & Stein, 1986). In a meta-analysis, Koelewijn and colleagues (2010) note that conversely to the narrow temporal window, spatial congruence is not always a prerequisite for multisensory integration to occur and might depend on the location of the visual stimulus. As such, research so far suggests that spatial congruence seems not be necessary for multi-sensory integration for stimuli that are located in the center of the visual field, while it plays an essential role for stimuli at the periphery of the visual field. Anecdotal evidence for multisensory binding of spatially incongruent stimuli may be provided by the ventriloquist effect.

Multimodal binding is obviously not a clear-cut mechanism that automatically and infallibly takes place, but rather depends on several factors, among which temporal synchrony appears to be the most important whereas the localization of the stimuli, in contrast, plays a minor role. Both factors, temporal and spatial congruence, are mainly localized in the superior colliculus. At this early stage of sensory processing, multisensory integration is probably automatic (Koelewijn, Thomas, Bronkhorst, & Theeuwes, 2010). Research suggests, however, that multisensory integration may be partially attention-driven and therefore task-dependent. In fact, researchers found early ERP-components peaking around 30 to 80 ms after stimulus onset and ERP-components that peaked much later, around 120 to160 ms after stimulus onset (Brown, Clarke, & Barry, 2006; Teder-Sälejärvi, McDonald, Di Russo, & Hillyard, 2002; Molholm, et al., 2006; Giard & Peronnet, 1999). Koelewijn and colleagues (2010) hypothesize that early ERPs reflect the automatic early stage of multisensory integration, whereas late ERPs reflect the attention-driven stage of multisensory integration (Koelewijn, Thomas, Bronkhorst, & Theeuwes, 2010). Temporal and spatial congruence, however, are not the only factors altering multimodal binding. The next chapter explores the third aspect of multimodal binding: semantic congruence.

2.3.3 The superior temporal sulcus: ensuring semantic congruence

Most studies agree that the superior temporal sulcus plays an essential role in combining stimuli that are semantically congruent[10] (Barraclough, 2005), and in particular in audio-visual speech (McGettigan, et al., 2012; Campbell, 2008; Klemen & Chambers, 2012). In support of this hypothesis, Barraclough presented actions with matching or mismatching

[10]Alvarado and colleagues (2007) have also reported cortical connection to the superior colliculus that mediate a response enhancement for bimodal stimuli. The authors suggest that these connections contribute to a context-dependent control of multisensory integration.

sounds to two monkeys while recording the responses of cells responding to their visual stimuli in the superior temporal sulcus. He found an 86% increase of firing rate in 23% of the recorded cells in a condition of audio-visual congruence compared to a condition of audio-visual incongruence. In some cells, however, the firing rate decreased even in those cases where the sound stimulus matched the visual stimulus (Barraclough, 2005).

McGettigan and colleagues (2012) measured activation in several sites of the brain while presenting audio-visual speech stimuli to the participants. The authors manipulated the clarity of the auditory and the visual signal and observed enhanced activation in the superior temporal sulcus in response to greater visual and auditory acuity, suggesting that the superior temporal sulcus contributes to audio-visual speech processing (McGettigan, et al., 2012). Willems, Özyürek and Hagoort (2009) reported increased activation in the superior temporal sulcus for speech and pantomimed actions. The activation was significantly higher, when the pantomime matched the auditory input than when both stimuli were incongruent. Moreover, recall accuracy was significantly higher for matched than for mismatched stimuli when participants were asked to judge, whether they had seen the pantomime before (Willems, Özyürek, & Hagoort, 2009). Similar findings were reported for co-speech gestures by Holle and colleagues (2008). In this study, participants were confronted with ambiguous sentences and matching or neutral accompanying gestures. The enhanced activations in the superior temporal sulcus to matching compared to neutral gestures corroborates further the hypothesis that the superior temporal sulcus is involved in the integration of semantically congruent auditory and visual information.

2.3.4 Feedback-mechanisms and further integration sites

Sensory information from different modalities is linked in many other ways. Neurophysiological studies in macaque monkeys and ferrets have revealed connections between modality specific cortici. For example, the

visual area V2 synapses to the secondary auditory cortex in the macaque monkey. These connections are believed to enable interactions between both sensory modalities. This would mean that sensory modalities are not only processed in a hierarchical way, but that they can modulate each other's response. (Klemen & Chambers, 2012; Noesselt, et al., 2007). Other areas that react to multimodal stimuli are the frontal eye fields, the intraparietal sulcus and the lateral parietal areas. The frontal eye fields might also contribute to the control of auditory attention, but further research is necessary to confirm and define the exact role of these areas. (Groh & Werner-Reiss, 2002; Klemen & Chambers, 2012).

Another part of the brain that may be involved in multisensory integration is the thalamus. The exact role of the thalamus in multisensory integration remains to date unknown. Some studies suggest that the thalamus plays a certain role in receiving and redistributing multisensory information across cortical areas. Moreover, there might be connections between different modality-specific cortical areas that cross the thalamus and that may be faster than connections between cortical areas that do not trespass the thalamus. Another possibility might be that the thalamus mainly serves to inhibit multisensory integration if inputs of different modalities need to be processed separately or to facilitate integration if it is expected that both inputs belong together (Klemen & Chambers, 2012). Expectation might also explain why some individuals are more sensitive to temporal asynchronies than others (Stevenson, Zemtov, & Wallace, 2012) or why attention modulates multisensory integration (Agnès, Möttönen, Sams, Soto-Faraco, & Tiippana, 2014)

2.3.5 Processing of audio-visual speech

The three aspects, temporal, spatial and semantic congruence, hold especially for audio-visual speech processing, where usually the auditory stream, e.g. what is being said, corresponds in general to the visual stream, the lip movements of the speaker. The most cited example for audio-visual speech integration is probably the McGurk-effect. Mc Gurk

and MacDonald (1976) asked their participants to tell whether they heard the sound /ba/, /ga/ or /da/, while watching a person pronouncing those syllables. The particularity of their experiment was that the auditory and visual inputs were partially mismatched. Participants did not always see the sound they heard. When participants heard the sound /ga/ while seeing the lip movements corresponding to the sound /ba/, thus two incongruent stimuli, they reported to hear “da”, rather than saying that both stimuli did not correspond to each other. McGurk and MacDonald interpreted this phenomenon as a fusion of two incongruent stimuli, a sort of compromise to make both stimuli compatible to each other (Mc Gurk & MacDonald, 1976). Again, this demonstrates how expectation can affect perception. The McGurk-effect is very persistent and occurs even if participants deliberately concentrate on one sensory input and try to ignore the other (van Wassenhove, Grant, Poeppel, & Halle, 2005), if they do not look directly at the mouth of the speaker (Paré, Richler, & Ten Hove, 2003) or if the stimuli are non-speech sounds (Brancazio, Best, & Fowler, 2006). Participants notice asynchronies or incongruences only if they are told to judge whether both stimuli correspond or not (Vroomen & Stekelenburg, 2011). Audio-visual speech integration seems thus to occur – at least in daily life – without voluntary effort.

Audio-visual speech integration facilitates listening comprehension because neither auditory nor visual speech inputs are unambiguous, or to put it in other words: no signal, whether it comes from the eye or from the ear, is perfect. This means, too, that auditory and visual information are never completely redundant. Some visual speech features match more than one sound, even if more features are visible than one would think (Munhall, Kroos, Jozan, & Vatikiotis-Bateson, 2004). For instance, closed lips match the sound /b/, /p/ or /m/. Similarly, auditory speech sounds can be very similar and hard to discriminate, especially in cases where the auditory input is not well perceived because of a distracting task (Mattys & Wiget, 2011) or in case of a hearing impairment (Kuchinsky, et al., 2013; Massaro & Cohen, 1999), background noise (Sumby & Pollack, 1953;

Hazan, Kim, & Chen, 2010; Bernstein, Auer, & Takayanagi, 2004; Benoit, Mohamadi, & Kandel, 1994) or a foreign accent (Anderson-Hsieh & Koehler, 1988). Complementary visual input, like lip movements, reduces the number of potential candidates that match the auditory input. As the number of candidates increase in adverse listening conditions because of the loss of auditory information, it stands to reason that visual input benefits most when the auditory input is poor and vice-versa (Massaro & Cohen, 1999).

To account for this reflection, Massaro and Cohen (1999) developed a *fuzzy logical model of perception*. They divide speech perception (and perception in general) in three steps: evaluation, integration and decision. In the first step, the incoming sensory units are compared to so-called *prototypes*. Prototypes are "descriptions of a perceptual unit of language. […] they include a conjunction of various properties called features" (Massaro & Cohen, 1999, p. 22). Massaro and Cohen assume that those *perceptual units* correspond to open syllables. Features can include articulatory movements, voicing or other properties, depending on the language specific phonetic and phonological rules necessary to classify incoming sensory input and to link it to the corresponding prototype. A fuzzy value expresses the extent to which the new sensory information corresponds to each prototype, e.g. the extent to which the sensory input contains the features of each prototype. The value 1, for instance, corresponds to a complete match, while the value 0 corresponds to a complete mismatch. In the second step, the evaluation results of different sensory inputs are integrated. This integration allows preparing the third step, the decision or identification of a prototype. One sensory input contributes to disambiguate the second one, if it provides complementary information. For example, if the auditory information is ambiguous and corresponds with a value of 0.6 to one existing prototype and with a value of 0.5 to another prototype, but the visual input is clearly assignable to the first prototype, a match of value 1, the decision is taken in favor of the prototype one and against prototype two (Massaro & Cohen, 1999).

Cognitive processing of stimuli

The *fuzzy logical model of speech perception* succeeds very well in explaining the interactions of different sensory modalities that have been revealed in physiological and behavioral experiments. Studies using fMRI or ERP have shown that visual enhancement is especially large when the auditory signal is blurred (McGettigan, et al., 2012; Ross, et al., 2007). Behavioral studies draw a similar picture: visual speech benefits particularly in adverse listening conditions. As early as 1954, Sumby and Pollack demonstrated that lip movements improve intelligibility in noisy environments (Sumby & Pollack, 1953). This finding was replicated (Hazan, Kim, & Chen, 2010; Bernstein, Auer, & Takayanagi, 2004; Benoit, Mohamadi, & Kandel, 1994), and extended to hearing-impaired subjects: Massaro and Cohen cite a study conducted by Erber with normal-hearing and hearing-impaired children. The children received auditory-only, visual-only or audio-visual input of a speaker pronouncing syllables and were asked to identify the consonant. Accuracy scores showed that hearing impaired children benefited the most from audio-visual input (Massaro & Cohen, 1999). Later, Massaro and Light (2004) demonstrated that audio-visual speech can be used to improve speech perception and production in individuals whose hearing is affected since their childhood. In a 1994 study, Benoit, Mohamadi and Kandel extended those findings to healthy adults and speech perception in noise. The beneficial effect of audio-visual speech can also be observed in undisturbed listening conditions and holds even when participants observed the speaker talking only during two minutes and perform a speech recognition task in undisturbed auditory-only speech (von Kriegstein, et al., 2008) or if only the oral parts of the face move and all others are artificially frozen with video cutting techniques (Thomas & Jordan, 2004). A similar model could also hold for other sensory modalities. Ernst and Banks (2002) demonstrated that the brain integrates visual and haptic information similarly to maximum likelihood estimation: The more the visual signal was blurred the more the haptic information dominated when participants were asked to judge the size of an object.

Cognitive processing of stimuli

One weakness of the *fuzzy logical model of speech perception* is that it does not assume any interaction with higher cognitive processes, like linguistic predictability or speech context that are known to modulate speech perception (see (Rosenblum, 2008; Spehar, Goebel, & Tye-Murray, 2015). In fact, all of the studies cited above contrasted meaningless syllables like /da/, /ga/ or /pa/. This, of course, does not correspond to natural speech or existing words. Iverson, Bernstein and Auer (1998) have demonstrated that audio-visual enhancement is larger for monosyllabic (English) words than for multisyllabic (English) words. The authors hypothesized that this effect is due to the larger competition among monosyllabic words compared to multisyllabic words. If competition is low, phonetic cues, even in degraded speech, can be sufficient to identify the target word. If competition is high, a visual signal may be required to further reduce perceptual noise (Iverson, Bernstein, & Auer, 1998). This approach could also hold for larger contexts, like whole sentences, that improve predictability and reduce the number of potential lexical candidates. Spehar, Goebel and Tye-Murray (2015) reported better word recognition when sentence context was provided, regardless of the modality (auditory or visual speech). To summarize, research so far suggests that audio-visual enhancement is persistent, but larger if the auditory stream is degraded (noise, hearing impairment). The facilitating effect is possibly reduced when competition among lexical candidates is low.

2.4 Multimodal input in simultaneous interpreting

The previous chapters dealt with multimodal binding and audio-visual speech processing in general. In this chapter, I will apply the theoretical bases to simultaneous interpreting. Particularly, I will take a look at the different sources of visual input in simultaneous interpreting and review the existing research on visual input and its effect on simultaneous interpreting.

2.4.1 Types of visual input in simultaneous interpreting

I imagine an interpreter in a booth in a conference room. What does she see? What kinds of visual information does she perceive? Probably, she will see some kind of documents on the desk in front of her like notes, speech manuscripts or glossaries. Maybe she has opened a web site on her laptop. Next to her, she may see her booth mate gesturing, mouthing or noting something. Looking out of the booth, she may see presentation slides, facial movements and gestures of the speaker, panelists and the listeners, the conference room and maybe even more. Each of those elements provides some information. Lip movements of the speaker enhance his auditory output and contribute to speech perception. Facial, head or hand movements[11] or gaze direction of the speaker can give cues to what the speaker refers to, how he feels about something or if he cedes the turn to another panelist. Facial, hand or head movements of the listeners are in general reactions to what is being said and shows how the listeners perceive the topic or if the speaker still captures their attention. A skeptical expression might indicate that they do not understand the translation; excited whispering might tell the interpreter that something unexpected is going on (technical issues, a fire break out…). The booth mate, in contrast, may not react directly to the speaker's output, but to the translation. As simultaneous interpreting is a highly demanding task, it is of general praxis to help each other during the conference. The booth mate may for example help out with translations or internet research. Visual input of the booth mate can include a simple pointing gesture on the correct translation, or a written note with some indications. Of course, the booth mate might also communicate other information, but for now I assume that she is meta-communicating on the speech and its translation.

[11] Hand movements or gestures can be categorized according to the function (Poyatos, 1984). As gestures are not the focus of this dissertation, I will not elaborate further on this subject.

Cognitive processing of stimuli

Nowadays, speaker often rely on presentations in order to guide through their speech. Presentation slides allow the speaker to show some structuring key words, to depict graphs or images or even to display a movie or a sound file. The speaker could also make use of a board to note some key words during his presentation. During a workshop or training, the speaker could gather the listeners around some kind of permanent installation, for example to explain the different parts of a machine. Documentation is essentially written information, like the speech manuscript, glossaries or related documents. The speech manuscript may help to follow the speaker. It has the potential to compensate for poorly intelligible auditory input, provided the speaker sticks to his manuscript. Glossaries, notes and further related documents provide either the necessary terminology or some background information that makes it easier to understand the speech content. The setting, finally, includes for example the conference room or the clothing of the conference participants. This information is not directly linked to the speech, but it informs the interpreter about the character of the event and the participants: is it a solemn event or a simple work meeting? A scientific conference? Who does attend? Politicians, scientists or collaborators?

Visual input sources can be grouped by different aspects. For instance, it is possible to look at the sources of visual input. Here, we would find for example

1) the speaker: oral, facial and head movements, gestures
2) the listeners: facial and head movements, gestures
3) the booth mate: gestures, head movements, notes
4) visual devices presented to the participants of the conference: presentation slides, notes on a board, permanent installations (machines, models..)
5) documentation: speech manuscripts, glossaries, notes, web sites…
6) setting: conference room, clothing…

Another way to look at visual input is the model of nonverbal communication in interpreting developed by Poyatos (1987; 1997). He described nonverbal and paralinguistic signs according to the information they carry: *space and time markers* (gestures that give indication about space or time), *deictics* that point to the location of the referent, *pictographs* or *kinotographs* that illustrate shapes or movements, *adaptors* that involve other objects or body parts, etc. and emphasized the importance of conveying the information carried by nonverbal or paralinguistic features in the target text.

These classifications have one important drawback: they are not very informative when it comes to the properties of its elements and its susceptibility to multimodal binding. Another approach is thus to check for each type of visual input if it could possibly integrate with the auditory stream of the speaker, the speech. The visual input type would thus have to fulfill the requirements of multimodal binding: temporal, spatial and semantic congruence. Table 1 summarizes the result:

source	**Visual information**	**Temporal synchrony**	**Spatial congruence**	**Semantic congruence**
Speaker	Lip movements	x	x	x
	Facial and head movements	x	x	partially
	Gestures	x	x	partially
Listener	Facial and head movements	x		partially
	gestures	x		partially

Booth mate	Facial and head movements	x		
	Gestures	x		
	Notes	partially		
Visual devices	Presentation slides	partially		x
	Notes on a board	partially		x
	Permanently attached material			X (provided the information is linked to the speech's topic)
Documentation	Speech manuscript			X (provided the speaker does not deviate from his manuscript)
	Glossaries			
	Notes			x
	Web sites			
setting	Conference room			

Table 1: Classification scheme for visual input in simultaneous interpreting

The only types of visual input that fulfill all three requirements and could integrate with the auditory input are the oral, facial, hand and head movements of the speaker. Lip movements have the highest integration potential as they have the highest degree of semantic overlap with the auditory stream. All other visual input sources may provide useful

complementary information, but they probably will not integrate with the speaker's auditory stream. This is why I chose lip movements for the present study (for a description of the experiment, see chapter 4). Interestingly, research on visual input in simultaneous interpreting has essentially focused on the speaker's visual information (see chapter 2.4.2).

2.4.2 Studies on visual information in simultaneous interpreting

A small number of studies have been conducted to elucidate the impact of visual input in simultaneous interpreting. Based on the facilitating effect of audio-visual speech and multimodal binding, it seems reasonable to assume that temporally, spatially and semantically concordant visual information, like the speakers oral, facial, head or hand movements, leads to better interpreting performance. Rennert (2008) opted for a rather direct approach: she asked student interpreters to translate two live speeches and deprived them from any visual information during one speech each. On the whole, she observed no difference in their interpreting performance, except of very few moments where visual input provided necessary complementary information. The author admitted:

> In many instances, visual information was quite redundant, since the information was contained in the verbal message as well. Here it was often difficult to judge the influence of visual input, as the information was conveyed by subjects from both groups. There are several cases where the group with visual contact and the blind booth conveyed information present in both the verbal and the nonverbal material, but it cannot be determined conclusively whether the visual nonverbal information was helpful. (Rennert, 2008, p. 214)

Despite the fact that participants delivered comparable renderings in terms of quality, they expressed a considerable unease when they had to interpret without visual input and rated the speech as being more difficult than when they had visual contact (Rennert, 2008).

Her results are in line with an earlier experiment conducted by Anderson (1994) who found that a video of the conference setting[12] did not improve intelligibility or the informational content of the translations. According to a review by Moser-Mercer (2005a), interpreters suffered from concentration difficulties and fatigue when interpreters where not in the conference room itself, but worked remote with video recordings of the conference room and thus with limited visual input. The subjective perception during remote interpreting stands in contrast to objective measures: in a very comprehensive study, Roziner and Shlesinger (2010) compared subjective ratings and objective measures of booth quality, ergonomic comfort and performance during remote and on-site interpreting. While objective measures revealed no difference between the remote and the on-site condition, subjective ratings were affected by the condition. For instance, interpreters were less satisfied with their performance, complained more often about headaches and drowsiness, and perceived the remote condition in general as being more stressful than the on-site condition.

A candidate for visual input that improves performance might be the written speech manuscript, provided that the speaker does not deviate from his manuscript (Lambert, 2004; but see De Laet & Plas, 2005 for the influence of preparation time on performance during simultaneous interpreting with text). Another visual help might be large numerals written in their Arabic form. In an eye-tracking-study, Seeber (2012) demonstrated that interpreters preferred written numerals to the speaker's face or gestures as soon as large numerals had to be interpreted. To sum up: contrary to what we might expect on the basis of multimodal binding, these studies did not reveal any differences between interpreting with or without visible input. However, they highlight the unease conference interpreters experience when they have to work without or with limited visual input and

[12] No further information about what exactly was visible on the video is provided.

show that interpreters deliberately search for visual information, at least in some cases like large numerals.

These counter-intuitive results might be due to multiple reasons. First, in most studies the sample is very small with a large variability between subjects that could have covered the effects of the independent variables (Anderson, 1994:108). Appropriate statistical techniques that account for this variability may provide a solution. Second, most researchers have given the priority to the ecological validity of their experimental setting. A more natural setting has certainly the advantage to allow drawing conclusions on the real world. But the researcher faces a complex interaction of many confounding factors which are not controlled for and which may influence the performance of the interpreter and thus the results. Simultaneous interpreting is a very complex process and visual input covers a range of different information of varying complexity. That is, while some kinds of visual information might facilitate source language comprehension or interpreting in general, for example lip movements of the speaker, others might require additional resources or processing capacities, even if they provide useful information, like presentation charts or additional written information. Researchers studying simultaneous interpreting need to be very careful in their experimental set-up and control for possible confounds in order to tear apart the effects of the various factors. Third, experiments with interpreters usually use the interpreter's performance, the source text, as dependent variable. This method has two drawbacks: 1) it does not provide a sensitive continuous measure during the whole time of the interpreting process, effects of more difficult passages in a text might therefore obscure the experimental effect; 2) common standards of how performances are to evaluate, are lacking. Consequently, target text evaluations might consider different aspects (intelligibility, information content, use of terminology, intonation, etc.) or use different methods (source text analysis, expert judgements, scales and subjective ratings) and therefore the studies might be not comparable.

Apart from these methodological issues, effects of visual input might be absent on a semantic or syntactical level because interpreters increase their cognitive effort to maintain interpreting quality even in adverse conditions. If this is the case, effects would either be visible at a more fine-grained level, for example in the richness of their vocabulary or in effective speech monitoring, or under higher cognitive load, for example when working in noisy conditions. To summarize, the main problems of research on visual input in simultaneous interpreting are insensitive methods to measure the impact of lacking visual input, confounding factors due to ecologically valid, but uncontrolled experimental settings and finally, inappropriate methods that do not take into account that interpreters can adjust their effort to maintain the overall quality. This leads to the next chapter on cognitive load and its effect on simultaneous interpreting.

3 Cognitive load and mental effort

In this chapter I take a closer look at the notion of cognitive load or mental load. The notion of cognitive load is often associated to working memory and the number of items that need to be maintained or manipulated concurrently for a given task. If I want to multiply two numbers, for example, I need to hold in my memory the two numbers while carrying out the necessary calculations. The task gets more and more difficult as I add more numbers: Multiplying three numbers requires the working memory to maintain the three numbers and the multiplication result of the first two numbers while multiplying with the third number.

As cognitive load is intrinsically linked to the working memory, I will first turn to models of working memory developed by Alan Baddeley and Nelson Cowan and look how these models can account for multimodal binding and its facilitation effect. In the second section, I will review different models and concepts of cognitive load and their application in experimental settings. The third section will cover different possibilities of measuring cognitive load and discuss what these methods tell us about

cognitive load. Finally, I will turn to simultaneous interpreting and how cognitive load has been approached so far in simultaneous interpreting.

3.1 Models of working memory

The notion of *working memory* became widely known in the 1970's when Baddeley and Hitch (1974) introduced a model of the memory that is not only supposed to hold information, but also to manipulate and process information and is seen as an "interface between perception, long-term memory and action" (Baddeley, 2003, p. 829). The working memory is described as having limited attentional capacity. Furthermore, Baddeley and Hitch assumed that all processes taking place in the working memory require attention. In its original form, their model consisted of a *phonological loop*, a *visual-spatial sketchpad* and a *central executive*. Later on, Baddeley added a fourth component, the *episodic buffer* (Baddeley, 1992; 2003).

The *phonological loop* encodes auditory and verbal information. It is important to note that the phonological loop is not limited to spoken words or sounds, but processes all kind of verbal information. Written words are transformed into their auditory form by articulating them silently. A similar process could take place for lip movements where the motoric information is transposed to auditory information. Auditory and verbal information is hold in a short-term buffer until the memory trace begins to decay after a couple of seconds. Baddeley and Hitch observed that subjects were able to expand the short-term buffer provided that they were allowed to repeat the items they were to recall. It also appeared that longer words were more difficult to recall than shorter ones. Based on these observations, they added a rehearsal component, the *articulatory loop* that repeats an item either by saying it aloud or by articulating it silently to prevent its memory trace from fading. One main reason for adding this component was the effect of articulatory suppression. Repeating aloud meaningless sounds blocks the "inner voice" and leaves the memory traces to decay (Baddeley, 2003). This also means that auditory input that enters the

phonological loop during simultaneous interpreting is not refreshed and fades after a couple of seconds. This is particularly problematic for items that induce a high memory load and cannot be derived from context, like large numbers (Gieshoff, 2012).

The visual-spatial sketchpad is thought to process visual and spatial information like color and shape of an object, location of an object and spatial motor planning. Analogous to the phonological loop, it is divided in two subcomponents: a visual short-term buffer holding information on shape and color and an inner scribe rehearsing visual and spatial information. Given the distinct visual pathways for the processing of information on “what” and “where”, it seems reasonable to assume distinct processes for visual and spatial information in working memory (Baddeley, 2003). Baddeley and Hitch did not include a short-term buffer for olfactory or tactile information, but did not exclude it either. In fact, they noted that the existence of further modality-specific slave systems is probable, but had not been confirmed to the date of publication (Baddeley, 2003).

In the beginning, the central executive was seen as a “pool of processing capacity” (Baddeley, 2003, p. 835). But this simplistic picture evolved when the importance of attentional control became apparent. Baddeley attributed some further tasks to the central executive: allocating, switching and dividing attention. Moreover, the central executive was thought to provide a link to long-term memory in order to enable the cooperation between the two memory systems. But the central executive alone was not sufficient to explain recall of sentences or stories that exceeded the storage capacity of the phonological loop. Baddeley assumed that grammatical information did not need to be stored in the phonological loop as it was provided by the long-term memory. Moreover, he saw the need to allow for an abstract representation in the working memory that is not bound to a specific modality. So Baddeley added a fourth component, the episodic buffer, a limited-capacity store that allows elements to interact with each other and to form larger chunks (Baddeley, 2003). In other

words, the episodic buffer allows us to connect words we hear to ideas and to relate these ideas to one another in order to achieve a full understanding of a text. In turn, if the working memory capacity is exceeded, text understanding may be shallow (for the role of working memory in simultaneous interpreting, see also Timarová, 2008).

When Baddeley and Hitch developed their multicomponent model of a working memory, they had a good reason to assume different modality-specific stores. In fact, it appeared at that time that memory tasks involving one modality did not impair memory tasks involving the other one (Baddeley, 1992). Subsequent research, however, revealed interferences between the auditory and the visual modality. For example, in a letter recall task, letters that sound similar are more prone to confusion than very distinct letters, even if the letters are presented visually (Cowan, 2009; 2000). Based on this finding, Cowan (2010; 2000) concluded that the need for modality-specific stores was scientifically not established and proposed an alternative model for the working memory. Instead of assuming modality-specific stores, he suggested that working memory is basically an activated part of memory that provides all items that are necessary for a given task: new sensory information, activated elements from the long-term memory or refreshed memory items. This *activated memory* is not limited in capacity, but in time with memory traces decaying after ten to thirty seconds (Cowan, 2010).

In addition to the activated memory, he postulated a *focus of attention*. It is capacity-limited, though not time-limited. Only three to four items[13], freshly incoming elements or re-activated items from the long-term memory, can

[13] The capacity of Cowan's focus of attention is much smaller than the "magical number seven" postulated by Miller in 1956 (Miller G. , 1956). The reason is probably a conceptual difference. Cowan's aim was to find out the number of items that are actually processed. This means that he considered only those items that receive attention, unlike Miller who used list recall to investigate memory span.

enter the focus of attention where they are processed, manipulated or combined to form larger chunks. Rehearsal refreshes the memory traces and prevents these elements from decaying. Like Baddeley's episodic buffer, the focus of attention enables us to relate words to form ideas and to put these into the text context. The third component of Cowan's working memory model is a central executive whose task is to search the long-term memory and to provide suitable items. Like in Baddeley's model, Cowan's central executive links long-term memory and working memory and plays a role in distributing attention by activating salient items (Cowan, 2009; 2010; 2004), for the implications of Cowan's model in simultaneous interpreting see Cowan, 2000/01).

Both models share some important assumptions:

1) Elements and processes taking place in working memory receive attention.
2) There is some sort of central buffer, an episodic buffer (Baddeley, 2003) or a focus of attention (Cowan, 2009) that serves to integrate or modulate memory items and that could (at least in theory) be involved in binding of semantically congruent stimuli from different modalities.
3) Working memory has a limited capacity, e.g. only a small number of items can be rehearsed, manipulated or combined.
4) Memory traces decay after a couple of seconds.
5) Central executive processes link working and long-term memory and coordinate the distribution of attention.

The main difference consists in the assumption of modality-specific stores. Baddeley proposed a phonological-verbal loop that transforms even visually presented verbal material into its auditory form, and a visual-spatial sketchpad with each a rehearsal function. This helped to explain why verbal (for example: repeating or reading a sentence) and spatial tasks (for example: tracking some stimulus with the computer mouse) do not interfere, while verbal material presented in one modality could affect

the other modality. For instance, printed letters which could be expected to strain only the visual-spatial sketchpad are more easily confused with letters that sound similarly than with letters that sound completely differently. On the contrary, Cowan proposed a modality-independent activated memory where elements with similar physical properties could interact to explain how interference between different modalities could occur. This account is much simpler, but encounters some difficulties in explaining why some tasks do not interfere while others do. Task interference (or the absence of interference) is also a key notion for Wickens (see chapter 3.2.3). Both models, however, agree on the fact that working memory is capacity-limited and that processes that take place in the working memory require attention. This understanding of working memory is crucial for the concept of cognitive load.

3.2 Models of cognitive load

The notion of cognitive load has been addressed in a number of ways and researchers have used different expressions to refer to working memory load: cognitive load or mental work-load, task-load, mental effort and others. Cognitive load, mental effort or working memory load, attentional capacities or resources are generally seen as being closely linked to each other.

3.2.1 Kahnemann's capacity model of attention

One of the first researchers to propose a distinction between these terms was Kahnemann. In his essay *Attention and effort* (1973), Kahnemann described *load* as being induced by the task, while *effort* reflects the amount of capacities that subjects use to solve a task. He further defined *arousal* as a physiological manifestation, such as pupil dilation, of task demands:

> "[…] physiological arousal varies second by second when a subject is engaged in a task, and [that] these variations correspond to momentary changes in the demands imposed by the task. Thus, arousal and effort

> are usually not determined prior to the action: they vary continuously, depending on the load which is imposed by what one does at any instant of time." (Kahnemann, 1973, p. 14)

Based on this distinction and numerous psychological studies, he proposed a model to describe how attentional capacities are distributed. He assumed that capacities are limited and that the amount of available capacity depends on the level of arousal. When the level of arousal is high, more capacities are available than when the level of arousal is low. He further stated that arousal and – indirectly - attentional capacity are a function of external factors. Increasing task difficulty (cognitive load) will lead to higher arousal and subjects will invest more capacities (mental effort) to solve the task. With increasing practice, the task can be done more efficiently and mental effort will decline. Inversely, mental effort will remain low if the task is simple (Kahnemann, 1973).

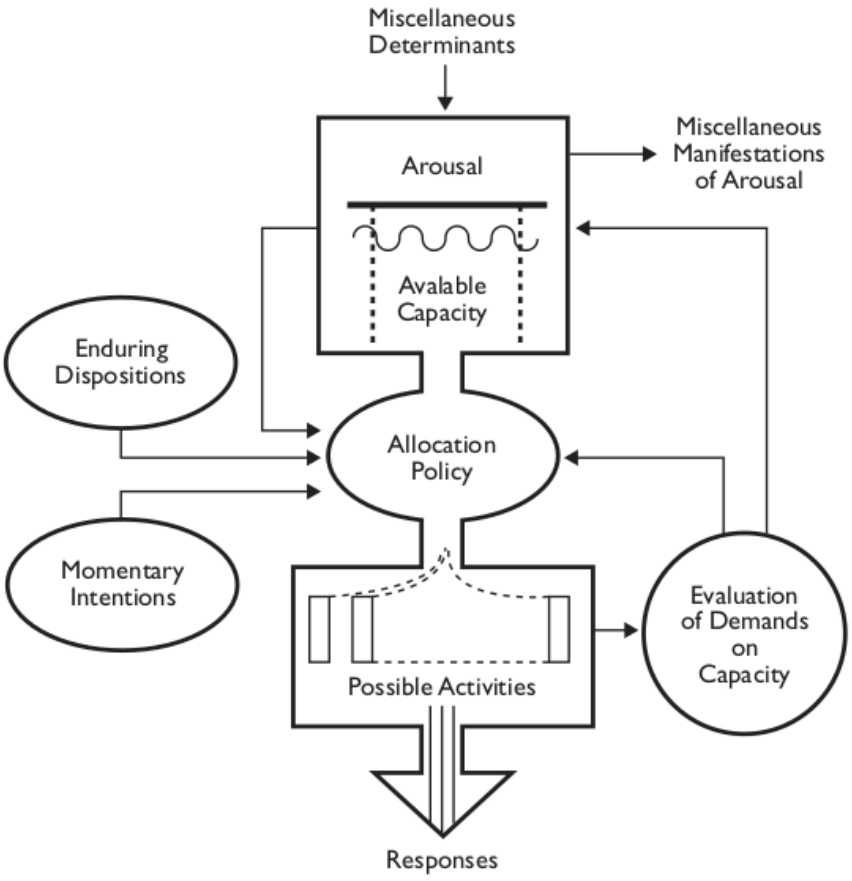

Figure 1: A capacity model of attention (Kahnemann, 1973, p. 10)

Cognitive load and mental effort

Attentional capacities are distributed according to the *allocation policy*. The allocation policy depends on several factors, like momentary intentions and goals, physiological features of the stimuli, their novelty and saliency (enduring dispositions of involuntary attention allocation), evolution of task demands and increase of arousal. It is important to note that arousal does not only depend on task demands, but also on a range of different states like fear, anger, anxiety, pain and other, that cannot be possibly controlled for in an experimental setting. As long as sufficient capacities are available, subjects can engage in several tasks concurrently. If the available capacities do not entirely cover task demands, task performance will suffer (Kahnemann, 1973).

At the time of its publication, the capacity model of attention was ground-breaking. The novelty of his model compared to former models like the bottleneck theory by Broadbent or the attenuation theory by Treisman (see (Anderson J. R., 2007, pp. 94-98) that were popular at that time, was the assumption of a pool of limited capacities. While Broadbent assumed that stimuli pass a filter before being processed, regardless of their meaning (Anderson J. R., 2007, pp. 94-96), Kahnemann proposed a model where stimuli receive attention depending on whether sufficient resources are available and whether the stimuli corresponds to the task goals. With his model he was able to explain a large part of the psychological findings at that time and prepared the grounds for capacity-limited working memory models. However, his model encounters some difficulties to explain why some tasks seem to suffer from stronger interferences than others. For instance, participants found it difficult to correctly recall digits pairs that were presented simultaneously on each ear, while they succeeded much better to recall digit pairs that were presented to the ear and the eye (Broadbent, 1956).

With regard to simultaneous interpreting, the notion of arousal and the link between arousal and capacities are highly interesting because it brings in more emotional aspects like stress. We can safely assume that

interpreters are very eager to deliver a good translation and that arousal is generally high. Maybe arousal is higher in younger and less experienced interpreters than in more experienced ones that got used to the situation. But what happens when the task of interpreting gets more complex, for example because parts of the speaker's auditory input become unintelligible, because the source text is particular dense or because the speech delivery speeds up? First of all, the task of interpreting as a whole becomes more difficult. The interpreter needs to make more effort to adapt to the new task demands. Additionally, the interpreter could feel more stressed when she has difficulties to follow the speaker which in turn leads to higher arousal. According to Kahnemann, the mental effort would rise up to the point where the task exceeds the interpreter's capacities. At this point, the target text quality will suffer from errors, omissions or stylistic imperfections.

3.2.2 Sweller's cognitive load theory

Maybe the expression cognitive load most commonly refers to the *cognitive load theory* by John Sweller. In the 1980's, Sweller wondered how learning takes place, how it could be facilitated and how learning material needs to be designed to make learning as easy as possible, "easy" meaning here to reduce as much as possible cognitive load. To start, he described learning as schemata building, which is just another expression for integrating information to form a new concept or a new procedure and to store it in the long-term memory. These schemata can be accessed and implemented automatically when needed (Sweller, 2010; van Merrienboer & Sweller, 2005). This happens for example when we learn how to play an instrument. At the beginning, the note and the corresponding fingering are two distinct elements. During practicing, they form one concept so that the fingering is automatically accessed when we see the note.

Sweller distinguishes three types of cognitive load, *intrinsic, extraneous and germane cognitive load*. Intrinsic cognitive load denotes the

interactivity of all elements that need to be processed simultaneously in order to build a schema. In the above example, we would need to hold in mind two elements, the fingering and the corresponding musical symbol. If all elements can be learned independently of each other, like learning musical notes, their interactivity is low and so intrinsic load is low. If the number of elements that need to be maintained simultaneously exceeds working memory capacity, like playing a tune which requires to process simultaneously notes, fingerings, rhythm and maybe further musical indications, the task is too difficult and the learning goal is not achieved. In this case, the task has to be chunked in easier steps in order to enable the learner to form schemata that lowers the memory burden and can be accessed automatically once he returns to the initial task. Intrinsic load depends thus on the learner's expertise: the more expertise increases, the more intrinsic load decreases (Sweller, 2010; Schnotz & Kürschner, 2007). In simultaneous interpreting, element interactivity is – given the simultaneous nature of the task – very high. Source text segments, target text segments, lexical candidates, visually presented information in form of charts or documents, text context and other associated information have to be processed in order to understand the source text, translate it into the target language and verify that the translation is correct.

Extraneous load is linked to the way learning material and instructions are presented. Extraneous load does not depend on element interactivity. On the contrary: it is cognitive load that can be removed without reducing element interactivity and without altering the learning goal. For example, redundant information increases the number of elements to process without increasing element interactivity (Chandler & Sweller, 1991; Schnotz & Kürschner, 2007). Of course, extraneous load depends highly on the learning goal that is specified (Sweller, 2010). If the learning goal is to play just a tune, dynamic indications (like forte or piano) printed in the scores would increase the extraneous load, because they are additional elements that are not necessary to achieve the learning goal. If, in contrast, the goal is to learn the dynamic indications and how to play a

tune louder or softer, then dynamic indications contribute to element interactivity and to intrinsic load. In simultaneous interpreting, extraneous load can be triggered by complicated technical installations (microphone and headset) or superfluous information sources that need to be processed at the same time, like a speech manuscript that is not followed by the speaker.

A somewhat special kind of cognitive load is germane load. Unlike intrinsic and extraneous load that are related to task performance because they depend on the number of elements to maintain in the working memory in order to accomplish a task (intrinsic load) or to cope with adverse conditions (extraneous load), germane load denotes all the working memory resources that are allocated to learning and schemata building. Examples are applying problem solving strategies or deriving patterns in order to facilitate problem-solving. At the same time, germane load is often correlated with intrinsic load because intrinsic load forces the learner to deepen his understanding and acquire new schemata. If extraneous load is high, the learner will use too much of his resources to handle extraneous load and consequently, the learning effect will be small. If extraneous load is low and element interactivity is high, the learner will be able to use nearly his entire resources to build up new schemata. Germane load also depends on mental effort, meaning that the learner mobilizes all of his resources for the task he faces (Sweller, 2010; Schnotz & Kürschner, 2007; van Merrienboer & Sweller, 2005). However, as de Jong points out it is difficult to distinguish these two concepts:

> The implications of the definitions of load and effort from Merriam-Webster's on-line dictionary are very straightforward: load is something experienced, whereas effort is something exerted. Following this approach, one might say that intrinsic and extraneous cognitive load concern cognitive activities that must unavoidably be performed, so they fall under cognitive load; germane cognitive load is the space that is left over that the learner can decide how to use, so this can be labelled as cognitive effort. (de Jong, 2010, p. 113)

According to Sweller (2010), it is not possible to distinguish these different kinds of cognitive loads with physiological or behavioral methods. The only possibility to disentangle the different types of loads is to analyze the task beforehand and to manipulate only one type of load and to check, whether there are any changes at a physiological or behavioral level (see also (Schnotz & Kürschner, 2007; Debue & van de Leemput, 2014; de Jong, 2010).

An important finding in regard to audio-visual speech is the redundancy effect. It predicts higher extraneous load if redundant elements are present because the learner would have to process additional elements that do not contribute to increase element interactivity. On the other hand, if two elements are complementary and need to be processed simultaneously for understanding, audio-visual presentation can reduce extraneous load (modality effect), because working memory can be used more efficiently (Sweller, 2010; Mousavi & Low, 1995). Moreno and colleagues (2002) reported facilitated learning when the text was simultaneously presented aurally and on screen rather than in an auditory-only condition (Moreno & Mayer, 2002). In the same line, Kalyuga (2012) reports in her review on cognitive load factors findings that indicate a facilitating effect on learning when written text accompanies the (same) spoken text, especially when the auditory input is lengthy or not well perceived (Kalyuga, 2012). One possible explanation for these contradicting findings is the experimental material. When Sweller (2010) refers to the redundancy effect, he considers graphs and accompanying text that is not necessary in order to understand the graph. Moreno and Mayer (2002), in contrast, used the same text in its spoken and written form. As described in chapter 2.3.5, audio-visual speech facilitates listening comprehension. This could also hold for speech that is simultaneously presented in a spoken and written form. Nevertheless, this example demonstrates that predictions from the cognitive load theory are not always straight forward.

3.2.3 Wickens' model of task interference

While Sweller looked at cognitive load from a learning perspective, Wickens based his model on the efficiency of dual-tasking. He analyzed about fifty studies on dual-tasking to find out when task-switching was effective and when different tasks interfered. The result was a three-dimension-model according to which tasks interfere if they use the same resources on any of the three dimensions: *stages of processing*, *codes of processing* and *modality*. *Stages of processing* refers to perceptual-cognitive actions, e.g. actions that take place in the working memory, as opposed to motor actions. An example would be to recall a list while tapping with a finger. Both actions are not expected to interfere because list recall involves perceptual-cognitive resources, while tapping with a finger needs essentially motor resources. *Codes of processing* reflects the phonological and visual components of Baddeley's working memory model and means that spatial and verbal-linguistic tasks exploit different resources. For instance, a visual tracking track would not be expected to impede a reading span task. Finally, *modality* denotes the sensory modality and indicates that auditory and visual perception do not share the same resources. Responding to single tones of a defined frequency should not encumber a visual search task (Wickens C. , 2009).

Based on these reflections, Wickens developed a computational model capable to predict mental work-load. He began by stating that every task generates some load according to the task complexity. An automated task has load factor 0, while a very complex task reaches the load factor 4. Two tasks overlapping in any of the three dimensions and competing for the same resources generate an additional conflict load. This conflict load is obtained by adding the individual load factors of the two tasks for each dimension that is shared by both tasks. The conflict load can thus range from 0 to 8 depending on the load factor for the completion of each task (Wickens C. , 2009). I will illustrate the computations with the following tasks: reading a text, answering questions, visual tracking. For visual

tracking, we assume a load factor of 1, for answering questions a load factor of 3 and for reading a text a load factor of 2. Now I look at each combination in turn: reading a text (load factor 2) and answering questions (load factor 3) are both perceptual-cognitive and verbal tasks, but differ in modality. For each dimension where both tasks compete for resources (the stage of processing and the code of processing), we add the sum of both load factors to obtain the conflict load factor (2+3=5). In order to obtain the overall load, we sum up the load factors for task completion (load factor 2 for reading the text and load factor 3 for answering questions) and both conflict load factors: 2+3+(2+3)+(2+3)=15. Both tasks combined would thus generate a load factor of 15. Answering questions (load factor 3) and tracking an item (load factor 1) are both perceptual-cognitive tasks, but differ in code and modality. We add thus once the load factor of each task (1 for tracking an item and 3 for answering questions) and the conflict load factor (1+3=4), and obtain the overall load: 1+3+(1+3)=8. Based on these calculations, I would expect that time-sharing is better between answering questions and visual tracking than between answering questions and reading a text.

Wicken's model is consistent with psychological research suggesting that two concurrent task can - to some extent and depending on the nature of the tasks in question - be accomplished (more or less) successfully (see for example (Broadbent, 1956; Horrey, Lesch, Garabet, Simmons, & Maikala, 2017). It is successful in explaining the high work-load in simultaneous interpreting where most processes that take place simultaneously, are verbal tasks and thus interfere on at least one modality. Indeed, Seeber has applied Wickens' model to simultaneous interpreting (see chapter 3.3.2). It encounters, however, difficulties in predicting facilitation effects. A cognitive load model focusing solely on task interference can predict an increase or the absence of cognitive load, but not a diminishment of cognitive load. It can explain, under which circumstances several concurrent tasks may disrupt each other, but it cannot tell us how information from different modalities complete each

other and contribute to achieve a certain goal. In this respect, it is incompatible with findings demonstrating a facilitation effect of lip movements for listening comprehension as described in chapter 2.3.5.

3.2.4 Lavie's load theory of selective attention and cognitive control

Lavie and colleagues early and the late selection: perceptual and cognitive attention (2004) addressed cognitive load from the perspective of attentional control and distractor interference. They defined attention as the selection of an item and opposed two mechanisms: Either an element is selected during perception, or it is selected during processing and response selection. The first option is supported by the finding that unattended information generally goes unnoticed. A famous example is an experiment conducted by Simon and Chabris (1999). The authors asked the participants to watch a basketball match on screen and to count the passes of the teams. Only half of the participants noticed the black gorilla crossing the screen (Simons & Chabris, 1999). The second option, the late selection, receives support from the finding that processing and response selection slow down if a distractor is present. Numerous examples are provided by paradigms using stroop or flanker tasks (Stroop, 1935; Eriksen & Eriksen, 1974). In psycholinguistic research, similar competition effects can be observed between two different languages. For instance, response times in picture naming in the second language (L2) slow down if a distractor word in the native language (L1) is presented that is phonetically similar (but semantically unrelated) to the picture's name in the L2 (Costa, Colomé, Gómez, & Sebastián-Gallés, 2003). Based on this debate, Lavie and colleagues postulated two mechanisms of selective attention that could explain both the early and the late selection: perceptual and cognitive attention (Lavie, Hirst, de Fockert, & Viding, 2004).

Elements are perceived automatically, but the attentional capacities are limited. As soon as the number of elements exceeds the attentional capacities, as it is the case in high perceptual load, they go unnoticed. If

perceptual load is low, all elements are processed and cognitive control is necessary to reject distractors. In a cognitive task with high load, fewer capacities are left to cope with distractors and distractor interference will increase. By manipulating both perceptual and cognitive load in a combined recall and perceptual distractor task (Eriksen flanker task), Lavie and colleagues were able to demonstrate the effect of working memory load on reaction times and accuracy in the perceptual task. Larger set sizes in the perceptual task, by contrast, decreased the distractor effect (Lavie, Hirst, de Fockert, & Viding, 2004).

In simultaneous interpreting, the conference interpreter has to maintain and process several information units simultaneously, which strains the working memory. As lip movements correspond to the auditory input, they are not expected to have a distractor effect. Still, they could increase perceptual load because more information needs to be processed. The additional perceptual load would drain attentional resources to the sensory input and leave fewer resources for cognitive processes. The notion of perceptual load could explain the anecdotal observation why some interpreters shut their eyes when they need to process “effortful” information like high-order numerals: They want to concentrate all their resources on the cognitive processing without “losing” attentional capacities on secondary perceptual information. Adding background noise to the auditory input, on the other hand, does not necessarily need to increase perceptual load. As noted in chapter 2.2.3, white noise makes auditory stimuli partially unintelligible because the different auditory streams interfere with each other (“energetic masking”). However, as will be discussed in more detail in chapter 3.3.4, white noise does not contain information that needs processing and may therefore not alter perceptual load but rather cognitive load because interpreters will need to rely on other cues like speech context to restore the missing information.

3.2.5 Barrouillet's time-based resource sharing model

Barrouillet and colleagues focused on the notion of time. They described cognitive load as the time during which an item receives attention. This theory is based on four assumptions:

1) Processing and maintaining items require attention. This statement is in line with working memory models that assume rehearsal mechanisms (Baddeley and Hitch, 1974) or some sort of activation to retrieve elements from long-term memory or to push them into the focus of attention (Cowan, 2010).
2) Memory traces decay if they are not refreshed (and refreshing requires attention, see 1). This also corresponds to working memory models. Especially the phonological loop and the effect of suppression have been extensively investigated. Based on recall performance under articulatory suppression, researchers estimate that memory traces fade within two to three seconds (Gathercole & Martin, 1996).
3) Items in the focus of attention receive activation, but begin to decay as soon as attention is switched away. This assumption builds on the first two ones.
4) Attention is a bottleneck: only one item at the time can capture attention. Attention sharing is thus rapid attention switching. This is opposed to Cowan (Cowan, 2009) who assumes that four elements can be processed and manipulated simultaneously.

If only one item at a time captures attention, processing of that item impedes all other activities. Long processing times will therefore result in a higher cognitive load. Cognitive load can thus be summarized as the following:

$$CL=a*N/T$$

where *N* is the number of retrievals that are required, *a* is the time during which an item captures attention and *T* is the total time of the task. The more "attention time" is available for processing and retrieval, the lower is

the cognitive load. If the duration during which an item receives attention is restricted or insufficient to ensure proper processing, performance suffers. “Attention time” is thus a function of the processing time of an item, of the number of retrievals and the total amount of time available to solve the task (Barrouillet, Bernadin, & Portrat, 2007).

Barrouillet and colleagues tested their model 2004 with two simple concurrent tasks: they asked participants to read digits and to memorize letters at the same time. The number of digits and the total amount of time were manipulated so that participants could allocate more or less time to refresh the letters. The results confirmed their model: the more digits were presented, the more recall suffered. A similar observation was made when manipulating the total amount of time: In the fast condition, recall was poorer than in the slow condition. According to the authors, the attention time available to refresh the letters is restricted in both conditions: in the first one, the longer reading task retains attention during a longer time; in the second one, the restriction of the total amount of time limits the time during which both tasks (reading and memorizing) receive attention (Barrouillet, Bernadin, & Camos, 2004).

In its simplicity, the time-based model is appealing. Moreover, time pressure is assumed to increase cognitive load. For instance, the NASA task-load index (Hart & Staveland, 1988) includes task pace as task-load dimension. Nevertheless, the model has some drawbacks. First of all, an attention switching policy is lacking. It cannot make any predictions about which task will suffer in a multitasking paradigm if processing of two tasks is too time-demanding to be accomplished within the total amount of time that is allocated. In the absence of an attention-switching policy, I have to assume that stimuli are processed in the order of appearance. As soon one item needs too much processing time, any other operation comes to halt. Second, Barrouillet tested his model with very simple tasks. Demonstrations with more complex tasks are lacking to date. If I carry on

the reasoning of the author, more complex tasks would simply imply an even more rapid switching between the items that need to be processed.

Given the high number of items and the time restriction, the time-based resource model would generally predict a high cognitive load during simultaneous interpreting and an even higher load when the speaker speaks faster or when the speech gets particular dense. If we add background noise and make the auditory stream more difficult to understand, processing of the auditory stream would slow down. Consequently, there would be less time available for other processes like target speech formulation or monitoring and performance would suffer. Audio-visual speech, in contrast, is known to speed up speech processing (see chapter 2.3.5). As audio-visual speech would demand less processing time, more time would be available for target speech formulation and monitoring and performance should benefit.

3.2.6 Cognitive load versus mental effort

The above examples reflect very different approaches to the concepts of cognitive load and mental effort that are essentially determined by the paradigms psychologists used in their studies. But they show that it is important to distinguish two concepts. The first one is the concept of *cognitive load*. It describes the amount of working memory capacities that a task requires or to what extent a task strains the attentional resources. Cognitive load is closely linked to the task and can be experimentally controlled. As outlined above, cognitive load results basically from the interactivity between the elements that need to be hold in working memory concurrently. Basically, all elements in relation to a task are processed. This means, too, that those elements, which are presented with the task but are not necessary to accomplish the task, impose an unnecessary load and leave fewer resources to meet the task requirements (extraneous load). Element interactivity can lead to interferences when elements share for example the sensory modality or the code of processing (visual-spatial vs auditory-verbal). Cognitive load is generally assumed to “sum up” in a

broader sense: more complex tasks or several concurrent tasks increase overall cognitive load, especially if task components interfere. Finally, cognitive overload leads to incapacity to reject distractor items, either at a perceptual or at a cognitive level.

The second one is the concept of *mental effort*. *Mental effort* – or *germane load* - is the amount of resources that are invested to solve a task or to learn a new concept. It is linked to task difficulty to the extent that mental effort generally increases with increasing task requirements up to the individual level of saturation, but it depends also on the state of arousal. Arousal is indicated by physiological reactions, like pupil dilation, and some researchers use these physiological reactions to determine the cognitive load experienced during a task. However, it is important to bear in mind that the link to cognitive load is only an indirect one and that arousal varies between subjects or even within subjects. Someone who is tired or bored will invest less effort than someone who is highly motivated, for example because she has been promised a reward for accomplishing the task. Likewise, there can be substantial differences within one participant at the beginning and the end of an experimental session.

At the heart of these two concepts are the notions of working memory and attention. Working memory processes and integrates elements that are – to borrow the words by Nelson Cowan – “in the focus of attention” (Cowan, 2000, p. 91). This holds also for Baddeley’s working memory model: the fact that phonological traces decay without rehearsal implicitly states that attention is needed to prevent the traces from fading. The more elements need to be combined, the more the working memory is strained and the more attentional resources are captured by the task. If two tasks compete for attention, for example multiplying double-digit numbers and pressing a button in response to some target stimulus, one task will suffer (more errors or longer reaction times) because the working memory and the attentional resources are limited in capacity. If a conference interpreter encounters difficulties in translating a technical term, she will maybe miss

the current source text segment because her attention has been captured by searching in her memory for the correct translation. Cognitive load and attention are thus tightly linked. If we want to determine cognitive load in simultaneous interpreting, it is crucial to determine which processing steps are automated and which components require attention.

3.3 Mental effort and cognitive load in simultaneous interpreting

Cognitive load has been a major focus in interpreting science ever since researchers have sought to understand how the brain deals with the simultaneity of listening and speaking. Researchers like Gerver (1975), Moser (1978) or Setton (1999) tried to decompose the interpreting process in sub-tasks in order to explain why simultaneous interpreting requires so much concentration. Others, like Gile (2009) and Seeber and Kerzel (2012), focused especially on the mental effort or cognitive load that is involved in different components of the interpreting process. As the latter two models are of special interest for the cognitive load or mental effort interpreters experience during their work they will be presented in the following sections.

3.3.1 Gile's effort model

One of the most influential models of simultaneous interpreting (though initially only developed for purely didactical purposes) is Gile's *effort model* (2009). Gile divides simultaneous interpreting in four *efforts*: the perception and analysis of the source text, production of the target text, and memory storage and retrieval. A fourth effort, the *coordination effort*, divides the attentional resources between the first three efforts. The hypotheses underlying his model state that a) each effort comprises processing steps that require attention, b) the different efforts "sum up" and c) interpreters work at their limit (*tightrope hypothesis*). If the overall effort exceeds the interpreter's resources or if the attentional resources are misbalanced, the interpreting quality suffers. Gile stresses that

because of the ear-voice span the cognitive load imported from the preceding sentence can affect the current sentence and lead to omissions and errors even if the current sentence does not contain any problem triggers. By the same token, the previous sentence can help to anticipate the following segment and thereby facilitate its processing (Gile, 2008).

Albeit Gile's effort model for simultaneous interpreting is – as he says himself - a "conceptual framework" (Gile, 2008, p. 60) that is difficult to falsify as it makes no predictions in the proper sense, researchers (and Gile himself) have conducted numerous studies to test the hypotheses underlying the effort model, and particularly the tightrope hypothesis. These studies focus mainly on problem triggers like numerals or names (Gile, 1985; Gieshoff, 2012), high delivery speed (Gerver, 2002; Lee, 2002) or accent (Sabatini, 2000; Mazzetti, 1999), and seem overall to confirm his hypothesis. Very few studies are dedicated to the effect of visual information in simultaneous interpreting. In fact, his model involves visual information only to the extent to which is facilitates or hampers source text comprehension but it does not allow to distinguish between different types of visual information.

It is interesting to note that, like Kahnemann, Gile speaks about an *effort* the interpreter needs to make in order to comprehend the source text, to retrieve information from memory or to produce the target text. He thereby adopts Kahnemann's terminology and implicitly his assumptions that the interpreter adapts her mental effort to the cognitive load that the task imposes (as long as her capacities are not exceeded and she is willing to give a high-quality rendition of the source text). In this light, the tightrope hypothesis takes on greater significance because it means that interpreters need to make use of strategies in order to cope with the task demands.

3.3.2 Seeber's model of cognitive load

Seeber developed 2011 a model of cognitive load in simultaneous interpreting based on Wickens' model of task interference (see chapter

3.2.3). The general idea which underlies his model is as follows: the task complexity of simultaneous interpreting resides not only in the fact that several processes (speech comprehension, speech production, memory storage or others) take place concurrently, but that they interfere and create thus an additional load. He decomposed the interpreting process in four sub-components: storage, perceptual auditory-verbal processing, cognitive-verbal processing of input and output and verbal response processing of output and assigned the load factor 1 to each component. For each combination of concurrent processes, he calculated the interference load. Then he identified for each moment during the interpreting process the processing components and the corresponding interference and calculated the global cognitive load at this particular moment (Seeber & Kerzel, 2012; Seeber, 2011).

Figure 2 illustrates see how cognitive load for the conference interpreter changes as the speaker pronounces his sentences “Wir glauben, die Delegierten treffen ihre Entscheidung nach einer langen Debatte” (English translation: “We believe that the delegates take their decision after a long debate”). In this example, the conference interpreter listens during the first two words “Wir glauben”. Consequently, the processing components involve storage of the source text segment, perceptual processing of the source text segment and cognitive processing of the source text segment. From the segment “die Delegierten” on, the interpreter starts her translation. From that moment on, the local cognitive load corresponds to the sum of the load induced by the storage of the first segment until this one is delivered, the storage the second segment, the perceptual auditory processing of the second source text segment and the first target text segment, the cognitive verbal processing of the second source text segment and the first target text segment, the verbal response processing of the first target text segment and the interference load factors. The amount of cognitive load remains more or less stable until the end of the source text sentence where cognitive load decreases while the interpreter finishes her rendering of the source text.

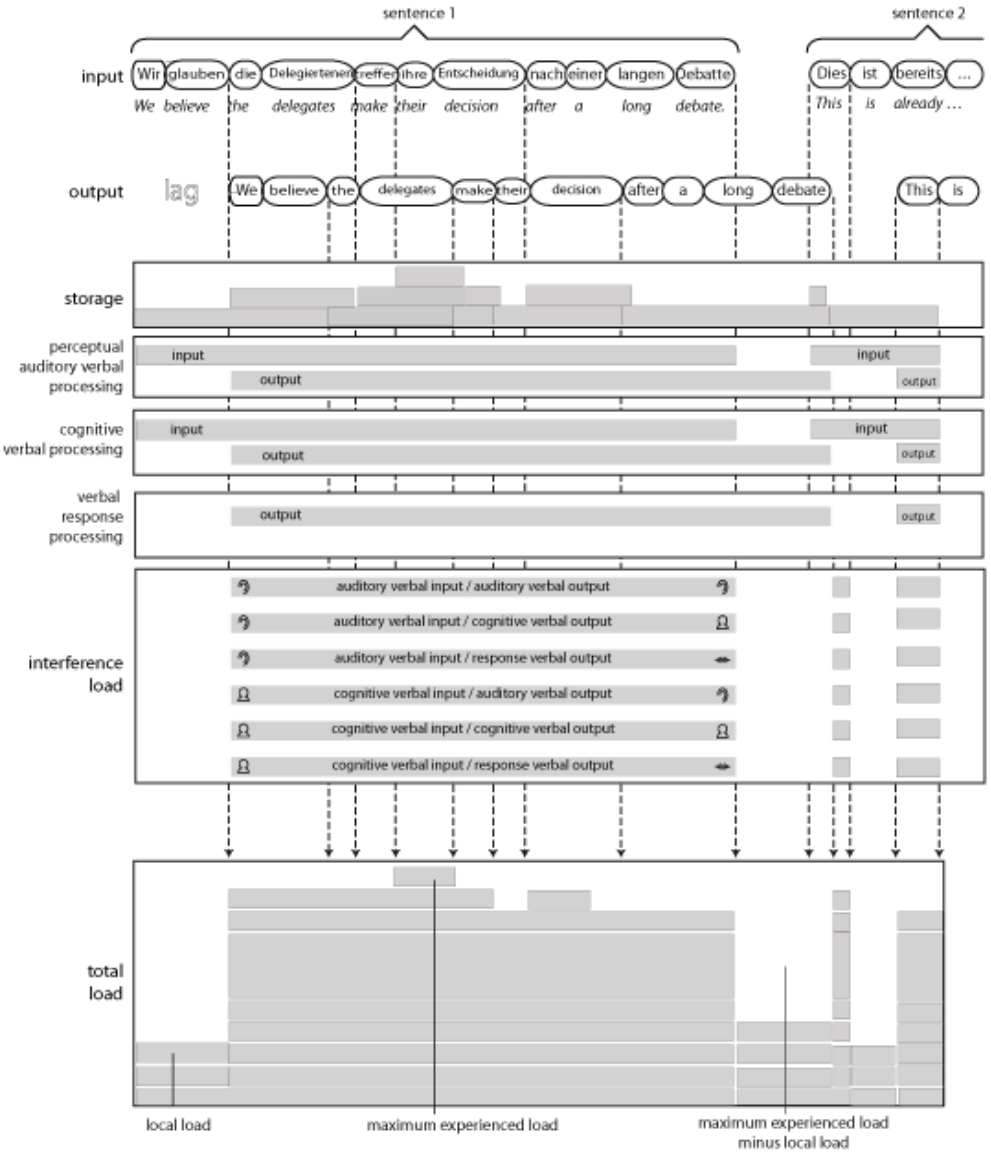

Figure 2: Seeber's cognitive load model (Seeber, 2011).

Seeber's model puts task difficulty in the center by indicating how many elements the conference interpreter has to process simultaneously. The more processes take place concurrently and the more attentional resources they need, the more complex the task gets. Any predictions on the mental effort the interpreter will make during his interpretation will only be based on the assumption that mental effort increases with cognitive load. It is in this sense complementary to Gile's effort model which focuses on the resources of the interpreter and their distribution rather than on the cognitive load. A strong point of Seeber's model is certainly that it allows

an online analysis of cognitive load during interpreting. It is therefore particularly interesting for researchers in interpreting science who study the impact of local problem triggers, like numerals or proper names. Moreover, it visualizes how cognitive load of the preceding segment can affect the processing of the current segment. One of the maybe most important conceptual differences between Gile's and Seeber's model concerns the assumptions about attentional resources. First, Gile assumes, similar to Kahnemann, one single pool of attentional resources whereas Seeber follows Wickens in assuming different resources according to the modality, the stage of processing or the code of processing. This differentiation makes it possible to make predictions about other sub-components of simultaneous interpreting, like visual information processing, even though his model at its present state does not include a visual information processing component. Second, Gile premises that interpreters always work close to saturation. As soon as one effort decreases, the attentional resources will be reallocated to other efforts improving for example speech production, recall or speech analysis. Seeber, on the contrary, assumes that cognitive load is only determined by task demands which means that cognitive load may vary during the speech. Reducing one load component might therefore not necessarily mean that performance or recall improves.

3.3.3 Visual information as a load factor in simultaneous interpreting

How does visual information affect simultaneous interpreting according to Gile's effort model? In order to take visual information into account, a fifth effort needs to be added that may be called *visual analysis effort*. This means that the limited pool of attentional capacity has to be divided on five instead of four efforts. Dealing with visual information means thus in general less attentional capacity for each single effort. Interpreters might decide whether the visual information they get really provide useful complementary information or whether it is superfluous and only represents additional load. Looking at a graph the speaker is explaining

might help to understand the message. If the speaker does not refer to the graph but talks about something else, it is surely more beneficial to simply ignore the visual input and concentrate on the auditory input. The case might be a little different for audio-visual speech that does not provide additional information but enhances the auditory speech input. In this particular case, it can be argued that the visual input – lip movements - does not concern the *visual analysis effort*, but the *listening and analysis effort*. If lip movements lower the listening and analysis effort, attentional capacities are freed up and can be invested in other efforts (storage, speech production) leading to a better interpreting performance or to better recall.

How can visual input be integrated into Seeber's cognitive load model and what would be the predictions? If I consider all we know about attention by now, I would expect a difference between processes that require attention and strain the working memory and those that do not. The question is thus whether information processing in simultaneous interpreting needs attention. The first thing to note is that perceptual processing in general occurs automatically. We see or we hear something without concentrating on it, but that does not mean that we are aware of what we see or hear. Attention is required as soon as a response or a decision is required[14]. Checking a glossary for terminology or following the speaker's presentation while interpreting, for example, are attention-demanding tasks because they could provide new and complementary information that the interpreter needs to bring in line with what the speaker is saying. In this case I would expect that the presentation or the glossary interfere with the speech, the auditory stream, as both elements are verbal and

[14] A quite impressive demonstration of the effect of attention is the 1999 study by Simon and Chabris. The authors instructed the participants to watch a movie of a basketball match and to count the passes of each team. Eventually, a person disguised as gorilla crossed the movie. Only half of the participants noticed the gorilla (Simons & Chabris, 1999).

cognitive-perceptual inputs. The case is different for gestures. Gestures accompany rather loosely the auditory input and provide additional cues to the meaning of what is being said. But they are perceived visually and do not take a verbal form. Hence, they should not interfere with the perceptual-auditory processing component or the cognitive verbal processing component. Gestures should therefore not increase cognitive load.

Lip movements could be seen as subcomponent of the perceptual verbal processing [15] . Lip movements differ from gestures, glossaries or presentations because the speaker's lip movements are perfectly congruent with his speech. They codify the same information and fill in the gaps that may occur in the auditory signal. Therefore, processing the speaker's lip movements does not need special attention, nor does the integration of auditory and visual speech inputs, at least as long as the lip movements and the auditory stream do not depart from each other and the interpreter has to decide which one fits better into the speech context and should prevail. But as we have seen in chapter 2.3.5, audio-visual speech is not only neutral when it comes to cognitive load. It has even a facilitating effect: it reduces the noise of the sensory signal and speeds up speech processing. With regard to Seeber's model, this means that I need to assume a facilitatory factor for audio-visual speech input that lowers the perceptual verbal processing load and thereby decreases the overall cognitive load during simultaneous interpreting.

3.3.4 Noise as load factor in simultaneous interpreting

Overlaying auditory speech input with other sounds can affect phoneme recognition and speech comprehension at various levels. If the second

[15] Perceptual verbal processing does not exclusively refer to auditory speech input, but also to other sensory speech input that take a verbal form, like lip movements, sign language or Braille.

auditory stream is likely to contain useful information for the task at hand, it can act as a distractor and drain attention from the first auditory stream to the second one. It does not necessarily mean that the first auditory stream is inaudible or not perceivable. For example, in dichotic listening tasks, participants receive simultaneously on one ear each a perfectly perceivable auditory input. Nevertheless, Broadbent demonstrated in his 1956 study that performance is worse for dichotic presentation than for monotochic or audio-visual presentation. His participants remembered only 62% of the presented digits correctly compared to 77% in an audio-visual and 92% in a monotochic presentation (Broadbent, 1956).

But sometimes a second auditory stream is used to cover partially the first auditory stream. In this case, researchers most commonly use white noise or talker babble. Talker babble is meaningless noise that sounds like several persons talking at the same time. It does not act as a distractor, because it is very distinct to the speaker's auditory input and contains no useful information for the task. Yet, it increases the load factor of speech perception because it interacts with the underlying auditory stream. Phonemes that are low in energy like fricatives, prosodic cues and word boundaries get lost (Lecumberri, Cooke, & Cutler, 2010). Because of these "holes", the auditory signal as a whole is more difficult to interpret: listeners have to consider a larger number of potential lexical candidates and have to make stronger lexical predictions. The larger the number of candidates, the larger is the benefit from visual cues that help to resolve the ambiguity (Iverson, Bernstein, & Auer, 1998). In a 1997 study, Kramer and colleagues used pupil dilation to measure the mental effort associated with speech perception in noise. Compared to the participant's individual detection threshold where 50% of the target stimuli were identified correctly, the authors observed a significant pupil dilation when the signal-to-noise ratio increased by 5 dB, e.g. the volume of the signal decreased by 5 dB while the volume of the noise was kept constant. According to the authors, this means that participants needed to make more effort to understand the speech signal when larger parts of the signal were

obscured by noise (Kramer, Kapteyn, Feesten, & Kuik, 1997). Similar results were obtained 2012 by Koelewijn and his team: pupils dilated during speech perception in noise according to two different detection thresholds (50% and 84%). Moreover, the participants' ratings of their own performance correlated negatively with pupil dilations (Koelewijn, Zekveld, Festen, & Kramer, 2012). If I consider once again Seeber's cognitive load model for simultaneous interpreting, addition of noise would mean that the load factor associated with the perceptual verbal processing component increases.

Noise does not only affect speech perception at the word level. In a 1974 study, Gerver asked conference interpreters to simultaneously interpret and to shadow speeches in noise at three different signal-to-noise ratios. He reported lower intelligibility and informativeness of the target text as rated by two independent judges and significantly more errors and omissions[16] compared to a condition without noise. He concludes: "As can be seen, significantly more of each passage was correctly shadowed than interpreted, and noise had a significantly adverse effect on performance of both tasks" (Gerver, 1974, p. 165). Another interesting observation in Gerver's study is that more self-corrections appeared as soon as noise was added to the source text, but that the number of self-corrections did not increase any further at more unfavorable signal-to-noise ratios (Gerver, 1974). Based on Gile's effort model (see chapter 3.3.1), this could suggest that with decreasing signal-to-noise-ratio, speech perception needs more attentional resources, leaving fewer resources for the target speech production, self-monitoring and corrections. To

[16] Word omissions are possibly not the best way to measure interpreting performance, for a word in the source text might be redundant and the interpreter might choose, for strategic reasons, to leave it out, to replace it by a pronoun or to reformulate the whole phrase in a very different way.

summarize: According to psychological and psycholinguistic research, listener's need to make more effort to understand speech in noise.

3.3.5 Effects of visual input and noise: predictions from psychological and interpreting research

If I consider the psychological and psycholinguistic research so far, the conclusion seems to be clear-cut: Audio-visual speech reduces ambiguity in the sensory input and narrows the number of lexical candidate that fits the sensory input. I would therefore expect to find the same pattern in interpreting research. Surprisingly, this is not the case. As described in chapter 2.4.2, researchers have not been able to demonstrate a beneficial effect of audio-visual speech or other visual input in simultaneous interpreting, nor to observe a loss of performance (Rennert, 2008; Anderson L. , 1994) or higher stress hormone levels (cortisol levels) when interpreting with limited visual access, as in remote-interpreting (Moser-Mercer, 2003; 2005b). Yet, these three studies report consistently anecdotal evidence that conference interpreters expressed a certain unease, fatigue, lack of concentration and sometimes even headaches when they worked without or with limited visual input. Comparing remote and on-site interpreting, Roziner and Shlesinger (2010) showed that great discrepancies exist between objective measures and subjective perception. Interpreters judged their interpreting performance as being inferior in the remote condition than in the on-site condition while objectively there was no significant difference. Similarly, sound quality or ergonomic comfort obtained lower ratings in the remote than in the on-site condition.

Apart from the methodological issues mentioned in chapter 2.4.2 it is of course possible that conference interpreters may do very well without visual input, but feel unsure, because they are not used to work without seeing the speaker, the audience or presentations. This explanation, however, seems a bit weak as interpreters continue to insist on visual contact. There is even an ISO-norm recommending a window for mobile

booths as to ensure the visual contact between the interpreters and the speaker (International Organization for Standardization, 1998). A far more reasonable explanation for these contradictory findings is that conference interpreters adapt their effort when working without visual input in order to maintain the interpreting performance. If this is true, I would expect an increase of mental effort during simultaneous interpreting without visual input.

For noise, however, the predictions from psychological and interpreting studies seem to converge. White noise, according to psychological research, augments the lexical ambiguity by making parts of the auditory input unperceivable, thereby increasing cognitive load related to speech perception. In 1974, Gerver has experimentally confirmed that noise affects the interpreting performance. One possibility to explain why simultaneous interpreting in noise is particularly prone to errors and omissions could be explained as follows: If the speech input is defective, listeners make greater use of speech context in order to identify the word. For interpreters, this would mean that they need to hold larger chunks of the source speech in memory before starting the interpretation. The memory storage load (and the ear-voice span) would therefore increase and result in memory overload that leads to information loss or an erroneous target speech production. Visual speech information could compensate at least partially for the defective source speech input by providing additional cues to identify the correct word. A 2014 study by Shahin and colleagues shows in fact that participants tolerate longer speech interruptions due to noise and still perceive the speech as continuous if lip movements are provided (Shahin, Kerlin, Bhat, & Miller, 2012).

In a nutshell, when audio-visual speech is provided, interpreters can rely partly on lip movements to resolve the ambiguity of the auditory input. Analyzing lips movements should not interfere (or only little) with analyzing the auditory input because auditory and visual stimuli are congruent.

Moreover, processing of lip movements and (auditory) speech call for different resources and can be processed separately. I can therefore expect that the cognitive load involved in listening and speech comprehension decreases. Noise, on the contrast, increases the ambiguity of the auditory input and interpreters need to concentrate harder on the auditory stream or alternatively, wait longer until they received sufficient context to understand what was said. The cognitive load involved in speech comprehension and possibly in memory storage should therefore increase. In this case, visual cues should be of particular use to the interpreter as they help to identify the correct lexical candidate.

3.4 Measuring work-load

Measuring task difficulty has been the aim of countless studies and experiments. But task difficulty can take on many different forms: the number of elements that need to be processed concurrently, the pace at which the task is presented, the sensory modalities that are involved, whether the task requires rather continuous vigilance or speeded reactions etc. Researchers have tried to map the different dimensions of task difficulty to theoretical concepts and came up with different models like the capacity model of attention (Kahnemann, 1973), the theory of cognitive load (Sweller, 2010), the model of task interference (Lavie, Hirst, de Fockert, & Viding, 2004) and others (see chapter 2.2), but on an experimental basis, it is often impossible to disentangle those concepts[17]. Therefore, I will use the more neutral expression of *work-load* throughout this chapter. Among the most common techniques to assess work-load, Galy, Cariou and Mélan (2012) name subjective ratings, physiological measures like pupillary response, galvanic skin response or heart rate variability, and task performance (Galy, Cariou, & Mélan, 2012). In the

[17] This is probably also the reason why *mental effort*, *cognitive load*, *mental load* or similar expressions are often used interchangeably.

following, I will explore different methods to investigate work-load along the lines of the three categories of subjective ratings, performance and physiological measures. The selection of the methods described below is limited to those that are relevant for the present study and is not exhaustive.

3.4.1 Subjective ratings

In subjective ratings, participants are asked to rate themselves the work-load they have experienced during the task. Examples that are commonly used are the NASA-Task-load Index (NASA-TLX) (Hart & Staveland, 1988) or the subjective task-load assessment technique (SWAT) (Reid & Nygren, 1988). The NASA-TLX, for instance, covers a whole range of load dimensions: mental and physical demand of a task, time pressure during task accomplishment, performance, perceived effort and feeling of frustration. The index was developed 1988 by Hart and Staveland on the basis of different tasks (simple and complex control tasks, aircraft simulation). The evaluation procedure consists of two steps: first, the participant indicates how much each of the six dimensions contributes to work-load. This weighting procedure allows accounting for individual differences in work-load perception. Then the participant rates each dimension on a 21-point scale from very low to very high (Hart & Staveland, 1988). The NASA-TLX is frequently used in research on aviation or similar fields (Hart S. , 2006), but also in many other cognitive tasks (Djamasbi, Mheta, & Samani, 2012; Rubio, Díaz, Martín, & Puente, 2004; Hart S. , 2006). The SWAT, a subjective rating scale developed 1988 by Reid and Nygren covers three dimensions: time load, mental effort load and psychological stress load. Each dimension has three levels. Like the NASA-TLX, it is a two-step procedure: the rater develops first a scale on the basis of 27 cards that contain all possible combinations of the three above mentioned dimensions and their load levels, and then rates the task according to the three dimensions (Reid & Nygren, 1988). The SWAT was originally designed to assess work-load in aviation, but today's

use encompasses many more applications like control room operations (Ikuma, Harvey, Taylor, & Handal, 2014), driving simulation (Baldauf, Burgard, & Wittmann, 2009), cognitive-perceptual dual tasking (Rubio, Díaz, Martín, & Puente, 2004).

The SWAT and NASA-TLX seem to provide essentially the same trends in ratings. Rubío and colleagues (2004) used a memory and a visual tracking task with each two different difficulty levels to compare both rating scales. Both scales distinguish reliably between a single and a dual task condition and reflect task difficulty. The SWAT was furthermore able to discriminate the memory and the tracking task, whereas the mean performance across all tasks did not differ between both scales. The authors conclude that both scales are equally valid and react very similar on task difficulty and single/dual task presentation. Both methods are widely accepted by the participants. As they are administered before (scale development procedure) and after the primary task (rating procedure), they do not affect the participant's performance during the task execution. But the authors also noted several drawbacks. First, the diagnosticity of rating techniques is limited. That is, they do not allow conclusions about the reasons of work-load changes, or only to a very limited extent (Rubio, Díaz, Martín, & Puente, 2004). Second, subjective ratings are not able to assess work-load changes online during task execution, for example during driving or interpreting. For this reason, rating techniques are often combined with physiological indictors (see chapter 3.4.3). Finally, due to the scale development, the NASA-TLX and the SWAT are both rather time-consuming. Researchers count approximately one hour to establish the scale for the SWAT or the NASA-TLX ratings (NASA Task Load Index (NASA-TLX), 2016; Subjective Workload Assessment Technique (SWAT), 2016). This drawback has led some researcher to refrain from the weighting procedure of the NASA-TLX (Hart S. , 2006) or to use their own rating system. For example, Ayres found the simple rating of task difficulty to correlate highly (r=0.85) with the difficulty level of arithmetical problems (Ayres, 2006).

Cognitive load and mental effort

A very interesting way to investigate work-load is to ask participants to estimate the time they spent on the task. Time perception of longer periods (several minutes) seems to depend mainly on temporal awareness. The more we are aware of the time passing, the longer it seems. Engaging in cognitively demanding tasks reduces the amount of attention that can be allocated to time perception. This is why time periods seem shorter during tasks execution (Schiffmann, 1996, pp. 497-499). Woehrle and Magliano (2012) asked their participants to judge time durations after solving an easy or a more difficult mathematical problem and during a control condition. Time judgements were significantly lower in the problem solving than in the control condition. Block, Hancock & Zankay (2010) confirm these results, but note also the effect of the paradigm that has been used, e.g. whether participants are aware beforehand that they will need to judge task duration or not. In a meta-analysis, they compared the results of 117 studies that used time judgements. They concluded that time judgements under high work-load are shorter if the participant is informed beforehand that he will have to judge the task duration. In contrast, if the participant is naïve, he will judge task duration to be longer under high work-load. The authors also note that effects of cognitive load on duration judgments only can be found if judgements are made immediately after stimulus presentation (Block, Hancock, & Zakay, 2010).

The present study uses ratings of source speech rate and of source speech complexity as indicators for work-load, as well as time judgements. This choice was initially based on two dimensions of the NASA-TLX, mental demand and time pressure that seemed most susceptible to reflect work-load changes in simultaneous interpreting. They have been adapted to simultaneous interpreting in an attempt to make it easier for the participants to rate these two dimensions. Other dimensions have been discarded to limit the duration of the whole experimental procedure and to avoid task disengagement effects. Details of the experimental procedure are given in chapter 4.

3.4.2 Performance

Performance based measures typically include task accuracy (errors) and reaction times. Task accuracy has been used as early as 1955 (Conrad, 1955) (and probably earlier) to assess work-load. The underlying assumption of this method is that the number of errors increases as a function of work-load. An example frequently used to assess concentration and attention is the d2-test. The test consists of several lines with similar letters (p and d) that are paired with one to four strokes. Participants have to cross out within twenty seconds per line all d's with two strokes (Brickenkam & Zillmer, 1998). In this example, performance can defined in terms of response accuracy, e.g. as the number of correctly identified targets[18]. Response accuracy as indicator for work-load has also been used in more complex paradigms, for example in driving (Horrey, Lesch, Garabet, Simmons, & Maikala, 2017), flight simulation (Peyasakhovich, Dehais, & Causse, 2015) or nuclear plant control (Reinerman-Jones, Metthews, & Mercade, 2016; Ikuma, Harvey, Taylor, & Handal, 2014). In learning research, primary task performance is often assessed via questions or multiple choice tasks. Participants are given a certain amount of time to learn some content from a text, a video or other material that is either presented in a way to reduce work-load during learning (for example by integrating visual and textual information or by omitting redundant subtitles in video material). Afterwards, they have to answer questions related to the learning material they were given. Typically, recall is better

[18] Two types of errors can be observed in the d2-test: either targets are left out (omission) or items are wrongly identified as targets (confusion). While the first type of error is related to attentional control and quality of performance, the latter one is rather related to inhibitory control. For test evaluation, the authors suggest to calculate the concentration performance that is defined as the number of correctly processed items minus the number of confusion errors. According to the authors, the concentration score reflects reliably the relationship between speed and accuracy (Brickenkam & Zillmer, 1998, pp. 11-15).

in low load or learning enhancing conditions (Chandler & Sweller, 1991; Moreno & Mayer, 2002). A particular case is the stroop task (Stroop, 1935). The stroop task does not directly measure performance but rather the inhibition of a wrong response. In stroop tasks, participants have to read color words and name the color of the ink instead of the color word. Since its first version, the task has been adapted to other modalities like hearing the words “high” and “low” in a high or low pitch, or it has been completed by other aspects like speed of presentation. Response accuracy decreases in the incompatible condition (McClain, 1983; Chuderski, Senderecka, Kalamala, & Kroczek, 2016; Renaud & Blondin, 1997).

Researchers can also make use of a secondary task to evaluate work-load if they fear that the primary task performance is not sensitive enough or not suitable for evaluation (Wickens, Kramer, Vanasse, & Donchin, 1983; Brown I. , 1978). Baldauf, Burard & Wittman (2009) for example, asked their participants to press a button every 17 seconds in a driving simulation task with different levels of task difficulty. The response time increased with higher task difficulty. Another example comes from Haji and colleagues (2015). The authors asked novice and expert surgeons to perform surgical nots while monitoring the heart rate of their (dummy) patients. Secondary task performance (heart rate monitoring) was lower for novice surgeons (Haji, et al., 2015). Finally, reaction times may also provide information about work-load. The higher the work-load, the more time participants need to make a response. An example for reaction times that vary as a function of work-load is provided by Hick (1952). Participants were asked to press a particular key in response to each of the maximally ten light bulbs that could light up. He noticed that response times increased with the number of light bulbs. Other examples are word naming and lexical decision for high frequency words and low frequency words. Response times for naming or lexical decision increase with decreasing word frequency (Schilling, Rayner, & Chumbley, 1998).

Cognitive load and mental effort

In simultaneous interpreting, performance is usually related to the errors and omissions that occur in the target text compared to the source text. Sometimes, errors are assigned to specific categories according to the research question the author investigates. The definition of performance, however, varies widely between the authors: in some cases it is each word that has not been translated correctly (Gerver, 2002), in others performance is evaluated by raters on a scale (Anderson L. , 1994), and often performance is the result of a qualitative analysis whose criteria are not described in detail (Rennert, 2008). How good performance works as an indicator of work-load depends thus highly on the experimental design and the definition of performance. Simultaneous interpreting performance can be highly informative with regard to specific features like numbers. Analyzing incorrect translations of numbers and assigning them to different categories depending on the number of syllables has revealed that interpreters can process numbers with up to four syllables without too many difficulties. Beyond this word length, the number of correctly rendered numbers declines rapidly (Gieshoff, 2012). On the other hand, performance can be completely inconclusive if the analysis is not fine-grained enough (see for example Rennert, 2008).

More general linguistic cues that are not confined to a specific word class but could nevertheless reveal changes in work-load during simultaneous interpreting are cognate translations (Oster, 2017) or the use of work-load reducing strategies like segmentation (splitting long source text sentences into several smaller and syntactically less complex target text sentences) (Yagi, 2000). Cognate translations are a very elegant technique to evaluate the effectiveness of monitoring during bilingual tasks like translation or interpreting. Cognates are words of different languages that are orthographically and/or phonetically very similar (English: *house* – German: *Haus*) and therefore more easily accessed during speech production (Costa, Satesteban, & Cano, 2005; Chistoffels, Firk, & Schiler, 2007; Dijkstra, Grainger, & van Heuven, 1999; Starreveld, de Groot, & Rossmark, 2014). This cognate facilitation effect has also been shown in

simultaneous interpreting (Dong & Lin, 2013; Gieshoff, 2017). The rationale behind this method is similar to the stroop task (Stroop, 1935): cognates are often considered to be “less natural” or less acceptable in the target language and therefore, translators and interpreters often seek to avoid them. In a 2017 study, Oster demonstrated that student translators use fewer cognates in written translation than in oral (sight) translation. She explains this finding with the higher time constraints in oral translations that leave fewer resources for monitoring (Oster, 2017). Segmentation – chunking long sentences in several shorter ones - is a strategy that is recommended in simultaneous interpreting during particular difficult or fast passages to lower memory burden[19] (Yagi, 2000). Lower utterance length and syntactic complexity is also observed in free speech under cognitive load (Cohen, Dinzeo, Donovan, Brown, & Morrison, 2015). In translation and post-editing, the number of syntactical changes has been found to correlate positively with reading times, indicating translation difficulty (Schaeffer & Carl, 2014; Carl & Schaeffer, 2017). Similarly, the number of possible translations for a word is associated with a higher number of fixation counts, longer fixation times and longer pauses during the translation process (Dragsted, 2012). However, in the present study, segmentation has not been considered as dependent variable, first, because it is very difficult to transform it in a way that it can be used as statistical variable, and second, because as a strategy, segmentation is voluntarily used by interpreters. The reasons for this choice do not necessarily need to be linked to task difficulty. It is also possible that under some circumstances, shorter sentences seem more appropriate to the interpreter to get the message across.

[19] The use of segmentation depends on a number of other factors, like source and target language, interactions with preceding sentences and speech rate (Goldman-Eisler, 2002) and certainly individual preferences, experience, mastering of the source and target language, memory capacity or others.

Another linguistic parameter that is interesting in regard to simultaneous interpreting is the ear-voice-span. The ear-voice-span corresponds to the lag between the source and the target text, e.g. between the speaker and the interpreter. I could expect that ear-voice-span is longer for segments that require a higher listening or speech analysis effort because interpreters might need to wait for more speech input. A similar idea has been explored for translation by Timerovà, Dragsted and Hansen (2011). They compared the time span from the first and last fixation to the moment where participants started typing (eye-key span) in trainee and professional translators and in translating and text copying. The authors found a larger variability of the eye-key span in trainees than in professional translators. Moreover, text copying was associated with lower eye-key spans than translating. The authors argue therefore that larger eye-key spans may reflect higher cognitive effort. In simultaneous interpreting, however, ear-voice-span is not a typical indicator of performance for different reasons. First, ear-voice span depends very much on the interpreter and on linguistic factors (syntactical similarity). For instance, Goldman-Eisler (2002) found larger ear-to-voice spans when interpreting from German to English than when interpreters worked from French to English. Second, as has been pointed out by several authors (Lee, 2002; Li, 2010; Seeber & Kerzel, 2012), interpreters can use strategies to catch up with the speakers such as summarizing, segmenting or anticipating. A short ear-voice span therefore does not necessarily mean that the load involved in processing the segment is low. Consequently, ear-voice-span is very difficult to operationalize as an indicator for cognitive load.

Finally, Lambert (1988) and Gerver (1974) have made use of recall to investigate work-load in listening, shadowing and simultaneous interpreting. Gerver (1974) defined recall as the number of correctly answered content-related questions. Lambert (1988) used apart from content-related (“semantic”) questions also lexical and syntactical recognition tests. Recall was defined as the mean percentage of correct

responses across the three tests. Both authors found independently from each other better recall for listening than for simultaneous interpreting, but better recall for simultaneous interpreting than for shadowing (Gerver, 1974; Lambert, 1988). Lambert (1988) interpreted these findings as indication for deeper level processing in listening compared to simultaneous interpreting. According to her, listeners can allocate all of their resources to speech analysis, while interpreters have to share their resources between speech analysis and verbal production. This interference may impede deep level processing (Lambert, 1988).

The present study tracks performance in simultaneous interpreting with cognate translations (for more detail on the stimuli and the data preparation see 4.3.1 and chapter 4.3.6.5) and translation accuracy. Translation accuracy, here defined as the number of correctly rendered proposition (see chapter 4.3.6.4 for more details), has been added to achieve a higher comparability with other studies in the field of interpreting research. In order to test memorization of speech content - or in the words of Craik and Lockhart (1972): “depth of processing “– participants were also asked to answer text-related questions after each speech (see chapter 4.3.6.3).

3.4.3 Physiological measures

High work-load means that the body must respond to unusually high demands in terms of attentional resources. In this respect, it is comparable to other situations where the body makes extra resources available, for example to run away from a tiger (or an exam). It seems therefore natural to expect similar physiological reactions during high work-load as during stressful situations. Peters and his colleagues (1998) differentiate two types of stress reactions: a) a sympatho-medullary system and b) a hypothalamic-piuitary-adrenal-cortical system. The first one copes with stressors: it increases heart rate and the blood pressure and releases epinephrine and norepinephrine. The muscles and the brain are flooded with oxygen and glucose and can adapt their performance to the

increased demands. The latter one, the hypothalamic-piuitary-adrenal-cortical system, reacts to the inability to cope with a stressor. It raises feelings of helplessness or uncontrollability and is responsible for the release of adrenocorticotropic hormone (ACTH) and cortisol (Peters, et al., 1998). Work-load is usually considered to be a stress reaction of the first type. It translates for example (and not exhaustively) in pupillary responses (see chapter 3.4.3.1), in higher voice frequency and intensity (see chapter 3.4.3.2), in a larger variability of the heart beat rate (HBR) (Brouwer, Hogervorst, Holewijn, & van Erp, 2014; Luque-Casado, Perales, Cárdenas, & Sanabria, 2016) or in a stronger galvanic skin response (Baldauf, Burgard, & Wittmann, 2009; Brouwer, Hogervorst, Holewijn, & van Erp, 2014). Further physiological effects that have been found in response to work-load are particular EEG-patterns (increased alpha-power bands, see for example Reinerman-Jones, Metthews, & Mercade, 2016; Ryu & Myung, 2005), increased cerebral blood flow (Horrey, Lesch, Garabet, Simmons, & Maikala, 2017), eye blink rates (Brouwer, Hogervorst, Holewijn, & van Erp, 2014) or increased gaze fixation durations (Djamasbi, Mheta, & Samani, 2012, for a review on different physiological reactions and their relation to work-load see Borghini, Astolfi, Vecchiato, Mattia, & Babiloni, 2014). An important advantage of physiological indicators is that they provide continuous measurements. Moreover, physiological reactions are largely involuntary, that is they cannot be altered or avoided by the participants. On the flip side, they often need costly technical equipment, like eye trackers, ECG- or EEG-apparatuses or EDA-meters that affect more or less the experimental setting. A participant sitting in front of an eye-tracker with his head fixated by a forehead and a chin rest may feel quite uncomfortable during the experiment. As physiological indicators are always a global result of a stress reaction, they do in general not inform us about the type of the stressor. It is therefore very difficult to distinguish different types of work-load on the basis of physiological reactions. The following sections will essentially concentrate on those measures that are of relevance for the

experiments described in chapter 3, e.g. pupillary responses and voice-related parameters, and briefly summarize how researchers have interpreted correlations between subjective ratings, task performance and various physiological measures.

3.4.3.1 Pupillary responses

The pupil size changes as a function of luminance or distance to an object, but the pupil reacts also to sympathetic activation (see chapter 2.1.3). Pupil responses occur with fear, anger (Kahnemann, 1973) and pain (Chapman, Oka, Bradshaw, Jacobson, & Donaldson, 1999), they indicate preferences (Goldwater, 1972), but also work-load (*task-evoked pupillary responses*). This large and undifferentiated range of factors that cause pupil dilations is the reason why Kahnemann (1973) attributed pupillary responses to a state of general arousal. In an experimental setting, this means that pupillary responses are not very informative with regard to the cause of the dilation. Therefore, the experiment needs to be carefully designed as to exclude all other factors that could have caused a pupillary response. Still, task-evoked pupillary responses (TEPR) provide useful insights in mental processes because the size of the pupil indicates the work-load level: The higher the work-load, the larger the pupils. As soon as the processing capacities of a participant are exceeded, a *task disengagement effect* can be observed. Granholm and colleagues (1996) found increasing pupil dilations with increasing work-load in a digit recall task, but a sudden pupil constriction as soon as the number of digits to recall exceeded the available resources and the participants failed to report the digit sequence. Another important distinction with regard to TEPR is the distinction between tonic and phasic pupil dilations. According to Gilzenrath and colleagues (2010), phasic pupillary responses reflect the process of exploitation and are associated with high performance and selective response to task-relevant stimuli. Tonic pupillary responses can be described as pre-test baseline pupil sizes and are linked to exploration, increased sensitivity to all kind of stimuli and degraded task performance

(for the mechanisms underlying tonic and phasic firing see also chapter 2.1.3). Tonic pupil dilations can also be modulated by monetary incentives or feedback (Heitz, Schrock, Payne, & Engle, 2008). Taken together, these findings corroborate Kahnemann's hypothesis (1973) that tonic pupil sizes indicate the state of general arousal, whereas phasic pupil dilation during the task reflects mental effort, e.g. the resources participants allocate to a task rather than the work-load itself (even if both often go hand in hand). He further states that the resources that are allocated to the task will not exceed the effort that is needed to solve the task (Kahnemann, 1973). A 2010 study by van der Meer and colleagues (2010) illustrates these relationships. They showed that individuals with a high IQ performed better on all task difficulty levels and had larger baseline pupil sizes than individuals with an average IQ, but the pupil dilation was only significantly larger when task difficulty was high. This demonstrates that participants only increase their mental effort if this is necessary in order to solve the task. Similar results were previously obtained by Heitz and colleagues (2008).

For experimental purposes, researchers usually measure the phasic pupillary response to a stimulus. During dark adaptation, the pupil can dilate up to a diameter of nearly 9 mm (60 mm^2) in approximately three seconds (Brown & Page, 1939), however, pupillary responses to work-load are typically lower than 0.5 mm (Klinger, 2010). Beatty (1982) for examples reported dilations between 0.2 mm and 0.4 mm in simple psychological experiments like digit recall, simple additions, same/different judgements and visual or auditory detection. Pupils dilated with increasing task difficulty (number of digits to recall, signal-to-noise ratio). Pupillary responses do not only allow researchers to differentiate between low and high work-load conditions, but also between auditory and visual presentation modes. Klinger and colleagues (2011) contrasted auditory and visual presentation of multiplication tasks and found aurally presented problems to evoke larger pupil dilations than visually presented problems (difference in peak dilation about 0.2 mm) and to lead to about 15% more

errors. Researchers made also use of pupil dilations to investigate the work-load during speech comprehension. Kramer and colleagues (1997) measured pupil dilations during speech perception in noise. Pupil dilations were largest when the stimuli sentences were presented at the detection threshold and decreased significantly when the signal was 10 dB louder as the noise. Similar findings were reported by Kuchinsky and colleagues (2013) and Zekveld and colleagues (2014). Other factors that elicit pupil dilations during speech comprehension are lexical competition (Kuchinsky, et al., 2013) or syntactical ambiguities (Engelhardt & Ferreira, 2010). In interpreting research, only two studies using pupil dilations are reported. The first one is a study conducted by Hyönä, Tommola and Alaja in 1995. They found significant differences in mean pupil sizes during listening, shadowing and interpreting (mean difference between shadowing and interpreting about 0.5 mm) (Hyönä, Tommola, & Alaja, 1995). The second one reports a study carried out by Seeber (2012). He contrasted the work-load generated by verb-initial and verb-final sentences during simultaneous interpreting and obtained a significant difference of mean pupil dilation of about 0.05 mm in the last segment of the target sentence (Seeber & Kerzel, 2012), which shows nicely that pupil dilations can even reveal very small changes in work-load (for an overview about the potential of pupillometry in simultaneous interpreting, see (Seeber, 2015). In the present study, pupillary reactions were tracked during simultaneous interpreting and listening during the whole time course of a speech.

3.4.3.2 Voice

Voice features of speech samples, like fundamental voice frequency, intensity or articulation rate, are an interesting and unobtrusive way to assess the stress level that a person experiences. Like pupil dilations, voice features may be caused by different stressors, emotional and cognitive ones. It is thus important to control for possible confounds. Voice features are chiefly determined by the supraglottal air pressure that causes the vocal chords to vibrate, and the laryngeal musculature

(Kreiman & Sidtis, 2011, pp. 32-49). The vibrations of the vocal chords are determined by their wave length and their amplitude. Fundamental voice frequency refers to the wave length: the faster the oscillations of the vocal cords, the higher the voice frequency. Voice intensity is related to the amplitude. Voice intensity increases with its amplitude (Kreiman & Sidtis, 2011, pp. 54-62). During stress reactions, the muscles of the body tense up, the respiration deepens and the supraglottal air pressure increases. Yap et al. (2015) have investigated the voice source under work-load and have found the opening and closing of vocal chords to be faster and less regular compared to speech without work-load. Moreover, the vocal chords were more tightened (Yap, Epps, Ambikairajah, & Choi, 2015). As a result, voice intensity (perceived as loudness) and fundamental voice frequency (perceived as pitch) increase and the vocal cords become more rigid which causes the speech output to become more monotonous (Scherer, 1989; Warren, 1999, pp. 155-163).

The most commonly investigated voice features in research are fundamental frequency (F0), variations of fundamental frequency, intensity, articulation rate and silent pauses (Laukka, et al., 2008; Darò, 1994; Streeter, Macdonald, Apple, Krauss, & Galotti, 1983; Chen, et al., 2011; Kreiman & Sidtis, 2011, pp. 318-329). As intensity and articulation rate cannot be applied to simultaneous interpreting because the interpreters depend on the source text and hence, cannot increase their speech rate or speak too loudly, I will concentrate on fundamental frequency and silent pauses. Increases of fundamental frequency as reaction to cognitive or emotional stressful situations have been reported during reading in a foreign language (Darò, 1994), during public speech (Laukka, et al., 2008), in simple psychological tests (like stroop tasks), think aloud protocols and during real life situations in control centers in power plants (Streeter, Macdonald, Apple, Krauss, & Galotti, 1983), air traffic control rooms, call centers or during bushfire control training (Chen, et al., 2011). From the examples, it seems that fundamental frequency is rather affected by emotional stress than by work-load. Still, high work-load

may be associated with emotional stress as it can trigger a feeling of being unable to cope with the task. This relationship is further corroborated by the fact that increases in fundamental frequency correlate with self-reported anxiety (Laukka, et al., 2008). The effects works also the other way round: listeners rate speakers of high pitch speech as being more nervous than speakers with low pitch speech, even after all verbal information has been removed with a low-pass filter (Laukka, et al., 2008). It is, however, important to bear in mind that the effects of stress on fundamental frequency differ between individuals (Streeter, Macdonald, Apple, Krauss, & Galotti, 1983; Scherer, 1989).

Decreased variability in fundamental frequency – though not as well documented – seems to be another reliable indicator of emotional and cognitive stress. Cohen and colleagues (2015) reported reduced variance in fundamental frequency in free speech when participants concurrently did a visual distractor task (respond to the target stimuli and inhibit the distractor stimuli) compared to a control condition without competing task. He was able to discriminate two levels of task difficulty (same target stimulus – alternating target stimuli) on the basis of the variance in fundamental frequency. Lively and colleagues (1993) found variability in fundamental frequency to decrease as a function of task difficulty in a tracking task. The situation might be different for simultaneous interpreting because the interpreter discovers the source text successively as the speaker delivers it. This dependency leads to unnatural prosody in the target speeche. As such, Ahrens (2004) observed more rising and constant intonation patterns in interpreted than in natural speech.

Research on work-load suggests that the percentage of silences in an uttering increases with work-load (Müller, Barbara, Jameson, Rummer, & Wittig, 2001) and nervousness (Laukka, et al., 2008). Müller and colleagues (2001) asked participants to ask questions during a navigation tracking task (simulation of a crowded airport) or without. Participants were instructed either to produce clear speech or to speak as quickly as

possible. The authors observed that the articulation rate decreased slightly under the high load condition (competing navigation task), regardless if participants were asked to produce clear or quick utterings. The number of silent pauses, however, increased only under the high load condition when participants were instructed to deliver quick speech (Müller, Barbara, Jameson, Rummer, & Wittig, 2001). Ahrens (2004) compared silent pauses between the source speech and three target speeches that were rendered by professional interpreters. She found fewer, but longer silent pauses in interpreted speech than in source speech. In the source speech, 1948 silent pauses occurred with a mean length of 0.61 seconds, whereas the number of pauses in the target speeches varied between 1670 and 1708 with a mean length between 0.81 seconds and 0.91 seconds. According to the author, interpreters used shorter pauses (< 0.4 seconds) in general to mark the end of an information unit. Longer pauses, in contrast, indicate comprehension and speech planning processes. She points out, however, that the interpreter depends on the source speech and on the pace at which it is delivered (Ahrens, 2004). Indeed, the interpreter sometimes needs to listen to the source text and to wait until she gets sufficient information to formulate the target sentence. In some cases, the speaker interrupts his speech, for example when the audience is laughing. Therefore, the number and duration of silent pauses in interpreted speech cannot be compared to natural speech to measure differences in cognitive load between both activities. Still, pauses might be a useful indicator to compare cognitive load during simultaneous interpreting in different conditions.

The present study uses fundamental voice frequency as indicator for work-load during simultaneous interpreting. It was hypothesized that higher work-load during simultaneous interpreting would increases interpreters' stress level which, in turn, should translate in higher fundamental voice frequency. According to Ahrens (2004) longer pauses during simultaneous interpreting can indicate speech planning or comprehension processes. The number and duration of pauses might therefore allow differentiating

low and high work-load conditions in simultaneous interpreting. For this reason, silent pauses were analyzed in the follow-up study.

4 Method and results

This chapter describes two studies, one pilot study and a larger follow-up study, that were conducted to investigate the impact of audio-visual speech in simultaneous interpreting.

4.1 Aims and general approach

The aim of my research project was twofold: first, I wanted to find empirical evidence whether audio-visual speech facilitates simultaneous interpreting or that the absence of visual input increases the work-load in simultaneous interpreting. Evidence for this is lacking today, despite some attempts to demonstrate the importance of visual information in simultaneous interpreting (see chapter 2.4.2). One of the main reasons why researchers might have failed so far to observe a clear effect of visual information is that most experiments took place in a very natural, thus uncontrolled setting, with many different factors that could have affected the results. In addition, different types of visual information were mixed up in those experiments, although our knowledge of sensory processing suggests that congruent and simultaneous visual and auditory stimuli trigger a stronger response and are processed faster than incongruent or subsequent stimuli (see chapter 2.3.1). I therefore opted for a more systematical and structured approach. Instead of a natural (mock) conference setting with all kinds of visual information, I chose a laboratory setting and limited the available visual information to the lip movements of the speaker. On the one hand, the drawback of such a controlled experimental setting is certainly that it is not a realistic environment. Interpreters might experience the presence or absence of visual input very differently in a real world conference. On the other hand, it is the only way to avoid interactions and to get a clear picture of how one type of visual information, in this case audio-visual speech, influences work-load during

simultaneous interpreting. As such, laboratory experiments can provide a first hint to effects that may also exist in the real world and motivate further research in more natural environments.

Second, it seemed to me that the work-load measures that have been used so far in interpreting research, analysis of interpreting performance or experts' judgements of interpreting performance, did not work well enough as indicator (see chapter 2.4.2). The fact that the interpreting performance seems to be stable independently of whether there was visual input or not, suggests that interpreters are able to adapt their attentional resources in order to maintain the quality of their translation. My idea was thus to substitute the traditional (more or less qualitative) analysis of the interpreting performance by physiological and narrowly defined linguistic indicators, like pupil dilation, fundamental voice frequency, silent pauses, cognate translations or text-related questions. Physiological indicators, in particular, provide a continuous and more sensitive measurement of work-load than errors or omissions in the interpretation and may thus unveil small effects that have been covered so far. As I was not sure whether the absence of audio-visual speech would increase work-load during simultaneous interpreting, I introduced a second "control variable": I overlaid the source speech with white background noise that is known to compromise simultaneous interpreting ((Gerver, 1974), see also chapter 3.3.4). This set-up would help to verify the effectiveness of the work-load indicators. If one of the work-load indicators I chose would neither react to the audio-visual speech nor to the addition of background noise, the indicator might not be suited to investigate simultaneous interpreting. If, however, it would react to the addition of white noise, but not to the audio-visual speech, it would rather suggest that the indicator works, but that the presence or absence of audio-visual speech has no effect on work-load during simultaneous interpreting (see chapter 3.3.3). The reader should bear in mind that work-load is a multidimensional phenomenon. The approach described above does not exclude the possibility that one indicator reacts only to specific types of

work-load, for example memory load, while another may react to a completely different type of stress, for example time pressure or emotional stress. In addition, both types are intertwined as high memory load can elicit more emotional aspects of work-load, like a feeling of frustration or pressure, anxiety and others which I do not cover in my work. Still, triangulating physiological and linguistic indicators measured during different experimental conditions should help to tap into the effect of audio-visual speech during simultaneous interpreting.

Based on the literature reviewed in the previous chapters, I expected that audio-visual speech would facilitate simultaneous interpreting or reduce work-load, while the addition of background noise should increase work-load. It should be noted that subjects usually adapt their resources to the demands: mental effort increases with work-load (see chapter 3.2.6). Hence, interpreters should invest more resources or show a larger effect of mental effort in interpreting when the speech was presented without lip movements and with white noise. The increased mental effort during the more difficult conditions (absence of visual input, addition of noise) should be visible at the physiological and linguistic scale. Consequently, I formulated the following hypotheses regarding audio-visual speech in simultaneous interpreting:

1) Pupil sizes are
    a. smaller during listening than during simultaneous interpreting[20] and
    b. smaller during simultaneous interpreting with audio-visual speech than without audio-visual speech. As described in chapter 3.4.3.1, pupils dilate when participants invest more effort to solve a task. If lip movements lower cognitive load by facilitating speech comprehension (see chapter 2.3.5) allowing participants to reduce their mental effort, then the

[20] The first part of this hypothesis only concerned the main study.

pupillary response should be smaller when visual cues are provided than without those cues.

2) Silent pauses are shorter during simultaneous interpreting with audio-visual speech than without audio-visual speech. The rationale behind this hypothesis goes along the same lines: Under cognitive load, speakers produce longer silent pauses. This difference can also be observed between a source speech and its corresponding orally translated target speech (see chapter 3.4.3.2). If audio-visual speech lowers cognitive load, silent pauses during simultaneous interpreting with audio-visual speech should be shorter than without audio-visual speech.
3) Fundamental voice frequency is lower during simultaneous interpreting with audio-visual speech than without audio-visual speech. As described in chapter 3.4.3.2, fundamental voice frequency may rise when participants are (emotionally) distressed. If audio-visual speech makes interpreters feel more comfortable (see chapter 2.4.2) or actually lowers their cognitive load (see chapter 2.3.5), fundamental voice frequency should be lower when interpreters work with audio-visual speech than without visual cues.
4) The number of correctly translated segments is higher during simultaneous interpreting with audio-visual speech than without audio-visual speech. As described in chapter 3.4.2, the more difficult a task, the more errors occur. If lip movements facilitate the task, there should be fewer errors than when visual cues are given.
5) The number of cognate translation is lower during simultaneous interpreting with audio-visual speech than without audio-visual speech. If lip movements enhance listening comprehension and reduce cognitive load (see chapter 2.3.5), interpreters can use more of their cognitive resources to monitor their output and avoid cognate translations (see chapter 3.4.2).

6) The number of correctly answered text-related questions is higher during simultaneous interpreting with audio-visual speech than without audio-visual speech. As described in chapter 3.4.2, enhancing listening comprehension by providing lip movements (see chapter 2.3.5) should leave more resources for deep level processing, thereby allowing for better recall of the speech content.
7) Speeches are perceived longer during simultaneous interpreting with audio-visual speech than without audio-visual speech. Under high cognitive load, participants perceive a task shorter than under low cognitive load (see chapter 3.4.1). If audio-visual speech reduces cognitive load (see chapter 2.3.5), interpreters should give shorter duration judgements for the speech than when lip movements are absent.

4.2 Pilot study

The pilot study included only a part of the hypotheses described above (chapter 4.1), in particular hypotheses 1b, 3 and 5. The aim was on one hand to test with a rather low number of participants the experimental design for potential flaws and on the other hand to ensure that some effect for audio-visual speech could be found at all given that previous research failed to demonstrate any effect for visual input in simultaneous interpreting (see chapter 2.4.2).

4.2.1 Experimental material used in the pilot study

The experimental material consisted of four speeches chosen from the basic level of the EU speech repository that makes test speeches available for candidates who prepare to be admitted as freelance interpreter at the European Union (see European Commission, 2017). The chosen speeches covered four different topics: air travel (hereafter: “air travel”), the Greek economic crisis (hereafter: “Greece”), work conditions (hereafter: “work”) and the demographic change (hereafter: “demographic change”). They were in great parts rewritten and edited in order to reduce

text complexity as far as possible and obtain a higher comparability between the speeches (for the full speeches, see appendix 7.2). The structure of the first paragraph[21] was the same across all speeches and served as "warm-up" for the interpreters. It contained the usual introductory expressions and announced twice the topic of the text. Words that did not belong to the 5000 most frequent words of American English (Davies, 2009) were substituted (word length *M*=4.63, *SD*=0.2). Potential problem triggers like numbers or proper names were omitted or replaced by a more general expression. Passive sentence constructions were omitted (with one exception: "born" in "Many children were born" was accepted, as it is the most frequent form of this verb). Long sentences were split up in order to obtain sentences with maximally one subordinate clause (words per sentences *M*=12.5, *SD*=2.2). The number of functional words (articles, prepositions and other words with a purely grammatical function) and type token relation served as indicator for information density. In every text, functional words made up approximately 40% of all words (ratio functional words: *M*=0.4, *SD*=0.03; type token relation: *M*=0.48, *SD*=0.05). Beyond this quantitative evaluation of information density, several measures were taken in order to reduce the information density at the textual level and to allow interpreters to catch up in case they missed the message. Essential messages were repeated in different words. Filler sentences or evident information without impact on the text coherence were introduced. All underlying logical relations within the text were made explicit by the use of conjunctions. Finally, each text was shortened to approximately 590 words (*M*= 588, *SD*=5.23).

[21] The structure of the first paragraph was as follows: "Ladies and Gentlemen, I am very pleased to be here at the international conference for [topic]. I am very honored to speak to so many distinguished guests who [short description with reference to the topic]. They are an example for all of us. Today, I want to talk about [topic]."

The speeches were read out by an American native speaker and recorded on video. All videos were .wmv-files with a resolution of 1920x1080 pixel. The speech rate was kept constant at a rate of 145 words per minute within and between texts. All videos were 3'30 to 3'50 minutes long. These videos were used to create a 2 x 2 factorial design: lip movements/ no lip movements x noise/no noise (see Table 2).

	Group 1		Group 2	
	Audio vs video		Audio vs video	
No noise vs noise	No lip movements – no noise Speech 1	Lip movements – no noise Speech 2	No lip movements – no noise Speech 2	Lip movements – no noise Speech 1
	No lip movements – noise Speech 3	Lip movements – noise Speech 4	No lip movements – noise Speech 4	Lip movements – noise Speech 3

Table 2: Experimental conditions.

The video (condition with visible lip movements, hereafter: video condition) showed the whole face of the speaker. In the audio condition (no visible lip movements), the video stream was replaced by a freeze frame of the speakers face (hereafter: audio condition). This method allowed to keep the screen brightness in all four conditions constant and to reduce light adaptations of the pupil. In the noise condition, white noise was added to the audio stream as to obtain a signal-to-noise ratio of 0 dB. In order to reduce potential speech-related effects, I created two counterbalancing groups and switched the speeches in the audio and the video condition. Moreover, I randomized the order of presentation of the conditions for each participant.

4.2.2 Participants of the pilot study

8 interpreting students in their final year at the Johannes-Gutenberg-Universität Mainz agreed to participate. Participants were assigned randomly to one of both groups and interpreted each of the four experimental texts. Participants received 10 € for participation. All

participants were Caucasian and had normal or corrected-to-normal vision and self-reported normal hearing. At the beginning of the experiment, none of them reported having taken any substance that could affect the pupillary reactions one hour prior to the experiment (eye drops, caffeine). One participant, though, said later on that she had been anesthetized by the dentist the same morning. Her data was not included in the analysis of the pupillary data. All participants confirmed to feel in good health.

4.2.3 Apparatus used in the pilot study

Pupillary data was recorded with a Tobii TX300 eye tracker with a collection rate of 120 hz. The eye tracker is fixed underneath a 23“-computer screen with a resolution of 1920x1080 pixels. The computer screen with the eye tracker was placed on a desk. Participants were seated in front of the eye tracker screen at approximately 60 cm distance where the eye tracker attains the highest accuracy according to the manufacturer. The chair was adjusted until the eyes of the participants appeared at the center of the calibration screen. No chin or forehead rest was used. The reasons for this choice were twofold: first, participants were required to speak during the experiment and a chin rest would compromise any articulatory movements. Second, previous experiences with chin or forehead rest at the laboratory of the Johannes Gutenberg-Universtät Mainz were unsatisfactory. Participants felt impaired by the chin and forehead rest, without any benefit for data accuracy or precision. According to Tobii Technology, the TX300 corrects online the raw pupil size for changes in gaze direction or distance to the eye tracker. On the basis of the pupil’s image, it constructs a 3D eye model for each eye that includes additional information about the distance between the eye tracker and the eye, the gaze direction, the horizontal and the vertical pupil size to estimate the pupil size (Tobii Technology, 2010; Tobii support, 2016). The pupil size is defined as the pupil diameter in millimeters.

The experiment was set up with Tobii Pro Studio, a software delivered by the manufacturer with the TX 300. The software allows the researcher to

record speech output with a microphone. But I found that the inbuilt recording function overlaid the source speech with the participant's translations so that in the end, neither the source text nor the target text was intelligible. Therefore, I used a second computer and recorded the translations with a separate microphone and the program audacity (Audacity, 2015). The ambient luminance was kept constant during the whole experiment and across participants.

4.2.4 Procedure of the pilot study

Participants received explanations about the procedure beforehand. Instructions guided them step by step through the experiment. Before starting the experiment, participants put on a headset. As it is usual for interpreters, participants covered one ear with the headphone and left the other one free so that they could hear their own voice. Participants were given the opportunity to adjust the volume to a level they felt comfortable with. The volume was kept constant throughout the entire experiment. The microphone was placed at approximately 5 cm to the participant's mouth and tested by pronouncing a few arbitrary sentences in order to ensure a good recording quality. Then, participants eyes were calibrated with a calibration method suited for Caucasian subjects. The experiment encompassed four blocks that each started with a black and white screen (30 seconds) to measure the pretrial pupil baseline. I was not sure how long the effect of work-load would last on. As the pupil reacts much stronger to luminance than to work-load, I used extreme luminance conditions, e.g. white and black screens, as pretrial baseline epoch in order to avoid potential effects of the preceding trial being carried on. Participants pressed a key to start the video and orally translated the speech they heard, while their pupil sizes were measured. Participants were instructed to look at the speaker's lips. After having orally translated the speech, participants were asked to rate the video and sound quality, the text difficulty, the speech rate and the clarity of the speaker's articulation on a scale from 1 (very good/easy/slow) to 4 (bad/difficult/fast).

For video quality, a fifth option was given: the number 5 on the scale stood for “static image”. The purpose of these last rating questions was to ensure that the experimental variables (audio/video and noise/no noise) were distinctive enough and that the speeches were not perceived as comparable in regard to their complexity and their speech rate, although both aspects were controlled for.

4.2.5 Data preparation and results of the pilot study

In the pilot study, I investigated participant’s ratings, pupil dilation, voice frequency and cognate translations as indicators for work-load during simultaneous interpreting in four experimental conditions: video – no noise, video – noise, audio – no noise, audio - noise. This chapter presents the results of the pilot study. The discussion of the results, especially with regard to the hypotheses described in chapter 4.1 follows in chapter 4.2.6. The results for the cognate translations are reported in Gieshoff (2017) and will not be described here. All statistical analyses were done with *R* (R Core Team, 2016), figures and graphs were done using the package *ggplot2* (Wickham, 2009).

4.2.5.1 Ratings

As there were only eight ratings for each experimental condition, no statistical analysis was conducted. The ratings are displayed in Table 3. They show that participants reliably distinguished both experimental variables (see video / sound quality). Text complexity was rated higher by five of the eight participants when noise was added to the source text. Whether this trend is statistically significant needs to be established with a larger number of participants.

	video / sound quality			**text complexity**			**speech rate**		
	mean	SD	mode	mean	SD	mode	mean	SD	mode
audio	5[22]	0	5	1,3	0,48	1	2,3	0,82	2
video	1,7	0,6	2	1,3	0,5	1	2,1	0,78	2
no noise	1,5	0,77	1	1	0	1	2,4	0,88	2
noise	3,9	0,32	4	1,6	0,52	2	2	0,67	2

Table 3: Participant's ratings of video and sound quality, text complexity and speech rate.

4.2.5.2 Pupil dilation

Pupil size measurements are prone to systematic errors because eye or head movements can alter the corneal reflection in a way that the eye-tracker gets a skewed image of the pupil (Brisson, et al., 2013). The Tobii TX 300 uses algorithms to correct these errors (see chapter 4.2.3). Nevertheless, I defined a region of interest around the lips of the speaker (visual angle 13°) and used only fixations within this area that obtained high validity scores and were successfully captured by the eye-tracker. The pupil size measures of both eyes were highly correlated ($r(1529500)=0.941$, $p<0.001$).

Next, I determined the baseline pupil size during the fixation period before each trial. Based on the pretrial baseline, I standardized each data point during the testing period, e.g. during simultaneous interpreting, for each participant and each trial using the following formula:

$$z=x-\mu/\sigma$$

[22] For video quality, a fifth option was given: the number 5 on the scale stood for “static image” (1=very good, 4=bad).

where x is the observed data point, μ the mean of all data points during the pretrial baseline epoch and σ the variance during the pretrial baseline epoch.

The standardization procedure reduces differences between participants and helps to obtain a near to normal distribution. Slight deviances to the normal distribution are probably due to the low number of participants. These data points were then fed into a linear mixed model. This approach is rather unconventional. Often, pupillary data is averaged over short epochs to determine the peak dilation (see for example Klinger, 2010). The reason for choosing a linear mixed model for the pupillary analysis was threefold: first, this method allowed me to take into account the whole testing period of about four minutes without giving too much weight to temporary changes in work-load that are common during simultaneous interpreting (see also Seeber & Kerzel, 2012); second, it can deal with repeated measures for participants; third, it can account for factors that cannot be controlled for by the experimental set-up, for example individual differences between participants in dealing with the different speeches or their physiological pupil size at rest.

I conducted a linear mixed model with the standardized pupil dilations as dependent variable[23]. Random effects included intercepts for speech and participant. One participant was excluded from analysis because of drug intake[24]. Fixed effects covered the visual presentation (audio/video), presence of noise, and the interaction of visibility of lip movements and presence of noise. P-values were approximated with maximum likelihood estimation. The model revealed main effects for the visibility of lip movements (*Estimate*=0.032, *SE*=0.001, *t*=26.08, *p*<0.001) and the

[23] Six participants are a rather low number for a linear mixed model. This is one of the reasons why I repeated the experiment with a higher number of participants (*N*=31).

[24] It appeared after the experiment that the participant had been anaesthetized by the dentist the same morning.

interaction of both variables (*Estimate*=0.027, *SE*=0.002, *t*=15.75, *p*<0.001). The main effect for the presence of noise failed to be significant. The model seems not to be very stable, as the effects plot (see Figure 3) suggests that there is no significant difference for any of the two experimental variables (noise/ no noise; audio/video) in contrast to the statistical model. This is probably due to the low number of items in each combination of variables.

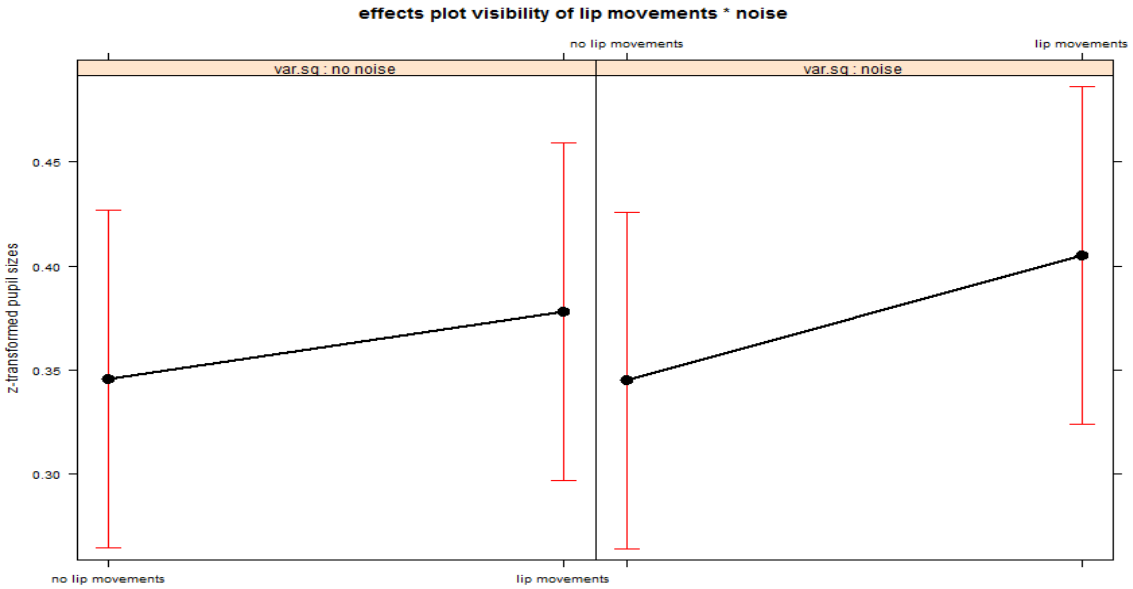

Figure 3: Fixed effects of both experimental conditions on pupil sizes

According to the model, the standardized pupil sizes were larger when interpreters worked with visible lip movements compared to the condition without visible lip movements. This increase was stronger when noise was added to the source speech than when no noise was added.

4.2.5.3 Voice frequency

Two participants were excluded due to missing recordings. Fundamental voice frequency was obtained with the program *praat* (Boersma & Weenik, 2013) and standardized the raw data for each participant and each trial. I conducted a linear mixed model on the standardized fundamental voice frequency as dependent variable. Random effects included intercepts for speech and participant. Fixed effects included visibility of lip movements,

presence of noise and the interaction of both variables. P-values were approximated with maximum likelihood estimation. The model revealed main effects for the presence of noise (*Estimate*=0.081, *SE*=0.036, *t*=2.249, *p*<0.05) and the interaction of both variables (*SE*=0.009, *Estimate*=0.04, *t*=4.128, *p<0.*05). The main effect for visibility of lip movements did not reach significance (*Estimate*= -0.016, *SE*=0.006, *t*= -2.573, *p*=0.81). According to the model, fundamental voice frequency rose by nearly 1 hz when white noise was added compared to an experimental condition without white noise (see Figure 4).

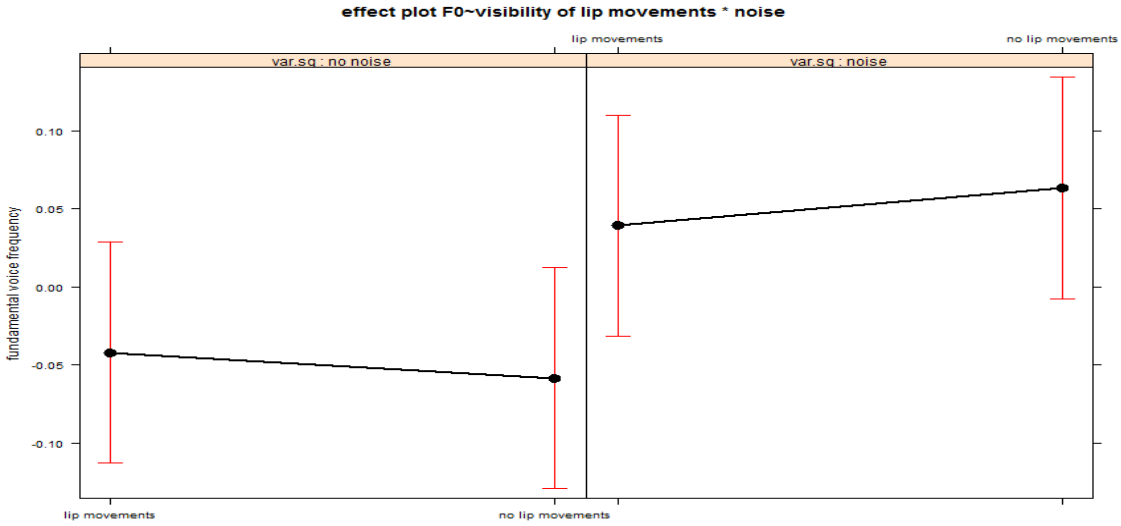

Figure 4: Fixed effects on the fundamental voice frequency of both experimental variables (lip movements, noise)

4.2.6 Discussion of the pilot study's results

Overall, the results of the pilot study are very heterogeneous and do not clearly support the hypothesis that audio-visual speech facilitates simultaneous interpreting. The presence or absence of audio-visual speech had no effect on fundamental voice frequency and only a very small effect on pupil dilation. If anything, the pupillary data shows rather the inverse trend than predicted: pupils dilated when lip movements were visible instead of constricting. Furthermore, pupil dilations during simultaneous interpreting with audio-visual speech were larger when white noise was added, which is in line with Kramer, Kapteyn & Feesten (1997).

At the same time, the number of cognate translations decreased with audio-visual speech, suggesting that the interpreters had more resources left for speech production (Gieshoff, 2017). It is possible that the pupil dilations are simply a reaction of arousal to the moving face. But this would not explain why the number of cognate translations went down when lip movements were visible. An alternative explanation that accounts for the effect on cognate translations could hold that interpreters used more of their resources for processes other than listening comprehension, like the monitoring of the speech production (cognate translations). Another potential candidate for pupil dilation is memory storage. For example, pupils become larger the more elements (letters, digits) a subject has to recall (Beatty, 1982). If audio-visual speech frees resources for storage, interpreters should be able to deliver a more accurate translation or to recall more precisely the content of the speech. Given the low number of participants it seems advisable to replicate these findings with more participants and to check whether a control group who just listens to the speeches, would show the same pattern.

The picture gets even more complex if we consider the effects of white noise. While the addition of background noise did not directly affect the pupil size or the number of cognate translations, it had a very significant effect on voice frequency. A maybe improbable though possible explanation is that the higher voice frequency during simultaneous interpreting with noise is caused by the Lombard effect. This effect denotes changes in speech parameters that occur when participants speak louder in a noisy environment (see Lombard, 1911). As a side effect of the increasing intensity, the fundamental voice frequency rises (Zollinger & Brumm, 2011; Lively, Pisoni, Summers, & Bernacki, 1993; Van Summers, Pisoni, Bernacki, Pedlow, & Stokes, 1988). In the case of the pilot study, the participants could have spoken up because they felt the need to drown out the background noise. This explanation is for several reasons improbable. First, interpreters are trained to control their voice and especially how loud they speak. They are aware of the fact that their

listeners (even if absent) would not hear the background noise and that there is no need to speak up. It can even be very unpleasant for the listener if the interpreter screams into the microphone. Second, the participants had no possibility to augment the volume during the experiment. If they had spoken too loudly, they would not have heard the source text anymore and would therefore have had to interrupt their interpretation. Still, it is not possible to completely rule out a Lombard effect as explanation for the increase of fundamental voice frequency.

Overall, the results suggest that voice frequency is sensitive enough to distinguish the condition noise/no noise. Yet, voice frequency did not show any main effect for audio-visual speech. It is possible that the effect of audio-visual speech is very small and will only reach significance with more participants. The overall volume participants chose at the beginning of the experiment might also play a role. Maybe a higher overall volume helps to discriminate speech sounds despite the background noise[25]. Another aspect that needs to be taken into consideration is individual differences in the way to deal with irrelevant sound. It might be that some participants had more difficulties ignoring the background noise and felt more perturbed by it than others. A solution to this problem may be to set the overall volume for all participants at the same level and to adapt the signal-to-noise ratio for each participant in a way that she recognizes 75% of the speech. Finally, previous research on speech related parameters (see chapter 3.4.3.2) associates voice frequency primarily to emotional stress. Even if the assumption that work-load triggers a stress reaction holds, the reaction to speech masking noise might be much stronger than to the absence of lip movements which provide less useful information for understanding the speech than the auditory input. Not understanding the source text may be very destabilizing for interpreters. It would be therefore interesting to check with a larger pool of participants if pupil dilation or

[25] Unfortunately, the overall volume was not recorded during the pilot study.

performance-based measures, such as cognate translations, show the same pattern as voice frequency and if changes in voice frequency can be ascribed to changes in work-load. This would, however, not rule out the possibility that changes in voice frequency are primarily induced by a more general feeling of emotional stress and that work-load is only one among many other causes. Another interesting question therefore, though beyond the scope of the present paper, would be the relationship of voice frequency and measures of stress perception or anxiety (for example by using testing methods like NASA-TLX, STAI).

In conclusion, the results did not support the hypothesis that audio-visual speech facilitates simultaneous interpreting, nor did they show a clear pattern. Pupils dilated more in audio-visual than in the auditory-only condition, whereas the number of cognate translations decreased when audio-visual speech was provided which could indicate that participants increased their effort to monitor their translation when audio-visual speech was provided. Overlaying the speech with background noise led to an increase in fundamental voice frequency suggesting that participants were more distressed when interpreting with than without background noise. If these tentative explanations are true, pupillary responses, cognate translations and fundamental voice frequency are mediated by different types of stress. In this case, pupillary responses would reflect mental effort and cognate translations would benefit from that additional effort, whereas fundamental voice frequency would depend on a general feeling of distress. However, neither pupillary responses nor cognate translations were affected by the addition of background noise that was expected to impede listening comprehension and thus to increase work-load in simultaneous interpreting (see chapter 4.1). In addition, the pilot study revealed some shortcomings which might provide alternative explanations for the results of the pilot study. These shortcomings include in particular a possible Lombard effect which may have caused the fundamental voice frequency to rise and differences within participants in dealing with noise. Furthermore, it is not clear if the effect of audio-visual speech on pupil

dilation is task-specific for simultaneous interpreting or if it holds for speech comprehension in general. Finally, the low number of participants and the noise in the pupillary and the voice frequency data (residual variance: 0.991) might have covered further effects. This makes it difficult to validate pupillary responses, fundamental voice frequency or cognate translations as indicators for work-load in simultaneous interpreting or to evaluate the impact of audio-visual speech on simultaneous interpreting.

In order to address those shortcomings, I decided to conduct a follow-up study (main study) with a larger number of participants to confirm the results of the pilot study and reveal further effects that might not have been visible in the pilot study. In order to further investigate the differences between mental effort, cognitive load and emotional stress, I decided to introduce two further performance-based measures for work-load in addition to cognate translation: 1) translation accuracy and 2) questions about the speech content that followed each speech (text-related questions). If performance based measures improve during simultaneous interpreting with audio-visual speech, e.g. if I observe an increase in translation accuracy and of the number of correct answers to text-related questions and a decrease in cognate translations, this suggests that audio-visual speech enhances the performance in simultaneous interpreting. In this case, larger pupil dilations during simultaneous interpreting with audio-visual speech mean that participants put more mental effort into the task, for example to monitor their output or to memorize the speech content, rather than simple arousal in response to a moving face. The same trend should be observed for simultaneous interpreting without background noise compared to simultaneous interpreting with noise added to the speech: Without background noise, translation accuracy and number of correct answers to text-related questions should increase, while the number of cognate translations should go down. However, if the performance-based measures react only to the auditory presentation (noise/no noise) and not to audio-visual speech, while pupillary responses are still larger during simultaneous

interpreting with audio-visual speech, the conclusion will have to be that audio-visual speech has no (or only a very weak) effect on cognitive load, but that participants are more aroused when they see a moving face. Moreover, I decided to add a control group of student translators who merely listens to the speeches and answers to the text-related questions instead of translating the speeches[26]. If audio-visual speech facilitates listening comprehension in general in a way that additional resources can be used for memory storage, effects on pupil dilation or response accuracy to text-related questions should be visible the same way in both groups, listeners and interpreters. However, if the effect of audio-visual speech is task-specific, it should be observed only in interpreters and not in translators.

Another issue that emerged in the pilot study was a possible Lombard effect that could be responsible for the increase in fundamental voice frequency. In order to prevent a Lombard-effect during interpreting in noise, I decided to keep the overall volume in the main study at a rather low level (approximately 40 to 50 dB) and constant for all participants. In addition, I decided to gather voice intensity data from each participant. A positive correlation of voice intensity and fundamental voice frequency would suggest that the increase in fundamental voice frequency can be explained with voice intensity rather than cognitive load or emotional stress. Keeping an overall low volume during the experiment and keeping it constant for all participants has another advantage: In the pilot study, it was not possible to exclude that some participants were less sensitive to noise because they had set the volume higher at the beginning of the experiment compared to others. Setting the same overall volume for all participants would ensure that the signal-to-noise ratio would have the

[26] Choosing interpreters as control group would have been, of course, the better choice in terms of comparability. Given the low number of interpreter students each year, this option did not seem realistic.

same masking effect for all participants. However, it is possible that some participants had fewer difficulties than others dealing with background noise, e.g. some participants might have been less affected by noise than others. I therefore decided to administer a word detection test with different levels of background noise to the participants before starting with the speeches to determine the speech-to-noise ratio where participants correctly identify 75% of all words and to apply the participant's specific speech-to-noise ratio to the speeches. This way, all participants should be affected to the same extent by the addition of background noise. Moreover, the word detection pretest would make it possible to test the effect of masking noise on pupil dilation and detection performance on single words, thereby confirming the effect of noise on listening comprehension.

4.3 Main study

In the follow-up study, I integrated the shortcomings that became apparent in the pilot study (see chapter 4.2.6). The main study was designed for 32 participants, 16 interpreters and 16 translators, however, only 31 participants could be recruited. It consisted of two parts: a pretest to confirm the effect of noise on listening comprehension and to adjust the noise level to the participant's individual 75% detection threshold and a main part with four speeches that participants either interpreted or listened to in order to investigate the hypotheses described in chapter 4.1.

4.3.1 Experimental material used in the main study

The speeches for the main part were the same as in the pilot study, but I slowed them down to obtain a video length of 4 minutes and a speech rate of about 140 words per minute.

For the noise pretest, I selected 64 stimulus words among the 5000 most frequent words of American English (Davies, 2009). All words were concrete, monosyllabic and did not start with a plosive or a fricative because tests beforehand had revealed that word boundaries, fricatives

and plosives were easier to detect than other phonemes. As it was not clear whether the cognate status of a word would have an effect on the ability to recognize a word, I decided to use a mixed list of stimuli with 32 cognates to their German counterparts and 32 non-cognates. For each word, I identified the most frequent translation on the online dictionary dict.cc (Hermetsberger, 2002) that shows how often a translation has been suggested by the online community (access: 29/07/2016). Appendix 7.1 lists all stimuli with their respective translations and how often this translation has been suggested by the online-community. Using the package *stringdist* (van der Loo M. J., 2014) in *R* (R Core Team, 2016), I calculated the Levenshtein distance between the (orthographic and the phonetic form of a) word and its translation e.g. the number of deletions, additions or transpositions necessary to obtain the translation. Finally, I compared the Levenshtein distance of the cognate and non-cognate stimuli. As the distance measurements did not meet the normality assumption of a one-way anova, I conducted a Kruskal-Wallis rank sum test. The mean ranks of cognates and non-cognates were significantly different (orthographic form: $F(1)=45.067$, $p<0.001$; phonetic form: $F(1)=42{,}135$, $p<0.001$). The Levenshtein distance for cognate and non-cognate pairs is depicted in Figure 5.

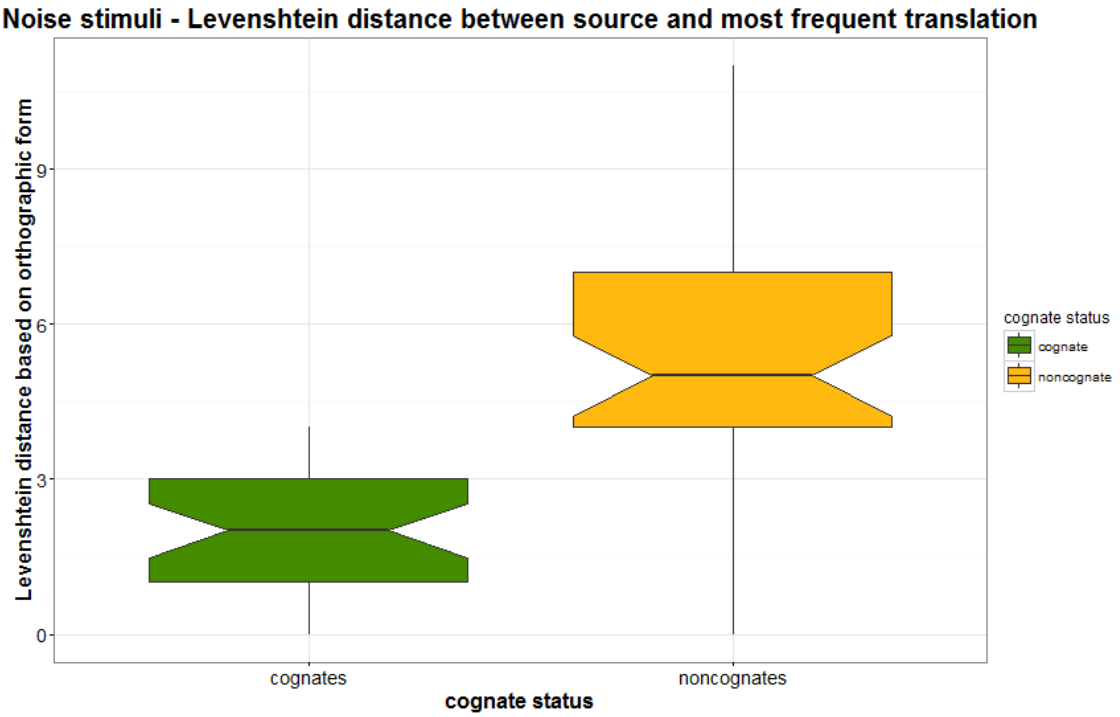

Figure 5: Levenshtein distance between the orthographic form of the English stimulus and its German translation of cognates and non-cognates. The Levenshtein distance expresses the number of deletions, additions or transpositions necessary to obtain the translation.

For most stimuli, the cognate status was unambiguous: the cognate translation was either inexistent or by far the most frequent translation. In some cases, however, a cognate translation existed, but was not the most frequent translation. Critical cases are displayed in Table 4. Given that the cognate translations were more frequent for the words *bar* and *neck*, these words were treated as cognates, whereas in the latter two cases, the words *wall* and *gate,* the cognate translation was less frequent than the non-cognate translation. These two words were thus treated as non-cognate stimuli.

English word	Most frequent translation (german)	Cognate translation (german)	Cognate status
bar	Balken (13, 669)	Bar (11, 316)	cognate
neck	Hals (13, 4768)	Nacken (11, 20)	cognate
wall	Wand (10, 2701)	Wall (11, 11)	non cognate
gate	Tor (8, 4048)	Gate (14, 40), Gatter (16, 0)	non cognate

Table 4: Critical cases of cognate pairs. The frequency class according to the Leipzig corpus (first number) and the number of suggestions in dict.cc (second number, accessed August, 18th 2017) are given in parentheses after each possible translation.

All stimulus words were read out by the speaker of the four speeches for the main part, an American native, and recorded on audio with *Audacity* (Audacity development team, 2015). Each sound stimulus was about one second long (*M*=1.05 seconds[27], range: 0.9-1.27 seconds). Finally, I normalized all word stimuli and all speeches to -1 dB using the program *Audacity* (Audacity development team, 2015).

4.3.2 Participants of the main study

Thirty-one subjects signed a written informed consent for participation and data collection. All participants were Caucasian. Seventeen were translator and fourteen were interpreter trainees at the Johannes Gutenberg-Universität Mainz. They were divided into two groups according to the cursus they followed (listeners and interpreters) and did a slightly different experiment. Whereas the interpreter group orally translated the four speeches in the main part, the listener group simply listened to the speeches, without any concurrent secondary task. The listener group worked as a control group in order to see if the results

[27] A mean duration of roughly one second seems long for single words. In fact, the sound clips were cut as to leave approximately 200 ms before and after the word.

obtained in the interpreter group would be task-dependent or not[28]. The translator trainees (listeners) were in their third or fourth year, while the interpreter trainees were in their final fifth year. Given that the interpreter trainees only specialize in conference interpreting from their fourth year on, interpreters had about as much experience in interpreting as the translator trainees in translating. Participants were randomly assigned to four groups. The purpose of these groups will be explained below. All participants confirmed to feel good and being in good health and received 10€ or ECTS points in exchange for their participation.

4.3.3 Apparatus used in the main study

The same eye tracker as in the pilot study, a Tobii TX300, was used, but the experiment was programmed with Psychopy 2 v1.83.01 (Peirce, 2007). Details about the experimental procedure are given in the next chapter below.

4.3.4 Procedure of the main study

Participants received explanations about the tasks (typing words, interpreting or listening to text, rating, estimating the speech duration, answering to text-related questions) and the data that was about to be collected (key presses, pupillary data, recordings) beforehand. When the participants were comfortably seated in front of the eye tracker, they put on the headset (one ear covered) and adjusted the microphone of the headset in front of their mouth (about 5 cm from the mouth). The microphone was tested by pronouncing a few arbitrary sentences to ensure a good recording quality. This time, the volume was kept constant across participants and during the whole experiment. After the calibration procedure of the eye tracker, I started the screen recording of the eye tracker and the actual experiment in Psychopy. Instructions in German

[28] Given the low number of interpreter trainees each year, it was not possible to recruit interpreting trainees for the control group.

guided the participants through the experiment. The first part was a noise pretest to adjust the noise level to the participant's individual 75% detection threshold. For this purpose, 64 stimulus words were aurally presented to the participant and masked with white noise at four different signal-to-noise ratios. The participant's individual 75%-threshold would later be used for the speeches. A threshold of 75% was chosen (instead of a 50%-threshold) to ensure that the interpreting performance was not totally interrupted. Indeed, speeches became nearly unintelligible during simultaneous interpreting at a 50%-threshold. Participants were only told that they would hear common English words. No further information about the type of word or context was given. I opted for words only, instead of sentences, in order to avoid any context effect that could help to identify a word. Each trial began with an instruction informing the participants about the keys to press to enter, correct or confirm their answer. When the participants pressed the space bar, a fixation cross appeared. Two seconds later, a sound file with the corresponding noise level was played. The fixation cross only disappeared after the sound file had finished. In Psychopy, the volume of a sound object is relative (as the actual volume depends for example on the speakers or the computer settings) and can be set on a scale from 0.0 (silent) to 1.0 where 1.0 is the maximal volume of the soundcard. For this reason, the signal-to-noise-ratio is not given in dB, but as ratio based on Psychopy's volume scale. While the volume of the stimulus was kept constant at level 0.1 (10% of the maximal sound level of the sound card), the volume of the noise augmented from 0.1 to 0.4 (thus 10% to 40% of the maximal sound level). Stimuli presented at a noise level of 0.1 could in general be identified without problems, whereas stimuli at a noise level of 0.4 were barely audible. In order to present the stimuli at different signal-to-noise ratios, I created four lists that assigned each stimulus randomly to one of the four noise levels. The order of presentation was the same for each participant, but as participants heard every word only once, the order of the stimuli and the order of the noise level were not predictable. So each participant heard each stimulus in the

same order, but at different noise levels. Participants responded by pressing the space bar and entered their answer that appeared on the screen while typing. In case of a spelling error, they could correct their answer using backspace.

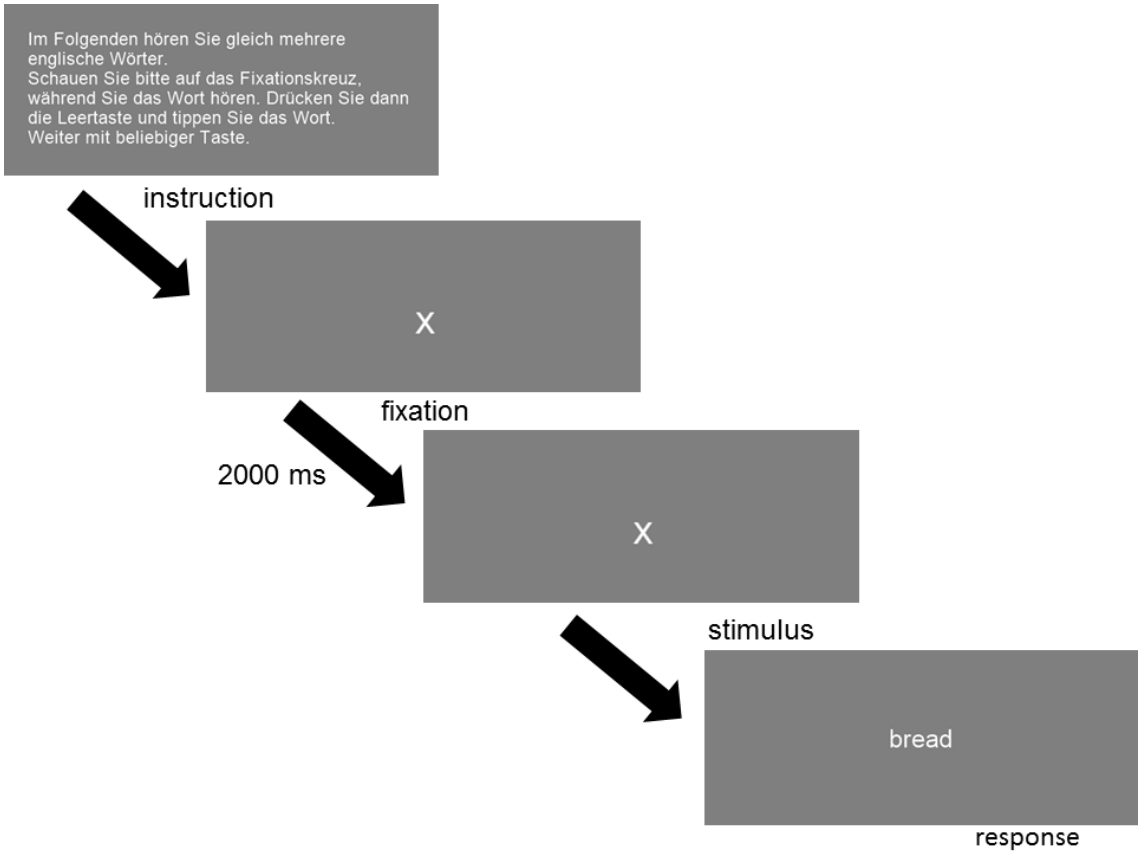

Figure 6: Procedure for the noise pretest. Participants read the instruction and fixated a cross at the center of the screen during 2000 ms. After the stimulus was played, participants entered their response.

After confirming the answer by pressing enter, an algorithm checked if the answer was correct and calculated the number of correct answers for each noise level and identified the noise level where the participant recognized 75% of the words correctly. The resulting noise level was applied to the speeches in the main part of the experiment. If the participant failed to recognize a sufficient number of stimuli, the lowest noise level was applied. The pretest took between seven and fifteen minutes. As the pilot study did not reveal a main effect of noise on pupil dilation, I collected pupillary data during the pretest in order to check whether decreasing

signal-to-noise-ratios, e.g. increasing noise levels would lead to larger pupil dilations in a word recognition task.

After the last stimulus, the main part of the experiment began. It consisted of four blocks that each started with a fixation to measure the baseline pupil size, a speech presented in one of the four experimental conditions, a question about the speech duration, four rating questions and five questions that were related to the speech. In order to account for potential differences between the speeches, I created two groups and switched the speeches in the audio/video-condition. It was not possible to completely randomize the conditions over the four speeches as this would have required a much higher number of participants. Participants (interpreters and listeners) were randomly assigned to one of the four groups. In all groups were one or two interpreters and two listeners. Table 5 shows the speeches in each experimental condition according to the group.

	Group 1		Group 2	
	Audio vs video		Audio vs video	
No noise vs noise	No lip movements – no noise	Lip movements – no noise	No lip movements – no noise	Lip movements – no noise
	Speech “demographic change”	Speech “air travel”	Speech “air travel”	Speech “demo-graphic change”
	No lip movements – noise	Lip movements – noise	No lip movements – noise	Lip movements – noise
	Speech “Greece”	Speech „work”	Speech “work”	Speech „Greece”

Table 5: Speeches in each experimental condition according to the group in the main part of the experiment.

Participants were first instructed to fixate for ten seconds both a cross on a black background and a cross centered on a picture of the speaker. The picture of the speaker instead of a white screen was chosen in order to avoid light adaption of the pupil during the speech. A second instruction followed and asked the participants according to the task they were assigned to, to either 1) interpret the following speech (interpreters) or 2) listen to the following speech (listeners). After pressing the space bar, the movie with the speech was displayed. Depending on the condition, the

movie was either a video with visible lip movements or an audio track with a freeze frame of the speaker (and thus without visible lip movements). In the noise condition, white noise at the predefined volume was played concurrently to the movie. For the interpreters, the translation was recorded during the movie. The recording ended 10 seconds after the movie so that participants had enough time to finish their sentence.

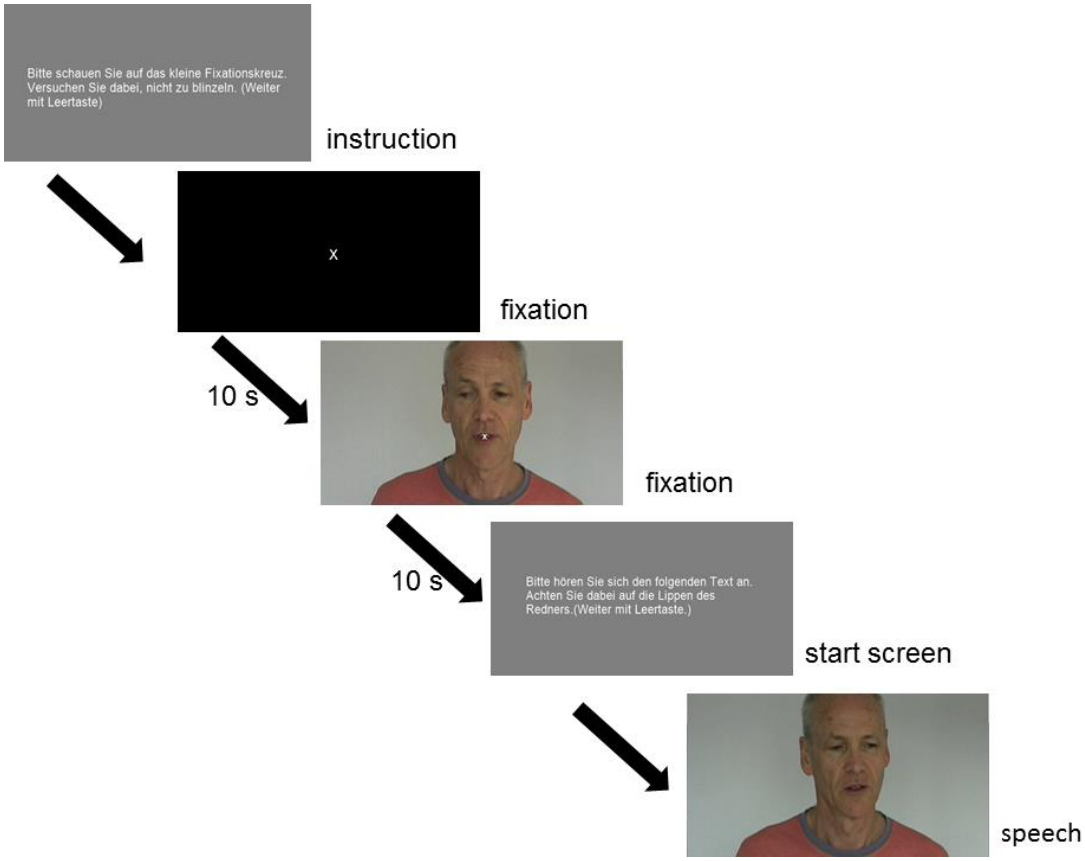

Figure 7: Procedure of the main experiment (first part). Participants fixated during 10 seconds a cross on black ground and on the speaker's image. Then, participants translated or listened to the speech.

Following the speech, participants were asked to estimate the duration of the speech (the duration of the speeches had not been told beforehand). For this purpose, they moved a cursor on a timeline from 0 to 8 minutes with the arrow keys until they reached the duration they deemed appropriate and confirmed their answer by pressing enter. The number of minutes and seconds according to the location of the cursor was displayed below the timeline. Research summarized by Block, Hancock and Zakay

(2010) in their review showed that effects of duration judgements were only found when the judgements were made immediately after stimulus presentation. Then, participants rated the video and audio quality, the speech complexity and the speech rate on a scale from 1 (very good/very easy/very slow) to 4 (bad/ difficult/ fast) by pressing the letter (a, b, c, d) that corresponded to the answer. All rating questions were presented in German.

Subsequently, participants answered questions that were related to the content of the speeches. Participants were told beforehand that they would be asked to answer to text-related questions after each speech. For each speech, five questions with three possible solutions, numbered from *a* to *c*, were given. The questions and solutions were presented in German for several reasons. First, all instructions were in German. Second, I expected that the linguistic transfer would prevent literal memory effects and that correct answers therefore should reflect real understanding – or deep processing - of the speech. The solutions were all conceivable and structurally similar, for example, they had the same number of items in an enumeration (for the questions and the choices, see appendix 7.3). Aside from the three possible solutions, a fourth choice (*d*) was given: “I don’t know.” The participant’s answers were again saved to a text file. When all questions were answered, a new trial started. Finally, after four trials, the final instruction appeared and participants closed the experiment by pressing escape. This key press also terminated the screen recording of the eye tracker. Before participants left the laboratory, they were asked to

report unexpected problems or difficulties and to give a general feedback.

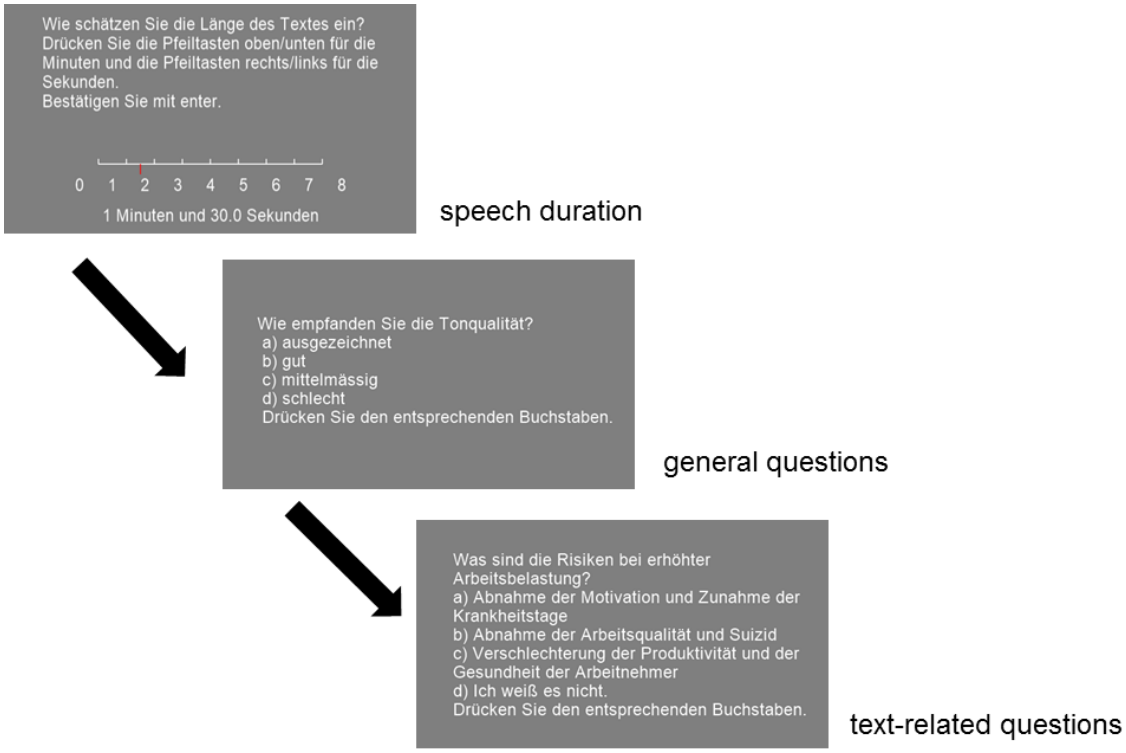

Figure 8: Procedure of the main experiment (second part). After translating (or listening to) the speech, participants estimated the speech duration, rated general parameters and answered to text-related questions.

4.3.5 Data preparation and results of the pretest

The aim of the pretest was twofold: first, to confirm that pupils dilate during word identification with increasing levels of masking noise; second, to set the volume of the background noise to a level where participants would recognize 75% of the stimulus words. Three variables were statistically investigated: response times, response accuracy and pupil dilation during trial. Statistical analysis was done in *R* version 3.3.2 (R Core Team, 2016), figures and graphs were done using the package *ggplot2* (Wickham, 2009). This chapter describes the data preparation and the results of the pretest and discusses the results. The general discussion of the complete results of the whole main study follows in chapter 4.4.

4.3.5.1 Response time

Each participant listened to 64 stimuli and pressed the space bar as soon as she or he identified the word. Response time was defined as the number of milliseconds between the moment when the sound file ended and the moment where the participant pressed the space bar to enter his response. The mean response time was 1726 milliseconds (*MD*=1301 ms, range: 228-14440 ms, *SD*=1.36). It appears that some trials lasted extremely long. This concerns mostly the first few trials and is largely due to the fact that participants still asked questions after the experiment had started. Therefore, I removed for each participant all trials that lasted longer than 1.5 standard deviation of the mean duration of all trials of that participant. On the average, about 4 trials were removed per participant (range: 1-11 trials, first quartile: 3 trials, third quartile: 5 trials). A boxplot of the response times of each participant suggested longer response times for participant P8T and a larger variance compared to the rest of the participants (see Figure 9). Indeed, an ANOVA on response time as dependent variable and participant as independent variable yielded significant differences between participants ($F(30, 1816)=16.68$, $p<0.01$). A post-hoc Tukey test confirmed this visual impression: P8T differed significantly from all other participants at $p<0.01$ (see appendix 7.4.1 for complete results). P8T was therefore removed from further analysis. Further significant differences between some of the participants were found in the Tukey test, but none was different from all other participants. In total, 127 trials (9.9%, 32% for participant P8T) were removed.

Method and results

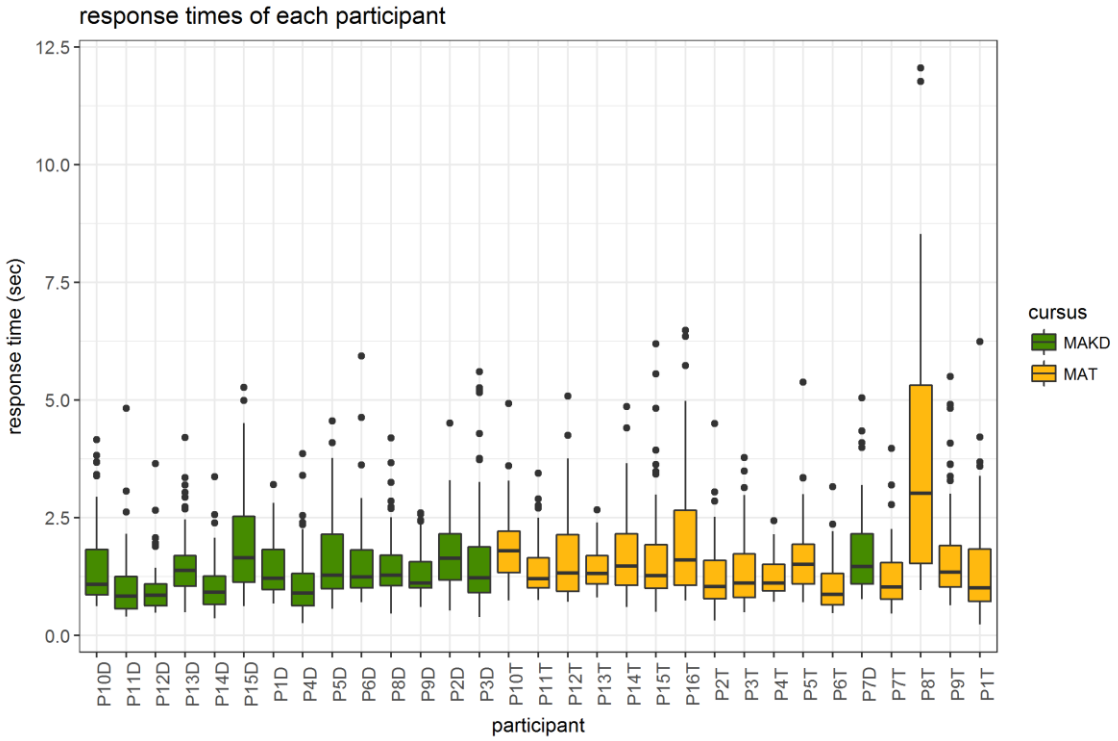

Figure 9: Boxplot of observed response times of each participant. Task is coded by color: Green boxes belong to interpreters (MAKD), yellow boxes stand for listeners (MAT).

After removal of participant P8T and trials with extremely long response times, the mean response time was 1502 ms (range: 228-6480 ms, *SD*=0.9087). Response times were subsequently log-transformed to reach a near-to-normal distribution.

A glance on Figure 9 also suggests that the difference between the response times of interpreters (green boxes, “MAKD”) and listeners (yellow boxes, “MAT”) is – apart from participant P8T – rather small. In fact, the observed mean response time is nearly the same for both groups: 1459 ms for interpreting trainees (*SD*=0.8875) and 1546 ms for listeners (*SD*=0.9375).

Figure 10 depicts the mean response time for each stimulus. The cognate status is coded by color: green boxes are cognate stimuli; yellow boxes are non-cognate stimuli. As can be seen, the inter-stimulus variance is much larger than the inter-participant variance. However, none of the stimuli seems to stand out particularly.

Method and results

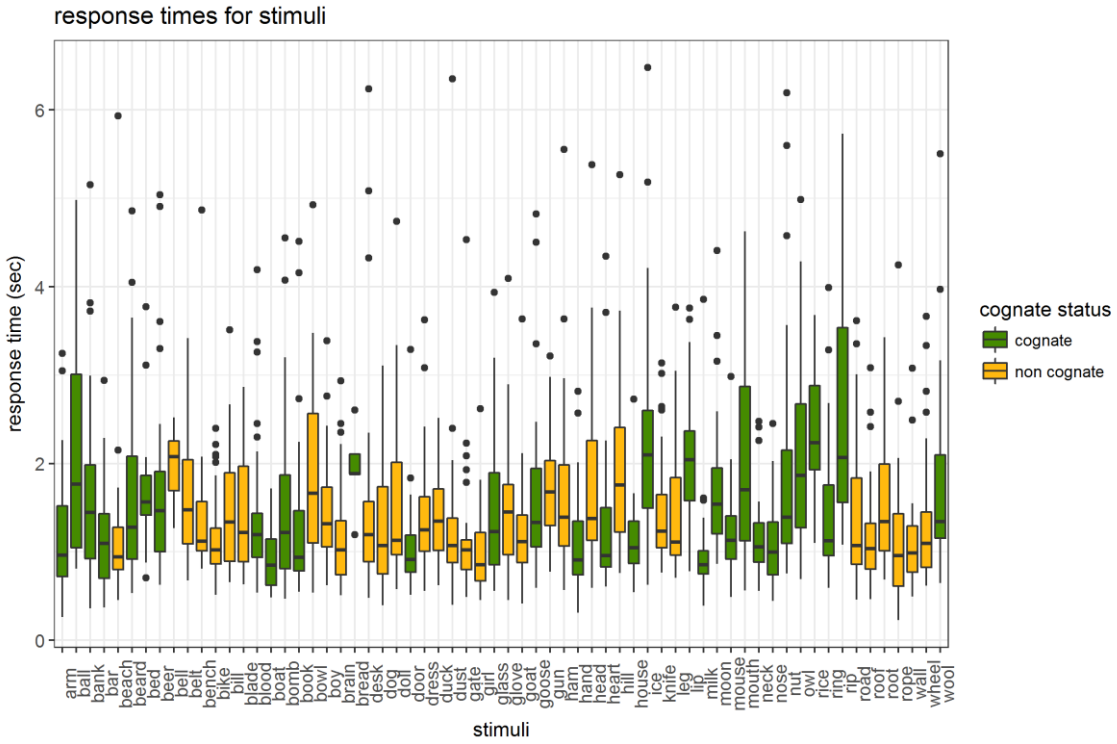

Figure 10: Boxplot of observed response times for each stimulus. The cognate status is coded by color: green boxes are cognate stimuli, yellow boxes are non-cognate stimuli.

The response times for cognate and non-cognate stimuli are depicted in Figure 11. The mean response time for cognate stimuli is 1570 ms (*SD*=0.9842). For non cognates, the mean response time is 1438 ms (*SD*=0.8241).

Method and results

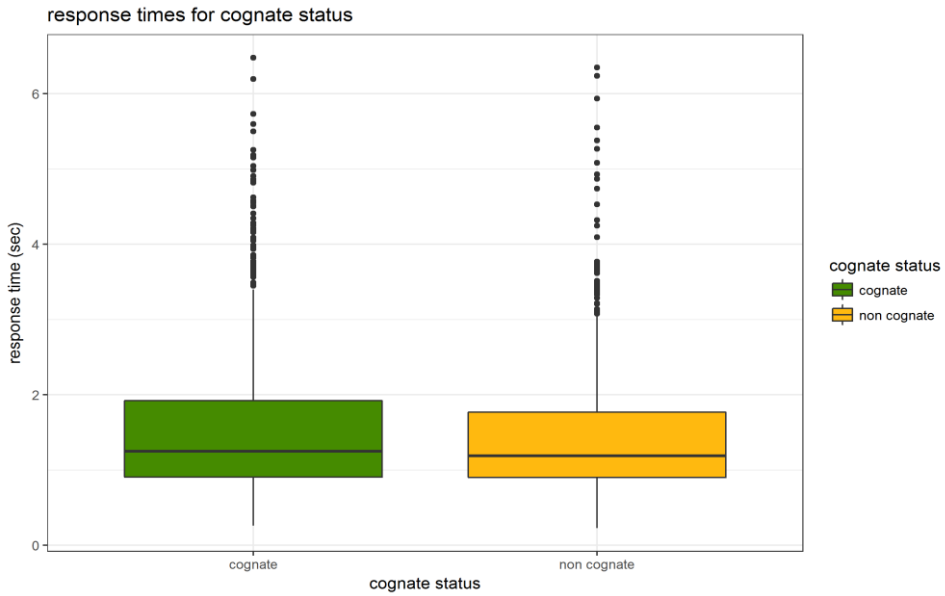

Figure 11: Boxplot of mean response times depending on cognate status. The boxplot depicts the observed mean response time for cognate stimuli (green box) and non-cognate stimuli (yellow box).

According to Figure 12, the mean response time increases with increasing background noise. At noise level 0.1, the mean response time is 1290 ms (*SD* = 0.7816). This value rises with increasing noise level until it reaches 1740 ms (*SD* = 0.9967) at noise level 0.4.

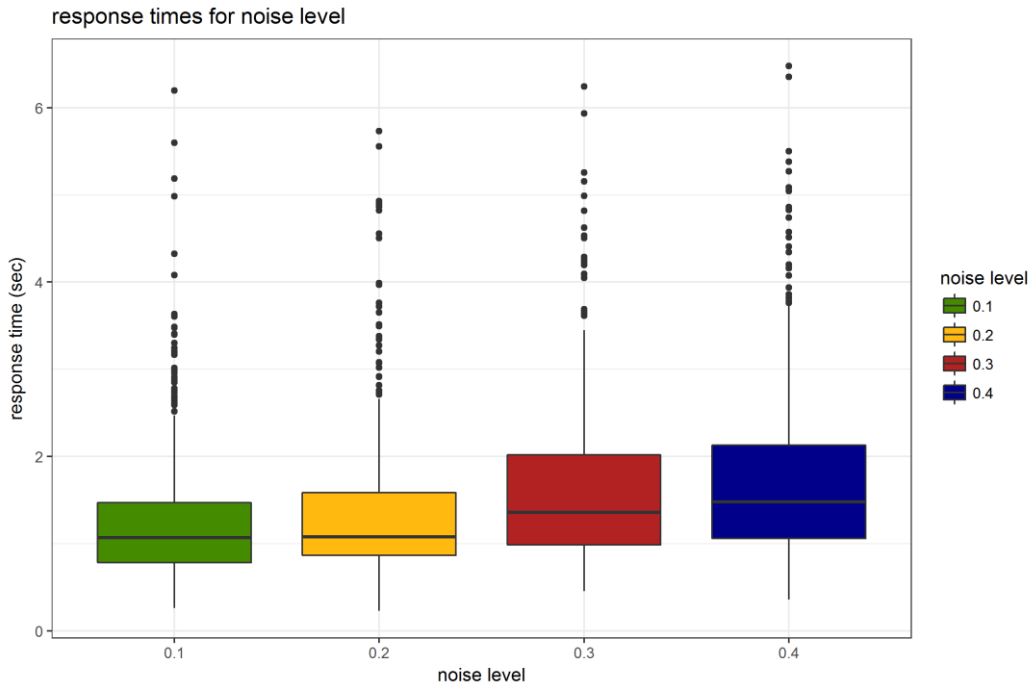

Figure 12: Boxplot of observed mean response times for each noise level.

Method and results

Response time was shorter for correctly identified stimuli than for wrong responses. Figure 13 shows the response duration for correctly and wrongly identified stimuli.

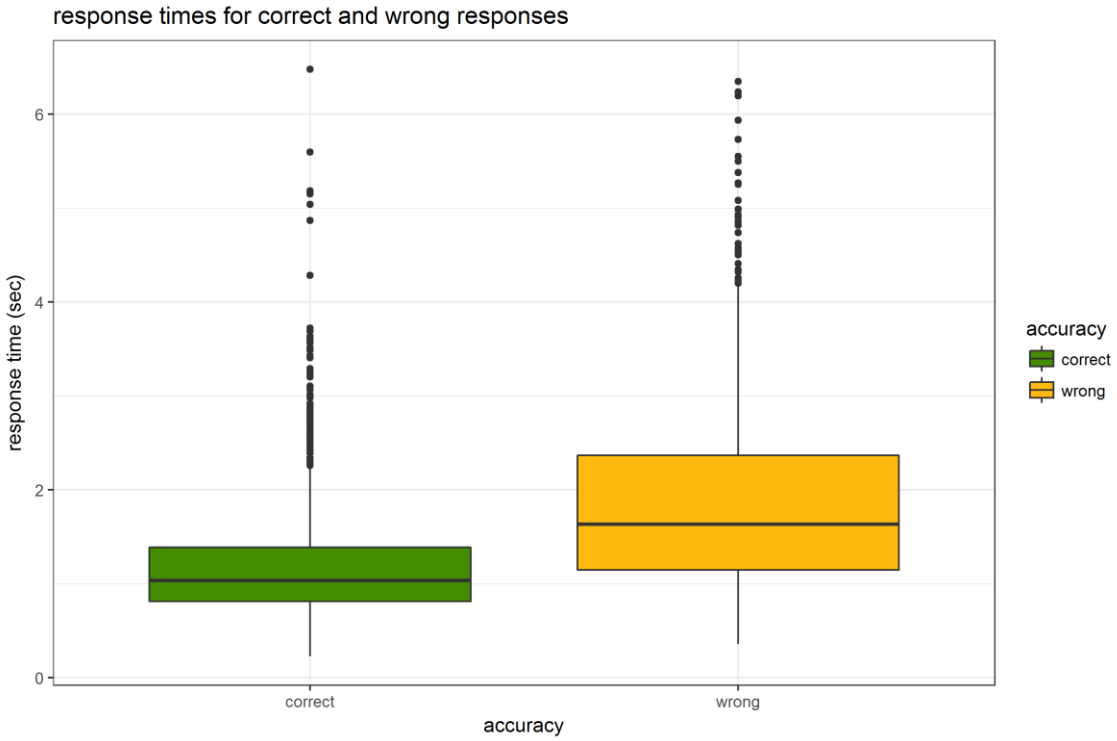

Figure 13: Boxplot of response duration for correct (green) and wrong (yellow) responses.

A linear mixed model on log-transformed response time was constructed using the R-package *lme4* (Bates, Maechler, Bolker, & Walker, 2015) and fit by maximum likelihood approximation. Random effects included intercepts for each participant and for each stimulus[29]. Fixed effects included noise level and response accuracy. Cognate status ($F(6, 1)=2.488$, $p=0.115$) and task ($F(6, 0)=0.00$, $p=1.0$) failed to reach significance when comparing models with and without those effects by maximum likelihood approximation.

The standard deviation was respectively $SD=0.167$ for stimuli and $SD=0.188$ for participants. P-values for fixed effects were obtained by

[29] The model failed to converge with trial slopes for each participant. This is not surprising as there was only one observation for each trial.

comparing the model with and without the effect in question using maximum likelihood approximation. Noise level was highly significant ($F(8, 2) = 31.379$, $p<0.001$). The response duration for stimuli presented at noise level 0.4 were 150 ms longer than for those presented at noise level 0.1. According to the model, the response time decreased with decreasing noise level. Response accuracy was also highly significant at $p<0.001$ ($F(5, 0) = 246.822$). After exponentiation of the log-transformed model estimate, the response time for correct responses decrease compared to wrong responses by approximately 852 ms. The model estimates are summarized in Table 6. The fixed effects of the model are depicted in Figure 14. The left facets shows the estimated decrease of the log-transformed response time for inaccurately and correctly identified stimulus words, the right facets shows the increase of the log-transformed response time with increasing noise level.

		Estimate	***Standard Error***	***t***
Noise level	0.1	-0.065	0.018	-3.628
	0.2	-0.055	0.017	-3.235
	0.3	0.044	0.017	2.57
Response accuracy	wrong	-0.161	0.013	-12.554

Table 6: Model estimates of the linear mixed model on response times. The estimates correspond to the estimated effect of noise level and response accuracy on log-transformed response time.

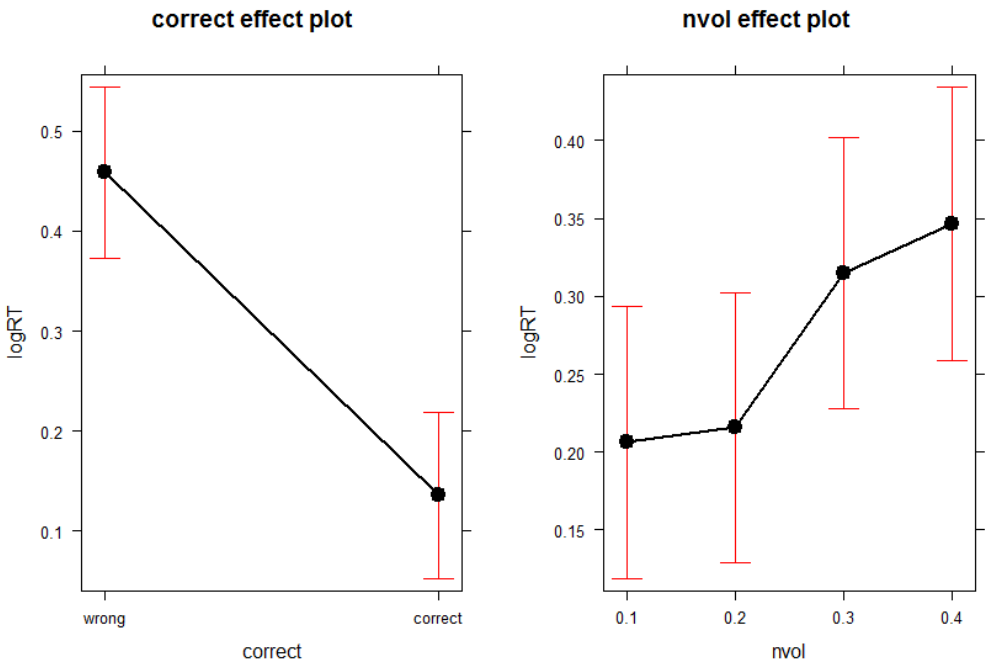

Figure 14: Effect plots for response time as a function of response accuracy and noise level. The left facets shows the estimated decrease of the log-transformed response time for inaccurately and correctly identified stimulus words, the right facets shows the increase of the log-transformed response time with increasing noise level.

4.3.5.2 Response accuracy

Each of the 31 participant listened to 64 stimulus words, which makes a total of 1984 observations. 1093 responses were correct, 891 responses were wrong. Participants identified about 50% of the stimuli correctly. This percentage was relatively stable across participants (*M*=0.5, *MD*=0.5, range: 0.35-0.66, *SD*=0.08). Participants were distributed among four groups. Stimuli were always presented in the same order, but the level of noise with which a stimulus was presented varied according to the group, e.g. each stimulus was presented (nearly)[30] equally often in all four noise levels. Again, in all four groups approximately 50% of the stimuli had been correctly identified (*M*=0.48, *MD*=0.46, range: 0.41-0.58, *SD*=0.06).

[30] In order to achieve a really equal distribution of the stimuli across the four noise levels, one more interpreting trainee would have been needed. Unfortunately, one interpreting trainee had to be excluded because of recording errors.

Nevertheless, the percentage of correctly identified stimuli varied largely between stimuli (*M*=0.5, *MD*=0.5, range: 0.0-1.0, *SD*=0.293). Some stimuli were rarely correctly recognized, whereas others were correctly identified by almost all participants. Stimuli that had been identified in less than eight participants (first quartile) are: ball (3)[31], bell (1), blade (5), bowl (4), bread (1), ham (5), head (6), leg (4), lip (5), moon (0), owl (7), rice (7), rip (5), wool (2). Stimuli that had been correctly identified by more than 23 participants (third quartile) are arm (24)[32], bench (25), bike (30), boat (31), book (25), door (31), dress (28), duck (26), dust (27), gate (26), girl (28), house (31), milk (30), neck (28), nut (27), roof (26), wall (25).

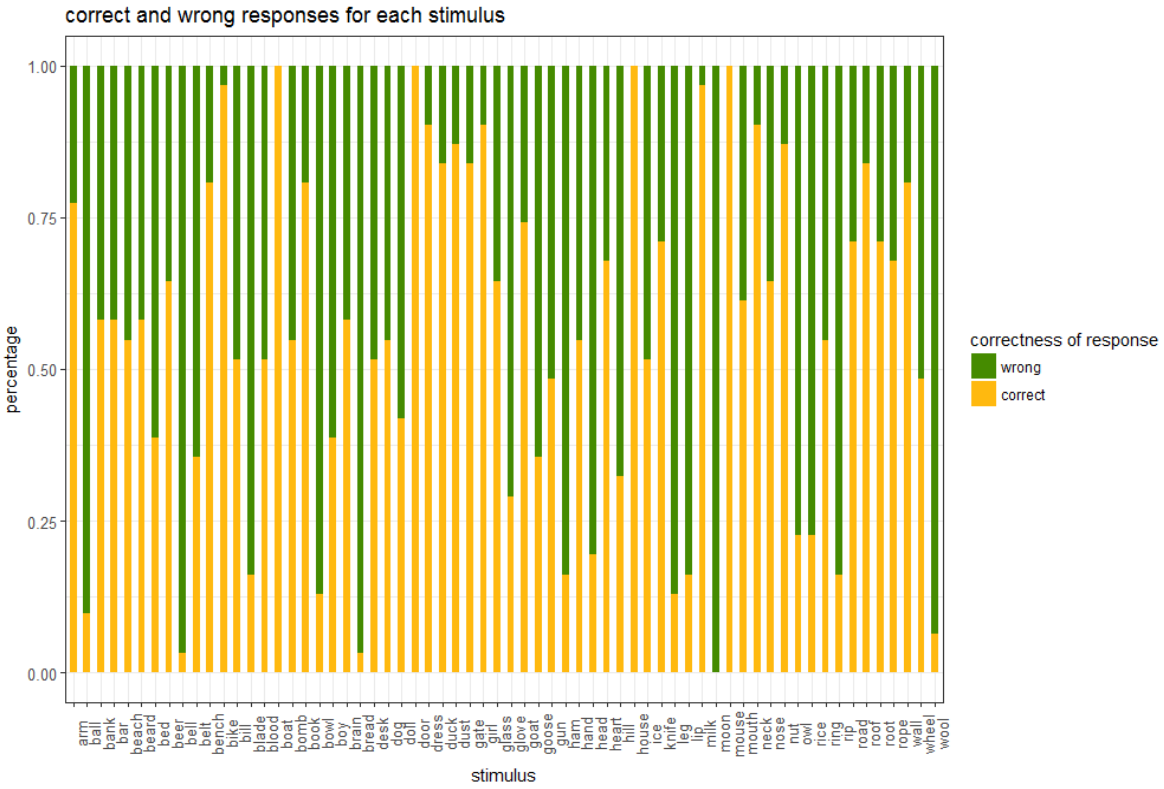

Figure 15: Percentage of correct (yellow) and wrong (green) responses to each stimulus.

[31] The numbers in parentheses correspond to the number of participants who correctly identified the stimulus.

[32] Numbers in parentheses correspond to the number of participants who correctly identified the stimulus.

Method and results

As is to be expected, the percentage of correctly identified stimuli varied also between noise levels. At noise level 0.1, 78% of the stimuli were correctly identified. This percentage went down to 66% at noise level 0.2, and further to 45% at noise level 0.3 and to 32% at noise level 0.4 (*M*=0.5, *MD*=0.5, range: 0.22-0.78, *SD*=0.2). The percentage of observed correct responses in each noise level is displayed below in Figure 16.

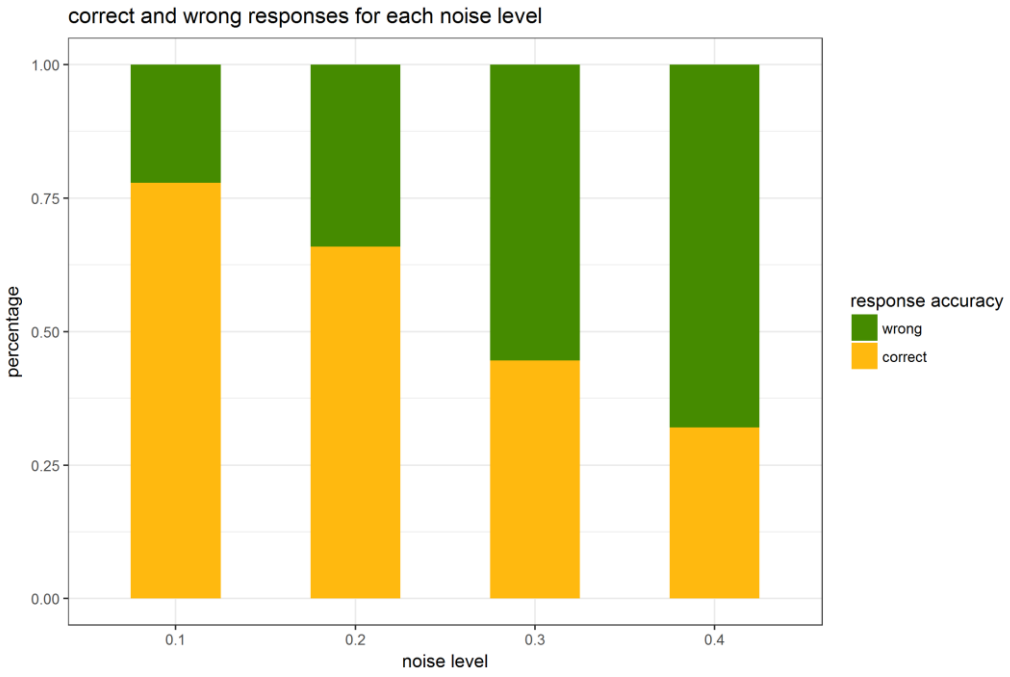

Figure 16: Response accuracy depending on noise level. The plot shows the percentage of correct (yellow) and wrong (green) responses for each noise level.

However, the number of correct or wrong responses to a stimulus did not change in relation to the cognate status. Whether cognate or not, participants identified about 50% of the stimuli correctly (*M*=0.5, *MD*=0.5, range: 0.45- 0.55). Figure 17 displays the percentage of correct and wrong responses for cognate stimuli words and non cognates.

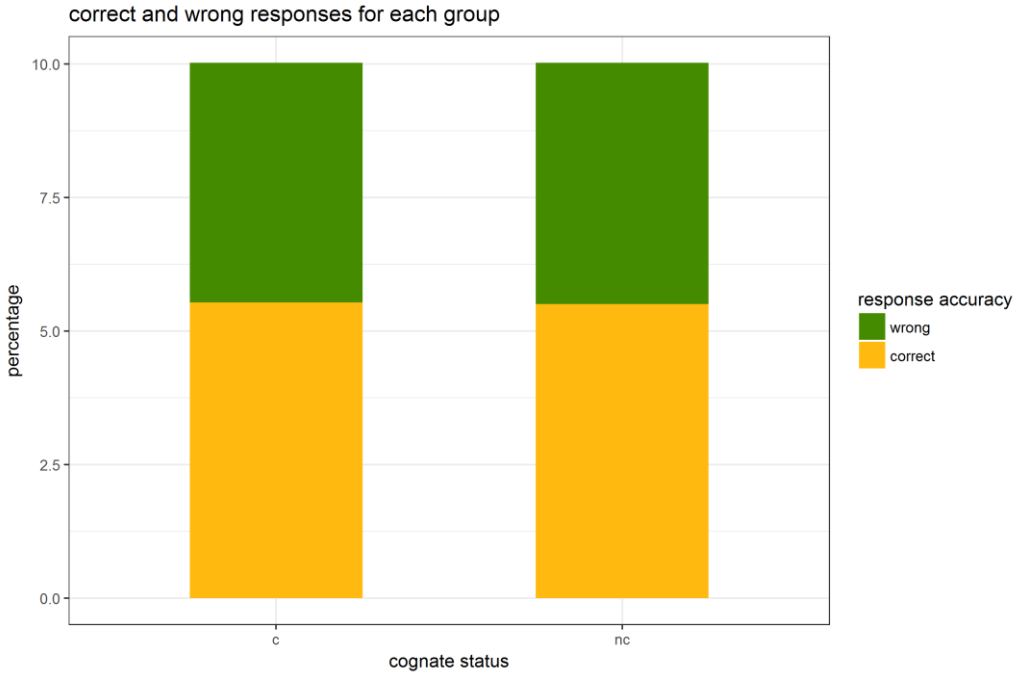

Figure 17: Response accuracy of cognates and non-cognates. The plot shows the percentage of correct (yellow) and wrong (green) responses for cognates (left bar) and non-cognates (right bar).

Prior to statistical analysis, all trials that lasted longer than 1.5 standard deviation of the mean duration of all trials per participant were removed. On the average, about 4 trials were removed per participant (range: 1-11 trials, first quartile: 3 trials, third quartile: 5 trials). For five participants, more than six trials had been removed. In total, 3.7% of all trials were removed. Three stimuli, “bell”, “bread” and “rice” were removed in more than half of the participants. These were the first three stimuli of the list. The video recordings of the experiments show that many participants asked questions or read the instructions more thoroughly than in successive trials which may account for the long trial duration[33]. All other stimuli were removed in fewer than 6 participants. The mean duration of all trials after removal of extremely long trials was 1580 ms (*MD*=1239 ms, range: 228-1206 ms).

[33] Participant P8T was not excluded from analysis as her response accuracy did not differ significantly from other participants.

Method and results

A generalized linear mixed model for response accuracy was constructed using the R-package lme 4 (Bates, Maechler, Bolker, & Walker, 2015). Random effects included intercepts for each participant and for each stimulus[34]. Fixed effects included noise level and response time. The predictor variable cognate status ($F(8,1)=0.303$, $p=0.582$) and task ($F(8,1)=1.999$, $p<0.157$) failed to reach significance when comparing models with and without the effects in question by maximum likelihood approximation.

As the descriptive statistics suggest, variance between participants was much lower (*SD*=0.345) than between stimuli (*SD*=1.999). All levels of the factor noise level were highly significant at $p<0.01$. The estimate expressed as log odd of noise level 0.2 decreased by 1.043 compared to the noise level 0.1 (*Estimate*= -1.043, *SE*=0.2104, *z*= -4.960, $p<0.01$). For noise level 0.3, the log odd estimate (still compared to noise level 0.1) even decreased by 2.488 (*Estimate*= -2.488, *SE*=0.2199, *z*= -11.317, $p<0.01$). At noise level 0.4, the log odd estimate fell by 3.315 (*Estimate*= -3.3157, *SE*=0.2298, *z*= -11.428, $p<0.01$). This corresponds to a drop from approximately 91% correct responses at noise level 0.1 to 28% at noise level 0.4 (for the effect plot see Figure 18). According to the model, response time is also a highly significant predictor (*Estimate*= -0.7996, *SE*=0.0879, *z*= -9.097, $p<0.01$). Figure 18 illustrates the estimated effects of noise level and response duration on response accuracy: the left facet shows that the probability to observe a correct response decreases with increasing noise level. The right facet shows the probability to observe a correct response decreases when the response duration increases. The longer participants take for the response, the smaller the probability that this response is correct.

[34] The model failed to converge with trial slopes for each participant.

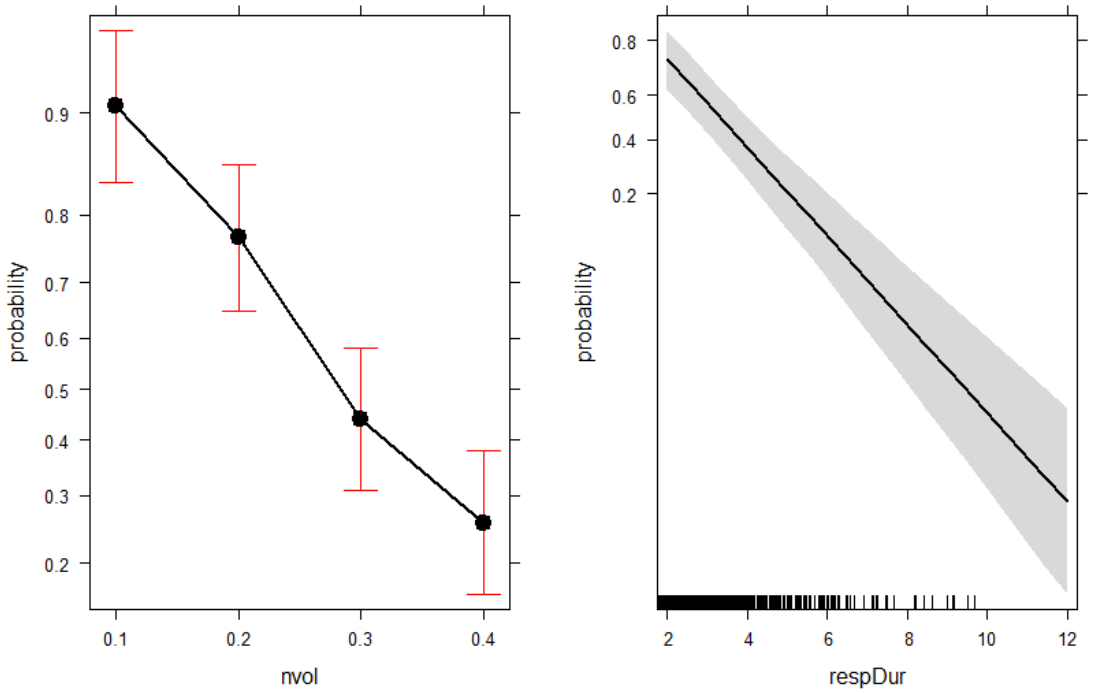

Figure 18: Effects plot for response accuracy. The figure depicts the effects of the predictor variables noise level and response time on response accuracy.

4.3.5.3 Pupil dilation

As the whole experiment had been programmed in Psychopy (Peirce, 2007) the eye tracking data did not contain any information about the stimulus or the trial. The first step therefore was to add the experimental information that had been saved to a file during the experiment to the eye tracking data. A comparison of the timestam*p* at which the first key press of the experiment was recorded showed that the eye tracking timestamp was systematically 0.06 seconds ahead. After having corrected the timing difference, I matched the experimental information by the means of the timestamp to the eye tracking data in order to construct the following variables: task (listener or interpreter), group (group 1 to 4, stimuli were presented with different noise levels according to the group), stimulus (name of the stimulus), trial (1 to 64), phase (baseline period, stimulus duration, response preparation, response), noise level (volume of background noise relative to the volume of the stimulus), and cognate

status (cognate, non-cognate). I also added the noise level and the cognate status of the preceding trial. Trials that lasted longer than 1.5 standard deviation of the mean duration of all trials per participant were excluded from analysis. On average, about 4 trials were removed per participant (range: 1 -11 trials, first quartile: 3 trials, third quartile: 5 trials). For five participants, more than six trials had been removed. In total, 3.7% of all trials were removed. Three stimuli, “bell”, “bread” and “rice” were removed in more than half of the participants. These were the first three stimuli of the list. The video recordings of the experiments show that many participants asked questions or read the instructions more thoroughly than in the successive trials which may account for the long trial durations. All other stimuli were removed in fewer than 6 participants. The mean duration of all trials after removal was 7.45 seconds (MD=7.15 seconds, range: 4.3-19.93 seconds).

The pupil size measures of both eyes were highly correlated (r(1528600)=0.934, p<0.001). All subsequent transformations were thus only done for the right eye. The second step was to remove blink artefacts and invalid measures. During blinks, the eye lid covers partially the pupil which biases the eye tracker’s estimation of the pupil size. Blink artefacts can therefore be described as sudden drops of the pupil size. For each participant and each trial, I calculated the differences of all pupil sizes to their preceding observation and detected outliers with the boxplot function in *R* (R Core Team, 2016). A “drop” corresponded thus to a difference from the preceding pupil size that exceeded 1.5 standard deviations of the mean over all differences. All observations that corresponded to a drop were replaced by NA (missing values). Invalid observations were defined according to the validity codes that are provided by the Tobii TX 300 eye-tracker. Validity codes range from 0 to 4. 0 and 1 indicate that both eyes (0) or at least one eye (1) were reliably found. Validity codes from 2 to 4 indicate that the validity of the gaze data is uncertain or that the eye tracker failed to track the eye. Data points with validity codes from 2 (validity of gaze data uncertain) to 4 (no eye found) were replaced by NA.

Method and results

Using the *R*-package *zoo* (Zeileis & Grothendieck, 2005), I substituted missing values up to a gap of 100 milliseconds with linear interpolation. In total, 473148 missing values (23.4%) were replaced. In eight cases, more than 26.5% (third quartile) of the data were interpolated (P10T: 41.06%, P11D: 45.29%, P14D: 56.20%, P14T: 26.68%, P17T: 48.70%, P3D: 53.97%, P8T: 28.00%, P9D: 29.00%).

The third step was to standardize the pupil sizes in order to remove differences between the participants and to obtain a (near to) normal distribution. As in the pilot study, I determined the pretrial pupil baseline and followed a standardization procedure to normalize the data using the formula:

$z=x-\mu/\sigma$

where x is the observed data point, μ the mean of all data points during the pretrial baseline epoch and σ the variance during the pretrial baseline epoch.

Performing standardization based on the pretrial baseline means that the normalized pupil sizes during the trial depend on the mean pupil size and its variance measured during a pretrial baseline period. I could not exclude that the preceding trial affected in some way the baseline pupil size. For instance, a higher noise level of the preceding trial could have led to higher baseline values. Standardization based on the pretrial baseline carries thus the risk that the pupil size during the trial is indirectly affected by the preceding trial. I therefore performed for each participant a second standardization based on the grand mean of all pretrial baseline phases of all stimuli and tested the values obtained by both methods for correlation. Both methods proved to yield nearly the same results ($r(2022900)=1$, $p<0.001$). In fact, only 1758 out of 1 541 687 values were different. The standardized pupil sizes based on the pretrial baseline periods were then used for statistical analysis.

Finally, in order to “smooth” out some of the noise in the data, I created time bins of 100 milliseconds and aggregated the data over these bins and

the four different groups by calculating the mean. The final data set used for the statistical analysis included the standardized pupil size, time bins of 100 milliseconds, phase, participants, task, stimulus, trial, noise level, noise level of the preceding trial, cognate status, response accuracy and response accuracy of the preceding trial. Figure 19 below displays the standardized pupil sizes as a function of noise level and cognate status during the first six seconds after stimulus onset. For better readability, blue lines were added to indicate when one phase ends and another begins. The variable “phase” codes rough timing information. For instance, “baseline period” indicates the pretrial period where participants fixated a cross in order to sample their baseline pupil size. “Stimulus duration” is the period during which the participants heard the stimulus words and corresponds thus to the recording duration of each stimulus. “Response preparation” corresponds to the time window between the moment where the recording ended and the moment where participants pressed the space bar to type in their response. “Response” is the period during which participants entered their response. As most of them needed to look down at the keyboard and interrupted thus for more than 100 ms the contact to the infrared interface of the eye-tracker, this phase included most of the missing data (missing data during “sound”: 5.7%, missing data during “response preparation”: 5.9%, missing data during “response”: 23.9%). The time course of the pupillary response from stimulus onset on is depicted in Figure 19. The left facet shows the time course for cognate stimuli, the right facet shows the time course for non-cognate stimuli. The different noise levels are coded by the color (green: 0.1, yellow: 0.2, red: 0.3, blue: 0.4). In Figure 19, the high amount of missing data becomes apparent through the larger variance during the response-phase. Nevertheless, a visual inspection of the data suggests an effect of noise level, as well as of cognate status. From the graph, it seems that pupil sizes increase as the level of noise increases and that pupil sizes are globally larger for non cognates than for cognates.

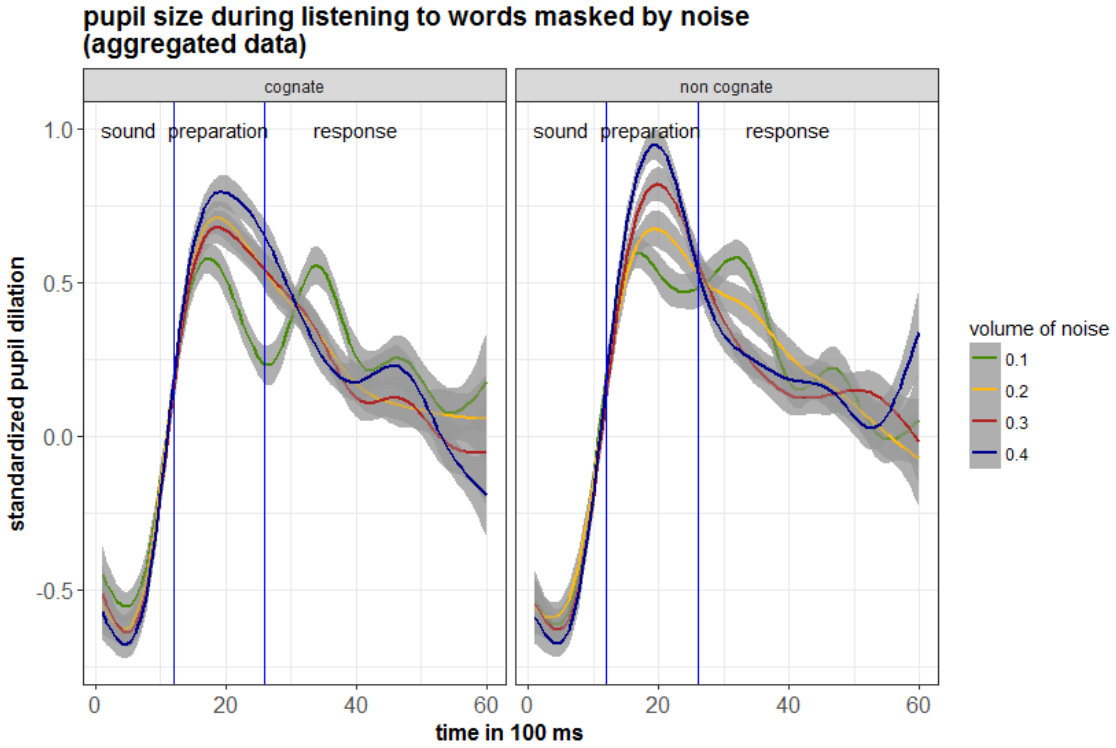

Figure 19: Pupil dilation depending on noise level and cognate status. The graph shows the evolution of the standardized pupil sizes as a function of cognate status and noise level during the first six seconds after stimulus onset during the duration of the stimulus, the response preparation phase and during the response. The blue lines indicate approximately where one phase ended and another one started.

Figure 20 below shows again the time course of the standardized pupil size during the first six seconds after stimulus onset during the different phases, but this time depending on whether the participant's response was correct or not. Again, the graph suggests an effect of noise level and response accuracy. Pupil sizes seem to be globally larger when the response was wrong than when the response was correct.

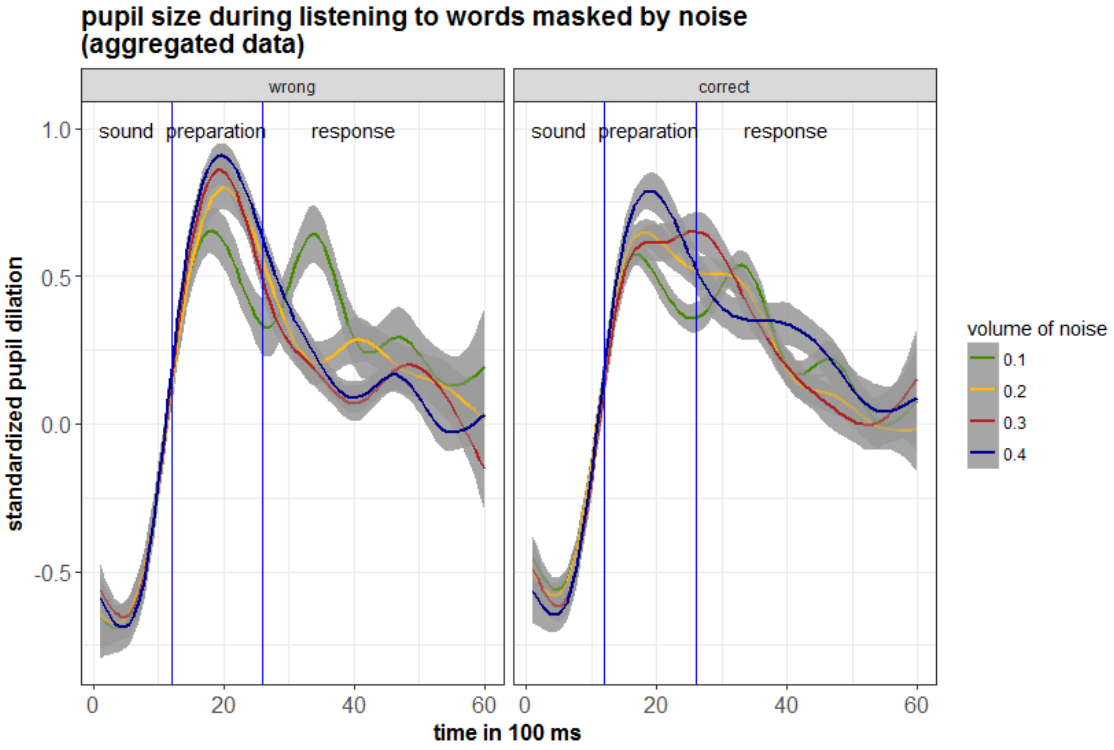

Figure 20: Pupil dilation depending on noise level and response accuracy. The graph shows the evolution of the standardized pupil sizes as function of noise level and correctness of the response during the first six seconds after stimulus onset during the duration of the stimulus, the response preparation phase and during the response. The blue lines indicate approximately where one phase ended and another one started.

Most importantly, both graphs clearly show that the effect of the predictors is not the same over the whole time course. In fact, the predictors seem to have an effect only during the response preparation phase. I thus tested the effect of the five predictor variables noise level, preceding noise level, cognate status, trial, correct, participant, stimulus and task on of the first three seconds. The latter three seconds covered the response entering and were due to the high amount of missing data not reliable and therefore excluded from analysis.

I conducted a growth curve analysis (Mirman, 2014) using the *R*-package *lme4* (Bates, Maechler, Bolker, & Walker, 2015) with the standardized pupil size on 200 ms time bins. A first attempt including the third order polynomial for the 100 ms time bins in the random effect structure failed consistently to converge. The overall time course was captured with third

order orthogonal polynomials. The random effect structure covered trial-by-participant random slopes on all time terms. Fixed effects included the level of noise, cognate status and response accuracy. Fixed effects were treated as contrasts. They were added one by one and p-values were approached using model comparison with maximum likelihood. Even though the effects of noise level and response accuracy on all time terms were significant, the model itself proved still to be a rather poor fit (see Figure 21).

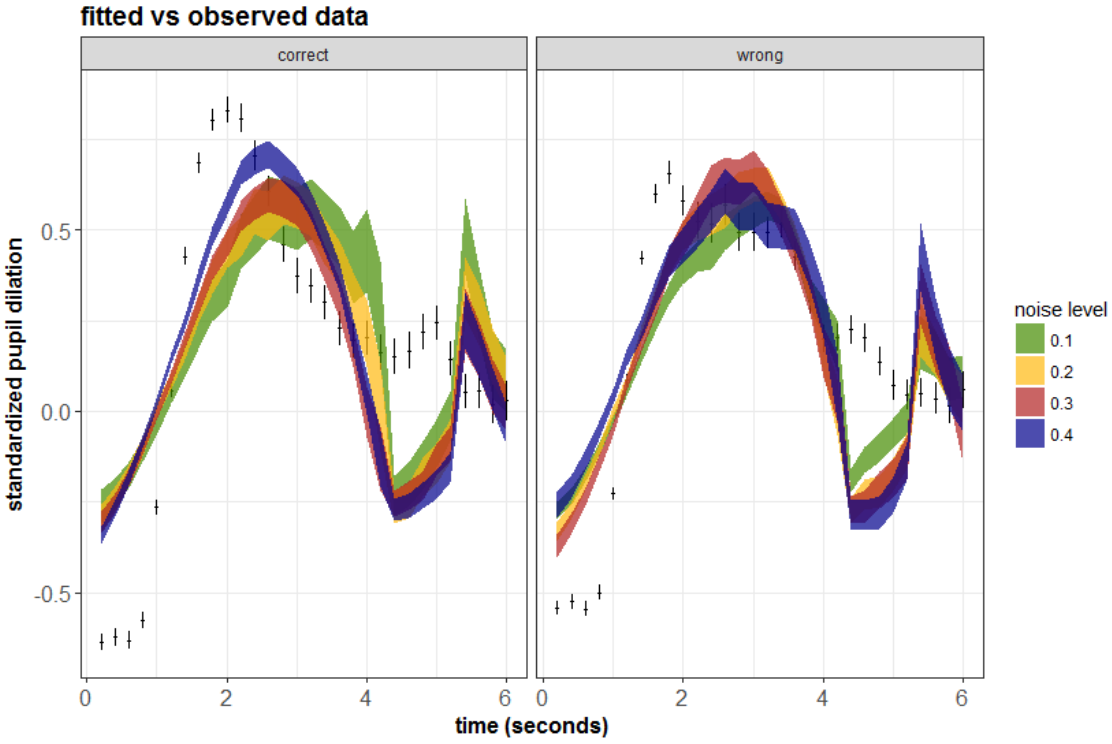

Figure 21: Fitted values against the observed standardized pupil dilation. The colors correspond to the different noise levels (green: 0.1, yellow: 0.2, red: 0.3, blue: 0.4). Facets show the fitted values (colored ribbons) against the observed standardized pupil dilation (black pointranges) for correct (left facet) and wrong answers (right facet).

I thus decided to run an analysis of variance for each time bin on the predictor variables which I also included in the mixed model: noise level, cognate status, correctness of response, participant, trial and stimulus. Not surprisingly, the analysis of variance revealed that “participant” was highly significant for all time bins (lowest value: $F(30, 1161) = 12.930$, $p<0.01$). “Noise level” was significant for time bins 15 to 26 (lowest value:

F(3, 1182) = 3.607, *p*<0.05), thus between 1.5 to 2.6 seconds after stimulus onset. “Cognate status” was significant for time bins 21 and 22 (lowest value: *F*(1, 1284)= 4.1557, *p<0.*05), thus 2.1 to 2.2 seconds after stimulus onset. “Response accuracy” was significant for time bins 1 to 9 (lowest value: *F*(1, 1661) = 3.886, *p<0.*05; time bins 5 and 6 only marginally significant with *F*(1, 1607) = 4.689, *p* <0.1) and from time bin 18 to 22 *(*lowest value*: F*(1, 1568) = 6.270, *p*< 0.05). “Trial” was significant for time bins 19 and 29 (lowest value: *F*(1, 1242) = 4.087, *p<0.*05). “Stimulus” was significant for time bins 10 and 12 (lowest value: *F*(61, 2186)=1.427, *p<0.*05). F-values and p-values for the predictors “noise level”, “cognate status”, “response accuracy”, “trial” and “stimulus” are displayed in Table 33 (see appendix 7.4.2).

The results of the analysis of variance suggest that stimuli or trial had (nearly) no linear effect on pupil dilation. The noise in the data is largely due to differences between participants. More importantly, the results suggest different time courses for the predictors noise level, cognate status and response accuracy as illustrated in Figure 22.

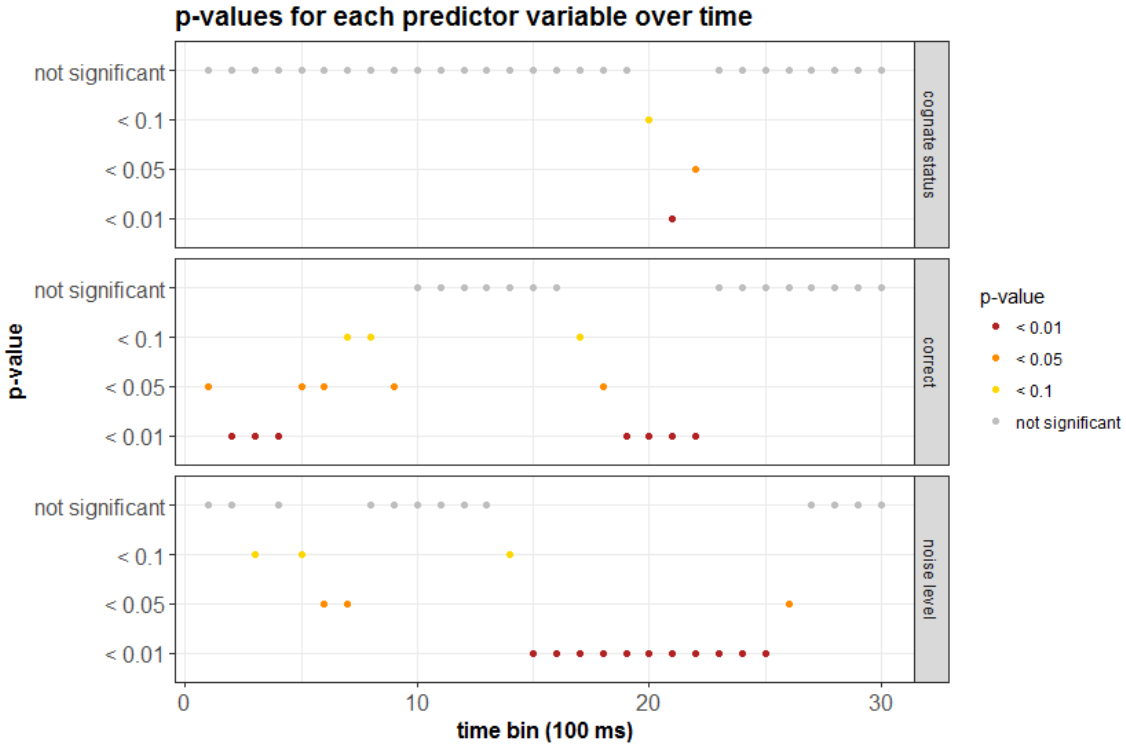

Figure 22: P-values for the predictor variables cognate status, response accuracy and noise level. Red and orange points indicate significant p-values, yellow points indicate marginal significance and grey points indicate not significant p-values.

The figures below (Figure 23 to Figure 25) depict the effect of the predictor variables noise level, response accuracy and cognate status in more detail. The effect of noise level is limited to 1.5 to 2.6 seconds after stimulus onset. In Figure 23 it becomes apparent that the higher the noise level, the more the pupil dilates. As such, pupil dilation is larger at noise level 0.4 than in all other noise level. However, the distance is not equal between each level of noise. A Tukey comparison contrasting the different noise levels in each bin revealed that only noise level 1 and 4 differed consistently in time bins 15 to 26 at a significance level of $p<0.05$. Noise level 1 to 3 showed a significant difference ($p<0.05$) in time bins 19 to 20 and in time bins 22 to 25. Noise levels 2 to 4 were different at a significance level of $p<0.05$ in time bins 17 and 21 and noise levels 1 to 2 were only different in time bins 24 and 25. The difference between noise levels 2 to 3 or noise levels 3 to 4 was not significant in any of the time bins 15 to 26 where the analysis of variance revealed a significant effect of

noise level. Figure 23 shows the effect of noise level in the first three seconds after stimulus onset. The part in grey highlights the section where a significant effect for noise level was found (1.5 to 2.6 seconds after stimulus onset).

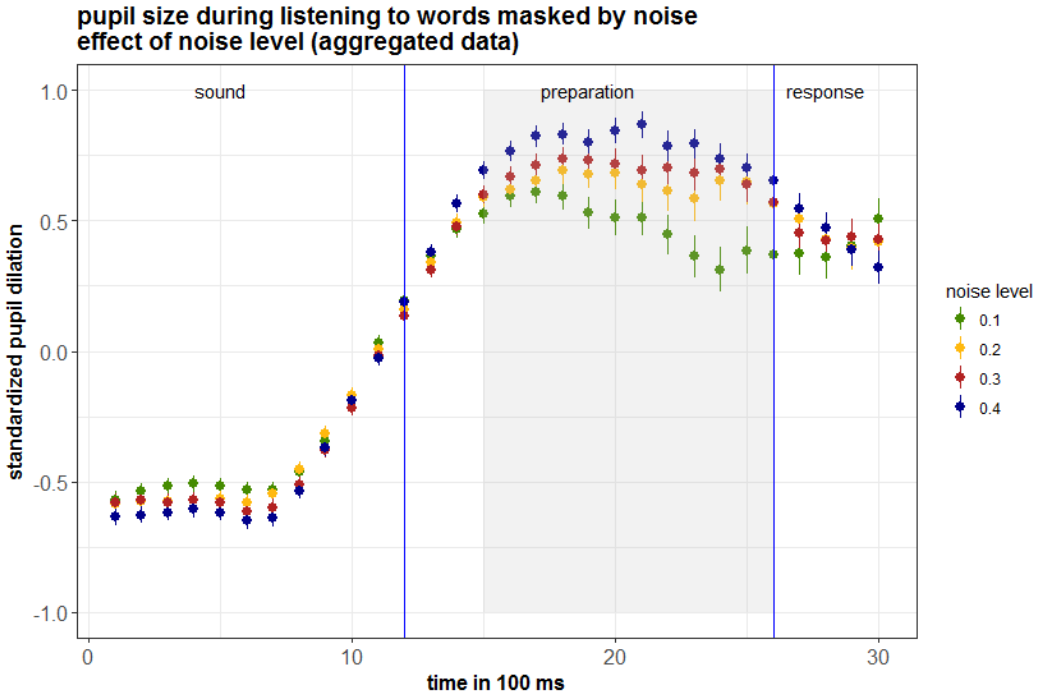

Figure 23: Time course of the effect of noise level in the first three seconds after stimulus onset. The part in gray highlights the time bins where a significant effect for noise level was found.

For response accuracy, a significant effect was found right after stimulus onset (time bins 1 to 9) and nearly a second later in time bins 18 to 22 where the pupil dilation reaches its maximum. As can be seen in Figure 24, pupil dilation in the early stage, during which participants heard the stimulus, is larger for correct responses. However, the effect is reversed at the later stage just before participants entered their response: here, pupil dilation is larger when participants gave a wrong answer.

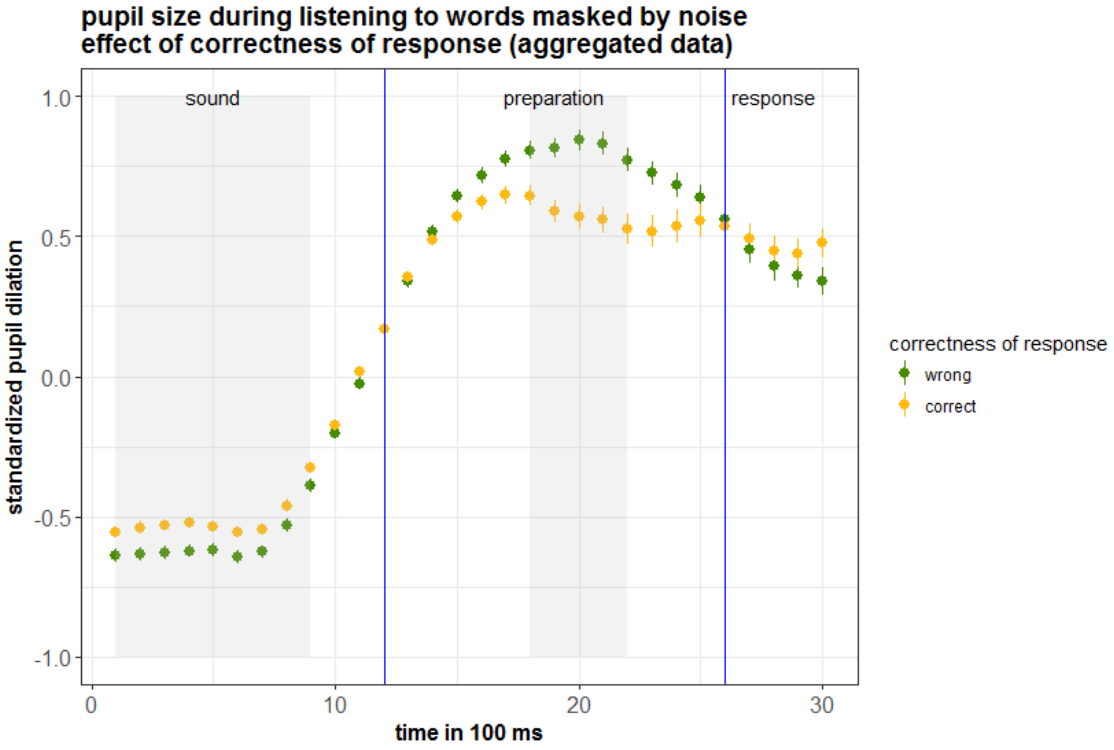

Figure 24: Time course of the effect of correct response in the first three seconds after stimulus onset. The part in gray highlights the time bins where a significant effect for response accuracy was found.

According to the analysis of variance, the effect of cognate status is limited to 300 ms, namely from time bin 20 to time bin 22 (even though the time bins immediately before time bin 20 and after time bin 22 revealed a marginally significant effect for cognate status). Figure 25 suggests that cognate status has only little effect on the pupil dilation as such: the maximum dilation for cognate and non-cognate stimuli is nearly the same. But non-cognate stimuli seem to trigger a more sustained pupil dilation compared to cognate stimuli. For non-cognate stimuli, the peak dilation declines only in time bin 22, whereas for cognate stimuli, is declines immediately after reaching its peak in time bin 17.

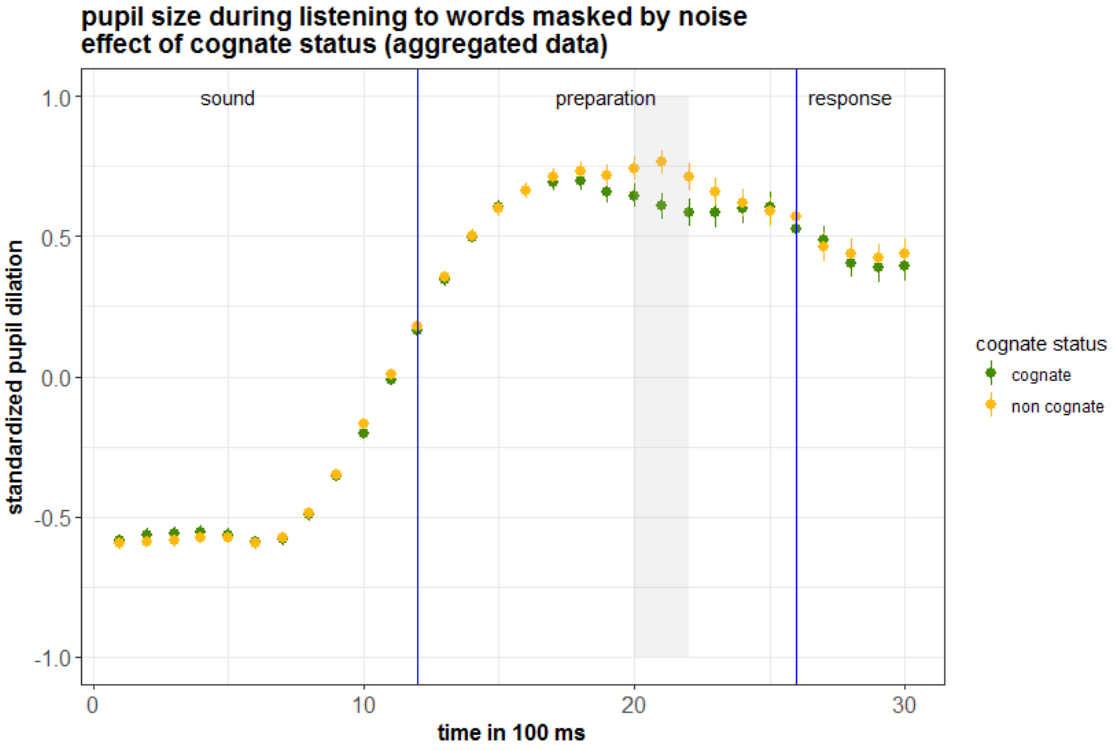

Figure 25: Time course of the effect of cognate status in the first three seconds after stimulus onset. The part in gray highlights the time bins where a significant effect for cognate status was found.

4.3.5.4 Discussion

The aim of the pretest was twofold: first, to confirm that pupils dilate during word identification with increasing levels of masking noise; second, to set the volume of the background noise to a level where participants would recognize 75% of the stimulus words. 64 stimuli at different noise levels (0.1 to 0.4 which corresponds to 10% to 40% of the maximum volume of the sound card) were presented to the participants. As the cognate status and differences in processing cognates or non cognates is not relevant for the investigation of visual input in simultaneous interpreting and only a by-product, the discussion of the pretest will mainly focus on noise level, response accuracy and response times.

Overall, the results are in line with the findings of Kramer and colleagues (Kramer, Kapteyn, Feesten, & Kuik, 1997) and Koelewijn and colleagues (Koelewijn, Zekveld, Festen, & Kramer, 2012) described in chapter 3.3.4: Pupil dilation increased with increasing levels of noise. The effect of noise

level could also be observed in response times and response accuracy: according to the statistical models, the response times for stimuli at noise level 0.4 was approximately 150 ms longer than for stimuli at noise level 0.1 and probability for correct responses decreased by 63% from noise level 0.1 to noise level 0.4. The cursus participants belonged to, or the cognate status of the stimuli seemed to make (nearly) no difference in none of the three dependent variables I investigated.

Despite the huge amount of noise in the data, the effects on pupil dilation for noise level are persistently significant during the response preparation phase. This demonstrates that the effect of noise level is actually quite strong, even though the largest amount of variance is explained by differences between participants and stimuli. Effects of trial and stimuli are not distinguishable because stimuli were always presented in the same order. Nevertheless and despite the careful selection of the stimulus words, stimuli seem to provide a better explanation to the data than trial. Indeed, neither pupil sizes or response times nor response accuracy increased or decreased systematically in each trial. Furthermore, the response times were in general rather high (mean response time about 1500 ms) compared to previous research using response times in word identification tasks[35]. In addition, the number of extremely long trials (outliers) was rather high (nearly 10% for response time, 3.7% for response accuracy). There are multiple explanations: either participants had to search for the space bar, or participants forgot to press the space bar and the response time which was recorded is erroneous. Also, participants were not instructed to react as fast as possible, but on the contrary, to take the time they needed. This problem could have been avoided by asking participants to repeat orally the word they had heard. But in this case it would not have been possible to automatically check

[35] For instance, Costa and colleagues (2003) reports reaction times around 750 ms in picture naming.

their answers which was necessary to set the noise level corresponding to the participant's individual 75%-threshold for the speeches in the main part of the experiment.

Another limitation of the pretest concerns interactions between noise level, response accuracy, and response times. Pupil dilation is larger for higher noise levels and for wrong responses, but wrong responses and higher noise levels tend to go – unsurprisingly – together. Even though response accuracy and noise level were both significant factors for modelling the time course of pupil dilation during word identification, it is impossible to know on a cognitive level whether pupil dilation is primarily due to the background noise or to the fact that the participant did not identify the word. The same holds for response times and response accuracy: response times for correct responses or low levels of noise are shorter than for wrong responses or high levels of noise. But again: Correct responses are much more frequent at low noise levels. One hint might indicate that noise level and response accuracy are indeed distinct: During the first 900 ms after stimulus onset, thus when participants listened to the stimulus, pupil sizes were larger for correct responses. This might suggest that participants identified the word even before they heard it completely. For the purpose of the present research, the exact distinction might not be of much importance. The results show that masking noise impacts listening comprehension (reduced response accuracy, longer response times for higher noise levels) and provokes larger pupil dilations which might be explained by increased work-load due to masking noise.

4.3.6 Data preparation and results of the main part

This chapter presents the data preparation and the results of the main part of the main study. During the main part, participants were asked to orally translate (interpreters) or listen to (listeners) four speeches, to rate general parameters and the speech duration and to answer to text-related questions. The data analysis for listeners and interpreters included the pupillary response during the speech, the ratings of the general

parameters and speech duration and the accuracy of the answers to the text-related question. Further performance-based and voice-related data that have been obtained from interpreters only are translation accuracy, cognate translations, voice frequency and silent pauses during the translation. All statistical analyses were done with *R* (R Core Team, 2016), figures and graphs were done using the package *ggplot2* (Wickham, 2009). Each analysis concludes with a brief discussion of the respective results. A more general discussion relating all results with each other follows in chapter 4.4.

4.3.6.1 Ratings of general parameters

After each speech, participants were asked to rate four general parameters: video quality, sound quality, text difficulty and speech rate. Video and sound quality ratings were done to ensure that the experimental conditions, the visual presentation (audio/video) and the auditory presentation (noise/no noise), were sufficiently distinct. Ratings of text difficulty and speech rate should ensure that participants regarded all speeches as equally complex and equally fast.

4.3.6.1.1 Video quality ratings

Participants rated the video quality according to four categories: very good, good, okay, and bad. The contingency table for the audio/video-condition is displayed in Table 7. Three ratings were invalid due to key presses that could not be related to any of the answers. As can be seen, no participant rated the video quality of any of the speech as “very good”.

Method and results

	audio	video
Very good	0	0
Good	6	24
Okay	14	34
Bad	40	3
Missing values	3	
Total	124	

Table 7: Video quality ratings: Counts for the different levels of video quality ratings (*N*=31).

A paired Wilcoxon signed rank test indicated that the distribution of the video quality ratings differed significantly between the video condition where the lips movements of the speaker were visible, and the audio condition with a freeze frame of the speaker (*V*=2817, *N*=121, *p*<0.001). The low p-value confirms that participants rated both conditions, the audio-visual and the auditory-only condition, differently. In order to detect further potential effects, I conducted an *ordinal* mixed effects model[36] with the probit distribution as link function and random intercepts for participants (*SE*=1.36) which confirmed this result. Fixed effects included visual presentation, auditory presentation, task, group and speech. P-values of the effect were estimated using likelihood ratio tests of cumulative link models to compare the model with and without the effect in question. All

[36] Analyses of variance or linear regressions are not suited to investigate ordinal data because ordinal data contains no information about the distance between each unit. Therefore, linear regression models or analyses of variance tend to be over-confident. Furthermore, the interpretation of model coefficients can be difficult, as values like 3.8 or 1.2 are not meaningful for ordinal data with distinct values (Christensen, 2015). Instead, packages in *R* (R Core Team, 2016) like *MASS* (Venables, 2002), *VGAM* (Yee T. W., 2017) or *ordinal* (Christensen, 2015) provide different methods for ordinal regression, including mixed effect ordinal regression methods in ordinal.

analyses were carried out in *R* version 3.3.2 (R Core Team, 2016) using the package *ordinal* (Christensen, 2015).

The predictor variable visual presentation (video/audio, reference level: audio condition) was highly significant (*Estimate*= -2.527, *SE*=0.410, *z*= -6.164, *p*<0.001). The odds of observing an increase of the rating by one unit (for example: good to okay; okay to bad) dropped by 92% in the video condition across all levels of rating compared to the audio condition which means in the end that video quality obtained better ratings in the video condition than in the audio condition. Further terms, like speech, trial, task or the auditory presentation of the speech (noise vs no noise) did not reach significance. The results for each term obtained by model comparison are reported in Table 8. The model fit is plotted in Figure 26.

	Likelihood ratio	DF	p-values
Video/audio	74.670	1	< 0.001
Noise/no noise	0.458	1	0.458
Task	0.587	1	0.446
speech	0.826	3	0.843
Group	2.959	3	0.398

Table 8: Video quality ratings: results of model comparison with likelihood ratio test (likelihood ration, degrees of freedom and p-value). Predictor variables were added one by one and compared against the model without the predictor in question.

Method and results

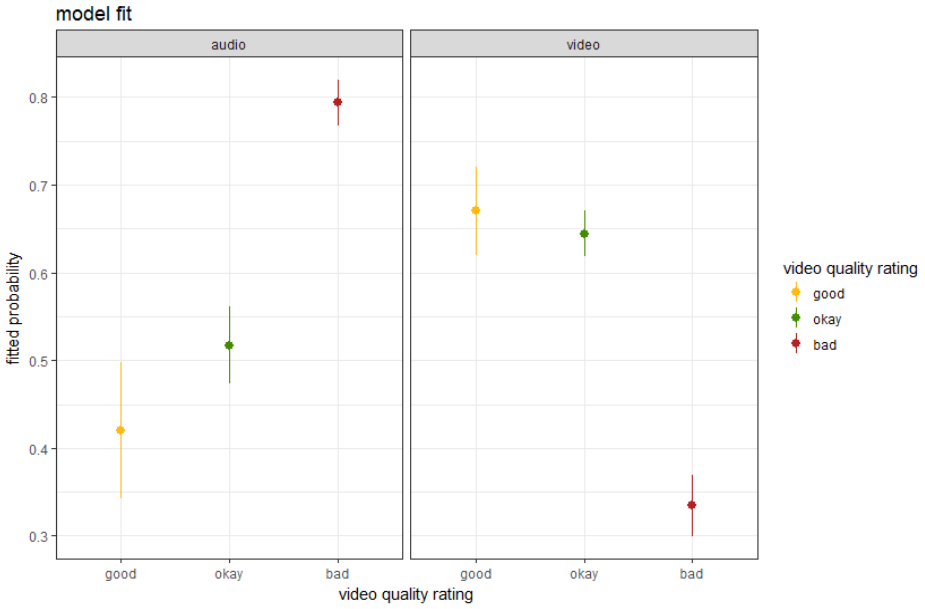

Figure 26: Fitted probabilities for video quality ratings in the video and the audio condition. The left facet shows the fitted probabilities in the condition without video, the right one shows the fitted probabilities in the video condition. The pointranges give the mean of the fitted probability for each rating unit with a 1.5 standard deviation. Yellow pointranges stand for “good”, green for “moderate” and red for “bad/freeze image”. The fitted probability for rating the video quality as “good” is much lower in the audio condition than in the video condition. Inversely, the fitted probability for rating the video quality as “bad” is much higher in the audio condition than in the video condition.

4.3.6.1.2 Sound quality rating

Participants were asked to rate the sound quality as either “very good”, “good”, “okay” or “bad”. There were 124 observations with no missing value. The rating distribution for the noise/no noise-condition is displayed in Table 9.

	Noise	**No noise**
Very good	0	10
Good	8	28
Okay	21	14
Bad	32	8
Total	124	

Table 9: Sound quality: counts of the different ratings in the noise and no noise condition

Method and results

A paired Wilcoxon signed rank test showed that the distribution of the ratings differed significantly between the condition where white noise was overlaid on the speech and the condition without white noise (*V*=4206, *N*=124, *p*<0.001). The low p-value confirms that both participants rated both conditions, speech with noise and without noise, differently.

In order to check if any other variable (task, visual presentation) influenced the ratings of the sound quality, I conducted an ordinal regression using the package *ordinal* (Christensen, 2015) with sound quality as response variable and random intercepts for participants (*SD*=0.92). P-values of the effects were estimated using Likelihood ratio tests of cumulative link models. P-values of each level were estimated using a Wald test. The result confirmed the Wilcoxon signed rank test: the predictor variable auditory presentation (reference level: noise condition) was highly significant (*Estimate*= -2.345, *SE*=0.345, *p<0.*001). The probability to observe a one-unit-increase (good to okay or okay to bad) of the sound quality rating when no noise was added to the speech decreased by 90% compared to the noise condition, or to put it in other words: Worse sound quality ratings were much less probable when no noise was added to the speech than when noise was added. Another significant effect was found for the predictor variable task (reference level: interpreters, *Estimate*= -1.912, *SE*=0.455, *p<0.*001). The probability to observe a one-unit increase (good to okay or okay to bad) of the listeners' ratings decreased by 85% compared to the interpreters which means that interpreters gave worse sound quality ratings than listeners. Finally, a third significant effect was found for the visual presentation (*Estimate*= -0.689, *SE*=0.238, *p*=0.004), indicating that probability for a one-unit increase of the sound quality rating (good to okay or okay to bad) drops by 50% in the video condition compared to the audio condition, e.g. speeches presented with video had a lower probability to get worse ratings. Table 10 displays the estimates of the ordinal regression.

	Estimate	Standard error	Z-value	p-value
Auditory presentation	-2.345	0.345	-6.797	<0.001
Task	-1.912	0.455	-4.203	<0.001
Visual presentation	-0.689	0.238	-2.982	0.004

Table 10: Sound quality ratings: model estimates, standard error, z-values and p-values for variables that significantly predict sound quality ratings.

Further terms, like speech, group or noise level, did not reach significance. The results for each predictor as obtained by comparing the model with and without the predictor in question are reported in Table 11. The model fit is depicted in Figure 27.

	Likelihood ratio	DF	p-value
Auditory presentation	66.252	1	< 0.001
Task	16.849	1	<0.001
Visual presentation	8.713	1	0.003
Speech	1.709	3	0.635
Group	4.317	3	0.178
Noise level	0.602	1	0.438

Table 11: Results of model comparison using likelihood ratio tests (Likelihood ratio, degrees of freedom and p-value)

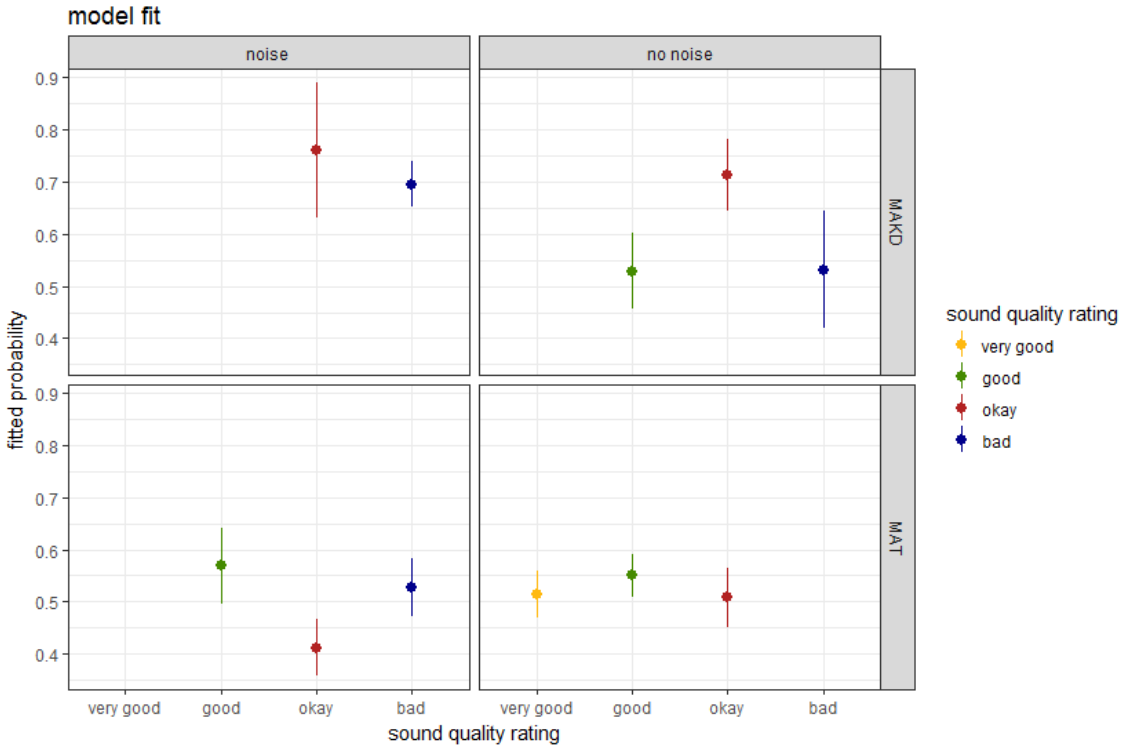

Figure 27: Fitted probabilities for sound quality ratings made by interpreters („MAKD“) and listeners („MAT“) in the noise and no noise condition. The colors indicate the different ratings (yellow: very good, green: good, red: okay, blue: bad). The fitted probability for rating the sound quality as “good” or “very good” was higher in the no noise condition than in the noise condition and on the whole higher for listeners than for interpreters. For listeners, the fitted probability for rating the sound quality as “very good”, “good” or “moderate” in the condition without noise was nearly equal. For interpreters, however, the fitted probability for rating the sound quality as “moderate” was higher than the fitted probability for rating the sound quality as “good”. The plot does not illustrate the effect of visual presentation (video/audio) which was much weaker than the effect of task or auditory presentation.

4.3.6.1.3 Text difficulty ratings

Participants rated the text difficulty according to the following categories: very easy, easy, moderately difficult, difficult. Each participant rated each of the four speeches once. The total number of observation was 123, with one missing value for the speech “work” in the interpreter group. The counts in each category according to the speech are reported below in Table 12.

Method and results

	air travel	demographic change	Greece	work
Very easy	3	5	2	4
Easy	21	21	16	13
Moderately difficult	7	5	11	9
Difficult	0	0	2	4
Total	31	31	31	30

Table 12: Counts of each text difficulty rating for each speech

I ran a non-parametric Friedman-test, made available in the package stats in *R* (R Core Team, 2016), to test if the median values were identical across the four speeches. As Friedman-tests need a complete block design, I replaced the only missing value with the rating that was most frequent for this specific speech and the task. The median of text difficulty ratings was different across the speeches: X^2 (3, N=124)=12.135, p=0.007. A post-hoc test with adjusted p-values revealed significant differences between the speeches “air travel” and “Greece” (Σ=1, p=0.035), “demographic change” and “Greece” (Σ=1, p=0.035) and a tendency indicating a difference between the speeches “demographic change” and “work” (Σ=3, p=0.070) (see Table 13).

Speech 1	Speech 2	Sigma (Σ)	p-value
air travel	demographic change	6	0.905
air travel	Greece	1	0.035
air travel	work	4	0.178
demographic change	Greece	1	0.035
demographic change	work	3	0.070
Greece	work	9	1.000

Table 13: Results of post-hoc comparison of text difficulty ratings between speeches

However, these differences might also be due to the effect of noise. Indeed, the speech “Greece” was always presented with noise, whereas the speeches “demographic change” and “air travel” were never masked by noise. In order to further explore this hypothesis, I conducted an *ordinal* regression using the package *ordinal* (Christensen, 2015) with text difficulty ratings as response variable and random intercepts (*SD*=1.05) for participants. Fixed effects included auditory presentation, visual presentation, task, noise level and group. 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 in *R* version 3.3.2 (R Core Team, 2016) with the package *ordinal* (Christensen, 2015).

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 increase of the rating (for example good-okay; okay-difficult) dropped by 17.9% in the condition without noise compared to the condition with noise. This confirms the hypothesis above that the differences between the speeches revealed by the Friedman-test were associated with the effect of noise. Ratings differed also between listeners and interpreters (reference level: listeners, *Estimate*= -0.7951, *SE*=0.433, *p<0.*066). The probability for a one-unit increase of the rating (very easy- easy, easy- okay or okay- difficult) dropped by 21.3% for the listeners compared to the interpreters. The predictor variable speech did not improve the model when the model already contained the predictor variable auditory presentation (noise vs no noise) (X^2(7,2)=1.3548, *p*=0.507). The results for each predictor obtained by comparing the model with and without the predictor in question are displayed in Table 14. The fitted probabilities are depicted in Figure 28.

Method and results

	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
Noise level	1.240	1	0.265

Table 14: Text difficulty ratings: likelihood ratio, degrees of freedom and p-values for each predictor variable that has been tested by comparing the model with and without the effect in question.

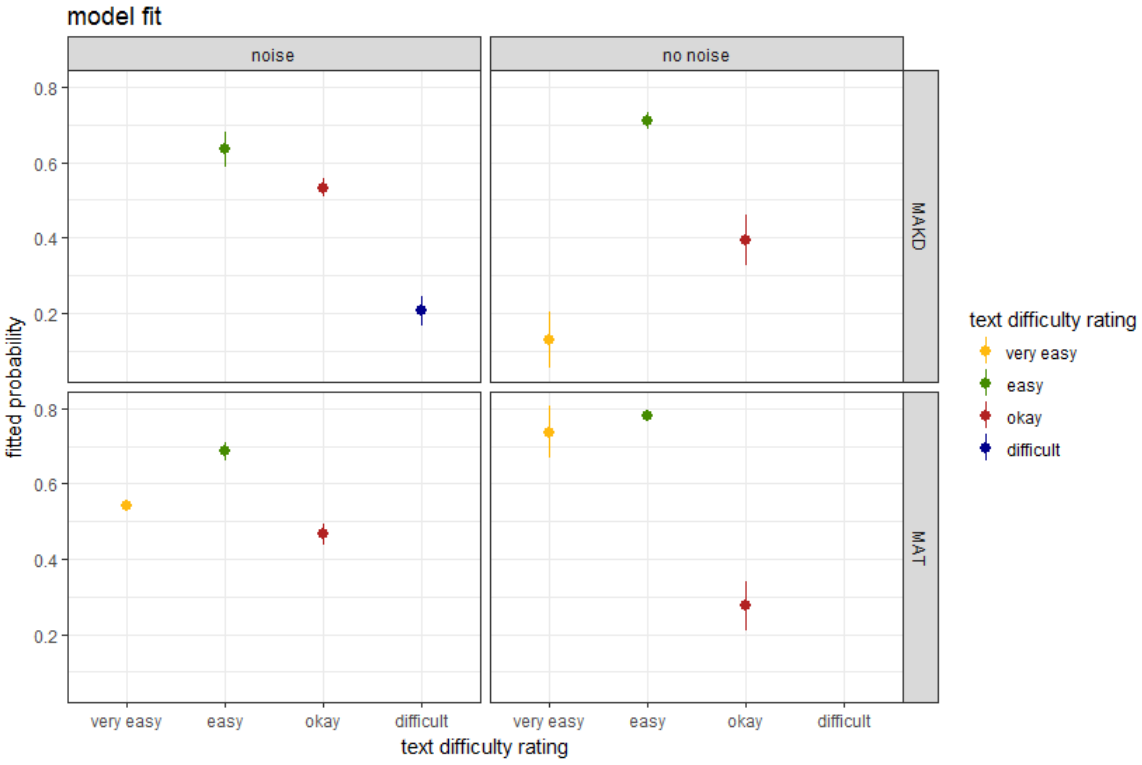

Figure 28: Fitted probabilities for text difficulty ratings (yellow: very easy, green: easy, red: okay, blue: difficult) in the noise (left facets) and no noise condition (right facet) for interpreters (top facets, “MAKD”) and listeners (bottom facets, “MAT”). The fitted probability for rating the text difficulty as “good” or “very good” was higher in the no noise condition than in the noise condition and higher for listeners than for interpreters.

4.3.6.1.4 Speech rate ratings

Speech rate was rated according to the following categories: very slow, slow, moderately fast, fast. Each participant rated each of the four speeches once. The total number of observation was 124, with no missing value. The counts in each category according to the speech are reported below in Table 15.

	air travel	**demographic change**	**Greece**	**work**
Very slow	4	2	3	2
Slow	22	27	22	25
Moderately fast	4	1	5	2
Fast	1	1	1	2

Table 15: Speech rate ratings: counts of each speech rate rating for each text

I performed a non-parametric Friedman-test, made available in the package stats in *R* (R Core Team, 2016) to compare the median of speech rate ratings across all texts. Speech rate ratings were equally distributed across all speeches: $X^2(7,3)=3$, $p=0.392$.

In order to reveal further potential predictors, I conducted an ordinal mixed effects regression on the speech rate ratings with random intercepts for participants. 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 in *R* version 3.3.2 (R Core Team, 2016) with the package *ordinal* (Christensen, 2015).

None of the predictor variables that have been tested reached significance. The results for each predictor obtained by model comparison are summarized in Table 16.

	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
Noise level	0.026	1	0.871
Speech	3.645	3	0.303

Table 16: Speech rate ratings: likelihood ratio, degrees of freedom and p-values for each predictor variable that have been obtained by comparing the model with and without the effect in question.

Finally, I tested whether text difficulty ratings and speech rate ratings were correlated. I therefore transformed all values to numeric values ranging from 1 to 4. Both variables showed a trend for a weak correlation ($r_T(2.594)=0.216$, $p=0.09$). When participants rated a text as being very easy, they also tended to give slower speech rate ratings. Conversely, speeches perceived as being difficulty had a tendency to be perceived as being faster. Figure 29 shows the proportion of each level of speech rate rating for each level of text difficulty rating.

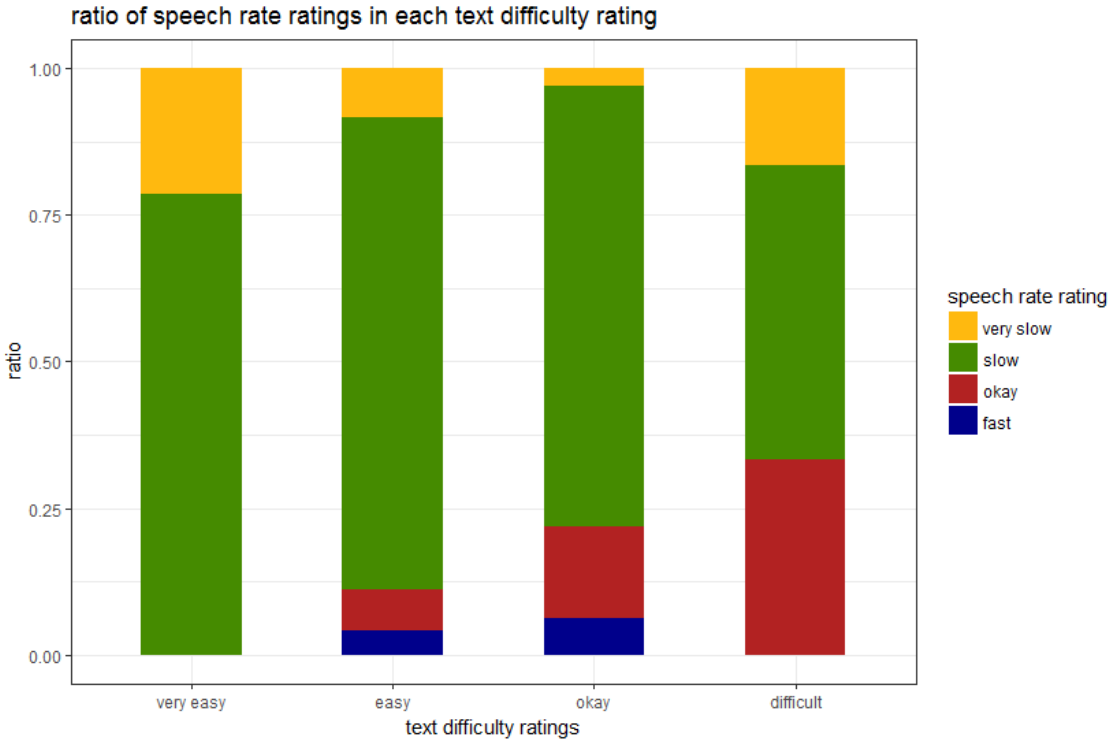

Figure 29: Proportion of speech rate ratings (yellow: “very slow”, green: “slow”, red: “okay”, blue: “fast”) for each rating of text difficulty (bars from left to right: “very easy”, “easy”, “okay”, “difficult”). The proportions of speech rate ratings as “slow” decreases slightly as the text difficulty is rated as “difficult”.

4.3.6.1.5 Discussion of the effects on participants’ ratings

Participant’s ratings of video and sound quality confirm that the experimental conditions were clearly distinct. As such, video quality ratings were significantly better when the speech was presented with the video of the speaker than with a freeze frame only. Still, none of the participants found the video quality “very good”. The reason is probably the rather low resolution of the videos (and freeze frames) of the speaker compared to commercial high quality movies that most people are used to nowadays. Nevertheless, participants reported after the experiment that the video quality was sufficient to perceive the lip movements. Sound quality also obtained better ratings when the speech was presented without noise than with noise. However, the model estimates for the sound quality ratings also suggest that the sound quality was not perfect, even in the condition without noise. Indeed, after the experiment some participants complained

about the eye-tracker making some noise. Furthermore, about half of the interpreters found that the overall volume was too low[37]. This could also have affected the sound quality ratings. Despite the low volume during the whole experiments, no participant disrupted the interpretation for more than 12 seconds, suggesting that the volume was still high enough to keep up with the translation. Interestingly, participants rated the sound quality more frequently as “good” or “very good” when the speech was presented in an audio-visual mode than in an auditory-only mode. This might suggest that audio-visual speech facilitated speech perception or at least that participants had the impression to better perceive the speech with audio-visual input than with auditory-only input.

Text difficulty ratings were not affected by the speech. This suggests that speeches indeed were perceived as being comparable with regard to the complexity of the content. On the whole, participants found the speeches “easy” to understand confirming that they were adapted to the participants’ level of English. Not surprisingly, text difficulty ratings depended on the auditory presentation (noise/no noise). This again shows that the speeches were perceived as being more complicated as soon as noise was added to the speech making it harder to understand the speaker. It validates background noise as a cognitive load factor on a larger scale in simultaneous interpreting and in speech comprehension as it has been observed for single words in the pretest. More interestingly, listeners rated text difficulty on the whole more often as “very easy” or “easy” than interpreters, which might indicate that interpreters indeed perceive work-load higher than listeners.

[37] As described in chapter 4.3.4, participants had no possibility to manipulate the volume before, during or after the experiment. The overall volume was the same for each participant and kept rather low in order to prevent a Lombard effect. The noise level, in contrast, was set according to the pretest.

Speech rate ratings did not differ significantly between the speeches indicating that speech rate was thoroughly controlled for. Most participants judged the speech rate as being “slow” indicating that the speech rate was adapted to the participants’ level of English and interpreting skills. It is interesting to note that neither auditory presentation nor the task influenced the ratings of speech rate although text difficulty ratings and speech rate ratings showed a trend to be weakly correlated. One reason might be that participants were able to perceive text difficulty and speech rate as two relatively independent parameters that affect cognitive load in simultaneous interpreting or listening understanding. Another aspect might be that only extreme values of text complexity affect the perception of speech rate. As there are – luckily – only very few “difficult”/”fast”-ratings, it is not possible to answer this question based on the data. More thorough investigations on how distinctly different potential stressors like speech rate, complexity of the speech’s content, numbers or other stressors are perceived would certainly be interesting in this respect.

Visual presentation (video/audio) did not affect text difficulty or speech rate ratings. This does not necessarily mean that visual presentation has no effect at all on work-load. Rather, its effect might be too weak or inconsistent compared to the auditory condition (noise/no noise) to reach significance in the statistical analysis, or the rating scale might be too gross to capture the subtle differences in text complexity or speech rate that participants might perceive thanks to visible lip movements. However, it is more probable that text difficulty or speech rate are simply not the right questions to ask to assess the effect of visible lip movements. Instead, participants might feel more comfortable when seeing the speaker’s face moving with the speech without necessarily relating this feeling to work-load. In this case, an interesting question would be how confident they feel or how satisfied they are with their performance. In this perspective, visual presentation would act on a more “emotional” aspect of stress, while auditory presentation affects a more “cognitive” aspect of stress.

4.3.6.2 Estimation of speech duration

Under high load conditions, time durations seem shorter to participants than under low load conditions. Duration estimations can therefore give interesting insights into work-load (see chapter 3.4.1). Participants were told before the beginning of the experiment that they would need to rate some general parameters like the speech duration. Directly after each speech, participants estimated the speech duration. The rating scale went from 0 seconds to 480 seconds (8 minutes). Each speech was 240 seconds long, but participants were completely naïve about the speech duration. 223 observations were made, with one missing value. The mean duration estimation was 225.2 seconds. From Figure 30 it becomes apparent that participants estimated the speech duration very differently (*SD*=66.14, range: 80-480 seconds). As can be seen in Figure 30, some participants vary largely in their judgements while others show much less variation. However, the estimations did not seem to depend on the auditory (noise/no noise) or visual presentation (audio/video) of the speeches for the estimations were nearly identical between the experimental conditions (*SD*=8.70). As the rating scale was limited from 0 to 480 and the measures therefore not infinite, I log-transformed all values before conducting further analyses.

Method and results

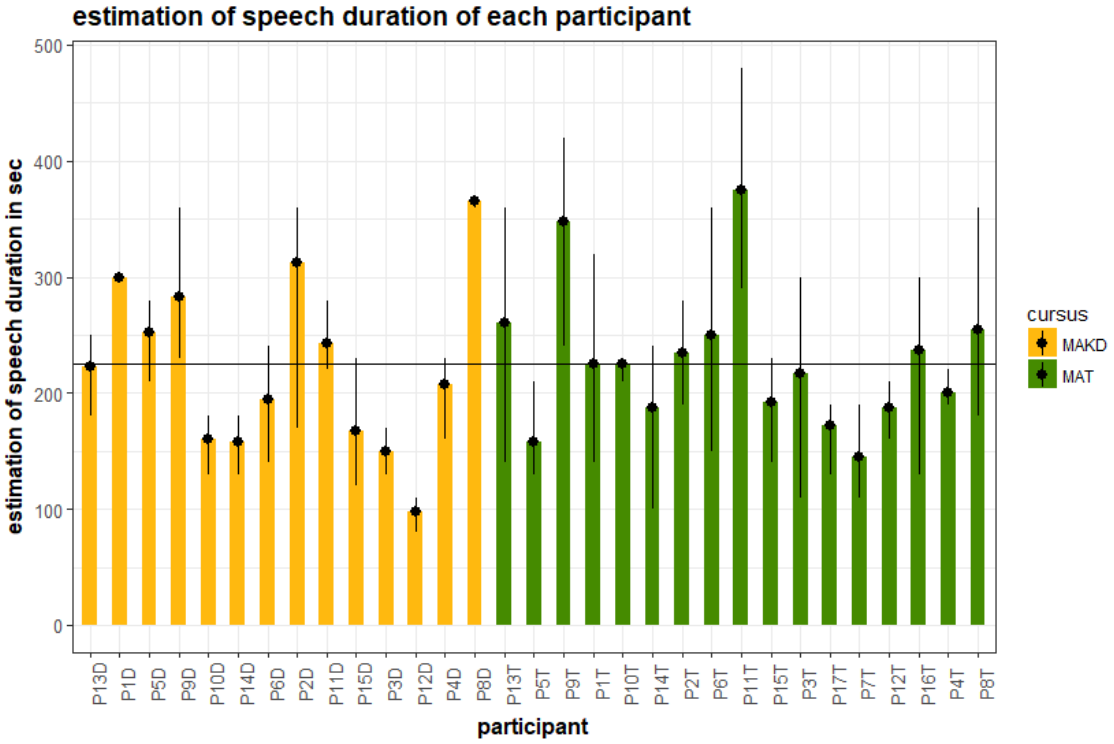

Figure 30: Speech duration estimations of each participant, including the range of their estimations. The black horizontal line corresponds to the overall mean of speech duration estimations.

A linear mixed model was constructed for speech duration estimation with random intercepts for participants. Fixed effects were participant, auditory presentation, visual presentation, trial, speech and noise level. Fixed effects were added one by one, e.g. comparing the model with and without the effect in question, and p-values were approached using maximum likelihood (results see Table 17:). All analyses were carried out in *R* version 3.2.2 (R Core Team, 2016) using the R-package *lme4* (Bates, Maechler, Bolker, & Walker, 2015).

Predictor variable	Likelihood ratio	DF	p-value
Participant	89.33	30	<0.001
Auditory presentation	1.097	1	0.295
Visual presentation	0.952	1	0.329
Trial	3.169	3	0.366
Speech	2.536	3	0.469
Noise level	2.423	1	0.120
Task	0.051	1	0.821

Table 17: Speech duration judgments: Degrees of freedom, likelihood ratio and p-value resulting for each predictor variable by comparing the model with and without the variable in question.

Despite the huge variability of participants in the random effects (*SD*=92.24), adding a participants-by-auditory presentation (noise/no noise) interaction still improved the model compared to a model with participant as only predictor variable (X^2(64,31)=46.22, p<0.039). This suggests that participants reacted differently on the auditory presentation. Some participants did not alter their estimation whether the speech was presented with or without noise (see for example participant P5D in Table 34, appendix 7.4.3: the estimate is in both cases the same). Others tended to estimate the speech duration shorter (see participant P9D in Table 34, appendix 7.4.3: her estimate in the noise condition is 20 seconds above the intercept[38], while it is 100 seconds above the intercept in the no noise condition) or longer (see participant P6D in Table 34, appendix 7.4.3: her intercept is 5 seconds below the intercept in the noise condition, but 50 seconds below the intercept in the no noise-condition) when it was presented with noise. The participant-by-visual presentation (video/audio)

[38] The intercept corresponds to the grand mean of speech duration estimations made by all participants.

interaction did not reach significance ($X^2(64,31)=28.949$, $p=0.572$). No further predictor contributed to improve the model. Complete model estimates are summarized in Table 34 (appendix 7.4.3). Figure 31 compares the fitted and the observed values for each participant in the noise- and no noise condition. Refitting the model to the ratio of estimation and real speech duration yielded comparable results.

Figure 31: Model fit for speech duration judgements. Fitted (green bars) versus observed (yellow bars) duration judgments made by each participant in the noise-condition (left) and no noise-condition (right). The model fit seems in most cases to be appropriate, but no overall pattern can be discerned from the observed or the fitted values. The variance in the data is largely explained by the variance between participants.

Time judgements can be moderated by the paradigm, e.g. if participants are aware prior to the experiment that they will be asked to judge the duration of the experiment or the trial (prospective paradigm) or if they are asked after the experiment to give an estimation of the speech duration (retrospective paradigm). Participants were informed beforehand that they would need to give their estimation about the speech duration. During listening to the first speech, however, they might have forgotten about the

speech duration judgments. In order to exclude that the paradigm affected participant's estimations, I conducted an analysis of variance on the interaction of auditory presentation and paradigm (trial 1: retrospective judgement, trials 2-4: prospective judgement) and contrasted the paradigms in a post-hoc test. The analysis of variance revealed no significant main effect of paradigm ($F(1,119)=1.171$, $p=0.281$) or auditory presentation ($F(1,119)=0.561$, $p=0.455$), nor an interaction between the paradigm ($F(1,119)=0.027$, $p=0.870$). Post-hoc comparisons using Tukey HSD revealed no significant interaction between any level of the variable paradigm and auditory presentation. Results of Tukey HSD-comparisons are summarized in Table 18.

interaction	difference	p-value
Retrospective:noise – prospective:noise	-21.576	0.789
Prospective :no noise - prospective noise	-12.392	0.885
Retrospective: no noise - prospective:noise	-28.365	0.666
Prospective: no noise – retrospective:noise	9.124	0.979
Retrospective:no noise – retrospective:noise	-6.849	0.996
Retrospective:no noise – prospective:no noise	-15.972	0.917

Table 18: Speech duration judgments: Results of the Tukey HSD post-hoc test contrasting each trial with all others.

Finally, I tested if there were any differences in variance between the experimental conditions or the listener and interpreter group. F-tests indicated no differences in variance between the noise and the no noise-condition ($F(61,60)=1.195$, $p=0.121$), the video and the audio condition ($F(61,60)=0.925$, $p=0.762$) or the listeners and the interpreters ($F(55,66)=0.973$, $p=0.922$).

4.3.6.2.1 Discussion of the effects on duration estimations

The hypothesis stating that participants would perceive speeches shorter during simultaneous interpreting with audio-visual speech than without audio-visual speech could not be confirmed: the absence or presence of visible lip movements did not influence the participants when estimating the speech duration. The most important predictor with regard to speech duration estimations was indeed the participants who varied considerably in their judgements and how they reacted on background noise. While some participants showed no difference whether noise was added to the speech or not others tended to indicate longer time durations for speeches with added noise. A third group of participants showed quite the opposite effect: They gave shorter estimations when the speech was presented with noise. It is not clear why the effect of noise was not consistent across participants. According to Block, Hancock and Zakay (Block, Hancock, & Zakay, 2010), duration estimations depend on how attention-demanding a task is. This effect is moderated by the paradigm: in prospective paradigms, e.g. participants are aware that they will be asked to assess the duration of the trial, the duration judgements decrease under high load, in retrospective paradigms, e.g. participants are not aware that they will need to estimate the duration, duration judgements increase under high load. If simultaneous interpreting is more attention-demanding than listening (as it requires more processes that are executed concurrently) interpreters should judge speech duration shorter than listeners. Moreover, simultaneous interpreting requires actively responding to the stimulus whereas listening is a passive task. Based on Block, Hancock and Zakay (2010), I expected interpreter's speech duration estimations to be shorter than those of the listeners. However, the statistical analysis revealed no task effect. Listeners and interpreters seemed to have estimated the speech duration similarly. A first thought might be that by the time the first speech started, participants might have forgotten about the duration judgment and might have become aware of it only with the second trial. If the awareness about the need to make a duration

judgements moderates the judgements, the first trial could have led to shorter duration estimations under high load (prospective paradigm) while all other trials should lead to longer duration judgements under high load (retrospective paradigm). This interaction could cover the effect of auditory or visual presentation. However, no significant main effect for paradigm or auditory presentation-by-paradigm interaction was found in the linear mixed model or the analysis of variance.

Interestingly, the mean of all duration judgements was rather accurate: the estimated duration was 225 seconds compared to the real duration of 240 seconds. In fact, as simultaneous interpreting is considered a very attention-demanding task by most interpreters[39], students start with very short speeches at the beginning of their training and gradually build up the speech duration throughout their training. At the University of Mainz, they start in general with 5 minutes of simultaneous interpreting in their first semester to end up with approximately 20 minutes at their final exam. Interpreting trainees are thus used to certain durations which might have helped them to give rather accurate duration judgments. In this case, there should be more variance in listeners' duration judgments compared to interpreters' judgments. But this was not the case: the F-test comparing the variance of listeners' and interpreters' duration judgements was not significant. A more probable explanation is a response bias because the scale was centered at the real speech duration (240 seconds). Participants could choose between 0 and 480 seconds, but the real speech duration was at the center of the scale. This could have influenced duration judgements, especially when participants were not too sure about the speech duration and therefore simply picked the "happy medium". In

[39] It is not without reason that researchers who study conference interpreting are often interpreters themselves and particularly interested in work-load (see for example Gile, 2009; Seeber & Kerzel, 2012).

further research it might be an option to use scales that are not centered at the real duration or that even change their center after each trial.

Another explanation might be that the participant-by-noise interaction accidentally captured some residual variance without being the real explanation. Maybe the duration participants were asked to assess was simply too long to be affected by work-load and participants responses therefore completely random. In their review, Block, Hancock and Zakay (2010) classify durations of one minute as “long”. In the present experiment, participants rated durations of four minutes. Block, Hancock and Zakay (2010) did not indicate the range of durations participants were asked to assess. Still, it cannot be excluded that a duration of four minutes is simply too long to allow for sensitive judgments.

4.3.6.3 Text-related questions

For each speech, all participants had to answer five questions on the speech content. All answers were structurally similar, for example the number of items in an enumeration was the same for all options, and were equally probable solutions with respect to the speech content. Questions and their answers for each speech are listed in the appendix (appendix 7.3). Participants were to pick to correct answer from three options that were given or to choose “I don’t know” by pressing the corresponding key. The participants’ answers were automatically checked. In total, each participant answered about 13 out of 20 questions correctly (*MD*=13, mode=13, range: 10-16). The number of correct answers was not distributed equally across all speeches. While the number of correct answers reached 131 for the speech “air travel” and 114 for the speech “Greece”, the number of correct answers was much lower for the speeches “demographic change” and “work” (93 and 79 respectively). It is important to bear in mind, however, that the latter two speeches were those presented with masking noise, whereas the first two were always presented without noise. Listeners gave 217 correct answers out of a total of 340 (63.8%); interpreters answered 192 out of 280 questions correctly

(68.6%). Figure 32 depicts the proportions of correct and wrong answers according to the task (interpreter/listener) and the auditory presentation (noise/no noise).

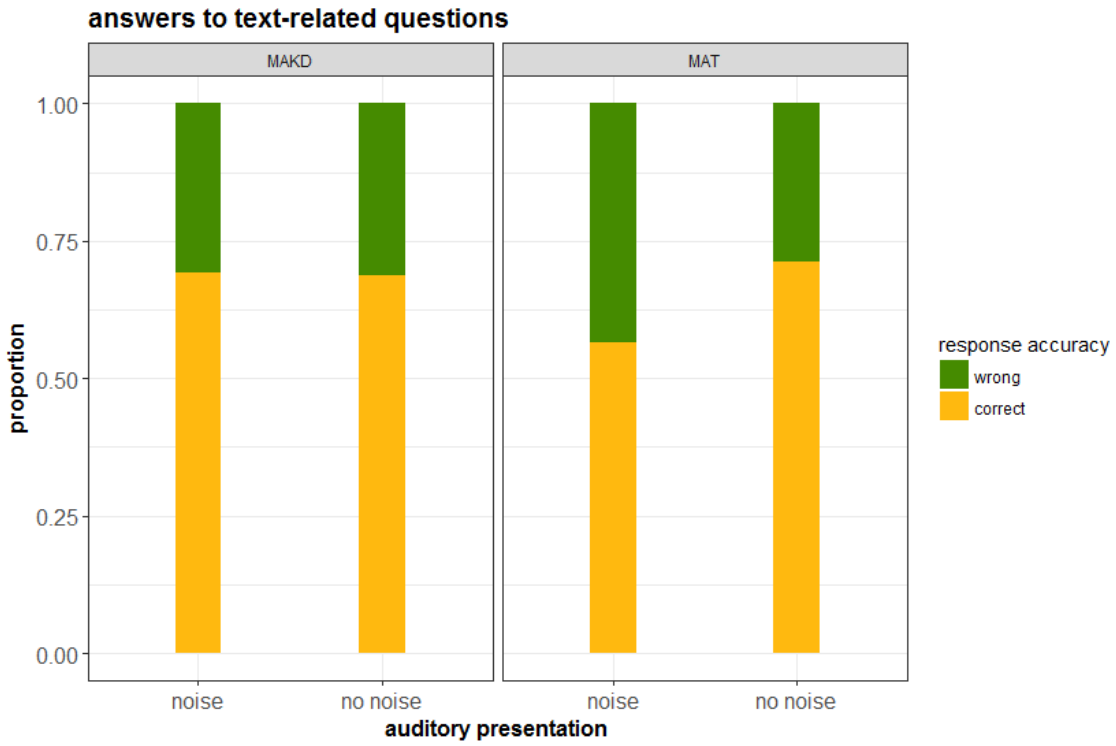

Figure 32: Text-related questions: Proportions of correct (yellow) and wrong (green) answers according to the task (facets: “MAKD”: interpreting, “MAT”: listening) and the auditory presentation (x-axis).

The data on response accuracy (correct/wrong) was fitted to a logistic generalized mixed model. Random effects included participant-by-trial intercepts and trial-by-question intercepts[40]. P-values for fixed effects were approximated with maximum likelihood by comparing the model with the effect in question against the model without the effect in question. The resulting p-values for each predictor variable are summarized in Table 19. P-values for levels of fixed effects were assessed using a Wald test. All

[40] The predictor variable *question* denotes the number of the question (1 to 5).

analyses were carried out in *R* version 3.3.2 (R Core Team, 2016) using the *R*-package *lme4* (Bates, Maechler, Bolker, & Walker, 2015).

Predictor variable	DF	Likelihood ratio	p-value
Auditory presentation	1	4.062	0.044
Visual presentation	1	0.005	0.943
Noise level	1	1.521	0.217
Task	1	1.832	0.176
Auditory presentation*task	3	9.519	0.0231

Table 19: Text-related questions: degrees of freedom, likelihood ratio and p-value for each predictor variable when compared to the base model without any predictor variable.

At first glance, there was a significant effect of auditory presentation (no noise/noise) on the intercept (see Table 19). But the main effect of auditory presentation disappeared after adding an interaction with the predictor variable task (main effect auditory presentation: *Estimate*= -0.018, *SE*=0.296, *z*= -0.061, *p*=0.952). In this model, the predictor variable task has a significant effect on the intercept (reference level: interpreter, *Estimate*= -0.647, *SE*=0.281, *z*= -2.304, *p*=0.021) suggesting that listeners were overall half as likely to give the correct response. When the predictor variable task is already included in the model, the interaction auditory presentation – task is still nearly significant (reference level: interpreter – noise, *Estimate*=0.771, *SE*=0.401, *z*=1.920, *p*=0.055) suggesting that compared to interpreters, listeners were about twice as likely to give a correct response when no noise was added to the speech. The auditory presentation-by-noise level-interaction was not significant when comparing the model with and without the interaction ($X^2(7,2)=2.088$, $p=0.352$). The model fit is depicted in Figure 33.

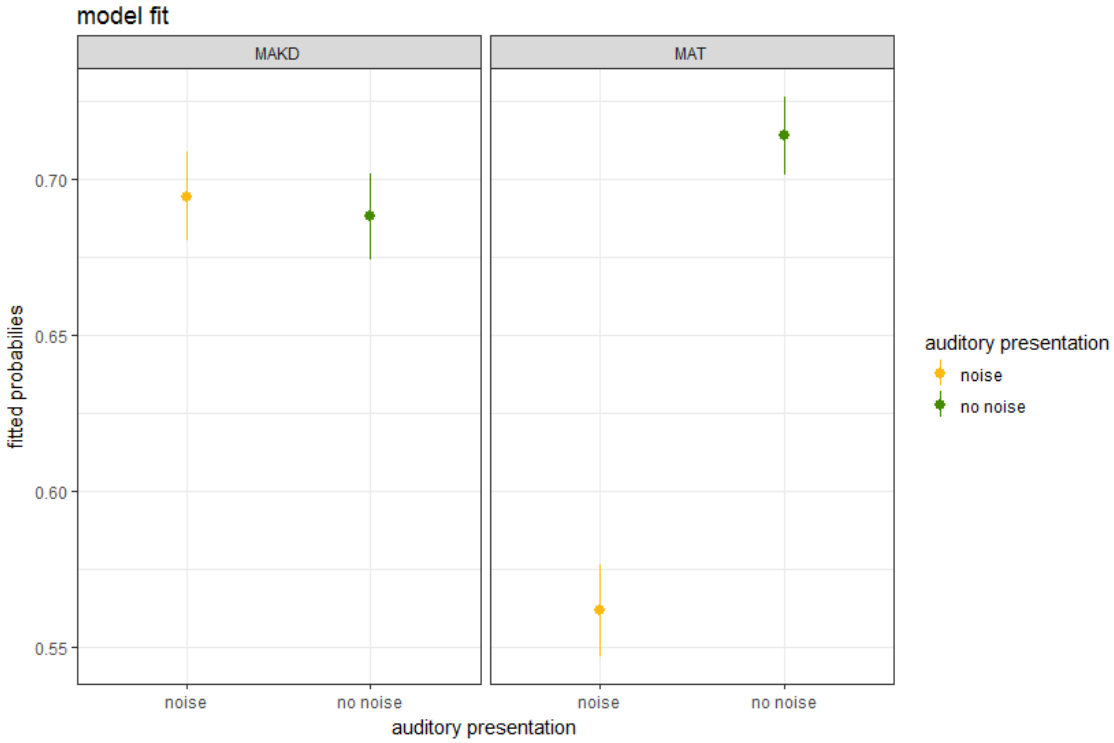

Figure 33: Fitted probabilities for correct answers to text-related questions. Yellow dots stand for the condition with noise, green dots for the condition without noise. The dot indicates the mean with a 1.5-standard deviation. The left facet shows the fitted probabilities for interpreters ("MAKD"), the right one those for listeners ("MAT"). In the condition without noise (green pointranges), the fitted probability to observe a correct answer is higher for listeners than for interpreters. The fitted probability to observe a correct answer in the condition with noise (yellow pointranges), however, was higher for interpreters.

4.3.6.3.1 Discussion of the effects on text-related questions

The statistical analysis does not confirm the hypothesis stating that participants would answer more text-related questions correctly in the video condition than in the audio condition. Visual presentation (audio/video) had no significant effect on the number of correctly answered text-related questions. However, background noise had an impact on response accuracy indicating that speech analysis and storage processes necessary to answer text-related questions, are generally sensitive to high load conditions and that text-related questions can work as an indicator for higher load conditions. With regard to the visual presentation, this means probably either that the effect of audio-visual speech is too weak to be noticeable in response accuracy or that audio-

visual speech does not enhance the storage of speech information at all. Furthermore, the task plays a significant role: In the condition without noise, listeners are slightly more probable to give a correct response than interpreters. This finding is in line with Lambert (1988) and Gerver (1974) who found higher scores in a semantic recognition test after listening compared to simultaneous interpreting (Lambert, 1988; Gerver, 1974). Lambert (1998) concludes that listeners can fully focus on the processing of the input speech without having to split their processing capacities or their attention on multiple tasks like in simultaneous interpreting.

Surprisingly, listeners were much more affected by noise than were interpreters who seem to be not affected at all. A first tentative explanation might be that listeners might have had overall higher noise levels compared to interpreters. Indeed, there were 6 listeners with noise level at 0.2 and one with noise level at 0.3, while there were 5 interpreters with noise level at 0.2 and none with a noise level at 0.3. Comparing the statistical model with and without an auditory presentation-by-noise level-interaction confirmed that the auditory presentation-by-noise level-interaction was not significant. The task-by-noise-interaction was thus not confounded with the predictor variable noise level. This raises interesting questions with regard to information processing in simultaneous interpreting and listening. It seems as if interpreters compensate better for the noise condition than listeners do. The need to anticipate during simultaneous interpreting could provide some explanation. Interpreters need to give an immediate translation of the speech they hear. Structural differences between the source and the target language (syntax, grammar) force the interpreter to constantly make predictions about how the speaker will finish his sentence. The interpreter has – so to say – to think along the lines of the speaker. In comparison, listening is a more passive task that does not require anticipating the speaker's intentions to the same extent as in simultaneous interpreting. The need to anticipate could contribute to make interpreters less dependent on the auditory presentation of the speech, e.g. if the speech is presented with or without

noise. Tightly related to this tentative explanation is the notion of depth of processing. The stronger need to anticipate when noise obscures parts of the source speech could have entailed a deeper processing of the speech's content and subsequently better recall performance. Another aspect might be task motivation. Of course, listeners can (in theory at least) make the same effort as interpreters which should give comparable performances. The listeners were made aware prior to the experiment that they would need to answer some text-related questions. But in contrast to interpreters, listeners were not forced to constantly make a response, e.g. to maintain the translation. When the noise made it harder to understand the speech, they could have lost interest and lowered their attention.

4.3.6.4 Translation accuracy

In order to prepare the data on translation accuracy for the statistical analysis, I divided all source texts into small segments and grouped them into different categories. Core segments are the "red thread" of the speech. They include in general the verb and its complements (subjects, objects, adverbial complements), e.g. all segments that are essential to understand the sentence or the speaker's intention. Secondary information includes information that could be left out without destroying the sentence structure, like adjectives or intensity particles. Repetitions correspond to segments that already appeared in the speech, even if the information was formulated differently. Information on time and place and – in a broader sense - logical sentential connectors or phrases that give information on how important the following segment will be (examples: "I will give you an example", "first" or "second", "because", "I point out", etc.) are labelled context information. Finally, *fillers* cover all segments that actually contain no information at all. The following example in Table 20 illustrates the segmentation of the source text sentences:

Source sentence		**Maybe I should note at this point that obesity is a very widespread disease in many countries of the western world.**
Segment 1:	Filler	Maybe
Segment 2:	Context information	I should note
Segment 3:	Secondary information	at this point
Segment 3:	Core information	(that) obesity is a (very widespread) disease
Segment 4:	Secondary information	very widespread
Segment 5:	Secondary information	in many countries
Segment 6:	Core information	in (of) the western world

Table 20: Example of segment categorization

The reason for this categorization was that interpreters might deliberately skip redundant or less important segments and concentrate on the core information in order to secure a translation of quality. By classifying information segments according to the importance with regard to the topic of the speech it was possible account for strategic choices in the statistical analysis. The number of segments in each speech is reported in Table 21.

Speech	Core information	Context information	Filler sentences	Repetitions	Secondary information	Total
Air travel	91	10	7	10	46	164
Demographic change	92	3	5	17	41	158
Greece	92	12	8	16	39	165
Work	93	12	0	24	50	179

Table 21: Translation accuracy: Number of segments of each category in each speech

For each segment, I checked if the translation was available and if it corresponded to the source segment. In the end, I obtained a table containing information about the participant, the text, the number of the trial (one to four), the condition (audio/video, with noise/without noise), the source segment and the translation (translation correct, translation available, but not correct, translation not available).

Evaluating interpreting performance bears an important pitfall: the evaluation is rarely objective and “quality” can quickly be mistaken as the way of speaking that seems most familiar to someone. For this reason, interpreting performance is most often evaluated independently by two judges. Due to practical constraints, however, this was not possible. The approach to investigate the primary task performance, e.g. the oral translation by solely verifying the translation segment by segment, is a compromise. It does not cover all aspects which could be associated to interpreting quality. For instance, it does not take into account terminological or grammatical errors, unfinished sentences or wrong intonations (which are very annoying for the audience). In this respect, this approach is of course incomplete. It has, nevertheless, the merit of being

largely objective and independent from one's own judgement. In this way, it compensates (at least partially) for the lack of judges.

A total of 55 recordings, four recordings per participant, were used for analysis. One recording was completely missing. On average, 39.75% of all translations segments were missing (*MD*=39.41, range: 16.07%-76.88%)[41]. Not taking into account missing translations, participants translated 55.33% of all segments correctly (*MD*=54.73%, range: 39.04-76.88%). With regard to the speeches, the proportions of correct translations was generally higher for the speeches "air travel" and "demographic change" (59.32% and 68.13% respectively) than for the speeches "Greece" and "work" (51.64% and 43.77% respectively). The latter two were also characterized by a higher proportion of missing translations (43.63% and 50.40% respectively compared to 36.58% for the speech "air travel" and 26.90% for the speech "demographic change"). It is, however, important to bear in mind that the speeches "Greece" and "work" were always presented with noise while the speeches "air travel" and "demographic" change were never masked by background noise. 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 are also those with the fewest proportion of missing translations (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 analyze translation accuracy in each speech. Missing values were treated as missing and not as a proper level of the response variable. The effect on translation accuracy was captured with participant-by-trial random intercepts (*SD*=0.26) and random

[41] Please recall that translation segments could be single words like intensity particles or adjectives that do not contain crucial information or repetitions; segments that can easily be left out without destroying the core information of the speech.

intercepts for participant (*SD*=0.44), and auditory presentation, visual presentation, noise level and segment category as fixed effects. Cursus was not tested as fixed effect as recordings were only available for interpreters. Speech was not tested as fixed effect because only two of the four speeches were presented with masking noise. Including speech as predictor variable would thus have led to rank deficiency and to probably wrong model estimates for the predictor variable speech. Including the time course captured as first order polynomial in the random effect structure did not improve the model (*F(*3,6)=6.004, *p*=0.112), indicating that performance kept at a constant level and did not deteriorate (or increase) at the end of the speech. 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. P-values for the levels of categorical predictor variables were estimated using a Wald test. Results of model comparisons are reported in Table 22. All analyses were carried out in *R* version 3.3.2 (R Core Team, 2016) using the package *lme4* (Bates, Maechler, Bolker, & Walker, 2015). Plotting was done using the R-package ggplot 2 (Wickham, 2009).

Predictor variable	DF	Likelihood ratio	p-value
Auditory presentation	4,1	13.312	<0.001
Visual presentation	4,1	0.047	0.828
Noise level	4,1	0.217	0.641
Segment category	7,4	17.289	0.002

Table 22: Translation accuracy (missing translations as missing values): Degrees of freedom, likelihood ratio and p-value of model comparisons. Missing translations are not taken into account. Fixed effects were added one by one and the resulting model was compared to a model without the effect in question.

There was a significant effect of auditory presentation on the intercept (reference level: condition without noise, *Estimate*= -0.4856, *SE*=0.124, *Z*=10.44, *p*<0.001) indicating that the probability to observe a correct

translation decreases in the condition with noise by approximately 3.7% compared to the condition without noise. There was also a significant effect of segment category (reference level: core information), however not all levels had a significant effect on the intercept. Significant effects were found for the category “context information” (*Estimate*=0.920, *SE*=0.313, *Z*=2.936, *p*=0.003) and “secondary information” (*Estimate*=0.296, *SE*=0.129, *Z*=2.301, *p*=0.021). According to the model, there is a higher probability to observe a correct translation in these two categories compared to the reference level “core information” (approximately 3.8% for the category “context information” and 1.6% for the category “secondary information”). Further variables, like visual presentation or noise level did not significantly improve the model (see Table 22). The effects on translation accuracy are depicted in Figure 34.

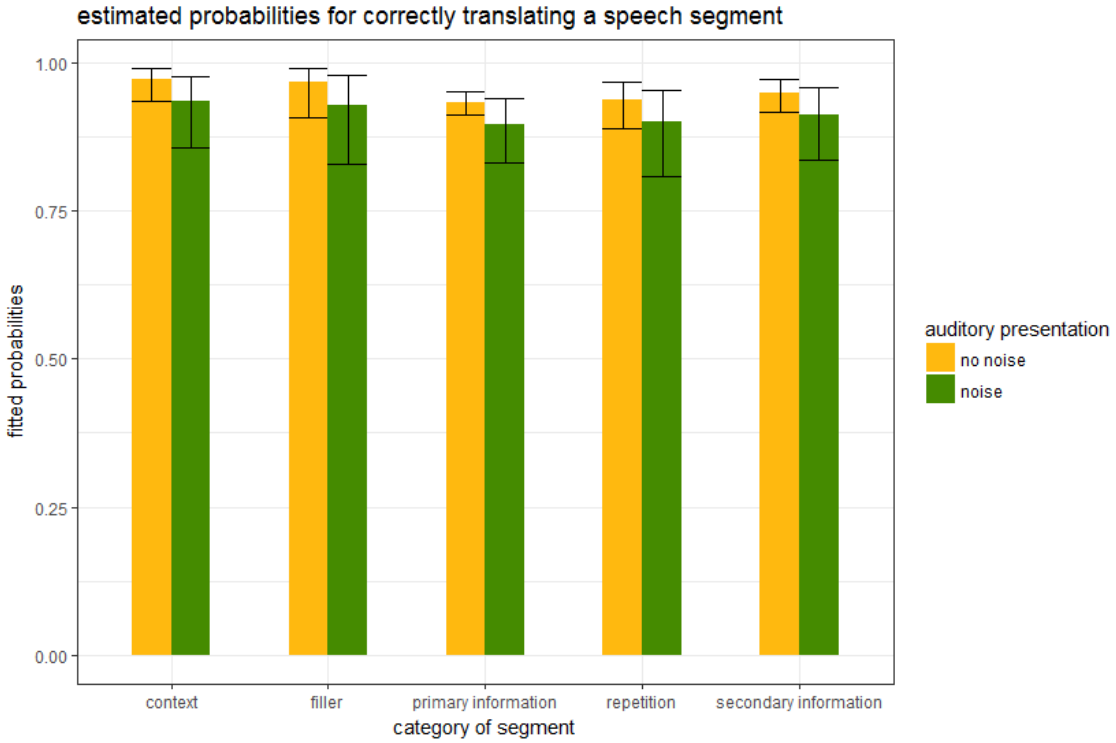

Figure 34: Model fit for translation accuracy (with missing translations treated as missing value). Yellow bars show the fitted probability to observe a correct translation in the condition without noise; green bars show the fitted probability to observe a correct translation in the condition with noise. For better readability, the log-odd estimates are recalculated as probabilities. In the condition with noise, the fitted probabilities to observe a correct translation are lower.

Even though the model confirms the effect of the auditory presentation on translation accuracy, the probabilities for correct translations as suggested by the model seem unrealistically high: about 90% of all speech segments are –according to the model – correct translations. The reason for these high probabilities is that missing translations have not been taken into account. In other words, the model says that if a segment is translated, it is mostly correctly rendered. However, missing translations are not innocent in that respect as missing translations can result of cognitive overload. The statistical analysis reported below considers missing translations as wrong translations. As above, it is a generalized mixed model with participant-by-trial random intercepts (*SD*=0.526) and random intercepts for participants (*SD*=0.331). The model failed to converge with time as random effect. Fixed effects cover auditory presentation (reference level: condition without noise) 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. P-values for the levels of categorical predictor variables were estimated using a Wald test. Results for each predictor as obtained by comparing the model with and without the predictor in question are reported in Table 23. All analyses were carried out in *R* version 3.3.2 (R Core Team, 2016) using the R-package *lme4* (Bates, Maechler, Bolker, & Walker, 2015). Plotting was done using the R-package *ggplot2* (Wickham, 2009).

Method and results

Predictor variable	DF	Likelihood ratio	p-value
Auditory presentation	4,1	18.941	<0.001
Visual presentation	4,1	2.059	0.151
Noise level	4,1	2.179	0.140
Segment category	7,4	182.657	<0.001

Table 23: Translation accuracy (missing translations as wrong translations): Degrees of freedom, likelihood ratio and p-value of model comparisons. Fixed effects were added one by one and the resulting model was compared to a model without the effect in question.

Again, a significant effect for auditory presentation (*Estimate*= -0.738, *SE*=0.148, *z*= -4.994, *p<0.*001) was found. According to the model, the probability for correctly translating a speech segment decreases across all segment categories by approximately 17.26% when noise is added to the speech. With regard to the segment category, all categories were significant when considering missing translations as wrong translations. 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 for secondary information by 15.56% (*Estimate*= -0.670, *SE*=0.052, *z*= -12.886, *p*<0.001) compared to core information segments. Further predictors like visual presentation or noise level failed to be significant (see Table 23). A visual-by-auditory presentation interaction did not reach significance neither (X^2(10,2)=3.887, *p*=0.143). The refitted probabilities of the model that treats missing translations as wrong translations are depicted in Figure 35. As can be seen, the probability for a correct translation is about 70% for core segments when no noise is added to the speech.

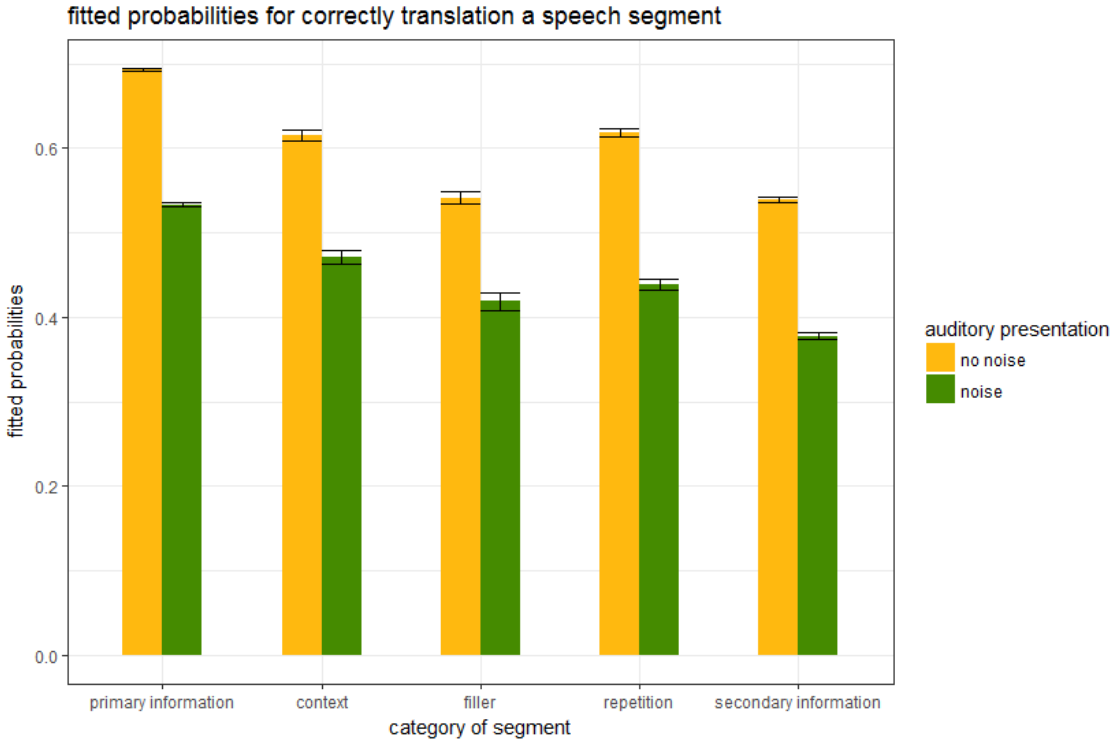

Figure 35: Model fit for translation accuracy (missing translations as wrong translations)for all segment categories. Yellow bars show the fitted probability in the condition without noise; green bars show the fitted probability in the condition with noise. For better readability, the log-odd estimates are recalculated as probabilities. The model treats missing translations as wrong translations. The probability to observe a correct translation is lower in the condition with noise (green bars) than in the condition without noise (yellow bars).

4.3.6.4.1 Discussion of the effects on translation accuracy

The hypothesis stating that the number of correctly translated segments is higher during simultaneous interpreting with audio-visual speech than without audio-visual speech was not confirmed. Visual presentation did not affect the translation accuracy. Two explanations are conceivable: a) translation accuracy is not affected by work-load; b) visual presentation has no (or only a weak) effect on work-load. I will consider both explanations in turn. The first explanation seems highly improbable, for several reasons. First, there is a large body of research suggesting the opposite (see Anderson L. , 1994; Gerver, 1974; Gerver, 2002), even though the authors' definitions of translation accuracy may diverge (see chapter 3.4.2). Second, the statistical analysis revealed a significant effect

for background noise on translation accuracy indicating that translation accuracy was affected by work-load and that the method chosen to evaluate and analyze translation accuracy was sensitive enough. The second explanation suggests that audio-visual speech has no or only a very weak effect on cognitive load. A statistical explanation could hold that the very huge effect of segment category covered the much weaker effect of visual presentation. This explanation seems not to hold given that the predictor variable visual presentation did not reach significance even when it was the only predictor variable added to the model.

At this point it is maybe important to recall the assumption that lip movements enhance speech perception by resolving the ambiguity in the auditory stream (see chapter 3.3.4). This would mean that participants may not necessarily benefit from lip movements when the auditory stream is perfectly intelligible but only when the auditory stream is defective, e.g. in the noise condition. The visual-by-auditory presentation interaction, however, failed to reach significance. If interpreters did not benefit from lip movements at all, neither in general, nor in the noise condition, then either because they could not see the lip movements properly or because they used strategies that made lip movements useless. Could participants see the lip movements properly? According to participant's ratings, the video quality was not perfect. Still, all participants reported that they had been able to see the speaker's lip movements without problems and according to the eye-tracking data, participants looked approximately 54% of the time on the region of interest defined around the speaker's lips. The percentage of fixations was not systematically higher in the conditions with noise than without noise (fixation percentage during speech with noise: 52.7%, fixation percentage during speech without noise: 55.5%). Despite the participants' reports, it is still possible that lip movements were not perceived well enough in order to extract any useful information. It is therefore not possible to completely rule out this explanation. However, most interpreters reported that it did not make any difference to them whether they translated with or without visible lip movements. Lip

movements may help to disambiguate auditory information the moment the speaker pronounces a word. But often, interpreters make predictions about what the speaker will be saying next. Hence, they might not depend as much on the speaker and benefit as much from visible lip movements as one could expect. If this is true then audio-visual speech should make a difference when no prediction is possible, for example in a shadowing task where anticipation is made impossible by using syntactically and semantically incoherent sentences.

With regard to interpreting strategies it is interesting to note that interpreters prioritize information. According to the statistical model, core segments had the highest chance to be rendered in the target language, while filler sentences or secondary information were most often omitted, independently of the auditory presentation. Interpreters tried to preserve the core structure and gave away on segments that did not seem essential to them. This finding also validates the scheme used to classify the speech segments.

4.3.6.5 Cognate translations

For analysis purposes, I extracted all words from the English source texts that showed considerable phonological or orthographical overlap with an existing German word. Some cognates appeared more than once in a speech. They were counted repeatedly for their (most appropriate) translation could change according to the context. Furthermore, I expected that participants would not translate each sentence and would therefore miss some cognates. Counting repeated cognates repeatedly allowed me to take into account missing translations. In total, 350 words were extracted (repetitions included). 145 cognates were repetitions. Table 24 shows the number of cognates in each speech.

	Air travel	Demographic change	Working conditions	Greek economic crisis
Number of cognate pairs with repetitions	83	69	94	104
Number of cognate pairs without repetitions	50	36	56	63

Table 24: Number of cognates in each text

I used the Levenshtein distance made available by the *stringdist* package in R (van der Loo M. , 2014) to compare the similarity of the cognate pairs. The Levenshtein distance is the number of changes (deletions, transpositions, insertions) that are necessary to obtain the target word. In order to take the word length into account, I calculated it as the percentage of the word that needed to be transformed: Levenshtein distance/number of characters = ratio Levenshtein distance. The mean ratio Levenshtein distance for the orthographical form of the cognate pairs was 0.33 (*MD*= 0.17, range: 0.0-1.4) which means that 33% of the words needs to change in order to obtain the cognate translation. As some cognate pairs differ considerably in their orthographical form, but are nevertheless phonologically very similar (for example: *techniques* – *Technik*), I converted additionally each word in its phonological code and recalculated the ratio Levenshtein-distance. Based on the phonological form, the mean ratio Levenshtein distance was 0.1 (*MD*= 0.0, range: 0.0-0.67), meaning that only 10% of the words needed to change phonologically in order to obtain the cognate translation. This confirms the high similarity between the cognate pairs, particularly at a phonological level. Cognate pairs that obtained a ratio Levenshtein distance over 0.167 (third quartile) are summarized in Table 35 (see appendix, chapter 7.4.4).

For each English cognate, all possible translations were checked on two online dictionaries (linguee, dict.cc) and their frequency class according to the corpus of the University of Leipzig (Quasthoff, Goldhahn, & Heyer, 2013) was noted. A cognate translation was considered as “high frequency cognate” if the German cognate was indeed the most frequent

translation or only one frequency class below the most frequent translation (example: *international* – *internationaal*). If another translation was considerably more frequent, e.g. at least two frequency classes higher, the cognate translation was considered to be a “low frequency cognate” (example: *to implement* – *implementieren,* the more frequent German translation in this context is *einführen* or *umsetzen*). If the meaning of the cognate translation did not correspond to the contextual meaning of the English source word, the cognate translation was categorized as “false friend” (example: *company* – *Kompanie:* the German word denotes a ballet group or a military unit). Table 25 displays the number of cognates in each category and each text.

Cognate category	**Air travel**	**demographic change**	**work conditions**	**Greek economic crisis**
High frequency cognate	34	38	49	62
Low frequency cognate	37	18	34	27
False friend	12	13	11	15

Table 25: Number of cognates in each category

Fifty-five recordings by fourteen participants where included in the analysis. One recording was missing. Overall, participants translated 35.1% of all candidates (including repeated cognates) as cognates (*M*=35.1%, *MD*=35.7%, *SD*=7.0%) and opted in 26.5% of the cases for non-cognate translations (*M*=26.5%, *MD*=26.9%, *SD*=4.4%). About 38.4% of all candidates were not translated at all (*M*=38.4%, *MD*=36.7%, *SD*=8.7%). In these cases, the whole phrase was missing, either because participants were not able to keep up or because they deliberately chose to strip or summarize information that seemed redundant or not essential to them.

Method and results

The highest percentage of cognate translations were observed in the video condition without noise (40.0%), followed by the audio condition without noise (36.1%), the video condition with noise (34.3%) and the audio condition with noise (31.5%). The audio condition with noise had also the lowest percentage of non-cognate translations (19.6%) and the highest percentage of missing values (49.0%), whereas the video condition without noise showed the highest percentage of non-cognate translations (32.5%) and the lowest percentage of missing values (27.4%). The video condition with noise (non-cognate translations: 24.8%, missing values: 40.8%) and the audio condition without noise (non-cognate translations: 31.7%, missing values: 32.2%) were situated in between. Figure 36 depicts the ratio of cognate and non-cognate translations of each cognate category (high and low frequency cognates, false friends) in each experimental condition, recalculated without missing values. In relative terms, fewer cognate translations were observed for high and low frequency cognates when participants translated the speech without noise than with masking noise. Still, the largest differences seem to be between the different cognate categories instead of the conditions.

Method and results

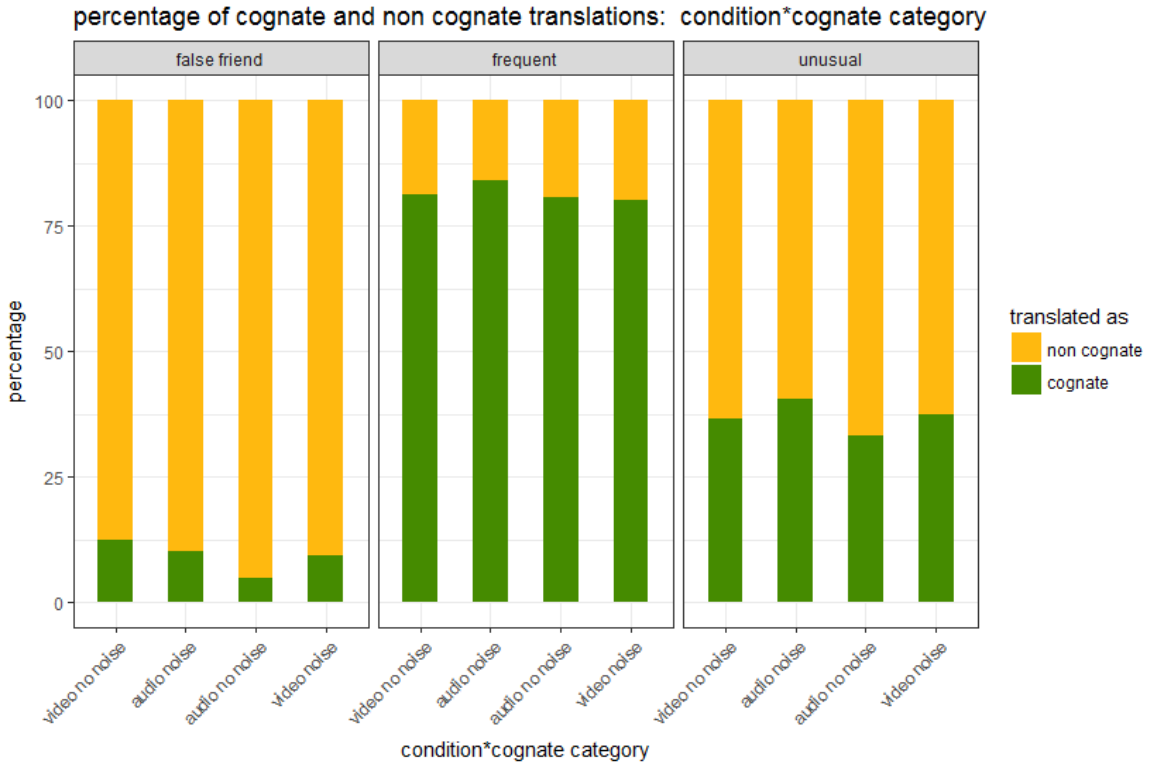

Figure 36: Proportions of cognate and non-cognate translations in all four experimental conditions, recalculated without missing values. Yellow bars stand for non-cognate translations, green bars for cognate translations. Facets correspond to cognate categories (left: false friends, center: high frequency cognates, right: low frequency cognates). High frequency cognates are more often translated as cognate than low frequency cognates or false friends: Differences between conditions seem to be rather low.

A generalized mixed model was fit to the data. I chose a logit distribution to accommodate the binomial nature of the data (candidate translates as cognate or not). Random effects included random slopes for participants (*SD*=0.297), word category (*SD*=0.244) and speeches (*SD*=0.630). Participant-by-trial intercepts did not contribute to improve the model (*SE*=0.000). The time course was not considered as the cognate candidates were not evenly distributed across the speech. Fixed effects included auditory presentation (noise/no nose), visual presentation (audio/video), cognate category (high frequency cognate, low frequency cognate, false friend), trial (one to four) and noise level (0.1 or 0.2). Fixed effects were added one by one and the resulting models were compared with likelihood ratio approximation. The results of the model comparisons

are summarized in Table 26. All analyses were carried out in *R* version 3.3.2 (R Core Team, 2016) using the package *lme4* (Bates, Maechler, Bolker, & Walker, 2015). Plotting was done with the R-package *ggplot2* (Wickham, 2009).

A significant effect was found for the predictor variables cognate category ($X^2(6,2)=1057.38$, $p<0.001$, reference level: false friends) and auditory presentation ($X^2(6,1)=5.79$, $p=0.016$, reference level: noise condition). When comparing the model with the predictor variable cognate category only and both predictor variables, e.g. cognate category and auditory presentation, however, the effect for auditory presentation was not significant anymore ($X^2(6,1)=0.276$, $p=0.599$). In the final model, all levels of the predictor variable cognate category were highly significant (false friends: *Estimate*=-2.337, *SE*=0.204, *z*= -11.455, $p<0.001$; low frequency cognates: *Estimate*=1.742, *SE*=0.181, *z*=9.617, $p<0.001$; high frequency cognates: *Estimate* =3.836, *SE*=0.180, *z*=21.343, $p<0.001$).

Predictor variable	DF	Likelihood ratio	p-value
Auditory presentation	1	5.79	0.016
Visual presentation	1	0.22	0.636
Cognate category	2	1057.38	<0.001
Speech	3	7.12	0.068
Trial	3	0.61	0.894
Noise level	1	0.02	0.880

Table 26: Cognate translations: Degrees of freedom, likelihood ratio and p-values for predictor variables compared to the base model without any predictors.

According to the model, it is about 5 times more likely to translate a candidate as cognate if the candidate is a low frequency cognate compared to when the candidate is a false friend. If the candidate is a high frequency cognate, a cognate translation is even about 46 times more likely than when the candidate is a false friend. Figure 37 compared the observed ratios versus fitted probabilities for a cognate translation within all cognate categories and the four experimental conditions. As can be

seen, neither the observed values nor the fitted values differ very much between the experimental conditions. In fact, the differences between the noise and the no noise condition or the video and the audio condition failed to be significant even when constructing a generalized mixed model with only high frequency cognates (auditory presentation: $X^2(1)=0.063$, $p=0.801$; visual presentation: $X^2(1)=0.397$, $p=0.529$; random intercepts for participants, word category and speech), low frequency cognates (auditory presentation: $X^2(1)=1.002$, $p=0.317$, visual presentation: $X^2(1)=0.083$, $p=0.773$; random intercepts for participants word category and speech) or false friends (auditory presentation: $X^2(1)=0.111$, $p=0.740$, visual presentation: $X^2(1)=2.228$, $p=0.107$, random intercepts for participants, word category and speech). Refitting the model without repeated cognates yielded comparable results (auditory presentation: $X^2(1)=0.88$, $p=0.348$; visual presentation: $X^2(1)=0.11$, $p=0.738$; random intercepts for participants, word category and speech). The same holds for a combined analysis of the pilot study (n=6) and the main study (n=14) (auditory presentation: $X^2(1)=0.88$, $p=0.348$; visual presentation: $X^2(1)=0.11$, $p=0.738$; random intercepts for participants, word category and speech). The only significant fixed effect in the combined analysis was again cognate category ($X^2(2)=929.84$, $p<0.001$).

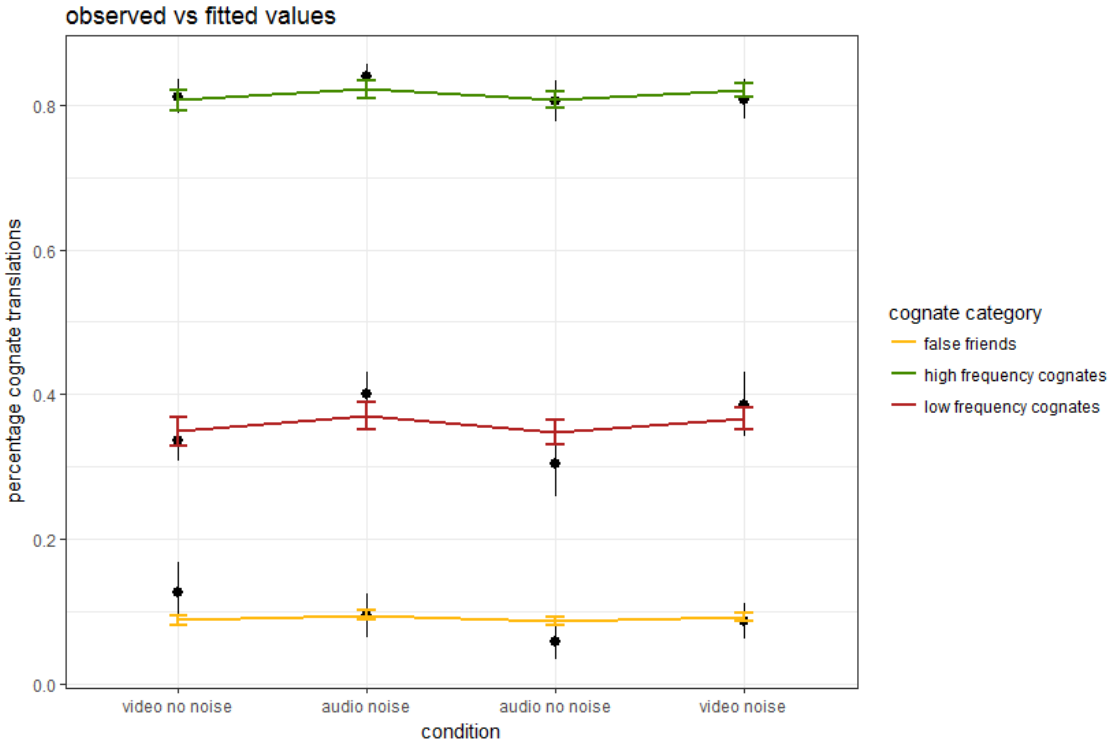

Figure 37: Cognate translations: Observed (black dots) versus model fit (colored lines). The green line depicts the fitted values for high frequency cognates, the red line those for low frequency cognates and the yellow line those for false friends. The line has only been chosen to ensure better readability.

4.3.6.5.1 Discussion of the effects on cognate translations

The analysis revealed a large effect for the cognate category. High frequency candidates have by far the highest chance to be translated as cognate, whereas low frequency candidates and in particular false friends have a less than 50%-chance to be translated as cognate. For false friends the fitted probability to be translated as cognates is even below 10%. This finding is in line with the pilot study and suggests that interpreters try to avoid translations that might sound awkward to the listener and are quite effective in doing so. This corresponds to a phenomenon described by Baker (Baker, 1996) called *normalization*, which denotes the “tendency to exaggerate features of the target language and to conform to its typical patterns” (Baker, 1996, p. 183).

The hypothesis stating that the number of cognate translations should be lower during simultaneous interpreting with audio-visual speech has not

been confirmed. This stands in contrast to the pilot study where visual presentation had a significant effect on cognate translations. Even in a combined analysis of the pilot study and the main study, however, neither the visual presentation nor the auditory presentation had any effect on cognate translations. This is surprising as the experimental material and the statistics used to investigate cognate translations were exactly the same. In both cases, participants were in their second year of training and studied English as their first or second foreign language so it can be assumed that their interpreting skills were comparable. One difference though is the number of participants: Six interpreting trainees participated in the pilot study compared to 14 interpreting trainees in the main study. The larger number of participants may have amplified the effect of the predictor variable cognate category. The fact that the main study did not replicate the results of the pilot study, despite a larger number of participants, questions the validity of cognate translations as indicator for work-load.

4.3.6.6 Silence

Research on work-load suggests that the percentage of silences in an uttering increases with work-load (see chapter 3.4.3.2). In order to investigate silences during simultaneous interpreting, I extracted silent pauses from each recording using *praat* (Boersma & Weenik, 2013). As short pauses are natural in speech, only silences longer than 500 milliseconds were considered. Further, all observations five seconds after the beginning and five seconds before the end of the recording were cut off as the interpreter needs to wait for the first few segments before he can start the interpretation of the speech. Silent pauses longer than 5 seconds were manually checked. Two observations were wrongly identified and needed manual correction. Silent pauses shorter than 5 seconds were only randomly checked, but appeared to be correct. One participant was completely excluded from analysis because the recording quality was insufficient for silence extraction. One further recording was missing. After

removal of invalid observations, the mean duration of silent pauses was 1.481 seconds (*MD* = 1.057 seconds, range: 0.500 -11.880 seconds). The mean duration of silent pauses varied between participants with a minimal mean duration of 1.095 seconds and a maximal mean duration of 2.222 seconds. Figure 38 depicts the silence durations for each participant.

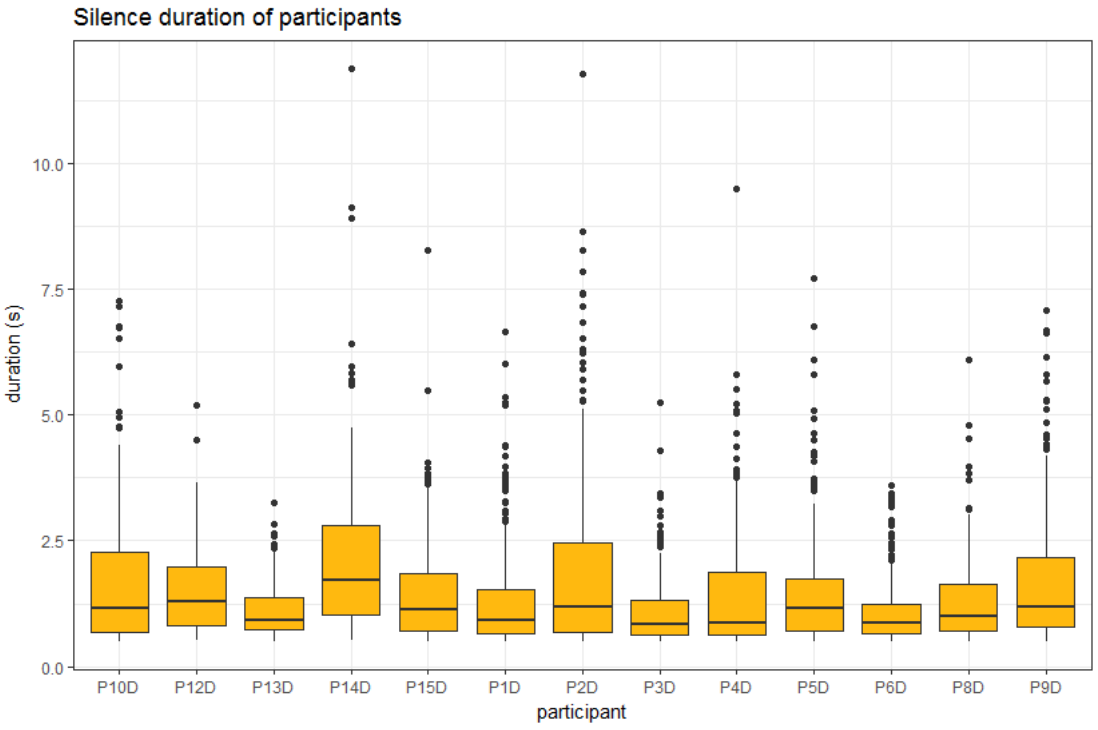

Figure 38: Boxplot of silence durations for each participant.

As can be seen in Figure 39, the mean duration of silences was slightly higher for the experimental condition with noise (*M*=1.638, *MD*=1.141) than without noise (*M*=1.346, *MD*=0.997) and also a little bit higher for the audio condition (*M*=1.577, *MD*=1.135) compared to the video condition (*M*=1.394, *MD*=0.973). However, these differences failed to be significant in a linear mixed model with random intercepts for participant and text ($F(13)=0.8014$, $p>0.1$) or a Levene's test for homogeneity of variance centered at the median ($F(3152)=1.135$, $p=0.2868$).

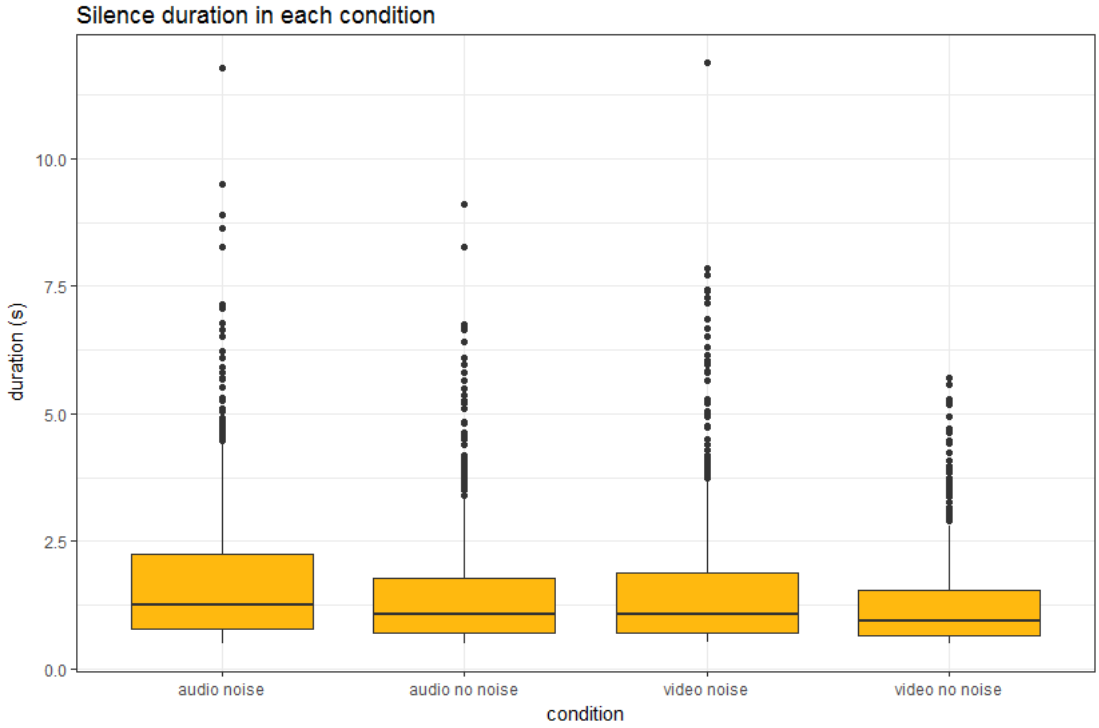

Figure 39: Silence duration in each condition.

However, this first analysis is inconclusive because it does not take into account the number of pauses. Fewer, but longer pauses lead to a similar mean duration as more, but shorter pauses. In order to reveal possible differences in the distribution of silence durations across the four experimental conditions, I grouped silent pauses according to their duration into five categories: very short silences (0.5 to 1 second)[42], short silences (1 to 2 seconds), moderately long silences (2 to 3 seconds), long silences (3 to 5 seconds) and very long silences (more than 5 seconds), as to obtain a categorization which is sufficiently detailed to explore differences in their distribution, but not too fine-grained as to blur potential differences. Though is seems plausible to assume that long pauses of several seconds indicate a disruption of the interpreting process, or more

[42] Silent pauses below 0.5 seconds were not taken into account. A rough analysis of the source speech showed that the shortest silent pause was 0.5 seconds long (*M*=0.845 s, range: 0.505-1.863 s). Furthermore, research by Ahrens (2008) suggests that pauses below 0.4 seconds are used to mark the end of an information unit (Ahrens, 2004).

generally that pauses indicate different underlying (cognitive) processes according to their length, it should be noted that this categorization contains no information about any linguistic function or cognitive process. Such an investigation would at least require knowing the context, e.g. the position of the pause with the target sentence and the source sentence. Nearly half of all pauses were between 0.5 and 1 second long. Pauses longer than 5 seconds accounted for only 2% of all observations. The number of silent pauses for each category is displayed in Table 27.

	Number	**Percentage**
0.5-1 s	1470	47.5%
1 -2 s	970	31.4%
2 – 3 s	347	11.2%
3 – 5 s	236	7.6%
> 5 s	69	2.2%
TOTAL	3092	100%

Table 27: Number and percentage of silent pauses in each category: 0.5 to 1 seconds, 1 to 2 seconds, 2 to 3 seconds, 3 to 5 seconds and longer than 5 seconds.

Table 28 displays the number of pauses in each experimental condition and their percentage relative to the total number of pauses in each condition. It seems that there are huge differences in the distribution of silent pauses between the video condition without noise and the audio condition with noise. For instance, pauses between 0.5 and 1 second accounted for 54.3% of all pauses in the video condition without noise, while they accounted for 39.3% in the audio condition with noise. The differences, however, are nearly negligible between the video condition with noise and the audio condition without noise: In the video condition with noise, silent pauses between 0.5 and 1 second accounted for 47.8% of all pauses, while their proportion was 46.5% in the audio condition without noise.

			0.5-1s	1-2s	2-3s	3-5s	>5s	TOTAL
Auditory presentation	**Noise**	Number	237	197	75	72	22	603
		Percentage	39.3	32.7	12.4	11.9	3.6	100
	No noise	Number	393	277	105	54	17	846
		Percentage	46.5	32.7	12.4	6.4	2.0	100
Audio-visual presentation	**Noise**	Number	382	236	87	69	25	799
		Percentage	47.8	29.5	10.9	8.6	3.1	100
	No noise	Number	458	260	80	41	5	844
		Percentage	54.3	30.8	9.5	4.9	0.6	100

Table 28: Number of silent pauses of different length in each experimental condition. The percentage is calculated based on the total number of silent pauses in each experimental condition.

Using the *ordinal*-package (Christensen, 2015), I constructed an ordinal mixed effects model on the silence category to analyze the distribution of silent pauses of different lengths across the different experimental conditions. As the proportional odds assumption did not hold, I refitted the model with flexible thresholds. Random effects included intercepts and trial slopes for each participant. Fixed effects were added one by one. P-values of the effect were estimated using Likelihood ratio tests of cumulative link models. The results are summarized in Table 29. P-values for each level were estimated using the Wald test. All analyses were carried out in *R* version 3.3.2 (R Core Team, 2016) using the package *lme4* (Bates, Maechler, Bolker, & Walker, 2015). Plotting was done with the R-package *ggplot2* (Wickham, 2009).

Predictor variable	DF	Likelihood ratio	p-value
experimental condition	2	16.003	0.001
auditory presentation	1	5.727	0.017
visual presentation	1	5.654	0.017
noise level	1	4.485	0.034
speech	3	6.194	0.103
group	1	2.376	0.123

Table 29: Silence duration: degrees of freedom, likelihood ratio and p-values resulting for each predictor variable when comparing the model wit and without the effect in question.

The effect visual presentation (reference level: audio, *Estimate*= -0.241*, SE*=0.057*, z*= -4.198, *p*<0.001) was significant suggesting that audio-visual input contributed to decrease the number of longer silent pauses. Furthermore, there was a significant interaction of noise level with the predictor variable auditory presentation (reference level: no noise and noise level at 0.1, *Estimate*=0.757*, SE*=0.181*, z*=4.195, *p*<0.001) suggesting that noise had a stronger effect on silent pauses at noise level 0.1 than at noise level 0.2. The main effect noise level (reference level: noise level at 0.1, *Estimate*= -0.245*, SE*=0.182*, z*= -1.347, *p*=0.178) lost significance after adding the interaction of noise level on auditory presentation suggesting that the variable noise level only made a difference when noise was added at all. The main effect of auditory presentation (reference level: no noise, *Estimate*= -0.030*, SE*=0.075*, z*= -0.406, *p*=0.685) disappeared, too, when adding the interaction with noise level which suggests that the effect of auditory presentation, e.g. if noise is added or not, is absent at noise level 0.1. The interaction of noise level on the predictor variable visual presentation did not reach significance (X^2(21, 3)=0.143, *p*=0.986). Effects of each predictor level are given in Table 30.

Method and results

The effects of speech or group failed to be significant ($p > 0.1$, see Table 29).

	log odd estimate (beta)	Standard error	z-value	p-value (Wald)
noise	-0.030	0.075	-0.406	0.685
video	-0.241	0.057	-1.347	< 0.001
noise level 0.2	-0.245	0.182	-4.198	0.178
noise*noise level 0.2	0.757	0.181	4.195	<0.001

Table 30: Silence duration: model estimates for the effects of experimental condition and noise level

The effects of the predictor variables are more easily visible in Figure 40. The barplot shows the fitted model values as probability of observing a silence of given length in different experimental conditions. Green bars stand for the condition with noise, yellow bars for the condition without noise. Facets at the top show the fitted values in the audio condition, at the bottom are the fitted values in the video condition. Facets on the left side show the fitted values at noise level 0.1, facets on the right side depict the fitted values at noise level 0.2. On the whole, the probability to observe longer pauses decreases more rapidly in the video condition (bottom facets) than in the audio condition (top facets). In the condition with noise level 0.2 (right facets), the “curve” is much flatter for the noise condition (yellow bars) than for the condition without noise (green bars) suggesting that the probability to observe longer pauses decreases more rapidly in the video condition than in the audio condition when noise at noise level 0.1 is added to the source speech. The case is different for noise level 0.1 (left facets): there is virtually no difference between the condition with noise and the condition without noise.

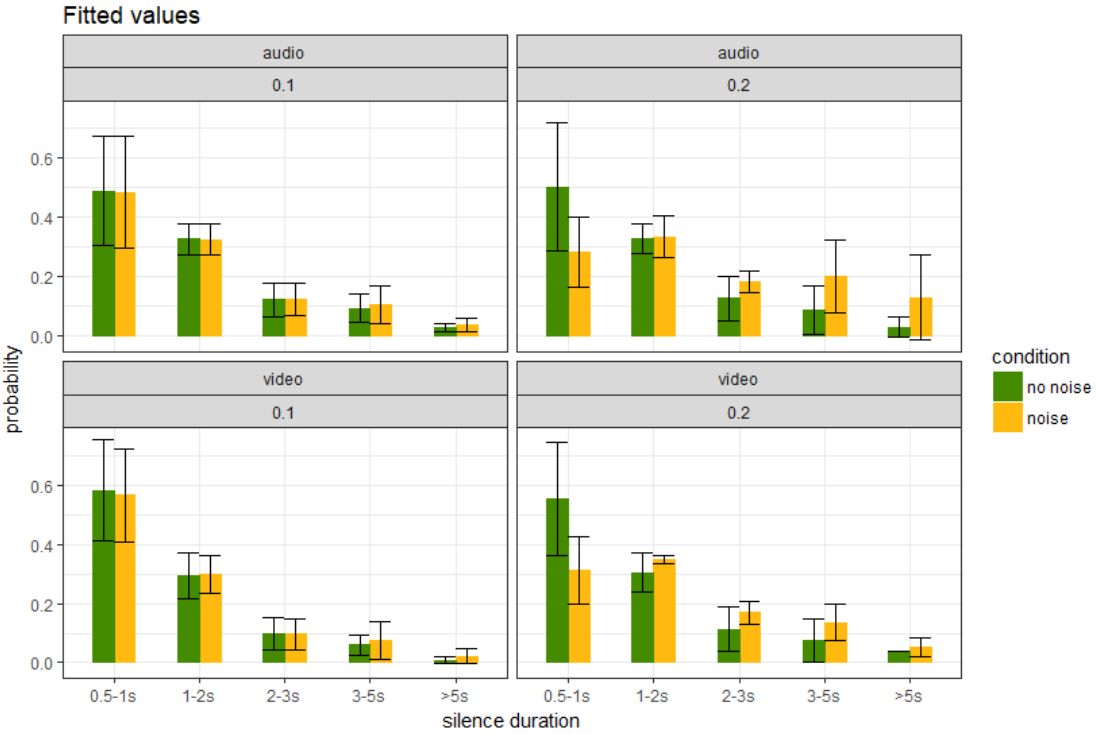

Figure 40: Silence duration: barplots of fitted values with error bars. Green bars stand for the condition with noise, yellow bars for the condition without noise. Facets at the top show the fitted values in the audio condition, at the bottom are the fitted values in the video condition. Facets on the left side show the fitted values at noise level 0.1, facets on the right side depict the fitted values at noise level 0.2. Error bars are estimated based on the fitted values and might be over-conservative.

4.3.6.6.1 Discussion of the effects on silent pauses

The results of the statistical analysis suggests that the classification of silent pauses into five categories (0.5-1 second, 1-2 seconds, 2-3 seconds, 3-5 seconds, more than 5 seconds) was effective. The hypothesis stating that silence durations during simultaneous interpreting with visual input should be shorter than without visual input was partially confirmed. Simultaneous interpreting with audio-visual speech led to significantly more short pauses than simultaneous interpreting without audio-visual speech, where silent pauses were less numerous, but longer. The auditory presentation (noise/no noise) in combination with the noise level also moderated the length of silent pauses: Long silent pauses were more frequent when the signal-to-noise-ratio was low, e.g. noise at level 0.2 was added to the source speech. The effect of noise, however, was

absent when the signal-to-noise-ratio was high, e.g. when noise level was at 0.1 or completely absent.

One possible explanation for the effect of visual presentation could hold that audio-visual speech helps the interpreter to identify more rapidly what the speaker said, especially when parts of the auditory stream are masked by noise and therefore unintelligible, so that the interpreter might be able to move on more quickly with her interpretation. Researchers observed that speech perception benefits from lip movements, especially in adverse listening conditions (cognitive load, background noise, hearing impairment) (Mattys & Wiget, 2011; Iverson, Bernstein, & Auer, 1998; Bernstein, Auer, & Takayanagi, 2004; Brancazio, Best, & Fowler, 2006; von Kriegstein, et al., 2008). To account for this observation, Massaro and Cohen developed a Fuzzy Logical Model of Perception (1999) described in chapter 2.3.5. He assumes that neither auditory speech nor visual speech inputs are unambiguous and that the combination of both signals helps to determine the phoneme by reducing the noise on both signals (Massaro & Cohen, 1999). When no visual input is provided and the auditory signal is too noisy, interpreters need to wait until the context provides sufficient information to disambiguate the auditory input. Hence, they make longer pauses. The fact, that noise only had a significant effect on the duration of silent pauses at noise level 0.2 suggests that the signal-to-noise ratio needs to be sufficiently low to affect the duration of silent pauses. On the basis of the Fuzzy Logical Model of Speech perception (Massaro & Cohen, 1999), I expected that silence durations during speech with high levels of noise would be shorter when audio-visual input was provided. In contrast to this hypothesis, no interaction of auditory and visual presentation was found. Interpreters did not benefit more from lip movements when the auditory stream was more severely degraded. Still, the effect of visual presentation (audio/video) was present in all conditions suggesting that interpreters benefitted from lip movements regardless of the signal-to-noise-ratio.

4.3.6.7 Voice frequency

Fundamental voice frequency (for details, see chapter 3.4.3.2) for each recording (only interpreters) was obtained with *praat* (Boersma & Weenik, 2013). For the present analysis, I took over the standard setting in *praat* (Boersma & Weenik, 2013,), e.g a pitch floor of 75 hz and a pitch ceiling of 600 hz. Within this range, *praat* searches for potential candidates and computes the fundamental voice frequency. As recommended in the praat manual (Boersma & Weenik, 2013), sampling rate was set at 10 ms to provide reliable samples in each analysis window. Due to this small window, a huge amount of the data is missing. On the average, 65% of the recording duration were silent pauses and fundamental voice frequency measures were thus missing (range: 57%-79%). One participant was excluded because the recording quality was not sufficient to sample the fundamental voice frequency reliably. For another participant, only three of four recordings were available. All other recordings were included in the analysis. Fundamental voice frequency for female participants ranged from 75 hz to 600 hz, with a mean at 210.60 hz (*MD*=216.10 hz, male participant: *M*=143.97 hz, *MD*=122.12).

Voice intensity may increase fundamental voice frequency. In order to avoid possible confounds in the statistical analyses later on, I used *praat* (Boersma & Weenik, 2013) to obtain the voice intensity for 100 ms time window. Voice intensity is measured in dB (decibel) which is related to the sound pressure level. As 0 dB is defined as the threshold of hearing and not as one could suppose as the absence of any sound pressure, negative or very low decibels in spoken speech are in theory possible. However, it is improbable that somebody would speak below the threshold of hearing or with a sound level that is often compared to a whisper or rustling leaves (Hormann, 2017; Sengpiel, 2017). For this reason, values lower than 10 dB were discarded (2.6% of all observations). After removal of these values mean voice intensity was at 52.23 dB (*MD*=56.6 dB, range: 10-91.1

dB). Voice intensity and voice pitch were only very weakly correlated ($r(434020) = 0.099$, $t=65.56$, $p<0.001$).

After assigning the voice frequency data to the condition and the speech during which it was obtained, I standardized the raw data using the formula:

$$z=x-\mu/\sigma$$

where x is the observed fundamental voice frequency, μ the mean of all data points during the trial and σ the variance during the trial. As there were – due to the ear-voice-span - very few observations during the first three seconds I removed all observations up to three seconds. In total, 0.5% of all data points were removed. Finally, I aggregated the data in 500 ms time bins by calculating the mean in each time bin for each participant and each trial. The graph below (Figure 41) shows the observed fundamental voice frequency and the trends observed in the noise/no noise condition for each participant (the data is averaged in 5 second time bins for better readability). It seems as if participants reacted differently to the noise condition: for some participants the slope seems steeper in the no noise-condition, for others it seems steeper in the noise-condition, for yet other participants there seems to be no difference.

Method and results

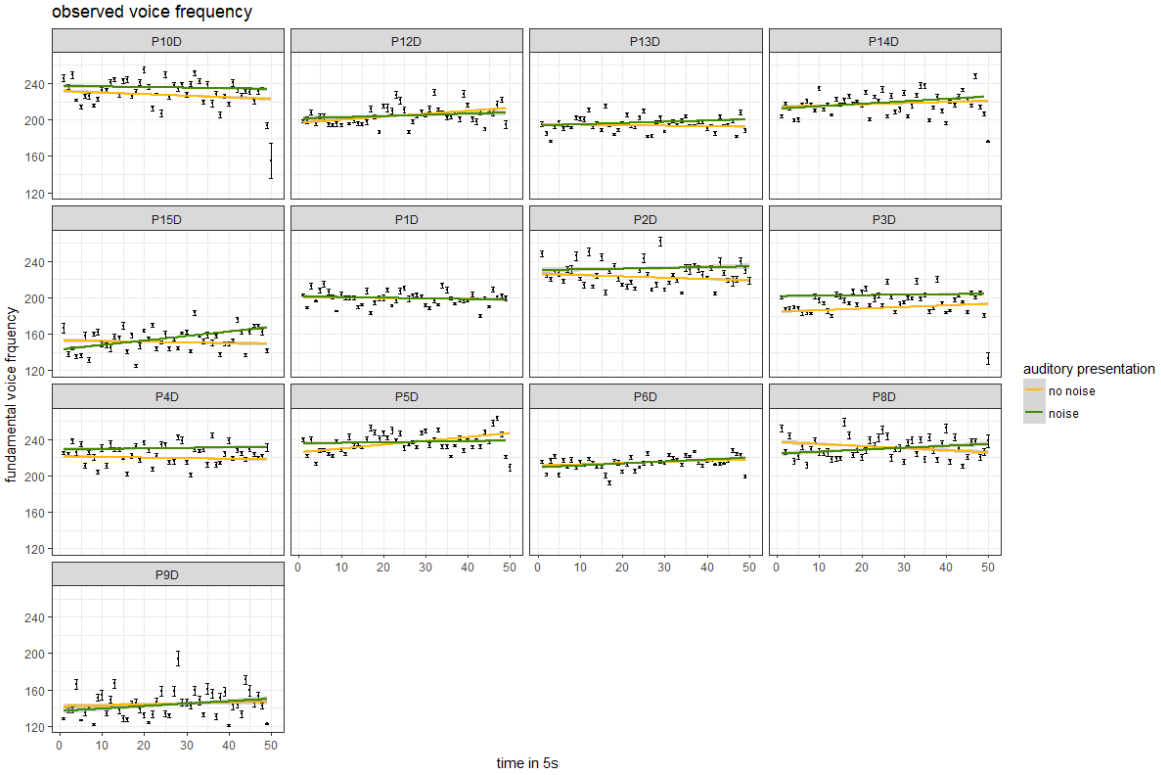

Figure 41: Fundamental voice frequency and trend line for the noise/no noise condition. For better readability, fundamental voice frequency has been averaged within 5 second time bins. The yellow line shows the trend for the no noise condition, the green line the trend for the noise condition. Each facet corresponds to one participant. Trend lines have been calculated with linear regression. Participants reacted differently on noise: for some participants the slope seems steeper in the no noise-condition, for others it seems steeper in the noise-condition.

I used linear growth curve analysis (Mirman, 2014) to analyze the time course of the standardized fundamental voice frequency[43]. The overall time course was captured with a first order polynomial with fixed effects for the auditory presentation (noise/no noise) and the visual presentation (audio/video). Random effects included participants-by-auditory presentation intercepts (*SD*=0.013) to account for the fact that participants could have reacted differently on the noise/no noise condition as is visible

[43] Another option would have been to use a log-normal distribution to fit the (non-transformed) data. But due to the large differences in scale between the dependent variable (75 hz to 600hz) and the time variable (0 to 250), the model failed to converge correctly.

in Figure 41. The model failed to converge with random text-by-participant slopes or random trial-by-participant slopes. P-values were approximated with likelihood ratio by dropping all predictors one by one.

There was no main effect of visual presentation (audio/video) (*Estimate*= -0.003, *SE*=0.003, *p*=0.322) or auditory presentation (noise/no noise) (*Estimate*= -0.0002, *SE*=0.006, *p*=0.972) on the intercept which means that neither noise, nor video led to an increase of voice frequency. However, there was a significant effect of auditory (*Estimate*=21.032, *SE*=1.995, *p<0.*001) and visual presentation (*Estimate*=14.520, *SE*=1.992, *p<0.*001) on the time term. This means that voice frequency increased faster when noise was added to the speech or when a video instead of a freeze frame was displayed. All results are summarized in Table 31. The model fit versus the observed data is depicted in Figure 42. The model failed to converge when adding a time-by-speech interaction. In order to check the effect of speech indirectly, I tested the effect of the predictor variable group (reversed order for the speeches in the video condition) on both experimental conditions. Neither the interaction of group by visual presentation interaction (X^2(14,6)=6.605, *p*=0.359) nor the interaction of group by auditory presentation (X^2(14,6)=5.338, *p*=0.501) reached significance.

Method and results

	Estimate[44]	SE	t-value	p-value
Intercept (audio – no noise)	0.005	0.004	1.043	
Time	4.458	1.705	2.615	
noise	-0.0002	0.006	-0.035	0.972
video	-0.003	0.003	-0.990	0.322
Time:noise	21.032	1.995	10.540	<0.0001
Time:video	14.520	1.992	7.288	<0.0001

Table 31: Fundamental voice frequency: estimates, standard error, t-values and p-values for the predictor variables.

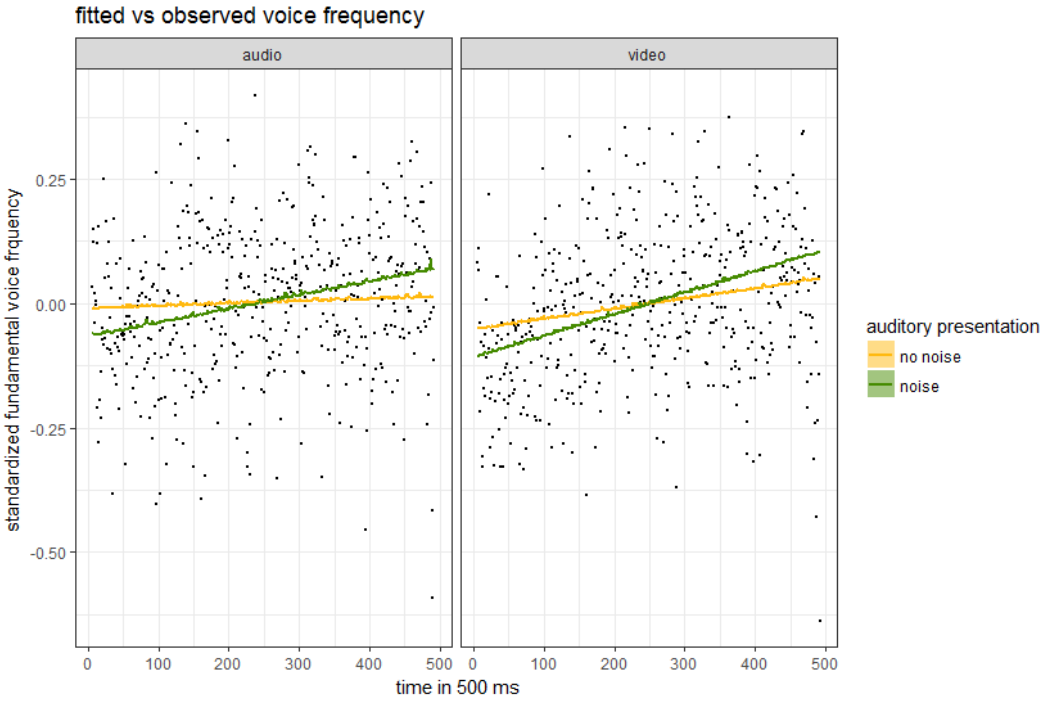

Figure 42: Fundamental voice frequency: Observed (black dots) versus fitted values (yellow and green line). The line corresponds to the condition without noise, the green line to the condition with noise added to the speech. The left facet shows the fitted values for the audio condition; the right facet shows the fitted values for the video condition (video of the speaker).

[44] It is not possible to transform standardized values back into real values. For this reason, no difference in hertz is given.

4.3.6.7.1 Discussion of the effects on voice frequency

The results described above suggest that audio-visual speech and background noise have an effect on the time course of fundamental voice frequency. According to the model, fundamental voice frequency rises faster during translation when visual input is provided or when noise is added to the source text. This finding confirms and extends the results of the pilot study where an increase of voice frequency was found in the noise condition. No effect, however, was found for the visual presentation in the pilot study. In contrast to the pilot study, the main effect of auditory or visual presentation was not significant indicating that voice frequency is not globally higher in the condition with noise or with audio-visual speech. This is probably due to differences in the statistical analysis: linear mixed models were used in the pilot study, whereas growth curve analysis, taking the time course into account, was used in the main study because the visual inspection suggested an increase of voice frequency over time. As suggested by Scherer (1989) and Warren (1999) (see chapter 3.4.3.2), the sympathetic stress response causes the vocal cords to tighten which in turn increases the fundamental voice frequency. Most studies have involved emotionally stressful stimuli (see chapter 3.4.3.2) and increases of fundamental voice frequency are associated with self-reported anxiety and the perception of nervousness (Laukka, et al., 2008)(see chapter 3.4.3.2). From this point of view, the steeper increase of fundamental voice frequency in the noise condition could suggest that participants felt more stressed and less able to cope with task. This feeling might have intensified throughout the speech as interpreters missed more and more segments and were less and less sure what the speaker was talking about. Alternatively, the effect on the time course could reflect tiredness of the vocal cords. The vocal cords might have tightend over time as a result of talking during several minutes. This hypothesis, however, is ruled out by the fact that the effect on time course is nearly absent in the audio condition without noise (see Figure 42).

Interestingly, the model revealed a faster increase of voice frequency in the condition with audio-visual speech compared to the auditory-only condition. This is in contrast to the hypothesis formulated in chapter 4.1 which predicts lower fundamental voice frequency during simultaneous interpreting with audio-visual speech compared to simultaneous interpreting without audio-only speech. Is audio-visual speech after all a stressor for interpreters, be it because it generates additional work-load, or for some other reason? This explanation seems not very probable. First, a large body of research suggests that audio-visual speech measurably facilitates speech comprehension: the detection threshold in noise is lower, reactions times to audio-visual stimuli are shorter compared to visual or auditory-only stimuli, recall is better for multimodal stimuli than for unimodal stimuli and the physiological response measured by EEG of fMRI is stronger (for a review of different studies see chapter 2.3.1). Second, interpreters insist on the need to have visual contact with the speaker. The German federation of conference interpreters VKD recommends placing the booth in a way interpreters can see the speaker and the conference room (Verband der Konferenzdolmetscher, 2017). The visual contact with the speaker is also regulated in the ISO-norm 4043 (International Organization for Standardization, 1998). It is not clear why interpreters would advocate something they feel stressed by.

When taking a closer look at the observed data, we can see different trend lines for the speeches (see Figure 43). It appears that participants who had a steeper increase of their voice frequency in the video condition (see Figure 44) were all assigned to group 2 and translated the speeches about the Greek economic crises and the demographic change (the green and red lines in the graph), whereas those in group 1 who translated the speech about air traffic and work in the video condition had a steeper increase for the audio condition. The interaction of group, where the speeches were switched in the video condition and the video condition was not significant. Still, given the pattern described above and the fact that the model failed to converge with the predictor variable speech it

cannot completely be excluded that the predictor variable speech accounted for the steeper increase in the video condition.

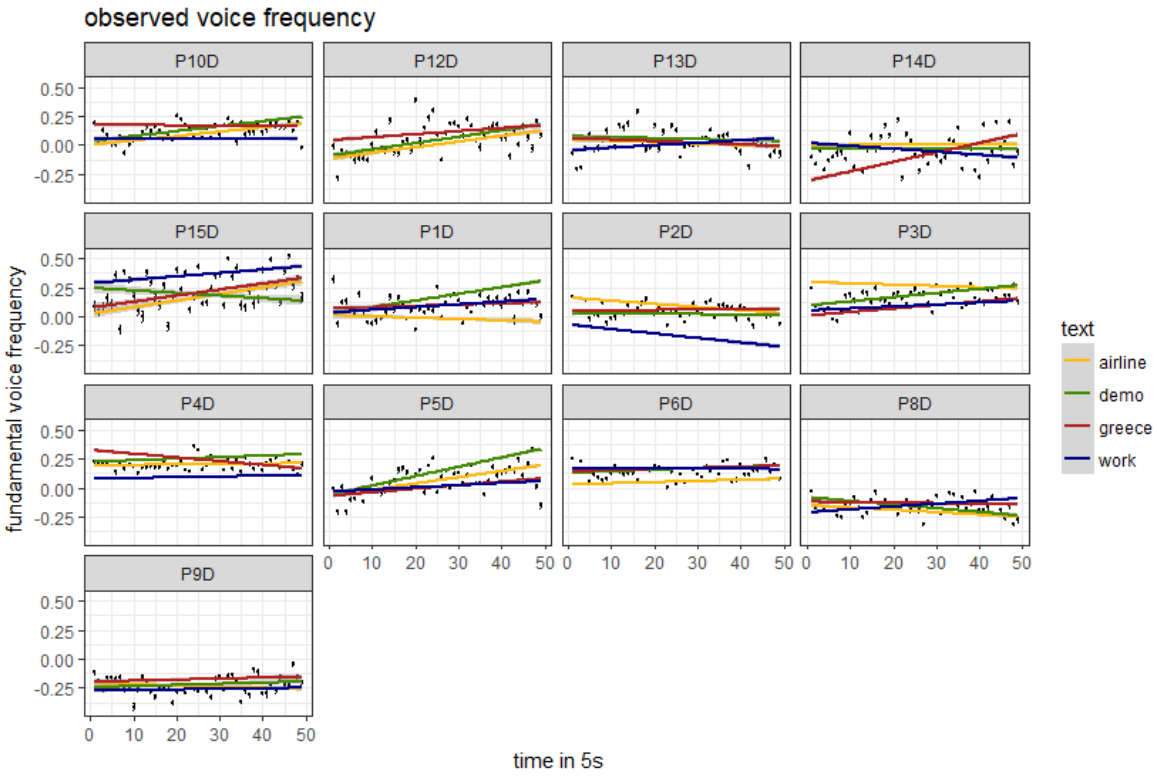

Figure 43: Fundamental voice frequency and trend line for each speech. For better readability, fundamental voice frequency has been averaged within5 second time bins. The yellow line shows the trend for the speech about air travel, the green line the trend speech about demographic change, the red line the trend for the speech about the Greek economic crisis and the blue one the trend for the speech about stress at work. Each facet corresponds to one participant. Trend lines have been calculated with linear regression.

Method and results

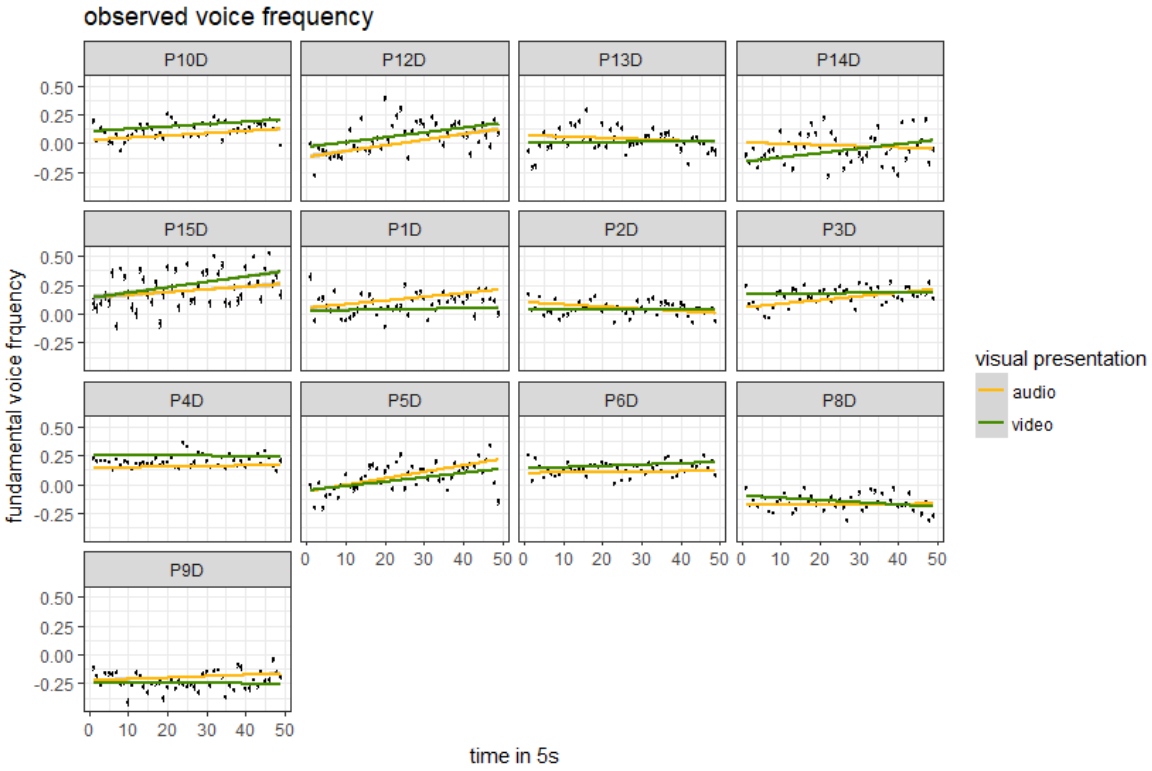

Figure 44: Fundamental voice frequency and trend line for each speech. For better readability, fundamental voice frequency has been averaged within5 second time bins. The yellow line shows the trend for the audio condition, the green line the trend for the video condition. Each facet corresponds to one participant. Trend lines have been calculated with linear regression.

A very interesting observation is that apparently, participants reacted very differently to the noise condition (see Figure 41). This is in line with previous research on the effect of stress on fundamental voice frequency (Streeter, Macdonald, Apple, Krauss, & Galotti, 1983; Scherer, 1989). Given the emotionally arousing stimuli used in these studies (for example stressful events in control centers in (Streeter, Macdonald, Apple, Krauss, & Galotti, 1983) and the positive correlation of fundamental voice frequency and self-reported anxiety (Laukka, et al., 2008), this finding might hint to individual differences in coping with stress and anxiety. One concept that might shed some light on these heterogeneous behaviors is the concept of repressors and sensitizers (Bruner & Postman, 1947; Byrne, 1964). According to this concept, two basic behavioral patterns are at the extreme ends of a continuum. The one extreme pole is repression, the other is sensitization. In a stressful situation a typical repressor will

report rather low states of anxiety while physiological measures, like heartbeat, respiration rate or pupil dilation will be rather high. A sensitizer, in contrast, will report higher levels of anxiety, whereas physiological indicators are rather low. Adopting this concept to the results of the present study, this would mean that repressors' voice frequency rise faster when noise is added to the source speech while sensitizers' voice frequency might keep at more or less the same level. A state-trait anxiety test (STAI) or some other method to collect self-reported anxiety prior to simultaneous interpreting with and without background noise might provide useful insights for further research.

Finally, it is important to bear in mind that speech samples over several minutes – as it was the case for this study – are characterized by prosody. This is certainly due to the fact that interpreters modulated their voices to mark the importance of particular segments or the end of a sentence which makes the target speech more enjoyable for the listener. It is remarkable that despite the extremely noisy data (residual variance of the model: 0.998) it was possible to find an effect of audio-visual speech and noise on the time course of fundamental frequency. This suggests that the effect is persistent. However, the effect is – due to the noise – very weak and it is difficult to say if a listener would perceive such a minimal effect at all.

4.3.6.8 Pupil dilation

As for the pretest, I prepared the data by first adding the experimental variables to the eye tracking data. I made use of the key presses in order to identify movie epochs: If, starting from a space bar, further key presses were absent for at least 240 seconds, the epoch was marked as "movie". The corresponding baseline epoch started from the preceding key press and lasted at least twenty seconds (black screen during ten seconds, picture of the speaker during ten seconds to avoid adaptation of the pupil during the movie). Plotting the data resulted indeed in eight neatly distinguished epochs, four baseline and four movie epochs, with

approximately the same number of observations (movie epoch: 240 seconds = 28805 observations, baseline: 10 seconds = 1200 observations). These epochs were subsequently matched with the condition (video-no noise, video-noise, audio-no noise, audio-noise), the speech (air travel, demographic change, Greece, work) and the trial (1 to 4). The variable condition was split up in two variables: visual presentation (video/audio) and auditory presentation (noise/no noise). Furthermore, I added the group (group 1 and group 2 with speeches presented in reversed condition as described in chapter 4.3.4) and the task to which each participant was assigned (interpreting, listening). As the matching process completely relied on key presses, no timestamp correction was necessary. Sizes of both pupils were highly correlated ($r(4148300)=92.02$, p <0.01). As there were slightly more missing values in the left pupil, all following transformations were done with the right pupil only. I removed blink artefacts and invalid data points following the same procedure as in the pretest: blink artefacts were defined as sudden drops in pupil size (a difference from the preceding data point that lies more than 1.5 times beyond the lower or the upper interquartile range of the mean of all differences); invalid data points were identified based on the validity codes provided by the eye-tracker. Gaps up to 500 milliseconds were replaced by linear interpolation (the approximate duration of a blink). 2% of the data was considered a blink artefact, an invalid observation or fell beyond the region of interest and was replaced by linear interpolation. The mean observed pupil size was 3.3 mm (*MD* = 3.4). Before interpolation, values ranged between 0.8 mm and 5 mm. A pupil diameter of 0.8 mm is very small and most probably due blinks artefacts. After interpolation, the mean pupil size was still 3.3 mm (*MD* = 3.4 mm), but values ranged between 1.5 mm and 4.9 suggesting that the interpolation procedure was successful.

Three participants, P9D, P11D and P14D, all interpreting trainees, were completely removed due to the high amount of missing data (52%, 82%, 84%). After removal of those participants, the percentage of missing data varied between 1.8% and 32.2% (*M*= 9.6%). During the baseline epochs,

Method and results

71% of all fixations were located on the fixation cross (range: 17-99%). The percentage was lower during the speeches: On average, 54% of all fixations were located in the region of interest defined around the speaker's lips (range: 12-92%, visual angle: 13°). For comparison: 14% of all fixations were located on the speaker's eyes. This shows that participants mostly followed the instructions to pay attention on the speaker's lips. The proportion of fixations was higher for listeners (59.6%) than interpreters (49.1%). There were also slight differences between the experimental conditions. The percentage of fixations was highest in the audio condition with noise (59.8%), followed by the video condition without noise (57.2%), the audio condition without noise (53.7%) and finally the video condition with noise (45.6%). Pupil sizes correlated only very weakly 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 correct pupil sizes for the gaze coordinates. All observations, even those falling outside the region of interest were thus included in the analysis as removing them would have led to high data loss – despite linear interpolation.

Pupil sizes are not normally distributed as pupil size is physiologically limited. In order to obtain a normal distribution, I standardized the interpolated pupil sizes separately for each participant and each trial using the formula:

$$z=x-\mu/\sigma$$

where x is the observed data point, μ the mean of all data points during the pretrial black screen baseline epoch and σ the variance during the pretrial baseline epoch. The black screen epoch was chosen in order to maximize as much as possible the differences of pupil sizes during the different conditions. The baseline epoch with the speaker's image helped to avoid luminance adaptation during the speech and let the participant time to familiarize herself with the speaker. Standardization based on the

pretrial baseline carries the risk that the pupil size during the trial is indirectly affected by the preceding trial. I therefore performed for each participant a second standardization based on the grand mean of all pretrial baseline phases of each speech and tested the values obtained by both methods for correlation. Both methods proved to yield nearly the same results ($r(4291600)=1$, $p<0.001$). In fact, only 30990 out of 5 410 992 values were different. A density plot did not reveal any obvious deviations for any of the two standardization methods from a normal distribution. Finally, I regrouped all observations in time bins of 500 ms and 1000 ms in order to remove some of the jitter. Time bins of 100 milliseconds were not practicable for statistical analysis for all attempts to model the data with 100 ms time bins failed.

Plotting the standardized data (see Figure 45), it seems as if pupil sizes differed between interpreters and listeners and between the video and the audio-condition over the time course. While pupil sizes in both tasks and in both conditions seem not to be much different during the first minute or so of the speech, the gaps between both groups and both condition become increasingly larger during the last three minutes of the speech. Differences between the pupil sizes during the speech with or without masking noise were not visible in a plot.

Method and results

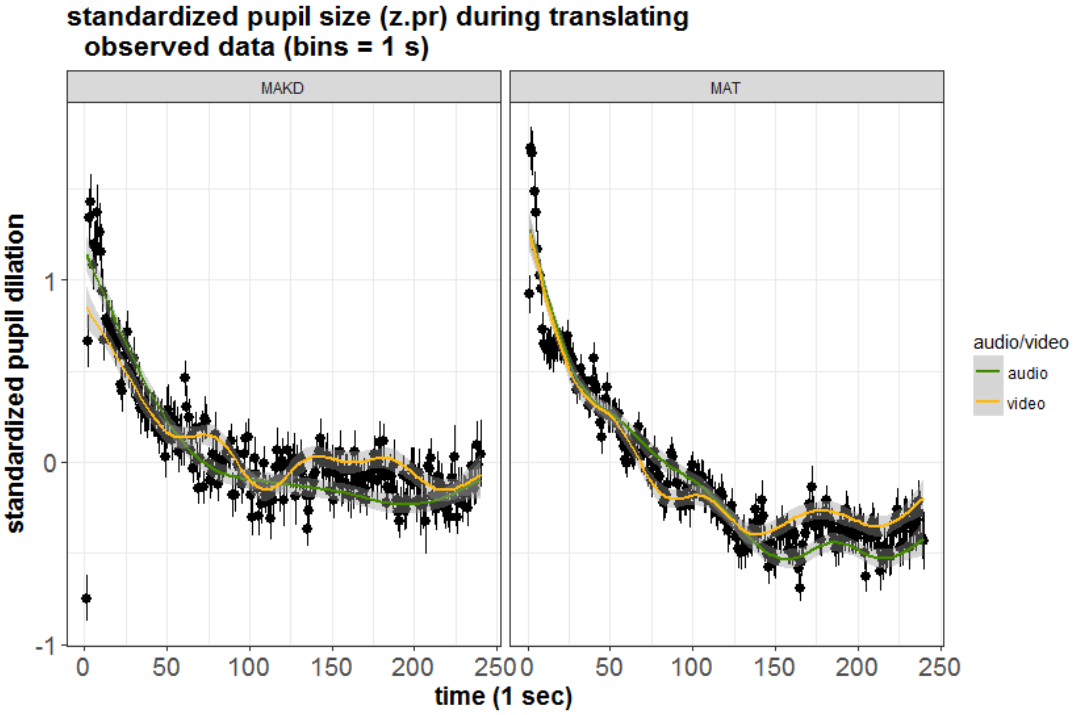

Figure 45: Pupil sizes during speeches. Interpreters (MAKD) translated the speeches, while listeners (MAT) merely listened to them. The green trendline corresponds to the audio condition, the yellow one to the video condition. The black dots are mean and standard deviation of the observed data regrouped per 1000 milliseconds.

I used growth curve analysis (Mirman, 2014) on the standardized pupil sizes to analyze the time-course of pupil dilation during the speeches. For the time course, time bins of 500 ms were chosen for the model failed to converge with smaller time bins. The overall time course was captured with a second order orthogonal polynomial. 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, auditory presentation, visual presentation, text difficulty ratings, speech rate ratings, speech and noise level. Fixed effects were added one by one and p-values were approached using maximum likelihood. All analyses were carried out in *R* version 3.2.2 (R Core Team, 2016) using the R-package *lme4* (Bates, Maechler, Bolker, & Walker, 2015). All plots were done with the R-package *ggplot2* (Wickham, 2009)

Method and results

The model revealed a significant effect for the linear (*Estimate*= -5.007, *SE*=0.860, *t*= -5.821, *p* <0.01) and the quadratic time term (*Estimate*=2.931, *SE*=0.566, *t*=5.181, *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, a significant effect of task on the intercept (*Estimate*= -0.148, *SE*=0.017, *t*= -8.800, *p* <0.01), the linear (*Estimate* = -2.628, *SE*= 1.109, *t*= -2.371, *p*<0.01) and the quadratic term (*Estimate*=0.808, *SE* = 0.729, *t*=1.107, *p* <0.01) was found, 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 audio condition than during the video condition (*Estimate*=0.032, *SE*=0.015, *t*=2.035, *p<0.*05). The effect of visual presentation on the linear time term was marginally significant (*F*(1,15)=3.342, *p*=0.067). Auditory presentation, text difficulty ratings, speech rate ratings, noise level, speech or trial did not improve the model. The results of the model comparisons for each of the aforementioned predictor variable are summarized in Table 32.

	X²	**DF**	**p-value**
Auditory presentation	0.007	15,1	0.913
Text difficulty ratings	0.415	16,2	0.813
Speech rate ratings	1.406	16,2	0.495
Noise level	0.010	15,1	0.919
Speech	1.666	17,3	0.645
Trial	0.198	17,3	0.978

Table 32: Pupil dilation: chi-squared, degrees of freedom and p-value when comparing the model with the predictor in question against the model without the predictor.

In Figure 46 the model fit (colored ribbons) is plotted against the observed data (black point-ranges). The left facet shows the model fit for the

interpreters (“MAKD”), the right facet the model fit for the listeners (“MAT”). The color codes the visual presentation (green: auditory-only speech, yellow: audio-visual speech).

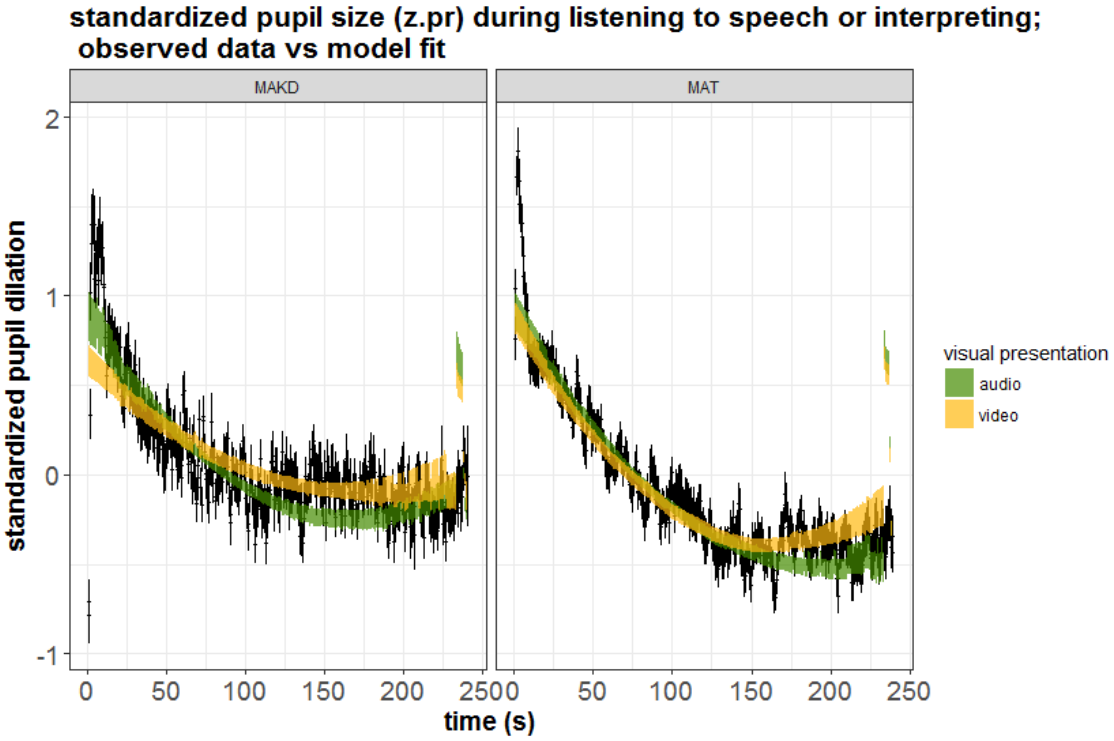

Figure 46: Model fit for a linear mixed model with visual presentation and task as fixed effects. The color coding corresponds to the visual presentation (green: auditory-only, yellow: audio-visual). Fitted values for interpreters (“MAKD”) are plotted in the left facet; fitted values for listeners (“MAT”) are plotted in the right facet. According to the model, pupil dilation declined faster during listening than during simultaneous interpreting and faster during the audio than during the video condition.

4.3.6.8.1 Discussion of pupillary responses

As expected, pupil dilation was larger during simultaneous interpreting than during listening and decreased more slowly during simultaneous interpreting than during listening. According to Kahnemann (1973; see also chapter 3.2.1), pupil dilation indicates a state of physiological arousal of a person. In the context of task solving it can be interpreted as mental effort or the effort participants exert to solve the task rather than task-induced cognitive load. The higher the state of arousal, the more effort a person is willing to put into a task. Even though participants in general will adapt their effort to the task demands, this is not always necessarily the

case: somebody who is bored, tired or unmotivated might do less effort to solve the task. It seems reasonable to assume that simultaneous interpreting triggers a higher work-load than mere listening for the simple reason that simultaneous interpreting involves more concurrent processes than listening (for a rough breakdown of sub-processes in simultaneous interpreting see chapter 1.3). Put in this light, this result would suggest that participants adapted their effort to the task demands increasing their efforts during simultaneous interpreting and reducing their effort during listening. However, another interpretation of this result could be that simultaneous interpreting elicited higher arousal because participants needed to respond continuously to the stimulus by translating the speech whereas listeners responded to the stimulus only after they heard the speech by rating general parameters and answering to text-related questions.

The hypothesis stating that pupil sizes will be smaller when audio-visual speech is provided compared to auditory-only speech could not be confirmed. In both tasks, simultaneous interpreting and listening, pupil dilation towards the end of the speech was larger in the video condition compared to the audio condition. One limitation to this finding is that all valid data points were used for the analysis. Valid observations outside the region of interest were not discarded in order to avoid too much data loss. About 54% of all fixations were within the region of interest, while 46% were outside the region of interest (though still on the screen). This may raise doubts with regard to the validity of the result. Data points beyond the region of interest may not have been affected by the facilitating effect of audio-visual speech. This result, however, replicates findings of the pilot study which took only observations within the region of interest into account. In the pilot study, audio-visual speech elicited larger pupil dilations than auditory-only speech. Even though contradicting the hypothesis, both, the pilot study and the main study, suggest that pupil sizes during simultaneous interpreting are larger when lip movements are visible than when this is not the case. It is not clear at this stage what this

finding exactly means. Several explanations are conceivable. First, audio-visual speech might cause higher work-load and participants adapt to it by increasing their mental effort. As mentioned in chapter 4.2.6 this explanation seems improbable given the vast body of research demonstrating the beneficial effect of audio-visual speech on speech comprehension (see chapter 2.3.5) and the fact that conference interpreters insist on the visual contact with the speaker (see chapter 1.2). Second, seeing a moving face may simply elicit higher arousal than a static face maybe because it triggers a stronger emotional response, without directly affecting the performance or relating to work-load. For instance, Courtney and colleagues (2010) found higher self-reported and physiological arousal (skin conductance, heart rate and eye blinks) in response to moving pictures of feared stimuli than to static pictures of the same stimuli (Courtney, Dawson, Schell, Iyer, & Parsons, 2010). This view would be consistent with the observation that both, listeners and interpreters, showed this effect[45]. Another indirect argument for arousal-elicited pupil dilation that is not related to task demands comes from the present study itself. The overall time course of the pupillary response during one speech shows a rapid decline at the beginning of the speech that flattens towards the end of the speech. This suggests that arousal is particular high at the beginning of the speech[46]. However, the introductory part of the speeches was most certainly not the most difficult part as it contained the usual introductory statements so that this pattern cannot be

[45] The effect appeared later for listeners than for interpreters. The reason why is not very clear. It may be that the arousal related to the experimental setting and to the task is higher than the arousal linked to the moving picture and covers the effect of the visual presentation until the arousal related to the experimental setting or the task has sufficiently decreased.

[46] Similar results though with different statistical methods have been obtained by Hyönä, Tommola and Alaja (1995): pupil dilation during simultaneous interpreting decreased over the time course of the speech (Hyönä, Tommola, & Alaja, 1995).

ascribed to cognitive load. It seems much more plausible to assume that this response reflects a state of general arousal. Participants did not know what to expect and might have been unsure about whether they could cope with the task. Under these circumstances, it seems only natural that participants were particularly alert at the beginning of the speech which caused their pupils to dilate. An elevated state of arousal in response to moving faces could also explain larger pupil dilations in participants when audio-visual speech was provided. Even though this explanation may seem a little bit far-fetched it cannot completely be excluded. Third, audio-visual speech does not increase work-load, but elicits higher arousal which in turn frees resources for other processes like memorization of the speech content or monitoring of the target speech (for interpreters). In this case, larger pupil dilations should correlate with improved performance, like a lower number of cognate translations, higher translation accuracy or a higher number of correctly answered text-related questions. The general discussion in chapter 4.4 will shed more like on this issue by linking pupillary response to performance-based measures like cognate translations, translation accuracy or text-related questions.

The fact that neither the predictor variable speech nor trial improve the model confirms once again that there was no significant difference between the speeches or trials. Participants did not seem to increase their effort in order to deal with one specific speech, nor did they suffer from fatigue at the end of the experiment. Indeed, the speeches were not very long and did by far not exceed the speech duration the participating interpreters were used to.

A rather surprising finding is that auditory presentation (noise/no noise) or the noise level had no effect on the pupil dilation although both predictors considerably affected pupil dilation in the pretest. This suggests that the effect of noise is task-specific. Noise has a considerable impact on sound discrimination because white noise interferes with the underlying auditory stream (see chapter 3.3.4). If the task is to identify single words, sound

discrimination plays an important role and noise substantially increases the task demands. In simultaneous interpreting or listening, however, the situation is different. Interpreters do not need to understand each word as many words can be anticipated with the aid of syntactical or semantic constraints. About 40% of the words that appeared in the four speeches used in the experiment had a purely grammatical function (like articles or prepositions) and could easily be predicted. Moreover, the noise level was fairly low (0.1 or 0.2) for most participants (only one participant reached noise level 0.3). This might not have been sufficient to compromise the listening task or the interpreting task in a way that participants needed to increase notably their effort. This could explain why noise and noise level did not affect pupil dilation during listening or simultaneous interpreting while it did in the pretest. It might be interesting to test whether it is possible to find a main effect for noise when the signal-to-noise ratio is higher, for example when the participants' individual threshold is lowered to 50% or even 30% of correctly identified stimuli in a word detection test that is administered prior to the translation task.

Maybe one might expect text difficulty ratings or speech rate ratings to affect pupil dilation because the ratings could be expected to reflect how much effort participants put into the task. This was, however, not the case. One reason is probably that those ratings were not sufficiently fine-grained (the scale had only four levels ranging from "very easy/slow" to "difficult/fast"). Most ratings ranged from "very easy/slow" to "easy/slow". The ratings might not have been sufficiently contrasted to explain significantly more residual variance in the linear mixed model. Furthermore, as mentioned in chapter 4.3.6.1.5, text difficulty ratings or speech rate ratings may not be the right questions to obtain an evaluation of work-load by the participants. Nevertheless, it is interesting to note that listeners who on the whole gave lower text difficulty ratings also had smaller pupil dilations than interpreters. Using more comprehensive work-load inventories with finer scales like the NASA-TLX and correlating those

ratings with pupil dilation might provide further insight on how pupil dilation can be interpreted with respect to self-experienced work-load.

4.4 General discussion and limitations

Conference interpreters face various sources of visual input during their work: the speaker's facial movements or gestures, reactions from the audience, presentations, charts, notes, etc. (see chapter 2.4 for a more complete description). Far from perceiving this additional visual information as increasing their work-load, interpreters insist on having access to visual information and in particular to the speaker (International Organization for Standardization, 1998). The question that arises is the following: What is the role of visual input in simultaneous interpreting? Beyond providing clues to the meaning of what the speaker is saying, does visual input reduce work-load in simultaneous interpreting? Previous research on visual input in simultaneous interpreting had failed to observe any facilitating effects of visual input on simultaneous interpreting possibly due to methodological issues (see chapter 2.4.2). The aim of this study was to investigate the impact of visual input and in particular visible lip movements on simultaneous interpreting using a more systematic and controlled approach. Based on previous research (see chapter 2.3.5 and chapter 3.3.4), I hypothesized that audio-visual speech should have a facilitating effect on speech comprehension and therefore should lower work-load in simultaneous interpreting. As I could not be sure whether audio-visual speech really has an impact on work-load in simultaneous interpreting or not, I introduced a second variable known to affect work-load in simultaneous interpreting, namely white noise that is added to the source speech and makes it partially unintelligible (for more details see chapter 4.1). The experiment contrasted thus simultaneous interpreting from English to German in four conditions: audio-visual speech without background noise, audio-visual speech with background noise, auditory-only speech without background noise and auditory-only speech with background noise.

In total, three experiments were conducted: a pilot study, a pretest and a main study. The aim of the pilot study was to test the experimental design with a limited number of participants (*N*=6) and to discover potential flaws in the experimental design (see the discussion of the pilot study in chapter 4.2.6). The main study (*N*=31, 17 listeners and 14 interpreters) was preceded by the pretest, a word recognition task, in order to determine the participants' individual signal-to-noise ratio where participants correctly identified 75% of the stimuli. Pupillary responses, response accuracy and response time during the pretest were statistically investigated. The signal-to-noise ratio that resulted from the pretest was then applied to the speeches during the main experiment. The main experiment followed roughly the same procedure as the pilot study: Participants orally translated four speeches. After each speech, participants estimated speech duration, rated video and sound quality, text difficulty and speech rate and finally answered five text-related questions. A control group of listeners matched for their level of English was included in order to control for possible task effects. The effects of audio-visual speech and white noise on pupil sizes, duration estimations, general ratings, text-related questions and (for interpreters only) fundamental voice frequency, silent pauses, translation accuracy and cognate translations were analyzed.

4.4.1 Effects of audio-visual speech

In line with the hypothesis, the main study revealed that audio-visual speech leads to fewer long silent pauses, but to more short silent pauses. Furthermore, audio-visual speech had an effect on sound-quality ratings: participants gave better ratings when the speech was presented in the audio-visual mode compared to the auditory-only mode. This suggests that audio-visual speech indeed speeds up or facilitates listening comprehension. Contrary to the hypothesis though, audio-visual speech led to larger pupil dilation during simultaneous interpreting compared to auditory-only speech. This effect was consistently found in the pilot and the main study, as well for interpreting as for listening. One first

explanation to reconcile these contradictory findings holds that the facilitating effect of audio-visual speech frees up resources that are reinvested in other sub-processes like deeper level processing. In this case, the larger pupil dilations in the video condition would reflect the additional effort caused by deeper level processing. Yet, this explanation does not seem to hold true. Despite eliciting larger pupil dilations audio-visual speech had no further effect on performance in the main study, neither on text-related questions, nor on translation accuracy or cognate translations[47].

Maybe it is necessary at this point to define speech comprehension more precisely. Studies claiming a beneficial effect of lip movements on speech comprehension used phoneme recognition tests and syllables as stimuli (see chapter 2.3.5). The task was thus actually not to comprehend a speech but to discriminate or identify different sounds. Yet, when listening to a whole speech, single phonemes and syllables are less important because the semantic, syntactical and phonological context allows the listener to make predictions about what the speaker might have intended to say. Comprehension goes thus beyond word or phoneme identification. In 1999, Setton published a model for simultaneous interpreting (Setton, 1999) which includes the idea of syntactical, semantic and even pragmatic predictions. For instance, determiners are commonly followed by nouns or by describing words like adverbs or adjectives. In many languages, including Germanic, Slavic or Finno-Ugric languages, linguistic cases enable predictions about how a noun relates to another. Similarly, specific formulas or phrases may be retrieved as a whole, once the interpreter heard the first word (“Ladies and …”). Finally, the speech context and the whole setting or the speaker’s attitude as they are perceived by the

[47] An effect of audio-visual speech on cognate translation was observed in the pilot study, however, this finding was not replicated in the main study despite the larger number of participants and might therefore be invalid (see chapter 4.3.6.5.1 for more details).

interpreter may constrain the number of lexical candidates that fit in the sentence. When the speaker talks about fishing and rivers, the word *bank* is certainly understood differently than when he talks about finances. The importance of predictability is also stressed by Chernov: he found the translation accuracy to be much higher for highly predictable utterances than for unpredictable utterances (Chernov, 2002). In a very elegant study, Ostrand, Blumenstein and Morgan demonstrated that visual speech information, like lip movements, is ignored if the stimulus word can be clearly identified. However, if the stimulus is a non-word, visual speech information is used to disambiguate the stimulus (Ostrand, Blumenstein, & Morgan, 2011). A similar explanation could hold at utterance level: if the utterance provides sufficient context information to substitute unintelligible segments or words, listeners or interpreters might first rely on the speech context. It is only when the context is not constraining enough, that listeners or interpreters might resort to visual speech information or to some other complementary input (documents, notes, colleague/ neighbor) that may fill the gap for audio-visual speech is not the only complementary source of speech information in a conference setting. Even though participants looked at the speaker's lips most of the time during the speeches, the speaker's lip movements may not have provided sufficient information for deeper level processing or – in a more general sense - to improve translation accuracy or recall of the speeches' content.

The considerations above might help to understand why performance did not improve with audio-visual speech input. Still, they do not explain why audio-visual speech was associated with shorter silent pauses and better ratings of sound quality. One account is that lip movements aid the prediction of the auditory signal. Vroomen and Stekelenburg (2010) conducted a compelling EEG-study demonstrating that multimodal integration occurs only when the auditory stimulus is reliably predicted by the visual stimulus. In the predictable condition, they used two disks as visual stimulus that approached a rectangle and "squeezed" it. On the moment the disks touched the rectangle, a sound was played. In the

unpredictable condition, the rectangle still changed its form when the sound was played, but the disks were absent and the sound could thus not be predicted. The authors observed an attenuation of the auditory N1 (a negative task-evoked potential peaking around 100 ms after stimulus onset) in the predictable condition. According to the authors, the preceding visual signal reduced the temporal uncertainty of the auditory signal (Vroomen & Stekelenburg, 2010). In other words: The leading visual stimulus tells the listener when to expect a sound. Similar results were obtained 2013 by Los and Van der Burg (2013). Van Wasserhove and colleagues (2005) replicated these findings using a McGurk-paradigm with leading, synchronous or lagging visual input: the auditory N1 was reduced in the audio-visual condition compared to the auditory-only condition, but only in those cases where the visual stimulus preceded the auditory one. Moreover, the authors found a longer facilitation effect for visual stimuli that were visually easy to identify (for example the phoneme /p/) compared to those that were more ambiguous (like the phoneme /k/) (van Wassenhove, Grant, Poeppel, & Halle, 2005). A similar effect might have occurred in the present study: the lip movements might have helped the interpreters to predict the next phoneme thereby reducing the duration of silent pauses.

So far I have discussed the effect of audio-visual speech on speech perception, but it is still unclear why pupil sizes increased in the video condition compared to the audio condition. In order to answer this question I first reconsider the notion of perceptual load developed 2004 by Lavie and colleagues (see chapter 3.2.4). Lavie and colleagues postulated two kinds of load: a perceptual load that increases with the number of stimuli and a cognitive load that increases with the number of responses. Both loads interact with one another: If perceptual load is high, fewer resources are left for cognitive processes and vice versa. Compared to auditory-only input, audio-visual input increase the perceptual load because more stimuli (phoneme and lip movement) seem task relevant and need processing. Perceptual load may have been the reason why pupils dilated.

It is interesting to note that white noise added to the speech did not elicit larger pupil dilations. This does not necessarily need to contradict the perceptual load approach. White noise may not increase perceptual load because it does not contain useful information for the task at hand and is therefore ignored at a perceptual level. White noise may still increase the load at the cognitive level because it increases competition between potential lexical candidates (see also chapter 3.2.4 and 3.3.4). Based on the studies discussed in chapter 3.4.3.1, however, this account seems problematic as pupil dilation has also been observed in tasks that clearly fall into the category of cognitive rather than perceptual load (lexical competition, syntactical ambiguities, and multiplication tasks). If white noise does not increase perceptual load but only cognitive load, I still should have expected larger pupil sizes in the noise condition which was not the case (see chapter4.3.6.8).

A second theory which could explain larger pupil dilations in the audio-visual condition refers to Kahneman's capacity model of attention published in 1973 (Kahnemann, 1973) (see also chapter 3.2.1). According to him, pupils dilate with increasing arousal. It is perfectly conceivable that seeing a moving face simply elicits higher arousal than a static face maybe because it triggers a stronger emotional response or just because of the movement, without directly affecting the performance relating to work-load. For instance, Courtney and colleagues (2010) found higher self-reported and physiological arousal (skin conductance, heart rate and eye blinks) in response to moving pictures of feared stimuli than to static pictures of the same stimuli (Courtney, Dawson, Schell, Iyer, & Parsons, 2010). Arousal may also be interpreted as "sense of presence". In her meta-review on remote interpreting, Moser-Mercer notes (2005) the importance of the interpreter's "feeling of presence" (p. 732) for their motivation during their assignment. If interpreters do not feel part of the conference setting, they lose their motivation and performance suffers. According to her, the speaker's image can help to increase the sense of presence (Moser-Mercer, 2005a). With regard of the present study, this

would mean that pupil dilations were larger because participants could more easily immerse themselves in the situation when they saw the speaker actually speaking. Another argument in favor of the arousal-based explanation is the observation in the present study that pupil dilations decreased over the time course of the speech, regardless if noise or visual input was added to the speech. This observation is in line with Hyönä, Tommola and Alaja (1995) and can easily be explained by an elevated state of nervousness at the beginning of the speech. Assuming higher cognitive load at the beginning of the speech seems less convincing as the introductory part of the speech was certainly not the most difficult and fairly repetitive (see also chapter 4.3.6.8.1).

The present study does not allow me to definitely decide which of these two approaches – the load theory of selective attention and cognitive control (Lavie, Hirst, de Fockert, & Viding, 2004) or the capacity model of attention (Kahnemann, 1973) - is accurate. Audio-visual speech may have elicited larger pupil dilations either because the visual input was an additional perceptual load (despite having some advantages for speech processing as discussed above) or because the moving face caused a higher state of general arousal that is not necessarily related to perceptual or cognitive load, but might be associated to other sources of arousal like nervousness as indicates the decreasing pupil dilations over the time course of the speech. For a more thorough investigation of pupil dilation during simultaneous interpreting it would be interesting to include in a post-trial questionnaire subjective ratings of performance satisfaction, anxiety or self-confidence or similar aspects.

In a nutshell, the present study suggests that audio-visual speech speeds up speech perception, possibly by guiding the interpreters' expectation of the oncoming auditory signal. Increased pupil dilations and voice frequency may be attributed to higher perceptual load or to a higher state of general arousal in the video-condition compared to the audio-condition.

4.4.2 Effects of noise

As white noise was not of main interest but only introduced as control variable, no concrete hypotheses were formulated with regard to the effects of noise added to speech. Based on the studies discussed in chapter 3.3.4 noise was, however, expected to increase work-load and to affect task performance and the perceived task difficulty. This was indeed the case. Participants judged speeches overlaid with noise as being more difficult than those without noise. Overlaying white noise on the speech decreased translation accuracy and response accuracy for text-related questions. These findings validate noise as a load increasing variable in simultaneous interpreting. One exception, though, were cognate translations. Cognates are words of different languages that are phonetically similar and therefore thought to be more strongly activated during translation (see chapter 3.4.2). Cognate translations were expected to increase during simultaneous interpreting with noise because monitoring is less effective (see chapter 3.4.2). Yet, no effect for noise was observed on cognate translations. The reason for the absence of effect may be that with noise masking the speech, the auditory input did not resemble its cognate translation sufficiently anymore. Hence, the cognate candidate did not receive more activation than other lexical candidates that fitted into the context.

Noise had also an effect on the duration of silent pauses: Interpreters made longer silent pauses in the condition with noise than when no noise was added to the speech. When audio-visual speech was provided, silent pauses were shorter than when no lip movements were visible suggesting that interpreters made effective use of complementary visual information to disambiguate the auditory input. Nevertheless, silent pauses were still longer when the speech was masked by noise – with or without audio-visual speech input. This could mean that if the auditory stream is defective, interpreters need to wait for more context information that might help them to fill the gaps. In this respect the notion of time seems

particularly interesting. Barrouillet and colleagues (2007) developed a time-based resource sharing model described in chapter 3.2.5. According to him, work-load comes essentially down to the amount of time that is needed to process a stimulus. If speech perception in noise takes longer than without noise, less time is available for in-depth-processing of the speech's content entailing poorer recall of the speech and lower translation accuracy. A similar account could hold for accented speech that has been found to affect the quality of simultaneous interpreting (Mazzetti, 1999).

The importance of sentence context and predictability has been demonstrated in a 2011 study by Ostrand, Blumenstein and Morgan (Ostrand, Blumenstein, & Morgan, 2011) and described for simultaneous interpreting by Chernov (Chernov, 2002) and Setton (Setton, 1999). Even though the authors speak essentially about "forward" predictions, it seems reasonable to assume the possibility of "backward" predictions that help to restore lost information. The fact that speech context can help to disambiguate the auditory signal raises the question whether according to the load theory of selective attention and cognitive control (Lavie, Hirst, de Fockert, & Viding, 2004)(see chapter 3.2.4) white noise acts at a perceptual or a cognitive level. In order to answer that question, it might be useful to recall the effect of noise on pupil dilation during the pretest (word recognition task) and the main part of the experiment (listening or translating). During the word recognition task in the pretest, pupils dilated reliably with increasing levels of noise. The lower the signal-to-noise ratio was, the larger the pupil size. In contrast to the pretest, though, no effect for noise on pupil dilation was observed during the main part of the experiment, e.g. during listening to or during oral translation of a speech. These contrasted findings might point to an effect of task. White noise may act on different levels according to the task that is given to the participants. During word recognition, stimuli were completely unpredictable. Participants needed thus to concentrate on what they hear if they wanted to respond correctly. In this case, noise acted according to Lavie and

colleagues' definition of perceptual load (see chapter 3.2.4) - essentially at a perceptual level[48]. It covered parts of the stimulus, making it more difficult to identify the phonemes. In a word recognition task, it is all about speech perception. The case is different for listening or oral translation. As the speaker moves on in his speech, interpreters and listeners can use more and more context information to understand what is being said. Listening comprehension goes beyond mere speech perception. Instead, a speech analysis takes place that includes not only perceptual features but also the syntactical, semantic or pragmatic structure (see for example Setton, 1999; Cutler & Clifton, 1999). When noise is added to the speech, certain parts become unintelligible and the number of lexical candidates increases, but it is still constrained by the speech context (and possibly by visual cues, see Massaro & Cohen, 1999). According the load theory of selective attention and cognitive control (Lavie, Hirst, de Fockert, & Viding, 2004, for details see chapter 3.2.4), the selection of the correct or at least fitting lexical candidate is a cognitive problem, not a perceptual one. During listening to or translation of a whole speech, white noise represents a cognitive load factor, not a perceptual one.

Does this mean that pupils react only on perceptual load and not on cognitive load given that pupils dilated with audio-visual input but not with noise? It seems unlikely that pupils exclusively respond to perceptual load because pupil dilation has also been observed in cognitive tasks (multiplication, syntactical ambiguities, lexical competition, see chapter 3.4.3.1). Instead, Kahnemann's capacity model of attention (1973) might provide a more suiting explanation by relating pupil dilation more generally to arousal. The signal-to-noise ratio was apparently sufficient for

[48] Noise would act exclusively at a perceptual level if non-words had been used, as (existing) words (and even pseudowords) still underlie phonological and lexical rules; this was, however, not the case. In fact, the purpose of the pretest was to set the word recognition level at the participant's individual 75% threshold. Using non-words did not seem practical for this purpose.

performance to suffer but white noise might not have increased arousal sufficiently to be noticeable in the statistical analysis. Arousal might have been primarily influenced by nervousness at the beginning of each trial or the sense of presence or immersion in the conference setting (see Moser-Mercer, 2005a), and much less by a feeling of anxiety that participants might have experienced when they translated the speech with noise. The results of the present study do not, however, support Kahnemann's idea that participants invest more capacities or mental effort when their arousal is high. If this was true performance should have decreased over the time course of the trial as pupils constricted. Yet, no effect of time on translation accuracy was found[49].

Interestingly, voice frequency behaved differently than pupil dilation. While pupil dilation decreased over the time course of the speech, voice frequency tended to increase. This increase was steeper when noise was added to the speech probably reflecting the stress that interpreters experienced when they were not able to perfectly perceive the auditory input. The fact that pupillary responses and voice frequency showed the opposite trend raises the question if pupillary responses and voice frequency are sensitive to different kinds of stress. Pupil sizes were particularly large at the beginning of each trial when participants were in a state of expectation. They did not really know what awaited them during the speech. As they moved on in the speech, the expectation dissolved gradually. Voice frequency, in contrast, was low at the beginning of the speech when participants translated the usual introductory statements and still felt confident. Over its course the speech moved from usual introductions to more specific content and voice frequency rose. The

[49] The time course of other performance-based measures used in this experiment, like text related questions or cognate translations, was not investigated, either because they were not related to time (text-related questions) or because they were not evenly distributed during the speech (cognate candidates).

stepper increase of voice frequency over the time course of the speech in the noise condition might suggest that participants found it increasingly difficult to follow the speech and may even have missed more segments towards the end of the speech (though no significant effect of time course on translation accuracy was found). The state of expectation seems to have a different quality than the feeling of anxiety that might come up in interpreters when they find it more and more difficult to follow the speech or might in general feel less confident to deliver a translation of good quality.

To sum up, the results of the present study suggest that white noise increases work-load during simultaneous interpreting or listening to a speech. As a result, the speech is perceived as being more difficult and performance suffers. The contrasted effects of white noise on pupillary responses on the pretest compared to the main part suggests that white noise has a different effect on work-load during a word recognition test than during listening to or oral translation of a whole speech. It seems in particular that white noise may not impair listening comprehension during a speech to the same extent as recognition of single words because listeners or interpreters can make use of the context to restore lost information.

4.4.3 Effects of task

Apart from effects of audio-visual speech and white noise, the study also revealed several task-related effects. For instance, listeners perceived speeches as being easier compared to interpreters. This finding is not surprising and suggests that simultaneous interpreting indeed causes higher work-load than listening. The higher work-load during simultaneous interpreting compared to listening has often been ascribed to the multitude of processes that take place simultaneously (see for example Gile, 2009). Wickens' model of task interference provides another view on work-load during simultaneous interpreting and listening (see chapter 3.2.3). According to Wickens, load in multitasking is caused by interference

between tasks that tap into the same resources either because both tasks share the same sensory modality, the same code (verbal or spatial) or the same stage (perceptual, cognitive, response). Applying Wickens' model to simultaneous interpreting, Seeber and Kerzel (Seeber & Kerzel, 2012) have demonstrated that a large amount of work-load is caused by interferences between the source and the target speech (see also chapter 3.3.2). As listening requires only processing one speech instead of two, the amount of interference is much lower during listening than during simultaneous interpreting. Yet another theoretical approach is provided by Sweller's cognitive load theory (see chapter 3.2.2). Sweller claims that cognitive load increases with the number of elements that interact with each other. It might be argued that interpreting represents higher element interactivity than listening, as listeners need to build only a mental model of the source speech whereas interpreters additionally need to maintain a mental model of their target speech in order to verify if both representations still correspond to each other.

Furthermore, listeners gave on the whole more correct answers to text-related questions. This has also been observed by Lambert (1998) and Gerver (1974) and might suggest that the higher work-load during simultaneous interpreting leaves fewer resources for in-depth-processing of the speech (see also chapter 4.3.6.3.1). One more interesting finding is that listeners were more affected by noise than interpreters. When noise was added to the speech, listeners showed poorer recall than did interpreters. One explanation might be provided by the need to anticipate in simultaneous interpreting. Interpreters need to give an immediate translation of the speech they hear. Structural differences between the source and the target language (syntax, grammar) force the interpreter to constantly make predictions about how the speaker will finish his sentence. This need to anticipate could contribute to make interpreters less dependent on the auditory presentation of the speech, e.g. if the speech is presented with or without noise.

Interpreters also showed overall larger pupil sizes than listeners. On a first thought, this effect may be ascribed to the higher work-load during simultaneous interpreting. However, if pupil dilation was solely explained by work-load, then we would at least expect an effect of noise on pupil dilation. Yet, no effect of noise on pupil dilation was found. Effects on pupil dilation were only found for visual presentation and task. It is also important to bear in mind the similarities. In both tasks, pupil sizes were large at the beginning of a trial and declined towards the end of the trial although the beginning of the speeches only contained the usual greetings and introductory statements that are expected at the beginning of a speech. Moreover, pupil sizes decreased in both tasks more slowly during speeches with audio-visual speech input compared to those with auditory-only input despite the facilitating effect of audio-visual speech for speech perception. This suggests that in the present study pupillary responses were rather related to arousal than to work-load. The sources of arousal, however, might be diverse and difficult to disentangle. Higher arousal at the beginning of a speech might be due to a state of expectation (see chapter 4.3.6.8.1). Audio-visual input may elicit higher arousal either because moving images are more interesting than static ones or because the face of the speaker and his lip movements strengthened the sense of presence (see chapter 4.4.1). In the case of the task effect, simultaneous interpreting may have led to higher arousal because simultaneous interpreting involves the need to constantly respond to the stimulus whereas listening does not require a constant response. The fact that listeners and interpreters reacted similarly on audio-visual input and that audio-visual input had no positive or negative effect on text-related questions or (for interpreters) on translation accuracy suggests, however, that the increase of pupil dilations during speeches in the audio-visual mode are probably task-independent. It does not seem as if lip movements represent an additional burden for interpreters compared to listeners nor does it seem as if interpreters use lip movements more extensively than listeners.

In conclusion, the findings of the present study indicate higher work-load and as a consequence lower in-depth-processing of the speech during simultaneous interpreting compared to listening. Notwithstanding the advantage of listening over simultaneous interpreting for in-depth-processing, interpreters seem to better compensate for adverse listening conditions like white noise. Moreover, simultaneous interpreting was associated with higher arousal than listening.

4.4.4 The impact of audio-visual speech input on work-load in simultaneous interpreting

The aim of this study was to find out whether lip movements reduce work-load in simultaneous interpreting. The results of the study as summarized in chapter 4.4.1 do not support this hypothesis. In particular, audio-visual speech did not contribute to enhance translation accuracy or recall of the speech content. The only beneficial effect observed on interpreting performance was the lower number of long silent pauses (see chapter 4.3.6.6) which might suggest that lip movements help to prepare the auditory processing of speech without necessarily reducing work-load. No adverse effect of audio-visual speech on work-load was found either. The only negative effect of audio-visual speech I found may be larger pupil dilations during listening or simultaneous interpreting with audio-visual speech but this effect may not necessarily be linked to work-load, but simply to higher arousal due to the moving face compared to the static face. These findings are difficult to explain with Gile's effort model (2009, see also chapter 3.3.1) or Seeber's model of cognitive load in simultaneous interpreting (2011, see also chapter 3.3.2) as in both cases the effort or cognitive load is expected to decrease with audio-visual speech. Considering Gile's model, lip movements were expected to reduce the *listening and analysis effort* and allow for a reallocation of attentional resources that should lead either to better speech production (translations accuracy, cognate translations) or to better storage (text-related questions) (see also chapter 3.3.3). Better speech production or

recall with audio-visual input, however, was not observed even though the lower number of long silent pauses suggests an *effect on the listening and analysis effort*. A similar conclusion holds for Seeber's model of cognitive load in simultaneous interpreting (see chapter 3.3.2). If lip-movements were to reduce the perceptual load component or at least were neutral with regard to the perceptual load component, the overall cognitive load during simultaneous interpreting with audio-visual input should have been lower than during simultaneous interpreting in the auditory-only condition (see chapter 3.3.3). Even though speeches in the audio-visual mode obtained easier ratings, pupillary responses were larger in the audio-visual condition than in the auditory-only condition which would suggest on the basis of Seeber's model that the perceptual load component actually increased. If audio-visual speech input really increases the overall work-load without having some benefit for the interpreter like enhancing the quality of the translation, it is difficult to conceive why conference interpreters wish to have visual contact with the speaker. In fact, it seems more likely that audio-visual speech leads to higher arousal in simultaneous interpreting (as well as listening) without affecting cognitive load or effort.

What does it mean if arousal is higher during simultaneous interpreting or listening with audio-visual input? According to Kahnemann (1973), arousal is a physiological manifestation of the effort a person makes to solve a task. Usually, a person will increase its effort as the task gets more difficult. There are, however, further factors that may influence mental effort, like fatigue or motivation. If pupil dilation reflects arousal and arousal is linked to the amount of mental effort that a person undertakes, the person should either be able to solve more difficult tasks or to perform better in a task. In the present study, both were not the case. All speeches were comparable with regard to their complexity. Speeches presented in the audio-visual mode were rated as being slightly easier than speeches in the auditory-only mode, but all speeches were presented in both conditions so that the rating differences can solely be attributed to the

condition and not to the speech itself. Furthermore, performance did not increase when audio-visual input was provided compared to the auditory-only mode. Still, pupils dilated more during simultaneous interpreting or listening with lip movements than without lip movements. In the case of a complex task like simultaneous interpreting or listening to a whole speech, Kahnemann's model does not seem to provide sufficient explanation to predict the impact of audio-visual speech on arousal in simultaneous interpreting.

Another interesting approach is the *sense of presence*. The concept of *sense of presence* emerged with the advent of technologies simulating virtual reality (Riva & Mantovani, 2014). *Presence* is the feeling of being part of an (mediated) environment. Factors that mediate the *sense of presence* are for example the ability to move within the environment or to modify the environment, the presence of others or the possibility to interact with others and especially the naturalness and vividness of perceptual sensations (Lessiter, Freeman, & Davidoff, 2001; Bystrom, Barfield, & Hendrix, 1999). The latter is not surprising as the ability to correctly integrate and interpret multisensory perceptions was (and still is) crucial for survival (Riva & Mantovani, 2014). This might be the reason why multisensory input like auditory or tactile cues in a virtual environment enhances the *sense of presence* (Dinh, Walker, Song, & Kobayashi, 1999). The *sense of presence* is associated with stronger task involvement. At its optimum, feeling present might even lead to a flow experience – a state of full presence that is associated with high levels of concentration, task engagement and satisfaction (Riva & Mantovani, 2014).

In interpreting research the concept of presence has found immediate application in remote interpreting. Remote interpreting means that the conference interpreter is not on site with direct view on the speaker and the audience but located on a remote site with a video transmission of the conference. Usually, the video transmission includes the speaker and the

conference room. It is used for example in cases when the attendants of a conference or a meeting cannot physically come together or when the number of interpreters that are required exceeds the number of available booths (Mouzourakis, 2006). In a study comparing on-site and remote interpreting, Moser-Mercer found that performance decreased more slowly during on-site than during remote interpreting. Interpreters working on site tended to feel less anxious and reported lower stress levels than when working remote (although cortisol levels did not differed significantly between the two conditions) (Moser-Mercer, 2005b). In studies conducted by the European Union and the United Nations, interpreters reported to feel more motivated and better able to anticipate when working on site than when working remote (Moser-Mercer, 2005a). Moser-Mercer (2005b), but also Mouzourakis (2006) explain the discrepancy between remote and on-site interpreting with a lack of sense of presence during remote interpreting. A similar explanation could hold for the findings reported by Seubert (2017): although the speaker's face was larger and therefore better visible on a second screen, fixation times revealed that interpreters preferred the "real" speaker to the screen transmission.

From the present study it seems that even the sole video of the speaker talking elicits higher arousal as evidenced by larger pupil dilations compared to an auditory-only condition. Given that participants perceived speeches with audio-visual input as being easier, it may be speculated that participants experienced this state of higher arousal as being present. In this sense, audio-visual speech has no direct effect on work-load during simultaneous interpreting. It is not even clear if lip movements facilitate speech perception during simultaneous interpreting. Even though the lower number of long silent pauses in the audio-visual condition might hint in that direction, it is also conceivable that higher arousal is at the origin of this effect and that a talking face prompts interpreters to continue their translation. The tight association between sense of presence, task engagement, concentration and satisfaction (see above) might also explain why interpreters insist on seeing the speaker even though direct

beneficial effects on performance or interpreting quality might not always be visible. This shows also that the idea of higher arousal increasing the amount of available resources that are automatically reinvested into other sub-components of a task might be a little bit simplistic. Considering the findings reported by Moser-Mercer (2005b) and in particular the faster decline of interpreting performance in the remote condition, it may be speculated if arousal instead contributes to maintain concentration and thus to ensure high interpreting quality over a longer duration.

4.4.5 Limitations

The present study suggests that interpreters might benefit from the talking face in terms of task engagement or sense of presence. This conclusion is, however, limited in several ways. First, visual input in the present study consisted either of a still of the speaker's face or of a movie of the speaker's talking face. No other visual input was provided. For this reason, the present study does not allow drawing conclusions on other types of visual input that might play a role in simultaneous interpreting like presentations, gestures or reactions from the audience with regard to their impact on simultaneous interpreting. These types of visual input might even have a very different effect on work-load during simultaneous interpreting as they can provide additional information that needs to be processed which is usually not the case for lip movements (at least as long as the speech is not degraded by noise, see chapter 3.3.4). Limiting the study to visual speech has, however, enabled a thorough and systematic investigation of this particular type of visual input in simultaneous interpreting and listening that would not have been possible to that extent if other types of visual input (gestures, presentations) were included in this paper. It was not only possible to exclude interactions between different kinds of visual input but also to use different methodologies that are not applicable to all kinds of visual input.

Second, it may be said that lip movements are actually not a typical case of visual input in simultaneous interpreting because interpreters often

enough are too far from the speaker to identify his lip movements. The experiment took place in a controlled environment that has little similarity with the usual conference setting interpreters are used to. Lip movements may not even be the reason why interpreters insist on having visual contact with the speaker. Instead, they may assign more importance to the speaker's gesture or his facial expressions. Still, it is remarkable that interpreters most often cite the speaker as a crucial source of visual input that induces the sense of presence (Moser-Mercer, 2005a). This may again be taken as an indicator that seeing the speaker enhances immersion and strengthens the sense of presence. This way, visual contact with the speaker may be crucial for task engagement and sustained concentration in simultaneous interpreting.

A third important limiting factor is the fact that only interpreter trainees of the University of Mainz (and translators in the listening group) participated in the study. Interpreter trainees at the University of Mainz may use the university's repository where they find all speeches that were recorded during class for exercise. Most often, only the lecturer's voice reading out the speech is recorded. Interpreting trainees are therefore used to work with auditory-only input. The case is different for conference interpreters who are accustomed to real life settings and may therefore react more strongly to the absence of visual input. On the other hand, experienced conference interpreters may have acquired techniques that may help them to deal more efficiently with visual input in general and potentially high load inducing visual elements, like presentations or graphs. Further research on how experienced interpreters deal with different types of visual information may provide interesting insights for teaching.

Finally, the study has not taken into account individual preferences with regard to visual input during simultaneous interpreting. Participants were not asked prior to the experiment whether they find it in general helpful to see the speaker's face or not. It cannot be excluded that some interpreters find visual speech input helpful or stress-reducing, while others experience

it as additional burden or do not see any difference in interpreting with or without seeing the speaker's face. Yet, differences between participants when dealing with visual input during simultaneous interpreting might have affected the results of the study and especially of physiological indicators like pupil dilation and voice frequency. In the same vein, there may be substantial differences between interpreters in coping with stressors in general. The effects on voice frequency, indeed, suggest that participants reacted differently to white noise during simultaneous interpreting (see chapter 4.3.6.7.1). While some interpreters may experience a high level of distress with low levels of physiological arousal (sensitizers), others may show the inverse pattern: they report comparatively low levels of distress but their physiological arousal may be rather high (repressors) (Bruner & Postman, 1947; Byrne, 1964). A state-trait anxiety test (STAI) or some other method to collect self-reported anxiety administered prior to the experiment combined with physiological methods might therefore give a more precise picture of how interpreters cope with stressors.

5 Conclusion

Simultaneous interpreting is a very complex task as it requires the interpreter to listen to and understand a speech in one language and to render it in real-time into another language. In addition, the interpreter faces various types of visual input that add to the auditory input in simultaneous interpreting. A study using physiological and performance-based measures and subjective ratings was conducted to investigate the impact of visual input and in particular visible lip movements on work-load during simultaneous interpreting. The findings suggest that seeing the speaker's face and his lip movements does not lower work-load during simultaneous interpreting but enhances the sense of presence and hence task engagement.

5.1 Methodological implications

One of the major interests of this study lies maybe in the combination of different physiological measures (pupillary responses, voice frequency), performance-based indicators (translation accuracy, cognate translation, silent pauses and response accuracy to text-related questions) and subjective ratings to investigate the effects of audio-visual speech and noise on simultaneous interpreting. Some of the methods have barely been used in simultaneous interpreting so far (voice frequency, pupil dilation, silent pauses) and may be interesting for further research. Performance-based measures have maybe the highest potential as work-load indicator in simultaneous interpreting. Performance-based measures are usually non-intrusive as they require only the recording of the interpreter's rendition. In particular, the duration of silent pauses proved to be a very interesting method that is sensitive to changes in work-load caused by visual input and white noise. Silent pauses are automatically recorded with the recordings of the interpreter's performance and can easily be detected using adequate programs like *praat* (Boersma & Weenik, 2013). Further methods, like the analysis of cognate translations or of translation accuracy[50], however, require manual tagging of the recordings and were not always successful in detecting differences in work-load between different conditions. Response accuracy to text-related questions was most strongly affected by the task (interpreting versus listening) and was not able to capture subtle differences between the audio-visual condition and the auditory condition, but it may provide valuable information with respect to deep processing or the degree of text comprehension (Craik & Lockhart, 1972). Translation accuracy was only affected by the addition of white noise to the speech; cognate translations did not show any effect at all. Most performance-based measures that

[50] Translation accuracy is defined in this study as the number of correctly rendered speech segments (see chapter 4.3.6.4).

have been used in the present study seem to be not fine-grained enough to capture te subtle differences in work-load that are triggered by the presence or absence of audio-visual speech in simultaneous interpreting.

Physiological measures are very interesting because they make it possible to track changes in work-load in real-time. Physiological measures can be very sensitive in a controlled experiment with a simple task. A very impressive example is provided by the pretest where pupil dilations reacted to the signal-to-noise ratio and to the cognate status. In a less controlled experiment or in a more complex task, the analysis of physiological measures can be cumbersome and difficult to interpret. As such, the effect observed on pupil dilations could not simply be traced back on work-load changes. Instead, it was necessary to relate all findings in order to understand the effect on pupil dilations and to associate it to arousal. Still, recording pupillary responses have the merit to interact very little with the experiment as the only requirement is that participants look at the screen. Often, it is recommended to use a head rest to obtain a better data quality when recording pupillary responses. While head rests in general certainly improve the data quality, they may also increase the intrusiveness of the method and may not be suited equally well for all kinds of experiments. For instance, they may hinder mouth or lip movements that are necessary during speaking. Combined with the appropriate statistical techniques, pupillary responses are a powerful tool to track arousal. They do not, however, inform about the origin of the arousal. It is therefore sensible to combine physiological measures with performance-based measures. A similar conclusion holds for voice frequency. Voice frequency is a completely non-obtrusive measure that can be collected over the whole duration of the experiment via the recordings of the interpreter’s rendition. The study has demonstrated, however, that the effects are minimal and difficult to ascribe to a specific factor. This is probably due to the modulation of the voice during speaking. Assuming that the intonation becomes more monotonous under stressful conditions when the voice is tense, another interesting approach may be

variance in voice frequency or intonation patterns during simultaneous interpreting.

Subjective ratings are interesting because they reflect the participants' perception of the task. As Roziner and Shlesinger (2010) have demonstrated, there can be huge differences between the subjective perception of work-load and physiological measures. Subjective ratings carry the risk that participants' responses are biased or confounded. In the present study, participants consistently rated speeches as being more difficult when noise was added. As all speeches were presented with noise, it was clear that these ratings did not depend on the speech and the speech difficulty but on the addition of noise. This example illustrates the importance of asking the right question. It might also be useful to include various aspects of work-load like it is the case in the NASA-TLX (Hart & Staveland, 1988) in order to obtain a differentiated view on work-load during an experiment. Another potential pitfall is the rating scale. As such, the estimations of speech duration did not show any effects, possibly because the participants centered their estimation on the "happy medium". Moreover, rating scales provide *ordinal* data which requires advanced statistical methods. *Ordinal* data can be bypassed by using just a line with two extremes where participants set the cursor at the place that corresponds to their rating. The strength of subjective ratings is certainly that they are non-intrusive. On the other hand, comprehensive questionnaires can take a lot of time and weaken the participant's motivation. Nevertheless, it is important to remember that in the end, work-load, stress or arousal are all defined by the participant's experience. If somebody experiences a task as being stressful or associated with high load, then the task in question is stressful or difficult for this person. In this respect, subjective ratings are a valuable tool to gain insight in how a task is experienced.

From the present study, it seems in particular that a combination of different methods can provide interesting insights. That is, each method

has the potential to compensate for the weaknesses of another method. By bringing together the results obtained with different methods and relating them to each other, it is possible to draw conclusions on how different experimental conditions affect simultaneous interpreting and to what extent. In the present study, the effect on each load indicator allowed me to draw conclusions not only on the impact of audio-visual speech and noise on simultaneous interpreting but also on how these effects were to be interpreted.

5.2 Practical implications

The present study dealt with the impact of the speaker's lip movements as a source of visual information during simultaneous interpreting. But in a real-life setting conference interpreters are often too far away too see the speaker's lip movements. An interpreter might therefore rightly ask: What are the practical implications for me? What can I learn from this study for my daily work? First of all, knowing about visual information, its informational value and its impact on work-load can contribute to improve work conditions during simultaneous interpreting. If interpreters know what types of visual input are particular important and for what reason, it is easier for them to insist on having access to those types of visual information. For example, the present study demonstrated that seeing the speaker's face increases task engagement and hence, may contribute to maintain the interpreting performance over longer periods, e.g. the effects of fatigue may set at a later stage (Moser-Mercer, 2005b). One very practical recommendation could be to locate the booth not only in a way that the interpreter has somehow visual contact with the speaker (International Organization for Standardization, 1998), but that she can clearly see the speaker's face from the front.

Seeing the speaker's face plays certainly an even more important role in remote interpreting where interpreters' task engagement suffers most from lacking visual input. As the findings of the present study suggest, showing the speaker's moving face on a screen can help to enhance the

interpreters' sense of presence and to immerse with the situation. This may also contribute to make remote interpreting less straining and more acceptable for conference interpreters. From earlier studies it seemed that the technical set-up – and especially the video transmission - may be more challenging in remote interpreting than in on-site interpreting [51] (Mouzourakis, 2006). Hence, remote interpreting may be particularly susceptible to degradations of sound or video quality. In this respect, a particularly interesting finding of the present study is that visible lip movements seem to facilitate lexical predictions as indicated by shorter silences during simultaneous interpreting with audio-visual speech. Audio-visual speech may not totally compensate for adverse listening conditions. Nevertheless, it may mitigate the devastating effect of bad sound quality a little bit.

What do the results of the present paper mean for training? At the University of Mainz (and most probably this is the case for other training institutes, too), speeches read out during the class are recorded and made available to the trainees in order to allow them to practice outside classes. The informational added value of lip movements may be negligible provided that the sound quality is excellent. Also, processing lip movements certainly does not need extra training as audio-visual speech the normal case in face-to-face communication. Still, providing not only an auditory record of the speech, but also a video of the lecturer's face may be a relatively simple and cost-efficient way to strengthen the trainee's task engagement during practice.

As described in chapter 2.4.1, visual input during simultaneous interpreting is very divers and reaches from lip movements that correspond exactly to the auditory stream to body movements without precise semantic value

[51] The technical constraints may have changed considerably over the last ten years and may not be an issue today to the same extent as it has been when the first studies on remote interpreting were conducted.

and written documentation that might even provide very different complementary information (glossaries, background information, etc.). It seems reasonable to expect that different types of visual information have very different practical implications for simultaneous interpreting. Some types of visual input may be processed without effort (lip movements) while others may increase work-load considerably (searching for a term in a glossary, for example). Some may clearly provide some additional information that may help to correctly interpret what the speaker is saying (a pointing gesture, for instance), while others may overlap to a larger extent with the auditory stream (lip movements or a speech manuscript followed by the speaker). Finally, some types of visual information may enhance the interpreter's feeling of presence (seeing the speaker's face, for example), while others may be completely irrelevant in this respect (for example written documentation). The present study is thus inconclusive with regard to the benefit of visual input in simultaneous interpreting; however, it contributes to shed light on the question how visual input affects interpreters.

5.3 Future research

Further research could center around two topics: 1) individual differences in coping with stressful situations in simultaneous interpreting and 2) the effect(s) of different types of visual input on simultaneous interpreting. The first topic holds that interpreters may react very differently on stressors during simultaneous interpreting (which is by itself a stressful task that is associated with high levels of arousal as evidenced by a slower decrease of pupil dilation during simultaneous interpreting than during listening, see chapter 4.3.6.8). Individual differences may in particular concern physiological measures, like voice frequency or pupil dilation. One indication for individual stress reactions in the present study comes from the effect of white noise on voice frequency. Voice frequency has been used as an indicator for emotional stress or anxiety (see chapter 3.4.3.2). While voice frequency was higher during interpreting with white noise than

in a condition without noise for some participants, it was lower for others and yet for others, voice frequency was similar in both conditions. This example demonstrates that individual stress reactions can sensibly influence physiological measures and that physiological data may show a very different pattern depending on how participants react to stressors. Future research that uses physiological measures to investigate simultaneous interpreting (though the following may be true for other tasks as well) should therefore take into account individual stress reactions in order to be able to correctly interpret physiological measures.

The second topic concerns the effects of different types of visual input. It is of particular relevance for remote interpreting and could contribute to improve visual input during remote interpreting. Interpreting research including the present study and anecdotal evidence suggests so far three categories of visual input during simultaneous interpreting. The first category comprises visual input that has informational value. This is for example the case for the audience. Rennert (2008) provides a good illustration of this category:

> "The speaker began his first speech with 'Good evening, ladies and gentlemen…ladies and gentleman' (followed by a little chuckle), using deictics to indicate first the audience as a whole and then the only male person in the audience. This is a good example of the importance of seeing the entire room and the audience, since the utterance only makes sense if the interpreter is aware that there is only one male person present." (Rennert, 2008, p. 213)

Other examples are the conference room that speakers often refer to, pointing gestures that show what the speaker means when he uses demonstrative pronouns or a smile on his face that indicates that he was making a joke. Visual input in this category provides complementary information that cannot be deduced from the auditory stream.

The second category includes visual input that affects work-load. One example is numbers. Numbers are a common "problem trigger" in simultaneous interpreting and rarely translated correctly (Gieshoff, 2012).

In an eye-tracking study on numbers in simultaneous interpreting, Seeber (2012) demonstrated that fixation durations on written numbers that were displayed on the screen were significantly longer during sentences with large numbers than during control sentences with small numbers where participants fixated mostly the speaker's face[52]. Written numbers may thus lower work-load caused by large numbers in simultaneous interpreting. The case may be different for speech manuscripts. For instance, DeLaet and Plas observed that interpreting performance decreased when interpreting trainees tried to follow the speech manuscripts during interpreting (De Laet & Plas, 2005). Speech manuscripts may thus be one type of visual input that increases work-load although the manuscript may also be helpful and provide complementary information, for example when the sound quality is severely degraded.

Finally, the third category contains visual input that has neither informational value nor an effect on work-load, but an effect on task-engagement during simultaneous interpreting. This seems to the case for the speaker's face as demonstrated by the present study. In contrast to the initial hypothesis, seeing the speaker's face did not lead to smaller pupillary responses indicating lower work-load, but to larger pupil dilations. In spite of larger pupil dilations, participants still rated speech with audio-visual input as being less difficult than speeches without audio-visual input. This discrepancy between subjective ratings or perceptions and objective measures has also been observed by Roziner and Shlesinger (2010). In a very comprehensive study, they showed that interpreters perceived the booth as less comfortable and the interpreting performance as less satisfying during remote interpreting with limited visual input than

[52] The study has some methodological issues, though. For instance, the regions of interest defined around the speakers face, his torso and the written numbers and used to calculate the fixation durations were not equal in size. Moreover, it is not clear if sentences were equal in length.

during on-site interpreting although there was no objective difference. Interpreters complained more often about headaches, eye strain, and concentration difficulties in the remote condition than in the on-site condition although no differences were found in interpreter's health. One explanation for this contrasted picture could be the fact that interpreters feel more alienated when they are not present on site or not in a real-life interpreting setting and engage less with the task. Finding out which types of visual input increase task-engagement may contribute significantly to improve the interpreters' well-being during remote interpreting. Similarly, interpreters and interpreting trainees may benefit from knowing about different types of visual input and which one provide informational value and at which costs. For instance, if following a speech manuscript during simultaneous interpreting increases work-load significantly, it only may prove useful in adverse listening conditions – or to put it differently: the manuscript may only have sufficient informational value when the auditory input is strongly degraded. Research on the effects of visual input may thus contribute to improve working conditions during remote interpreting and may inform interpreting trainees what types of visual input helps them best and under which conditions.

6 References

Agnès, A., Möttönen, R., Sams, M. E., Soto-Faraco, S., & Tiippana, K. (2014). Effect of attentional load on audiovsisual speech perception: evidence from ERPs. *Frontiers in psychology, 5*, pp. 1-9.

Ahrens, B. (2004). *Prosodie beim Simultandolmetschen* (Vol. 41). (K. Pörtl, Ed.) Frankfurt am Main: Peter Lang.

Alvarado, J. C., Stanford, T., Vaughan, W., & Stein, B. (2007). Cortex Mediates Multisensory, But Not Unisensory Integration in Superior Colliculus. *Journal of Neuroscience, 27*(47), pp. 12775-12786.

Anderson, J. R. (2007). *Kognitive Psychologie* (6. ed.). (J. Funke, Ed., & G. Plata, Trans.) Berlin, Heidelberg: Springer.

Anderson, L. (1994). Contextual Aspects and Translation Aspects of SI. In B. Moser-Mercer, & S. Lambert, *Bridging the Gap* (pp. 101-120). Amsterdam/Philadelphia: John Benjamin's.

Anderson-Hsieh, J., & Koehler, K. (1988). The Effect of Foreign Accent and Speaking Rate on Native Speaker Comprehension. *Language Learning, 38*(4), pp. 561-613.

Andres, D., Behr, M., & Dingfelder Stone, M. (2013). *Dolmetschmodelle - erfasst, erläutert, erweitert.* Frankfurt a.M.: Peter Lang.

Audacity development team. (2015). *Audacity*. Retrieved 05 01, 2015, from http://audacityteam.org/copyright

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction, 16*, pp. 389-400.

Baart, M., Stekelenburg, J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia, 53*, pp. 115-121.

References

Baddeley, A. (1992). Working Memory. *Science, 255*, pp. 556-559.

Baddeley, A. (2003). Working Memory: Looking Back and Looking Forward. *Nature Reviews, 4*, pp. 829-839.

Baddeley, A., & Hitch, G. (1974). Working Memory. In G. Bower, *The psychology of learning and motivation: Advances in research and theory.* (Vol. 8, pp. 47-89). New York: Academic Press.

Baker, M. (1996). In H. Somers, *Terminology, LSP and Translation. Studies in language engineering in honour of Juan C. Sager* (pp. 175-186). Amsterdam/Philadelphia: John Benjamin's.

Baldauf, D., Burgard, E., & Wittmann, M. (2009). Time perception as a workload measure in simulated car driving. *Applied Ergonomics, 40*, pp. 929-935.

Barraclough, N. e. (2005). Integration of Auditory and Visual Information by Superior Tempora Sulcus Neurons Responsive to the Sight of Action. *Journal of Cognitive Neuroscience, 17*(3), pp. 377-391.

Barrouillet, P., Bernadin, S., & Camos, V. (2004). Time Constraints and Resource Sharing in Adult's Working Memory Spans. *Journal of Experimental Psychology: General, 133*(1), pp. 83-100.

Barrouillet, P., Bernadin, S., & Portrat, S. (2007). Time and Cognitive Load in Working Memory. *Journal of Memory and Language, 33*(3), pp. 570-585.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), pp. 1-48.

Beatty, J. (1982). Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources. *Psychological Bulletin, 91*(2), pp. 276-292.

References

Benoit, C., Mohamadi, T., & Kandel, S. (1994). Effects of Phonetic Context on Audio-Visual Intelligibility of French. *Journal of Speech and Hearing Research, 37*, pp. 1195-1203.

Bernstein, L., Auer, E., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lip-reading. *Speech Communication, 44*, pp. 5-18.

Bhatnagar, S. (2013). *Neuroscience for the Study of Communicative Disorders.* Philadelphia, Balitimore: Lippincott, Williams and Wilkins.

Blank, H., & von Kriegstein, K. (2013). Mechanisms for enhancing visual-speech information by prior auditory information. *NeuroImage, 68*, pp. 109-118.

Block, R. A., Hancock, P. A., & Zakay, D. (2010). How cognitive load affects duration judgments: A meta-analytic review. *Acta Psychologica, 134*, pp. 330-343.

Boersma, P., & Weenik, D. (2013). *Praat. Doing Phonetics by Computer[Computer program]. Version 5.3.51*. Retrieved june 02, 2013, from http://www.praat.org

Bonhage, C., Mueller, J., Friederici, A., & Fiebach, C. (2015). Combined eye tracking and fMRI reveals neural basis of linguistic predictions during sentence comprehension. *Cortex, 68*, pp. 33-47.

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., & Babiloni, F. (2014). Measuring neurphysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience and Biobehavoral Reviews, 44*, pp. 58-75.

Brancazio, L., Best, C., & Fowler, C. (2006). Visual Influences on Perception of Speech and Non-Speech Vocal Tract Events. *Language and Speech, 49*(1), pp. 21-53.

References

Brickenkam, R., & Zillmer, E. (1998). *d2 - Test of Attention* (1st U.S. edition ed.). (D. Emmans, Trans.) Cambridge/Toronto/Oxford/Bern/Göttingen: Hogrefe.

Brisson, J., Mainville, M., Mailloux, D., Beaulieu, C., Serres, J., & Sirois, S. (2013). Pupil diameter measurement errors as a function of gaze direction in corneal reflection eyetrackers. *Behavorial Research, 45*(4), pp. 1322-1331.

Broadbent, D. E. (1956). Successive responses to simultaneous stimuli. *Quartely Journal of Experimental Psychology, 8*(4), pp. 145-152.

Brouwer, A.-M., Hogervorst, M., Holewijn, M., & van Erp, J. B. (2014). Evidence for effects of task difficulty but not learning on neurophysiological variables associated with effort. *International Journal of Psychophysiology, 93*, pp. 242-252.

Brown, C., Clarke, A., & Barry, R. (2006). Inter-modal attention:ERPs to auditory targets in an inter-modal oddballtask. *International Journal of Psychophysiology, 66*, pp. 77-86.

Brown, I. (1978). Dual Task Methods of Assessing Work-load. *Egonomics, 21*(3), pp. 221-224.

Brown, R., & Page, H. (1939). Pupil dilation and dark adaptation. *Journal of Experimental Psychology, 25*(4), pp. 347-360.

Bruner, J. S., & Postman, L. (1947). Emotional Selectivity in Perception and Reaction. *Journal of Personality, 16*(1), pp. 69-77.

Bühler, H. (1985). Conference Interpreting. A multichannel communication. *Meta: Journal des Traducteurs, 30*(1), pp. 49-54.

Byrne, D. (1964). Repression-Sensitization as a Dimension of Personality. *Progress in Experimental Personality Research, 72*, pp. 169-220.

Bystrom, K., Barfield, W., & Hendrix, C. (1999). A Conceptual Model of the Sense of Presence in Virtual Environments. *Presence: Teleooperators and Virtual Environments, 8*(2), pp. 241-244.

References

Calvert, G., & These, T. (2004). Multisensory integration: Methodological approaches and emerging principles in the human brain. *Journal of Physiology-Paris, 98*, pp. 191-205.

Calvert, G., Hansen, P., Iversen, S., & Brammer, M. (2001). Detection of Audiovisual Integration sites in Humans by Application of Electrophysiological Criteria of the BOLD-Effect. *NeuroImage, 14*, pp. 427-438.

Campbell, R. (2008). The Processing of Audio-Visual Speech: Empirical and Neural Bases. *Pholosophical Transactions: Biological Sciences, 363*(1493), pp. 1001-1010.

Carl, M., & Schaeffer, M. J. (2017). Why Translation is difficult: A Corpus-Based Study of Non-Literality in Post-Editing and Scratch Translations. *HERMES - Journal of Language and Communication in Business, 43*, pp. 43-57.

Carlyon, R. (2004). How the brain separates sounds. *Trends in Cognitive Science, 8*(10), pp. 466-471.

Catford, J. (1965). *A Linguistic Theory of Translation. An Essay in Applied Linguistics* (5th ed.). Oxford: Oxford University Press.

Chandler, P., & Sweller, J. (1991). Cognitive Load Theory and the Format of Instruction. *Cognition and Instruction, 8*(4), pp. 293-332.

Chapman, R., Oka, S., Bradshaw, D. H., Jacobson, R. C., & Donaldson, G. W. (1999). Phasic pupil dilation response to noxious stimulation in normal volunteers: Relationship to brain evoked potentials and pain report. *Psychophysiology, 33*(1), pp. 44-52.

Chen, F., Ruiz, N., Choi, E., Epps, J., Khawaja, M. A., Yin, b., & Wang, Y. (2011, May). Mulitmodal Behavior and Interaction as Indicators of Cognitive Load. *ACM Transactions on Interactive Intelligent Systems (TiiS), 2*(4), pp. 39-72.

References

Chernov, G. V. (2002). Semantic Aspects of pscholinguistic research in simultaneous interpretation. In F. Pöchhacker, & M. Shlesinger, *The Interpreting Studies Reader* (pp. 99-109). London/New York: Routledge.

Chistoffels, I., Firk, C., & Schiler, N. (2007). Bilingual language control: A event-related brain potential study. *Brain research, 1147*, S. 192-208.

Christensen, R. H. (2015). *Analysis of ordinal data with cumulative link models - estimation with the R-package ordinal.* Retrieved may 01, 2017, from r-project.org: https://cran.r-project.org/web/packages/ordinal/vignettes/clm_intro.pdf

Chuderski, A., Senderecka, M., Kalamala, P., & Kroczek, B. (2016). ERP correlates of the conflict level in the multi-response Stroop task. *Brain research, 1650*, pp. 93-102.

Cohen, A. S., Dinzeo, T. J., Donovan, N. J., Brown, C. E., & Morrison, S. C. (2015). Vocal acoustic analysis as a biometric indicator of information processing: Implications for neurological and psychiatric disorders. *Psychiatry Research, 226*, pp. 235-241.

Conrad, R. (1955). Some effects on performance of changes in perceptual load. *Journal of Experimental Psychology, 49*(5), pp. 313-323.

Costa, A., Colomé, À., Gómez, O., & Sebastián-Gallés, N. (2003). Another look at cross-language competition in bilingual speech production: Lexical and phonological factors. *Bilingualism: Language and Cognition, 6*(3), pp. 167-179.

Costa, A., Satesteban, M., & Cano, A. (2005). On the facilitatory effect of cognate words in bilingual speech production. *Brain and Language, 94*, pp. 94-103.

Courtney, C. G., Dawson, M. E., Schell, A. M., Iyer, A., & Parsons, T. D. (2010). Better than the real thing: Eliciting fear with moving and

static omputer-generalted stimuli. *International Journal of Psychophysiology, 78*, pp. 107-114.

Cowan, N. (2000). The magical Number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, pp. 87-185.

Cowan, N. (2000/01). Processing limits of selective attention and working memory. Potential implications for interpreting. *Interpreting, 5*(2), pp. 117-146.

Cowan, N. (2009). Sensory and Immidiate Memory. In W. Banks, *Encyclopedia of Consciousness* (Vol. 2, pp. 327-339). Oxford: Elsevier.

Cowan, N. (2010). Multiple Concurrent Thoughts: The Meaning and Developmental Neuropsychology of Working Memory. *Developmental Neuropsychology, 35*(5), pp. 447-474.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning ad Verbal Behavior, 11*(6), pp. 671-684.

Cutler, A., & Clifton, J. C. (1999). Comprehending spoken language: A blueprint of the listener. In C. Brown, & P. Hagoort, *The neurocognition of language* (pp. 123-166). Oxford, New York: Oxford University Press.

Darò, V. (1994). Non-Linguistic Factors Influencing Simultaneous Interpretation. In S. Lamber, & B. Moser-Mercer, *Bridging the Gap. Emperical Research in Simultaneous Interpretation* (pp. 249-271). Amsterdan/Philadelphia: John Benjamnin's.

Davies, M. (2009). *Word frequency data. Corpus of Contemporary American English*. (B. Y. University, Editor) Retrieved 03 15, 2014, from http://www.wordfrequency.info/

References

de Jong, T. (2010). Cognitive Load Theory, Educational Research: Some Food for Thought. *Instructional Science, 38*, pp. 105-134.

De Laet, F., & Plas, R. V. (2005). La traduction à vue en interprétation simultanée: quelle opérationnalité ambitionner? *Meta: journal des traducteurs, 50*(4), p. no page indicated.

Debue, N., & van de Leemput, C. (2014). What does germane load mean? An empirical contribution to the cognitive load theory. *Frontiers in Psychology, 5*, pp. 1-12.

Dijkstra, T., Grainger, J., & van Heuven, W. (1999). Recognition of Cognates and Interlingual Homographs: the Neglected Role of Phonology. *Journal of Memory and Language, 41*, pp. 469-518.

Dinh, H., Walker, N. H., Song, C., & Kobayashi, A. (1999). Evaluating the importance of multi-sensory input on memory and the sense of presence in virtual environments. *Proceedings of the IEEE Virtual Reality*, (pp. 222-228).

Djamasbi, S., Mheta, D., & Samani, A. (2012). Eye Movements, Perceptions, and Performance. *Proceedings of the eighteenth America's Conference on Information Systems (AMCIS)*, pp. 1-7.

Dong, Y., & Lin, J. (2013). Parallel processing of the target language during source language comprehension in interpreting. *Bilingualism: Language and Cognition, 16*, pp. 682-692.

Dragsted, B. (2012). Indicators of difficulty in translation - correlating product and process data. *Across Languages and Cultures, 13*(1), pp. 81-98.

Ehrensberger-Dow, M., Göpferich, S., & O'Brian, S. (2015). *Interdisciplinarity in Translation and Interpreting Process Research.* Amsterdam/Philadelphia: John Benjamin's.

References

Engelhardt, P. E., & Ferreira, F. (2010). Pupillometry reveals processing load during spoken language comprehension. *The Quarterly Journal of Experimental Psychology, 63*(4), pp. 639-645.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon identification of a target letter in a nonsearch task. *Perception & Psychophysics, 16*(1), pp. 143-149.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*, pp. 429-433.

European Commission. (2017, 11 24). *What is it?* (European Commission, Editor) Retrieved 11 24, 2017, from Speech Repository. Interpretation: https://webgate.ec.europa.eu/sr/content/what-it

Friederici, A. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Science, 6*(2), pp. 78-84.

Friederici, A. (2009). Pathways to language: fiber tracts in the human brain. *Trends in Cognitive Science, 13*(4), pp. 175-181.

Friederici, A. (2012). the cortical language circuit: from auditory perception to sentence comprehension. *Trends in Cognitive Sciences, 16*(5), pp. 262-268.

Friederici, A., & Gierhan, S. (2013). The language network. *Current Opinion in Neurobiology, 23*, pp. 250-254.

Friederici, A., & Singer, W. (2015). Grounding language processing on basic neurphysiological principles. *Trends in Cognitive Sciences, 19*(6), pp. 329-338.

Galy, E., Cariou, M., & Mélan, C. (2012). What is the relationship between mental workload factors and cognitive load. *International Journal of Psychophysiology, 83*, pp. 269-275.

References

Gathercole, S., & Martin, A. (1996). Interactive processes in phonological memory. In S. Gathercole, *Models of Short-term memory* (pp. 73-100). East Sussex: Psychology Press.

Gerver, D. (1974). Simultaneous listening and speaking and retention of prose. *Quarterly Journal of Experimental Psychology, 26*(3), pp. 337-341.

Gerver, D. (1974). Simultaneous listening and speaking and retention of prose. *Quarterly Journal of Experimental Psychology, 26*(3), pp. 337-341.

Gerver, D. (1974). The effects of noise on the performance of simultaneous interpreters: Accuracy of Performance. *Acta Psychologica, 38*(3), pp. 159-167.

Gerver, D. (1975). A Psychological Approach to Simultaneous Interpretation. *Meta: Journal des Traducteurs, 20*(2), pp. 119-128.

Gerver, D. (2002). The effects of source language presentation rate on the performance of simultaneous conference interpreters. In F. Pöchhacker, & M. Shlesinger, *The interpreting studies reader* (pp. 53-66). London, New York: Routledge.

Gerver, D. (2002). The effects of source language presentation rate on the performance of simultaneous conference interpreters. In F. Pöchhacker, & M. Shlesinger, *The interpreting studies reader* (pp. 53-66). London/New York: Routeledge.

Ghose, D., Maier, A., Nidiffer, A., & Wallace, M. (2014). Multisensory Response Modulation in the Superficial Layers of the Superior Colliculus. *Journal of Neuroscience, 34*(12), pp. 4332-4344.

Giard, M., & Peronnet, F. (1999). Auditory-visual Integration during Multimodal Object Recognition in Humans: A Behavorial and Electrophysiological Study. *Journal of Cognitive Neuroscience, 11*(5), pp. 473-499.

References

Gieshoff, A. C. (2012). *Aus 92 wird zwölf - Zahlen im Simultandolmetschen (unpublished MA-thesis).* University of Mainz, Germersheim, Germany.

Gieshoff, A. C. (2017). Audiovisual speech decreases the number of cognate translations in simultaneous interpreting. In S. Hansen-Schirra, O. Czulo, B. Meyer, & S. Hoffmann, *Empirical modelling of translation and interpreting* (pp. 147-166). Berlin: Language Science Press.

Gile, D. (1985). Les termes techniques en interprétation simultanée. *Meta: journal des traducteurs, 30*(3), pp. 199-210.

Gile, D. (2008). Local cognitive load in simultaneous interpreting and its implications for empirical research. *Forum, 6*(2), pp. 59-77.

Gile, D. (2009). *Basic Concepts and Models for Interpreter and Translator Training.* Amsterdam/Philadelphia: John Benjamin's.

Gilzenrath, M. S., Niewenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive Affective Behavioral Neuroscience, 10*(2), pp. 252-269.

Goldman-Eisler, F. (2002). Segmentation of input in simultaneous translation. In F. Pöchhacker, & M. Shlesinger, *The interpreting studies reader.* London/New York: Routledge.

Goldwater, B. (1972). Psychological significance of pupillary movements. *Psychological Bulletin, 77*(5), pp. 340-355.

Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resource limitations. *Psychophysiology, 33*, pp. 457-461.

Groh, J., & Werner-Reiss, U. (2002). Visual and Auditory Integration. In *Encyclopeadia of the human brain* (Vol. 4, pp. 739-752). Elsevier Science.

References

Haji, F. A., Khan, R., Regehr, G., drake, J., de Ribeaupierre, S., & Dubrowski, A. (2015). Measuring cognitive load during simulation based psychomotor skills training: sensitivity of secondary task performance and subjective ratings. *Advances in Health Science Education, 20*, pp. 1237-1253.

Hart, S. (2006). NASA-Task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and ergonomics Society 50th annual meeting*, pp. 904-908.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology, 52*, pp. 139-183.

Hazan, V., Kim, J., & Chen, Y. (2010). Audiovisual perception in adverse conditions: Language, speaker and listener effect. *Speech Communication, 52*, pp. 996-1009.

Heitz, R. P., Schrock, J. C., Payne, T. W., & Engle, R. W. (2008). Effects of incentive on working memory capacity: Behavioral and pupillometric data. *Psychophysiology, 45*, pp. 119-129.

Hermetsberger, P. (Ed.). (2002). *dict.cc. Deutsch-Englisch Wörterbuch*. (dict.cc GmbH) Retrieved 12 10, 2014, from dict.cc. English-German Dictionary: http://www.dict.cc/

Hick, W. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology, 4*(1), pp. 11-26.

Holle, H., Gunter, T., Rüschemeyer, S.-A., Hennenlotter, A., & Iacobini, M. (2008). Neural correlates of the processing of co-speech gestures. *NeuroImage, 39*, pp. 2010-2014.

Hormann, M. (2017, 08 25). *Hörlabor HTW Berlin*. Retrieved from Lärmpegel: http://hearing.htw-berlin.de/wordpress/?page_id=56

Horrey, W., Lesch, M. F., Garabet, A., Simmons, l., & Maikala, R. (2017). Distraction and task engagement: How interesting and boring

information impact driving performance and subjective and physiological responses. *Applied ergonomics, 58*, pp. 342-348.

Hyönä, J., Tommola, J., & Alaja, A.-M. (1995). Pupil Dilation as a Measure of Processing Load in Simultaneous Interpretation and Other Language Tasks. *The Quartely Journal of Experimental Psychology Section A: Human Experimental Psychology, 48*(3), pp. 598-612.

Ikuma, l. H., Harvey, C., Taylor, C. F., & Handal, C. (2014). A guide for assessing control room operator performance using speed and accuracy, perceived workload, situation awreness, and eye tracking. *Journal of Loss Prevention in the Process Industries, 32*, pp. 454-465.

International Organization for Standardization. (1998). Booths for simultaneous interpretation. General characteristics and equipment. *4043*, 13. Vernier, Geneva, Switzerland.

Iverson, P., Bernstein, L., & Auer, E. (1998). Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition. *Speech Communication, 26*, pp. 45-63.

Jacobs, D. (2013). Conceptual Base Found? - Eine Übertragung von Mosers Modell auf das Konsekutivdolmetschen. In D. Andres, B. Martina, & M. Dingfelder-Stone, *Dolmetschmodelle - erfasst, erläutert, erweitert* (pp. 89-104). Frankfurt a. M.: Peter Lang.

Kahnemann, D. (1973). *Attention and Effort.* Englewood Cliffs, New Jersey: Prentice-Hall Inc.

Kalyuga, S. (2012). Instructional benefits of spoken words: A review of cognitive load factors. *Educational Research Review, 7*, pp. 145-159.

Klemen, J., & Chambers, C. (2012). Current perspectives and methods in studying neural mechanisms of multisensory inteactions. *Neuroscience and Biobehavioral Reviews*(36), pp. 111-133.

References

Klinger, J. (2010). Fixation-aligned Pupillary Response Averaging. *ETRA*, (pp. 275-281). Austin.

Klinger, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology, 48*, pp. 323-332.

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Processing Load Induced by Informational Masking Is Related to Linguistic Abilities. *International Journal of Otolaryngology*, pp. 1-11.

Koelewijn, Thomas, Bronkhorst, A., & Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta Psychologica, 134*, pp. 32-384.

Kramer, S., Kapteyn, T., Feesten, J., & Kuik, D. (1997). Assessing Aspects of Auditory Handicap by Means of Pupil Dilation. *Audiology, 36*, pp. 155-164.

Kreiman, J., & Sidtis, D. (2011). *Foundations of Voice Studies. An Interdisciplinary Approach to Voice Production and Perception.* Chichester: Wiley-Blackwell.

Kuchinsky, S. E., Ahlstrom, J. B., Vaden, K. I., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology, 50*, pp. 23-34.

Lambert, S. (1988). Information Processing among Conference Interpreters: A Test of the Depth-of-Processing Hypothesis. *Meta: Journal des traducteurs, 33*(3), pp. 1492-1421.

Laukka, P., Linnman, C., Åhs, F., Pissiota, A., Örjan, F., Faria, V., . . . Furmark, T. (2008). In a Nervous Voice: Acoustic analysis and Perception of Anxiety in Social Phobics' Speech. *Journal of Nonverbal Behaviour, 32*, pp. 195-214.

References

Lavie, N., Hirst, A., de Fockert, J., & Viding, E. (2004). Load Theory of selective Attention and Cognitive Control. *Journal of Experimental Psychology: General, 133*(3), pp. 339-354.

Lecumberri, M. L., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication, 52*, pp. 864-886.

Lee, T.-H. (2002). Ear Voice Span in English into Corean Simultaneous Interpretation. *Meta: journal des traducteurs, 47*(4), pp. 596-606.

Lessiter, J., Freeman, J. K., & Davidoff, J. (2001). A Cross-Media Presence Questionnaire: The ITC-Sense of Presence Inventory. *Presence, 10*(3), pp. 282-297.

Levin, L., Nilsson, S., Ver Hoeve, J., & Wu, S. (Eds.). (2011). *Adler's physiology of the eye* (11th ed.). Edinburgh, London, New York, Oxford, Philadelphia, St Louis, Sydney, Toronto: Sauders Elsevier.

Lewandowski, L., & Kobus, D. (1989). Bimodal Information Processing in Sonor Performance. *Human Performance, 2*(1), pp. 73-74.

Lewandowski, L., & Kobus, D. (1993). The Effects of Redundancy of Bimodal Word Processing. *Human Performance, 6*(3), pp. 229-239.

Li, C. (2010). Coping Strategies for Fast delivery in Simultaneous Interpretation. *The Journal of Specialised Translation, 13*, pp. 19-25.

Lively, S. E., Pisoni, D. B., Summers, V. W., & Bernacki, R. H. (1993). Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences. *Journal of the Acoustical Society of America, 93*(5), pp. 2962-2973.

Lombard, E. (1911). Le signe de l'élévation de la voix. In *Annales des maladies de l'oreille, du larynx, du nez et du pharynx* (Vol. 37:2, pp. 101-119).

References

Los, S., & Van der Burg, E. (2013). Sound Speeds Vision Through Preparation, Not Integration. *Journal of Experimental Psychology: Human Perception and Performance, 39*(6), pp. 1612-1624.

Luque-Casado, A., Perales, J. C., Cárdenas, D., & Sanabria, D. (2016). Heart rate variability and cognitive processing: The autonomic response to task demands. *Biological Psychology, 113*, pp. 83-90.

Maier, J., Di Luca, M., & Noppeney, U. (2011). Audiovisual Asynchrony Detection in Audiovisual Speech. *Journal of Experimental Psychology: Human Perception and Performance, 37*(1), pp. 245-256.

Malmkjær, K. (2011a). Linguistic Approaches to Translation. In K. Malmkjær, & K. Windle, *The Oxford Handbook of Translation* (pp. 57-70). Oxford, New York: Oxford University Press.

Malmkjær, K. (2011b). Meaning and Translation. In K. Malmkjær, & K. Windle, *The Oxford Handbook of Translation* (pp. 108-122). Oxford, New York: Oxford University Press.

Malmkjær, K., & Windle, K. (2011). *The Oxford Handbook of Translation Studies.* Oxford, New York: Oxford University Press.

Massaro, D., & Cohen, M. (1999). Speech perception in perceivers with hearing loss: Synergy of multiple modalities. *Journal of Speech, Language and Hearing Research, 42*(1), pp. 21-41.

Massaro, D., & Light, J. (2004). Using Visible Speech to Train Perception and Production of Individuals with Hearing Loss. *Journal of Speech, Language and Hearing Research, 47*(2), pp. 304-320.

Mattys, S., & Wiget, L. (2011). Effects of cognitive load on speech recognition. *Journal of Memory and Language, 65*, pp. 145-160.

Mattys, S., Brooks, J., & Cook, M. (2009). Recognizing speech under processing load: Dissociating energetic from informational factors. *Cognitive Psychology, 59*, pp. 203-243.

References

Mattys, S., Caroll, L., Li, C., & Chan, S. (2010). Effects of energetic and informational masking on speech segmentation by native and non-native speakers. *Speech communication, 52*, pp. 887-899.

Mazzetti, A. (1999). The influence of segmental and prosodic deviations on source-text comprehension in simulltaneous interpretation. *The Interpreter's Newsletter, 9*, pp. 125-147.

Mc Gurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, pp. 746-748.

McClain, L. (1983). Stimulus-response compatibility affects auditory Stroop interference. *Perception & Psychophysics, 33*(3), pp. 266-270.

McGettigan, C., Faulkner, A., Altarelli, I., Obleser, J., Baverstock, H., & Scott, S. (2012). Speech comprehension aided by multiple modalities: Behavourial and neural interactions. *Neuropsychologia, 50*, pp. 762-776.

Meredith, A., & Stein, B. (1986). Visual, Auditory, and Somatosensory Convergence on Cells in Superior Colliculus Result in Multisensory Integration. *Journal of Neurophysiology, 56*(3).

Meredith, A., Nemitz, J., & Stein, B. (1987). Determinants of Multisensory Integration in the Superior Colliculus Neurons. I. Temporal Factors. *Journal of Neuroscience, 7*(10), pp. 3215-3229.

Miller, G. (1956). The Magical Number seven, plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review, 63*, pp. 81-97.

Miller, J. (1982). Divided Attention: Evidence for Coactivation with Redundant Signals. *Cognitive Psychology, 14*, pp. 247-269.

Mirman, D. (2014). *Growth Curve Analysis and Visualization Using R.* Boca Raton, London, New York: CRC Press.

Molholm, S., Sehatpour, P., Mehta, A., Shpaner, M., Gomez-Ramirez, M., Ortigue, S., . . . Foxe, J. (2006). Audio-visual Multisensory Integration in Superior Parietal Lobule Revealed by Human Intercranial Recordings. *Journal of Neurophysiology, 96*, pp. 721-26.

Moreno, R., & Mayer, R. (2002). Verbal Redundancy in Multimedia Learning: When Reading Helps Listening. *Journal of Educational Psychology, 94*(1), pp. 156-163.

Moser, B. (1978). Simultaneous Interpretation: A Hypothetical Model and its Pratical Application. In D. Gerver, & H. Sinaiko, *Language Interpretation and Communication* (Vol. 6, pp. 353-368). Bosten, MA: Springer.

Moser-Mercer, B. (2003). *Remote interpreting: assessment of human factors and performance parameters.* Retrieved 2 8, 2014, from aiic: http://aiic.net/page/1125/remote-interpreting-assessment-of-human-factors-and-performance-parameters/lang/1

Moser-Mercer, B. (2005a). Remote-Interpreting: Issues of Multisensory Integration in a Multilingual Task. *Meta: Journal des Traducteurs, 50*(2), pp. 727-738.

Moser-Mercer, B. (2005b). Remote interpreting. The crucial role of presence. *Bulletin VALS-ASLA, 81*, pp. 73-97.

Mousavi, S., & Low, R. S. (1995). Reducing Cognitive Load by Using Auditory and Visual Presentation Modes. *Journal of Educational Psychology, 87*(2), pp. 319-334.

Mouzourakis, P. (2006). Remote interpreting. A technical perspective on recent experiments. *Interpreting, 8*(1), pp. 45-66.

Müller, C., Barbara, G.-H., Jameson, A., Rummer, R., & Wittig, F. (2001). Recognizing Time Pressure and Cognitive Load on the Basis of Speech: En Experimental Study. In J. Vassileva, P. Gmytrasiewicz,

References

& M. Bauer, *UM 2001, User Modeling: Proceedings of the Eigth International Conference.* Berlin: Springer.

Munhall, K., Kroos, C., Jozan, G., & Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perception & Psychophysics, 66*(4), pp. 574-583.

NASA Task Load Index (NASA-TLX). (2016, 05 07). Retrieved from Human Performance Repository: https://www.eurocontrol.int/ehp/?q=node/1583

Noesselt, T., Rieger, J., Schoenfeld, M. A., Kanowski, M., Hinrichs, H., Heinze, H.-J., & Driver, J. (2007, Oktober). Audiovisual Temporal Correspondance Modulates Human Multisensory Superior Temporal Sulcus Plus Primary Sensory Cortices. *The Journal of Neuroscience, 27*(42), pp. 11431-11441.

Oster, K. (2017). The influence of self-monitoring on the number of cognates in translations. In S. Hansen-Schirra, O. Czulo, B. Meyer, & S. Hoffemann, *Empirical modelling of translation and interpreting.* Berlin: Language and Science Presse.

Ostrand, R., Blumenstein, S., & Morgan, J. (2011). When Hearing Lips and Seeing Voices Becomes Perceiving Speech: Auditory-Visual Integration in Lexical Acces. *Proceedings of the Annual Meeting of the Cognitive Science Society, 33*, pp. 1376-1381.

Paré, M., Richler, R., & Ten Hove, M. (2003). Gaze behaviour in audiovisual speech perception: the influence of ocular fixations on the McGurk effect. *Perception and Psychophysics, 65*(4), pp. 553-567.

Peirce, J. W. (2007). Psychopy - Psychophysics software in Python. *Journal of Neuroscientific Methods, 162*(1-2), pp. 8-13.

Peters, M. L., Godaert, G. L., Ballieux, R. E., van Vliet, M., Willemsen, J. J., Sweep, . . . Heijnen, C. J. (1998). Cardiovascular and endocrine

responses to experimental stress: effects of mental effort and controllability. *Psychoneuroendocrinology, 23*(1), pp. 1-17.

Peyasakhovich, V., Dehais, F., & Causse, M. (2015). Pupil diameter as a measure of cognitive load during auditory-visual interference in a simple piloting task. *Procedia Manufacturing, 3*, pp. 5199-5205.

Pöchhacker, F. (1995). Simultaneous Interpreting: A Functionalist Perspective. *Journal of Linguistics, 14*, pp. 31-53.

Pöchhacker, F. (1999). Situative Zusammenhänge. In M. Snell-Hornby, H. G. Hönig, P. Kußmaul, & P. A. Schmitt, *Handbuch Translation* (2. ed.). Tübingen: Stauffenburg.

Pöchhacker, F. (2011). Consecutive Interpreting. In K. Malmkjær, & K. Windle, *The Oxford Handbook of Translation* (pp. 294-306). Oxford, New York: Oxford University Press.

Pöchhacker, F. (2011). Simultaneous Interpreting. In K. Malmkjær, & K. Windle, *The Oxford Handbook of Translation* (S. 275-293). Oxford, New York: Oxford University Press.

Pöchhacker, F., & Shlesinger, M. (2002). *The interpreting studies reader.* New York: Routledge.

Poyatos, F. (1984). The Multichannel Reality of Discourse: Language-Paralanguage Kinesics and the Totality of the Communicative System. *Language Sciences, 6*(2), pp. 307-337.

Poyatos, F. (1987). Nonverbal Communication in Simultaneous and Consecutive Interpretation: A Theoretical Model and New Perspectives. *TEXTconTEXT, 2*(2/3), pp. 73-108.

Poyatos, F. (1997). *Nonverbal Communication and Translation. New Perspectives and Challenges in Literature, Interpretation and the Media.* Amsterdam/Philadelphia: John Benjamin's.

Poyatos, F. (1997). The reality of multichannel verbal-nonverbal communication in simultaneous and consecutive interpretion. In F.

Poyatos, *Nonverbal Communication and Translation. New Perspectives and Challenges in Literature, Interpretation and the Media* (pp. 249-282). Amsterdam/Philadelphia: John Benjamin's.

Putzar, L., Goerendt, I., Heed, T., Richard, G., Büchel, C., & Röder, B. (2010). The neural basis of lip-reading capabilities is altered by early visual deprivation. *Neuropsychologia, 48*, pp. 2158-2166.

R Core Team. (2016). *R: A language and environment for statistical computing.* Retrieved from R Foundation for Statistical Computing: https://www.R-project.org

Rackow, J. (2013). Dolmetschen als Kommunikation - verbale und nonverbale Informationsverarbeitung im Dolmetschprozess. In D. Andres, M. Behr, & M. Dingfelder Stone, *Dolmetschmodelle - erfasst, erläutert, erweitert* (pp. 129-152). Frankfurt a. M.: Peter Lang.

Reid, G. B., & Nygren, T. E. (1988). The subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload. *Human Mental Workload, 52*, pp. 185-218.

Reinerman-Jones, L., Metthews, G., & Mercade, J. E. (2016). Detection tasks in nuclear power plan operation: Vigilence decrement and physiological workload monitoring. *Safety Science, 88*, pp. 97-107.

Reiß, K., & Vermeer, H. J. (1991). *Grundlegung einer allgemeinen Translationstheorie* (2nd ed.). Tübingen: Niemeyer.

Renaud, P., & Blondin, J.-P. (1997). The stress of Stroop performance: physiological and emotional responses to color-word interference, task pacing, and pacing speed. *International Journal of Psychophysiology, 27*, pp. 87-97.

Rennert, S. (2008). Visual Input in Simultaneous Interpreting. *Meta: Journal des Traducteurs, 53*(1), pp. 204-217.

References

Riccardi, A. (2002). Translation and Interpretation. In A. Riccardi, *Translation Studies. Prspectives on an Emerging Discipline* (pp. 75-91). Cambridge: Cambridge University Press.

Riccardi, A. (2002). *Translation Studies. Perspectives on an Emerging Discipline.* Cambridge: Cambridge University Press.

Riccardi, A., Marinuzzi, G., & Zecchin, S. (1998). Interpretation and stress. *The Interpreters' Newsletter, 8*, pp. 93-106.

Riva, G., & Mantovani, F. (2014). Extending the Self through the Tools and the Others: a General Framework for Presence and Social Presence in Mediated Interactions. In G. Riva, J. Waterworth, & D. Murray, *Interacting with Presence: HCI and the Sense of Presence in Computer-mediated Environments* (pp. 9-31). De Gruyter Open.

Rosenblum, L. (2008). Speech Perception as a Multimodal Phenomenon. *Current Directions in Psychological Science, 17*(6), pp. 405-409.

Ross, L., Saint-Amour, D., Leavitt, V., Molholm, S., Javitt, D., & Foxe, J. (2007). Impaired multisensory processing in schizophrenia: Deficits in the visual enhancement of speech processing in noisy environmental conditions. *Schizophrenia Research, 97*, pp. 173-183.

Roziner, I., & Shlesinger, M. (2010). Much ado about something remote. *Interpreting, 12*(2), pp. 214-247.

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods. *Applied Psychology: An International Review, 53*(1), pp. 61-86.

Ryu, K., & Myung, R. (2005). Evaluation of mental work load with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics, 35*, pp. 991-1009.

References

Sabatini, E. (2000). Listening comprehension, shadowing and simultaneous interpretation of two ‘non-standard’ English speeches. *Interpreting, 5*(1), S. 25-48.

Schaeffer, M., & Carl, M. (2014). Measuring the Cognitive Effort of Literal Translation Processes. *Workshop on Humans and Computer-assisted Translation* (pp. 29-37). Gothenburg, Sweden: Association for Computational Linguistics.

Scherer, K. (1989). Vocal Correlates of Emotional Arousal and Affective Disturbance. In H. L. Wagner, & A. S. Manstead, *Handbook of social Psychophysiology* (pp. 165-197). New York: Wiley.

Schiffmann, H. R. (1996). *Sensation and Perception. An Integrated Approach* (4 ed.). New York, Chichester, Brisbane, Toronto, Singapore: John Wiley and Sons.

Schilling, H. E., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: word frequency effects and individual differences. *Memory & Cognition, 26*(6), pp. 1270-1281.

Schmidt, R., Lang, F., & Heckmann, M. (2010). *Physiologie des Menschen mit Pathopyhsiologie* (31. Auflage ed.). Heidelberg: Springer Verlag.

Schnotz, W., & Kürschner, C. (2007). A Reconsideration of Cognitive Load Theory. *Education and Psychological Review, 19*, pp. 469-508.

Seeber, K. (2011). Cognitive load in Simultaneous Interpreting. Existing theories - new models. *Interpreting, 13*(2), pp. 176-204.

Seeber, K. (2012). Multimodal Input in Simultaneous Interpreting: An Eyetracker Experiment. In L. N. Zybatow, *Translationswissenschaft: Alte und neue Arten der Translation in Theorie und Praxis, Translation Studies: Old and New Types of Translation in Theory and Practice* (pp. 341-347). Frankfurt am Main: Peter Lang.

References

Seeber, K. (2012). Multimodal input in simultaneous interpreting: An eye-tracking experiment. *Translata. Conference proceedings*, pp. 341-347.

Seeber, K. (2015). Cognitive load in simultaneous interpreting. Measures and Methods. In M. Ehrensberger-Dow, S. Göpferich, & S. O'Brian, *Interdisciplinarity in Translation and Interpreting Process Research* (pp. 19-34). Amsterdam/Philadelphia: John Benjamin's.

Seeber, K., & Kerzel, D. (2012). Cognitive Load in Simultaneous Interpreting - Model meets data. *International Journal of Bilingualism, 16*(2), pp. 228-242.

Sengpiel, E. (2017, 08 25). *sengspielaudio.com*. Retrieved from Decibel table - SPL - sound pressure level chart: http://www.sengpielaudio.com/TableOfSoundPressureLevels.htm

Setton, R. (1999). *Simultaneous Interpretation: a cognitive-pragmatic analysis.* Amsterdam, Philadelphia: John Benjamin's.

Seubert, S. (2017). Simultaneous Interpreting is a Whole Person Process. In M. Behr, & S. Seubert, *Education is a Whole-Person Process. Von ganzheitlicher Lehre, Dolmetschforschung und anderen Dingen* (pp. 271-303). Berlin: Frank Timme.

Shahin, A., Kerlin, J., Bhat, J., & Miller, L. (2012). Neural restoration of degraded audiovisual speech. *NeuroImage, 60*, pp. 530-538.

Simons, D., & Chabris, C. (1999). Gorilla in our midst: sustained inattentional blindness for dynamic events. *Perception, 28*, pp. 1059-1074.

Somers, H. (1996). *Terminology, LSP and Translation. Studies in language engineering in honour of Juan C. Sager.* Amsterdam/Philadelphia: John Benjamin's.

Spehar, B., Goebel, S., & Tye-Murray, N. (2015). Effects of Context Type on Lipreading and Listening Performance and Implications for

Sentence Processing. *Journal of Speech, Language and Hearing Research, 58*, pp. 1092-1102.

Starreveld, P., de Groot, A., & Rossmark, B. v. (2014). Parallel language activitation during word processing in bilinguals: Evidence from word production in sentence context. *Bilingualism: Language and Cognition, 17*(2), pp. 258-276.

Stenzl, C. (1983). Simultaneous Interpretation: Groundwork towards a Comprehensive Model. Unpublished M.A. thesis, University of London.

Stevenson, R., Zemtov, R., & Wallace, M. (2012). Individual Differences in the Multisensory Temporal Binding Window Predict Susceptibility to Audiovisual Illusions. *Journal of Experimental Psychology: Human Perception and Performance, 38*(6), pp. 1517-1529.

Streeter, L., Macdonald, N., Apple, W., Krauss, R., & Galotti, K. (1983). Acoustic and perceptual indicators of emotional stress. *Journal of the Acoustical Society of America, 73*(4), pp. 1354-1360.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psycholgy, 18*, pp. 643-662.

Subjective Workload Assessment Technique (SWAT). (2016, 07 05). Retrieved from Human Performance Repository: https://www.eurocontrol.int/ehp/?q=node/1588

Sumby, W., & Pollack, I. (1953). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, pp. 212-215.

Sweller, J. (2010). Element Interactivity and Intrinsic, Extraneous and Germane Cognitive Load. *Educational Psychology Review, 22*, pp. 123-138.

Teder-Sälejärvi, W., McDonald, J., Di Russo, F., & Hillyard, S. (2002). An analysis of audio-vidual crossmodal integration by means of event-

related potention (ERP) recording. *Cognitive Brain Research, 14*, pp. 106-114.

Thomas, S., & Jordan, T. (2004). Contributions of Oral and Extraoral Facial Movement to Visual and Audiovisual Speech Perception. *Journal of Experimental Psychology: Human Perception and Performance, 30*(5), pp. 873-888.

Timarová, Š. (2008). Working Memory and Simultaneous Interpreting. In P. Boulogne (Ed.), *Translation and Its Others. selected Papers of the CETRA Research Seminar in Translation Studies 2007*, (pp. 1-28). Leuven.

Timarovà, S., Dragsted, B., & Hansen, I. G. (2011). Time lag in translation and interpreting. In C. Alvstad, A. Hild, & E. Tiselius, *Methods and Strategies of Process Research: Integrative approaches in Translation Studies* (pp. 121-148). Amsterdam, Philadelphia: John Benjamnins.

Tobii support. (2016, 07 20). Personal communication.

Tobii Technology, A. (2010). *Tobii Eye Tracking. An introduction to eye tracking and Tobii Eye Trackers.* Retrieved from https://de.scribd.com/document/26050181/Introduction-to-Eye-Tracking-and-Tobii-Eye-Trackers

Trepel, M. (2012). *Neuroanatomie. Struktur und Funktion.* München: Urban & Fischer.

Ungerleider, L., & Pessoa, L. (2008). What and where pathways. *Scholarpedia, 3(11)*, 5342.

van der Loo, M. (2014). The stringdist-package for approximate string matching. *R Journal, 6*(1), pp. 111-122.

van der Loo, M. J. (2014). The stringdist package for approximate string matching. *R journal, 6*(1), pp. 111-122.

van der Meer, E., Beyer, R., Horn, J., Foth, M., Bornemann, B., Ries, J., . . . Wartenburger, I. (2010). Resource allocation and fluid intelligence: Insights from pupillometry. *Psychophysiology, 47*, pp. 158-169.

van Gerven, P., Paas, F., van Merrienbroer, J., & Schmidt, H. (2004). Memory load and the cognitive pupillary response in aging. *Psychophysiology, 41*, pp. 167-174.

van Merrienboer, J., & Sweller, J. (2005). Cognitive Load Theory and Complex Learning: Recent Developments and Future Directions. *Educational Psychology Research, 17*(2), pp. 147-178.

Van Summers, W., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., & Stokes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *Journal of the Accoustical Society of America, 84*(3), pp. 917-928.

van Wassenhove, V., Grant, K., Poeppel, D., & Halle, M. (2005). Visual Speech Speeds Up the Neural Processing of Auditory Speech. *Proceedings of the National Academy of Sciences of the United States, 4*, pp. 1181-1186.

Venables, W. B. (2002). *Modern Applied Statistics with S.* (Forth edition ed.). New York: Springer.

Verband der Konferenzdolmetscher, (. (2017, 07 21). *VKD: Verband der Konferenzdolmetscher imm BDÜ.* Retrieved 2 2014, 8, from Planung einer Veranstaltung mit Dolmetschern: http://vkd.bdue.de/fileadmin/verbaende/vkd/Dateien/PDF-Dateien/VKD-Infoblatt_fuer_Veranstalter.pdf

Vermeer, H. J. (1992). *Skopos und Translationsauftrag - Aufsätze* (3rd ed.). Frankfurt a.M.: Verlag für Interkulturelle Kommunikation.

Viaggio, S. (1997). Kinesics and the simultaneous interpreter. The advantages of listening with one's eyes and speaking with one's body. In F. Poyatos, *Nonverbal Communication and Translation. New Perspectives and Challenges in literature, Interpretation and*

the Media (pp. 283-293). Amsterdam/Philadelphia: John Benjamin's.

von Kriegstein, K., Dogan, Ö., Grüter, M., Giraud, A.-L., Kell, C., Grüter, T., . . . Kiebel, S. (2008). Simulation of Talking Faces in the Human brain Improves Auditory Speech Recognition. *Proceedings of the National Academy of Sciences of the United States of America, 105*(18), pp. 6747-6752.

Vroomen, J., & Stekelenburg, J. (2010). Visual Anticipatory Information Modulates Multisensory Interactions of Artifical Audiovisual Stimuli. *Journal of Cognitive Neuroscience, 22*(7), pp. 1583-1596.

Vroomen, J., & Stekelenburg, J. (2011). Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition, 118*(1), pp. 75-83.

Ward, J. (2010). *The Student's Guide to Cognitive Neuroscience, 2nd edition.* Hove, New York: Psychology Press.

Ward, J. (2010). *The Student's Guide to Cognitive Neuroscience, 2nd edition.* Hove, New York: Psychology Press.

Warren, R. (1999). *Auditory Perception. A New Analysis and Synthesis.* Cambrigde, New York, Melbourne: Cambridge University Press.

Wickens, C. (2009). Multiple Resources and Mental Workload. *Human Factors, 50*(3), pp. 449-455.

Wickens, C., Kramer, A., Vanasse, L., & Donchin, E. (1983). Performance of concurrent tasks: a psychophysiological analysis or the reciprocity of information-processing resources. *Science, 221*(4615), pp. 1080-1082.

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

Willems, R., Özyürek, A., & Hagoort, P. (2009). Differential roles for left inferior frontal and superior temporal cortex in multimodal

integration of action and language. *NeuroImage, 47*, pp. 1992-2004.

Woehrle, J., & Magliano, J. P. (2012). Time flies faster if a person has a high working-memory capacity. *Acta Psychologica, 139*, pp. 314-319.

Yagi, S. M. (2000). Studying Style in Simultaneous Interpretation. *Meta: journal des traducteurs / Meta: Translator's Journal, 45*(3), pp. 520-547.

Yap, T. F., Epps, J., Ambikairajah, E., & Choi, E. H. (2015). Voice source under cognitive load: Effects and classification. *Speech Communication, 72*, pp. 74-95.

Yee, T. W. (2010). The VGAM package for Categorical Data Analysis. *Journal of Statistical Software., 29*(6), pp. 1-34.

Yee, T. W. (2015). *Vector generalized Additive Models: With an Implementation in R.* New York, USA: Springer.

Yee, T. W. (2017). *VGAM: Vector Generalized Linear and Additive Models. R package version 1.0-3.* Retrieved from https://CRAN.R-project.org/package=VGAM

Zeileis, A., & Grothendieck, G. (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. *Journal of Statistical Software, 14*(6), pp. 1-27.

Zollinger, A., & Brumm, H. (2011). The Lombard effect. *Current Biology, 21*(16), pp. R615-R615.

7 Appendix

7.1 Pretest stimuli

The table shows i) the English stimuli words used for the pretest, ii) the most frequent German translation, iii) the cognate status based on the phonetic form, iv) how often the most frequent translations was suggested in the dict.cc community and v) how often the second translation was suggested by the dict.cc community.

English	German	Cognate status	most frequent translation	second translation
bread	Brot	cognate	32702	21
book	Buch	cognate	5572	83
rib	Rippe	cognate	2988	24
owl	Eule	cognate	854	10
bank	Bank	cognate	1156	683
bar	Bar	cognate	294	581
beard	Bart	cognate	4405	16
bed	Bett	cognate	2298	67
beer	Bier	cognate	30795	15
blood	Blut	cognate	1726	8
boat	Boot	cognate	2147	165
bomb	Bombe	cognate	908	11
glass	Glas	cognate	3683	1

hand	Hand	cognate	4792	75
heart	Herz	cognate	3904	2653
house	Haus	cognate	4997	62
ice	Eis	cognate	664	10
lip	Lippe	cognate	3611	40
milk	Milch	cognate	30070	0
mouse	Maus	cognate	1475	17
neck	Nacken	cognate	537	4665
nose	Nase	cognate	4995	53
nut	Nuss	cognate	663	903
rice	Reis	cognate	32147	1
door	Tür	cognate	1845	29
mouth	Mund	cognate	4934	750
arm	Arm	cognate	4812	80
ball	Ball	cognate	1017	355
ring	Ring	cognate	1157	56
wool	Wolle	cognate	729	1
goose	Gans	cognate	1145	1
brush	Bürste	cognate	994	441
ear	Ohr	cognate	4270	44

egg	Ei	cognate	1886	7
bridge	Brücke	cognate	1239	44
bench	Bank	non-cognate	1136	153
bike	Fahrrad	non-cognate	526	93
duck	Ente	non-cognate	990	31
bill	Rechnung	non-cognate	32136	369
glove	Handschuh	non-cognate	1420	1
belt	Gürtel	non-cognate	2015	177
goat	Ziege	non-cognate	2594	1
dog	Hund	non-cognate	2711	15
bell	Glocke	non-cognate	472	293
knife	Messer	non-cognate	4977	1
root	Wurzel	non-cognate	943	167
gate	Tor	non-cognate	3990	540
hill	Hügel	non-cognate	1379	107
blade	Klinge	non-cognate	2587	139
rope	Seil	non-cognate	1445	240
wheel	Rad	non-cognate	2087	167
gun	Pistole	non-cognate	2854	2686
dust	Staub	non-cognate	1925	43

bowl	Schüssel	non-cognate	4463	2901
doll	Puppe	non-cognate	958	36
brain	Gehirn	non-cognate	3803	86
girl	MädBhen	non-cognate	28569	29
desk	Schreibtisch	non-cognate	2291	231
wall	Wand *Wall: false cognate	non-cognate	2524	596
dress	Kleid	non-cognate	2079	144
roof	Dach	non-cognate	3284	5
boy	Junge	non-cognate	28550	42
leg	Bein	non-cognate	5413	217
ham	Schinken	non-cognate	5742	20
road	Straße	non-cognate	346	767
head	Kopf	non-cognate	4838	361
beach	Strand	non-cognate	2839	50
horse	Pferd	non-cognate	1858	152
duck	Ente	non-cognate	990	31
lunch	Mittagessen	non-cognate	3595	30
meal	Mahlzeit	non-cognate	1242	509
wrist	Handgelenk	non-cognate	5708	1

7.2 Speeches

7.2.1 Speech “Greece”

Ladies and Gentlemen,

It is a real pleasure for me to be here at the international conference for economic development. I’m very honored to speak to so many distinguished guests who promote the economic development in their respective countries. Today, I want to talk to you about the economic situation in Greece.

Since a couple of years now, Greece has been in a crisis situation. The economy is in a terrible state. There has been a major banking crisis. Many people have lost their jobs. They have huge loans to repay. There are no foreign investments at the moment. There is no profit being made. People have almost no hope at all. They are not doing very well, and it does not look as if much will change in the very near future. Indeed, the future does not look very bright.

However, Greece has some assets. For example, it produces some rather good products, like oil. Everybody knows this, and many people buy Greek oil. Even though the oil is very good, Greek oil has a very small market share. The Greek producers do not sell many of their oil bottles. How does this happen? What is going wrong?

There are mainly two reasons. First, the market for oil is dominated by Italy and by Spain. Why is that? The agricultural market in Greece is well behind. For example, the Greek market is disorganized and there is no quality label. Spain and Italy introduced several years ago a quality label for oil. Bottles with this label satisfy standard quality criteria. Buyers prefer these products because they can trust them.

Second, the presentation of the bottle counts as well. Often you will find that the Greek cheese or oil is not well presented. The bottle is a bit

scratched, the packaging is damaged, or the label sticks on badly. The product doesn't look good and so consumers don't buy it.

So, these are two points where Greece could boost its production. But would this really help? I do not think so. These proposals would essentially hel*p* farmers. But there are only very few people who work as farmers. Most Greeks work in the service sector. They work as hairdressers, waiters, or secretaries. What would hel*p* them?

Maybe, we should remember one thing: Greece has a beautiful coast and it is one of the sunniest places in Europe. A perfect destination for tourists. Unfortunately, so far, Greece has been focusing on tourists with little money. These are mostly young people who travel throughout Europe. They stay only a day or two at the Greek beaches and spend the night in a youth hostel. Of course, they do not bring much money into the country.

Instead, Greece should focus on rich tourists and adapt its services to their needs. Hotels could offer extra services like car rental car or boat hire. Tourism agencies could organize guided tours to the historical sites in Greece or offer special events like Greek theater or boat tours to the many islands of Greece. This way, tourists would stay longer and enjoy the fine weather and the beach, and would spend their money in Greece. This, in turn, would hel*p* the economic development to pick up. But, of course, this will not be possible without any investments. I think Europe will have to support Greece till investors trust Greece again. This is the only way to drain investments and to support the economic development in the country. Thank you.

7.2.2 Speech "demographic change"

Ladies and Gentlemen,

It is a real pleasure for me to be here at the international conference for demography and demographic challenges. I'm very honored to speak to so many distinguished guests who deal with these challenges everywhere

in the world. They are an example for us all. Today, I want to talk about the demographic shift and its impact on pensions.

We are all facing a big demographic shift, particularly in the western countries. We are feeling the effects of this shift everywhere and ever more as an ever higher number of people are reaching retirement age. You may wonder why so many people are reaching retirement age. Just after the Second World War, the baby boomers were born. It was a time when Europe was in ruins and young people needed to rebuild our continent. The baby boomers could also finance their parents because they were far more in number than the previous generation. But now the situation has changed, and it is exactly the opposite. The baby-boomers are more numerous than the young generation. In addition, we live longer, because we have a healthier life style, much better medical care, and a better diet. On to*p* of that, the birth rate has dropped significantly over the last decades. So, things look very different. More and more people retire. Their number is increasing every year. But there are fewer and fewer younger people who work to support them.

Is this all negative? Yes and no. We do live longer because we have a better life style. But it also means that we work longer. When we reach retirement age, many are still in good health. There is no particular need to retire. Of course, some people do hard labor and will be very glad to sto*p* when they have reached retirement age. But others have to retire, while they find themselves perfectly able to continue. In fact, in most countries retirement age is being increased. This has essentially financial and budgetary reasons. Many people are happy to stay at work for numerous reasons. First, we would be bored if we had to stay at home. Second, we have gained a lot of experience. Third, people expected to receive a given pension, but they receive a lot less than they have been promised. In fact, the pensions have been reduced in many countries because of the economic crisis.

I think that an arbitrary retirement age is really out of date now because people age differently. Some people will need hel*p* in their daily life very soon. Others are still perfectly able to work or to live on their own. Also, pensions cost the state a lot of money. We all know that. It is a big financial burden and it is one of the reasons why states have been increasing retirement age over the last years. So, why should people not continue to work, when they reach retirement age?

Some people say that those older people will be taking the jobs of young people. But this is wrong. At retirement age, most people have gained a lot of experience. They can pass on their experience. So, the younger people can benefit from their experience. Older people can continue to work and will not feel useless. Today, many people are working much over retirement age. And these people are quite happy to do so. They do not necessarily need to work for large sums of money or to work fulltime. They are just happy to kee*p* active. Thank you.

7.2.3 Speech “air travel”

Ladies and Gentlemen,

It is a real pleasure for me to be here at the international conference for passenger rights in air travel. I’m very honored to speak to so many distinguished guests who fight for passengers rights everywhere in the world and who are an example to all of us. Today, I want to talk about a new policy for overweight passengers in planes.

Many people are annoyed if they take the plane and have to sit next to a large person who builds over to their seat. In order to solve this problem, air companies have recently decided to implement a new policy. Their plan is as follows: as long as a passenger can’t fit comfortably into a normal standard seat on an airplane, he or she will have to buy two seats. That would be the rule.

Of course, this plan has caused a certain amount of uproar, as you may imagine. Many people have criticized this decision, others have approved

of it. The critics say that this policy discriminates against obese people. Therefore, it should not be allowed. After all, obesity is an illness. It is no-ones fault, but something we should at least be a little sorry about.

But the air companies say that they have received countless complaints from passengers who have paid good money to have a seat and to fly in comfort. When they sit next to an obese person, these passengers find themselves crushed during their tri*p* just because the person in the next seat has eaten too many cheeseburgers. In addition, the fuel consumption is directly linked to the weight of the passengers. The heavier the aircraft, the more fuel the plane needs. Since overweight people increase the weight of the aircraft, they should pay more for the flight.

Maybe, I should note at this point that obesity is a very widespread disease in many countries of the western world. This is the case in the U.S., in the UK, or in Germany, where many important air companies are based. Overweight people make up a large share of their clients. If the air companies really introduce the overweight policy, they will lose those clients. If this happens, they will sell less flight tickets and their profit will go down. So, it would be better for them not to realize their plan.

Furthermore, overweight people are not the only ones who have those problems. There are many more examples. Just think of tall people who take the plane and who stretch out their legs to their neighbors because there is not enough room in front of their own seat. Some passengers read newspapers in the plane and use the space of nearly two seats. Finally, there are parents who travel with their young children on their lap. They, too, need a lot more space, especially when their children are excited.

In my opinion, an extra charge for overweight people is not the solution. Air companies should better adapt the seat size to the needs of their passengers and offer larger seats with more room for legs. This could also be a competitive edge for the air companies because tall or overweight passengers would prefer airlines which offer extra-large seats. Sure, air companies could transport only a smaller number of passengers. But

those tall or overweight passengers, who chose for an extra-large seat, would be willing to pay a bit more for their flight. In this way, the air company could compensate for the higher fuel costs, and everybody would fly in comfort. Thank you.

7.2.4 Speech "work"

Ladies and gentlemen,

It is a real pleasure for me to be here at the international conference of work and health. I am very honored to speak to so many distinguished guests who fight for better working conditions everywhere in the world. They are an example for us all. Today, I want to talk about one of the main risk factors of health at work which is stress.

Over the last decades, our working environment has completely changed. We live now in a globalized world. As more and more players become part of the global economy, western companies struggle to preserve their competitiveness. The work load has increased and most employees feel that they have far too much work. The increasing work load is a great problem that affects the quality of the work and the employee's health.

The impact is tremendous. More and more employees suffer from depression. They feel exhausted and tired. They lose any interest in their work. Sometimes, a treatment can hel*p* them to cope with the stress. But sometimes, it is too late. Every year, a large number of people commit suicide because they cannot stand it any longer. We cannot continue to ignore this fact. Every company and every employee must understand that stress management is the key to productivity at work and a healthy life.

Today, I want to present some solutions and initiatives to this problem. The first ste*p* is time management. Several companies offer their employees workshops on time management. These workshops focus especially on skills like planning and prioritizing. The participants learn how to make to-do lists and how to arrange enough time for every task. But there is something even more important. They learn how to set

priorities and how to focus on them. Some approaches may seem a little drastic. Some recommend, for example, shutting down all e-mailing-programs or communication devices because they interrupt the work flow too often. But the most drastic measures are also the most efficient.

I will give you an example. A British company observed that the masses of e-mails decreased the productivity of its staff. Too, the employees complained that they could not concentrate on their tasks anymore because they constantly had to answer an e-mail. So the company decided to launch e-mailing-programs only after lunch. Since then, the productivity has markedly increased. In addition, employees experience less stress and feel more motivated at work.

The second ste*p* is stress management. Stress management workshops typically include different relaxation techniques and conflict solving strategies. Participants learn what helps them best and what they can do in stressful situations.

Consider teachers, for instance. They are under huge pressure when they have to teach very unmotivated, sometimes even naughty students. In order to support the teaching staff, schools organize weekly meetings. During these meetings, the teachers discuss the most difficult cases in a group. Together, they develo*p* strategies to manage those students. Difficult students are no longer a problem for one person, but a challenge for the whole team. These weekly meetings have reinforced cooperation and solidarity among the team and the number of sick leaves has diminished.

I find these two examples very encouraging. IN both cases, the measures are simple but effective. They clearly show the positive impact of stress management on health and productivity. I hope that we will see more inspiring examples in the future.

7.3 Text-related questions

Correct answers are printed in bold type. The English translation is given in italic typ.

7.3.1 Speech “Greece”

1. Griechenland erlebt zurzeit eine Krise. Wie drückt sich die Krise aus?
Greece is currently experiencing a crisis. How does the crisis manifest?
a. Arbeitslosigkeit, Zurückstellen von wichtigen Investitionen, Massenauswanderung
Unemployment, deferral of important investments, mass emigration
b. Arbeitslosigkeit, Streichung von Sozialleistungen, Aufbegehren der Bevölkerung
Unemployment, cuts to social benefits, uprisings
c. Arbeitslosigkeit, Zusammenbruch der Banken, Hoffnungslosigkeit bei der Bevölkerung
Unemployment, collapse of banks, despair among the population
d. Ich weiß es nicht.
I don't know.

2. Welche Produkte und Dienstleistungen könnte Griechenland vermarkten? *What products and services could Greece market?*
a. Wein, Feigen, Kulturangebote
Wine, figs, cultural activities
b. Olivenöl, Schafskäse, gehobener Tourismus
Olive oil, sheep's cheese, upscale tourism
c. gefüllte Weinblätter, Thunfisch, Kreuzschifffahrten
Stuffed vine leaves, tuna, cruises
d. Ich weiß es nicht.
I don't know.

3. Woran scheitert die Vermarktung griechischen Öls?

Why is Greek oil not selling well?

a. Die Ölflaschen sind in einem schlechten Zustand.

The oil bottles are in poor condition.

b. Die Qualität des Öls ist unzureichend.

The quality of the oil is inadequate.

c. Es wird zu wenig Olivenöl produziert.

Too little olive oil is being produced.

d. Ich weiß es nicht.

I don't know.

4. Wie könnte Griechenland seine Tourismusangebote ausbauen?

How could Greece develop its tourism industry?

a. Mit Wanderausflügen und Weinproben.

Hiking trips and wine tastings.

b. Mit Besichtigungen antiker Stätten und Bootsausflügen.

Tours of antique towns and boat trips.

c. Mit Stadtbesichtigungen und Vermietung von Liegestühlen am Strand.

City tours and beach recliner rentals.

d. Ich weiß es nicht.

I don't know.

5. Was ist ausserdem nötig, damit die Krise in Griechenland beendet wird?

What else does Greece need to end the crisis?

a. Die Hilfe der EU und ausländische Investitionen.

The help of the EU and foreign investors.

b. Die Hilfe der EU und des IWF.

The help of the EU and the IMF.

c. Die Hilfe der EU und höhere Steuern.

The help of the EU and higher taxes.

d. Ich weiß es nicht.

I don't know.

7.3.2 Speech “demographic change”

1. Welche Folgen hat der demografische Wandel?

What are the consequences of demographic change?

a. Die Menschen arbeiten immer länger.

The population has to work longer.

b. Es müssen immer weniger Berufstätige für immer mehr Rentner aufkommen.

Fewer professionals have to pay for more and more pensioners.

c. Die Wettbewerbsfähigkeit der betroffenen Länder sinkt.

The respective countries' ability to compete decreases.

d. Ich weiß es nicht.

I don't know.

2. Die Lebenserwartung ist gestiegen. Woran liegt das?

Life expectancy has increased. Why?

a. an einem gesünderem Lebensstil mit mehr körperlicher Bewegung.

Healthier lifestyles and more exercise.

b. an dem intensiveren beruflichen und ehrenamtlichen Engagement.

Increased professional and voluntary commitments.

c. an der besseren medizinischen Versorgung und an der gesünderen Ernährung.

Improved medical care and healthier diets.

d. Ich weiß es nicht.

I don't know.

3. Welche Massnahme wird vorgeschlagen, um das Rentensystem zu entlasten?

What measure to relieve the pension system does the speaker suggest?

a. Anheben des Renteneintrittalters

Raising the age of retirement.

b. stärkere Einwanderung

Increased immigration.

c. Senkung der Renten

Lower pensions.

d. Ich weiß es nicht.

I don't know.

4. Warum freuen sich viele ältere Menschen, wenn sie länger im Beruf bleiben können?

Why do many older people like the idea of working past retirement age?

a. Ihre Rente steigt, wenn sie länger arbeiten.

Their pension claims increase the longer they remain in the work force.

b. Sie helfen somit, die Rentenleistungen auf für zukünftige Generationen zu erhalten.

They can contribute to safeguarding pension claims for future generations.

c. Sie möchten ihre Berufserfahrung weitergeben und langweilen sich weniger.

They want to pass on their experience and are less prone to boredom.

d. Ich weiß es nicht.

I don't know.

5. Welche Ängste werden gegen die Erhöhung des Renteneintrittalters angeführt?

What fears are associated with raising the age of retirement?

a. Man befürchtet, dass durch zu viele ältere Arbeitnehmer die Innovationsfähigkeit der Unternehmen sinkt.

The fear that too many older employees will decrease the company's potential for innovation.

b. Man befürchtet, dass es weniger Arbeitsplätze für junge Menschen gibt.

The fear that there will be fewer available positions for younger people.

c. Man befürchtet, dass ältere Arbeitnehmer zu viele Fehler machen.

The fear that older employees will make too many mistakes.

d. Ich weiß es nicht.

I don't know.

7.3.3 Speech “air travel”

1. Wieso stellen übergewichtige Passagiere die Fluggesellschaften vor ein Problem?

What problems do overweight passengers pose for airlines?

a. Sie haben auf Flügen ein höheres gesundheitliches Risiko.

They have an increased health risk during flights.

b. Sie verlangen größere Portionen bei den angebotenen Snacks und Mahlzeiten.

They demand larger portions of the in-flight snacks and meals.

c. Sie schränken den Flugkomfort anderer Fluggäste ein.

They encroach on the comfort of other passengers.

d. Ich weiß es nicht.

I don't know.

2. Was planen Fluggesellschaften in Bezug auf übergewichtige Passagiere?

What do airlines plan to do about overweight passengers?

a. Übergewichtige Passagiere sollen zwei Plätze bezahlen.

Overweight passengers will be required to buy two seats.

b. Übergewichtige Passagiere werden nur noch in die Businessclass gebucht.

Overweight passengers will only be booked in business class.

c. Übergewichtige Passagiere werden bei der Belegung des Flugzeugs zurückgestellt.

Overweight passengers will be considered last in seat allocation.

d. Ich weiß es nicht.

I don't know.

3. Warum wird Idee, dass Übergewichtige zwei Tickets zahlen sollen, kritisiert?

Why do some people criticize the idea of forcing overweight passengers to buy two seats?

a. Diese Richtlinie würde zu hohen Kosten führen.

This guidelines would result in high costs.

b. Diese Richtlinie würde übergewichtige Menschen diskriminieren.

This guideline would discriminate against overweight passengers.

c. Diese Richtlinie würde die Auslastung der Flugzeuge senken.

This guideline would decrease the plane's overall capacity.

d. Ich weiß es nicht.

I don't know.

4. Welche Passagiere benötigen ebenfalls mehr Platz?

Which other groups of passengers also require more space?

a. große Menschen, Eltern mit Kindern und Zeitungsleser.

Tall passengers, parents with children, passengers with newspapers.

b. große Menschen, Menschen mit Behinderung und Laptop-Nutzer.

Tall passengers, passengers with disabilities, laptop users.

c. große Menschen, Menschen mit Gips, Linkshänder.

Tall passengers, passengers with plaster casts, left-handed passengers.

d. Ich weiß es nicht.

I don't know.

5. Welche Lösung schlägt der Redner vor?

What solution does the speaker propose?

a. Die Fluggesellschaften sollten den Sitzabstand vergrößern.

Airlines should increase the distance between the seats.

b. Die Fluggesellschaften sollten die Businessclass ausbauen.

Airlines should increase the number of business class seats.

c. Die Fluggesellschaften sollten breitere Sitze anbieten.

Airlines should offer wider seats.

d. Ich weiß es nicht.

I don't know.

7.3.4 Speech "work"

1. Was sind die Risiken bei erhöhter Arbeitsbelastung?

What are the risks associated with a heavy workload?

a. Abnahme der Motivation und Zunahme der Krankheitstage

Decreased motivation and increased number of sick days.

b. Abnahme der Arbeitsqualität und Suizid

Decreased quality of work and suicide.

c. Verschlechterung der Produktivität und der Gesundheit der Arbeitnehmer

Deteriorated productivity and health.

d. Ich weiß es nicht.

I don't know.

2. Welche Schlüsselqualifikationen für den Umgang mit Stress werden in Workshops vermittelt?

Which key qualifications for dealing with stress are discussed in the workshops?

a. Motivationsfähigkeit und Kommunikationstechniken

Motivation and communication techniques.

b. Zeiteinteilung und Umgang mit Konflikten

Time management and dealing with conflicts.

c. Stressresistenz und Entspannungstechniken

Resistance to stress and relaxation techniques.

d. Ich weiß es nicht.

I don't know.

3. Was stellt eine große Belastung für Arbeitnehmer in einem Arbeitsprozess dar?

Which of the following is a major stress factor in the work process?

a. ständige Arbeitsunterbrechungen durch Kollegen

Constant interruptions caused by colleagues.

b. ständige Arbeitsunterbrechungen durch elektronische Nachrichten

Constant interruptions caused by electronic messages.

c. ständige Arbeitsunterbrechungen durch Telefonate

Constant interruptions caused by telephone calls.

d. Ich weiß es nicht.

I don't know.

4. Was ist eines der größten Stressfaktoren in Schulen?

What is one of the main stress factors in schools?

a. Konflikte mit schwierige Schüler

Conflict with difficult students.

b. Konflikte mit Eltern

Conflict with parents.

c. Konflikte mit Kollegen

Conflict with colleagues.

d. Ich weiß es nicht.

I don't know.

5. Was wird Lehrpersonal angeboten, um Konfliktsituationen anzugehen?

What conflict management tools are available to teachers?

a. Ratgeber und Leitfäden zum Konfliktmanagement

Guidebooks and guidelines on conflict management.

b. psychologische Betreuung

Psychological support.

c. gemeinsame Fallbesprechungen

Joint case discussions.

d. Ich weiß es nicht.

I don't know.

7.4 Tables

7.4.1 Results of the tukey comparison of response accuracy between participants during the pretest

	difference	lower	upper	Adjusted p-value
P11D-P10D	-0.465665142	-11.416.602.835	0.210329999	0.7253225
P12D-P10D	-0.511411169	-11.846.071.463	0.161784808	0.5085112
P13D-P10D	0.038779230	-0.6400978272	0.717656286	10.000.000
P14D-P10D	-0.463837595	-11.343.135.770	0.206638388	0.7169288
P15D-P10D	0.492391059	-0.1754407215	1.160.222.839	0.5787387
P1D-P10D	-0.082856163	-0.7617332201	0.596020893	10.000.000
P4D-P10D	-0.366993527	-10.429.886.679	0.309001614	0.9750672
P5D-P10D	0.120238871	-0.5586381853	0.799115928	10.000.000
P6D-P10D	0.067801180	-0.6110758764	0.746678237	10.000.000
P8D-P10D	0.023664998	-0.6612396130	0.708569609	10.000.000
P9D-P10D	-0.187057276	-0.8659343331	0.491819780	10.000.000
P2D-P10D	0.233186793	-0.4456902633	0.912063850	0.9999930
P3D-P10D	0.214851760	-0.4732068464	0.902910367	0.9999992
P10T-P10D	0.392026593	-0.2811693841	1.065.222.571	0.9414541
P11T-P10D	-0.096945089	-0.7674210719	0.573530893	10.000.000
P12T-P10D	0.221657441	-0.4601880901	0.903502972	0.9999979
P13T-P10D	-0.064421076	-0.7432981332	0.614455980	10.000.000
P14T-P10D	0.247133293	-0.4347122379	0.928978824	0.9999776
P15T-P10D	0.225072835	-0.4481231422	0.898268812	0.9999961
P16T-P10D	0.686250921	0.0102557796	1.362.246.062	0.0410785
P2T-P10D	-0.235271379	-0.9141484355	0.443605678	0.9999915
P3T-P10D	-0.101762625	-0.7777577660	0.574232516	10.000.000
P4T-P10D	-0.243247184	-0.9250927153	0.438598346	0.9999840
P5T-P10D	0.187114015	-0.5110235472	0.885251577	10.000.000
P6T-P10D	-0.430623753	-11.066.188.945	0.245371388	0.8539805
P7D-P10D	0.249119046	-0.4213569360	0.919595029	0.9999624
P7T-P10D	-0.279442013	-0.9526379906	0.393753964	0.9996634
P8T-P10D	2.386.181.958	17.101.868.167	3.062.177.099	0.0000000
P9T-P10D	0.240252552	-0.4386245047	0.919129609	0.9999866

P1T-P10D	-0.074663353	-0.7451393358	0.595812629	10.000.000
P12D-P11D	-0.045746026	-0.7217411675	0.630249115	10.000.000
P13D-P11D	0.504444372	-0.1772085199	1.186.097.264	0.5701537
P14D-P11D	0.001827548	-0.6714589067	0.675114002	10.000.000
P15D-P11D	0.958056201	0.2874028678	1.628.709.535	0.0000363
P1D-P11D	0.382808979	-0.2988439128	1.064.461.871	0.9617813
P4D-P11D	0.098671616	-0.5801111460	0.777454377	10.000.000
P5D-P11D	0.585904014	-0.0957488780	1.267.556.906	0.2360747
P6D-P11D	0.533466323	-0.1481865691	1.215.119.215	0.4381024
P8D-P11D	0.489330141	-0.1983259749	1.176.986.256	0.6584074
P9D-P11D	0.278607866	-0.4030450257	0.960260758	0.9997489
P2D-P11D	0.698851936	0.0171990441	1.380.504.828	0.0359571
P3D-P11D	0.680516903	-0.0102806458	1.371.314.452	0.0603229
P10T-P11D	0.857691736	0.1816965946	1.533.686.877	0.0007767
P11T-P11D	0.368720053	-0.3045664015	1.042.006.508	0.9720978
P12T-P11D	0.687322583	0.0027132531	1.371.931.913	0.0475216
P13T-P11D	0.401244066	-0.2804088259	1.082.896.958	0.9339781
P14T-P11D	0.712798435	0.0281891053	1.397.407.765	0.0289841
P15T-P11D	0.690737978	0.0147428365	1.366.733.119	0.0376328
P16T-P11D	1.151.916.063	0.4731333015	1.830.698.825	0.0000001
P2T-P11D	0.230393764	-0.4512591281	0.912046656	0.9999951
P3T-P11D	0.363902518	-0.3148802442	1.042.685.279	0.9788741
P4T-P11D	0.222417958	-0.4621913721	0.907027288	0.9999980
P5T-P11D	0.652779158	-0.0480579581	1.353.616.273	0.1136200
P6T-P11D	0.035041389	-0.6437413727	0.713824151	10.000.000
P7D-P11D	0.714784189	0.0414977344	1.388.070.643	0.0218054
P7T-P11D	0.186223129	-0.4897720119	0.862218270	10.000.000
P8T-P11D	2.851.847.100	21.730.643.386	3.530.629.862	0.0000000
P9T-P11D	0.705917694	0.0242648026	1.387.570.586	0.0312744
P1T-P11D	0.391001789	-0.2822846654	1.064.288.244	0.9432531
P13D-P12D	0.550190399	-0.1286866582	1.229.067.455	0.3576039
P14D-P12D	0.047573574	-0.6229024081	0.718049557	10.000.000
P15D-P12D	1.003.802.228	0.3359704474	1.671.634.008	0.0000076
P1D-P12D	0.428555006	-0.2503220511	1.107.432.062	0.8656315
P4D-P12D	0.144417642	-0.5315774989	0.820412783	10.000.000
P5D-P12D	0.631650040	-0.0472270163	1.310.527.097	0.1148824
P6D-P12D	0.579212349	-0.0996647075	1.258.089.406	0.2500817
P8D-P12D	0.535076167	-0.1498284441	1.219.980.778	0.4422119
P9D-P12D	0.324353893	-0.3545231641	1.003.230.949	0.9960993
P2D-P12D	0.744597962	0.0657209057	1.423.475.019	0.0131009
P3D-P12D	0.726262929	0.0382043226	1.414.321.536	0.0237616
P10T-P12D	0.903437762	0.2302417849	1.576.633.739	0.0002002

P11T-P12D	0.414466080	-0.2560099029	1.084.942.062	0.8900149
P12T-P12D	0.733068610	0.0512230788	1.414.914.141	0.0179887
P13T-P12D	0.446990092	-0.2318869643	1.125.867.149	0.8057294
P14T-P12D	0.758544462	0.0766989310	1.440.389.993	0.0103565
P15T-P12D	0.736484004	0.0632880267	1.409.679.981	0.0136550
P16T-P12D	1.197.662.090	0.5216669486	1.873.657.231	0.0000000
P2T-P12D	0.276139790	-0.4027372665	0.955016847	0.9997710
P3T-P12D	0.409648544	-0.2663465971	1.085.643.685	0.9099607
P4T-P12D	0.268163984	-0.4136815464	0.950009515	0.9998812
P5T-P12D	0.698525184	0.0003876217	1.396.662.746	0.0496462
P6T-P12D	0.080787415	-0.5952077256	0.756782557	10.000.000
P7D-P12D	0.760530215	0.0900542329	1.431.006.198	0.0074072
P7T-P12D	0.231969156	-0.4412268217	0.905165133	0.9999925
P8T-P12D	2.897.593.127	22.215.979.857	3.573.588.268	0.0000000
P9T-P12D	0.751663721	0.0727866642	1.430.540.778	0.0112142
P1T-P12D	0.436747816	-0.2337281669	1.107.223.798	0.8223943
P14D-P13D	-0.502616824	-11.787.967.391	0.173563091	0.5598210
P15D-P13D	0.453611829	-0.2199462760	1.127.169.935	0.7675733
P1D-P13D	-0.121635393	-0.8061463807	0.562875595	10.000.000
P4D-P13D	-0.405772756	-10.874.256.483	0.275880136	0.9253879
P5D-P13D	0.081459642	-0.6030513460	0.765970630	10.000.000
P6D-P13D	0.029021951	-0.6554890371	0.713532939	10.000.000
P8D-P13D	-0.015114231	-0.7056035946	0.675375132	10.000.000
P9D-P13D	-0.225836506	-0.9103474937	0.458674482	0.9999971
P2D-P13D	0.194407564	-0.4901034239	0.878918552	0.9999999
P3D-P13D	0.176072531	-0.5175454336	0.869690495	10.000.000
P10T-P13D	0.353247364	-0.3256296931	1.032.124.420	0.9859586
P11T-P13D	-0.135724319	-0.8119042339	0.540455596	10.000.000
P12T-P13D	0.182878211	-0.5045769236	0.870333346	10.000.000
P13T-P13D	-0.103200306	-0.7877112939	0.581310682	10.000.000
P14T-P13D	0.208354063	-0.4791010714	0.895809198	0.9999996
P15T-P13D	0.186293606	-0.4925834512	0.865170662	10.000.000
P16T-P13D	0.647471691	-0.0341812008	1.329.124.583	0.0921597
P2T-P13D	-0.274050608	-0.9585615961	0.410460380	0.9998314
P3T-P13D	-0.140541855	-0.8221947464	0.541111037	10.000.000
P4T-P13D	-0.282026414	-0.9694815488	0.405428721	0.9997305
P5T-P13D	0.148334786	-0.5552825041	0.851952075	10.000.000
P6T-P13D	-0.469402983	-11.510.558.749	0.212249909	0.7259856
P7D-P13D	0.210339817	-0.4658400981	0.886519732	0.9999993
P7T-P13D	-0.318221243	-0.9970982996	0.360655814	0.9971351
P8T-P13D	2.347.402.728	16.657.498.363	3.029.055.620	0.0000000
P9T-P13D	0.201473322	-0.4830376654	0.885984310	0.9999998

P1T-P13D	-0.113442583	-0.7896224979	0.562737332	10.000.000
P15D-P14D	0.956228654	0.2911388053	1.621.318.502	0.0000300
P1D-P14D	0.380981431	-0.2951984836	1.057.161.346	0.9602277
P4D-P14D	0.096844068	-0.5764423868	0.770130522	10.000.000
P5D-P14D	0.584076466	-0.0921034489	1.260.256.381	0.2269748
P6D-P14D	0.531638775	-0.1445411400	1.207.818.690	0.4272332
P8D-P14D	0.487502593	-0.1947287064	1.169.733.892	0.6493325
P9D-P14D	0.276780318	-0.3993995966	0.952960233	0.9997417
P2D-P14D	0.697024388	0.0208444732	1.373.204.303	0.0333581
P3D-P14D	0.678689355	-0.0067082417	1.364.086.952	0.0566014
P10T-P14D	0.855864188	0.1853882054	1.526.340.170	0.0006760
P11T-P14D	0.366892505	-0.3008524029	1.034.637.413	0.9708978
P12T-P14D	0.685495035	0.0063348575	1.364.655.213	0.0443363
P13T-P14D	0.399416518	-0.2767633968	1.075.596.433	0.9313907
P14T-P14D	0.710970888	0.0318107097	1.390.131.065	0.0268124
P15T-P14D	0.688910430	0.0184344473	1.359.386.412	0.0348852
P16T-P14D	1.150.088.515	0.4768020607	1.823.374.970	0.0000001
P2T-P14D	0.228566216	-0.4476136990	0.904746131	0.9999951
P3T-P14D	0.362074970	-0.3112114849	1.035.361.424	0.9779868
P4T-P14D	0.220590410	-0.4585697677	0.899750588	0.9999980
P5T-P14D	0.650951610	-0.0445635052	1.346.466.725	0.1080436
P6T-P14D	0.033213841	-0.6400726134	0.706500296	10.000.000
P7D-P14D	0.712956641	0.0452117330	1.380.701.549	0.0200052
P7T-P14D	0.184395581	-0.4860804011	0.854871564	10.000.000
P8T-P14D	2.850.019.552	21.767.330.978	3.523.306.007	0.0000000
P9T-P14D	0.704090147	0.0279102317	1.380.270.062	0.0289441
P1T-P14D	0.389174241	-0.2785706668	1.056.919.149	0.9409041
P1D-P15D	-0.575247222	-12.488.053.276	0.098310883	0.2481726
P4D-P15D	-0.859384586	-15.300.379.193	-0.188731252	0.0006186
P5D-P15D	-0.372152187	-10.457.102.928	0.301405918	0.9687716
P6D-P15D	-0.424589879	-10.981.479.839	0.248968227	0.8674062
P8D-P15D	-0.468726061	-11.483.588.951	0.210906774	0.7230039
P9D-P15D	-0.679448335	-13.530.064.406	-0.005890230	0.0446730
P2D-P15D	-0.259204265	-0.9327623708	0.414353840	0.9999230
P3D-P15D	-0.277539298	-0.9603504799	0.405271883	0.9997742
P10T-P15D	-0.100364466	-0.7681962462	0.567467315	10.000.000
P11T-P15D	-0.589336148	-12.544.259.965	0.075753700	0.1831529
P12T-P15D	-0.270733618	-0.9472835358	0.405816299	0.9998330
P13T-P15D	-0.556812135	-12.303.702.408	0.116745970	0.3142113
P14T-P15D	-0.245257766	-0.9218076837	0.431292152	0.9999776
P15T-P15D	-0.267318224	-0.9351500043	0.400513557	0.9998321
P16T-P15D	0.193859862	-0.4767934718	0.864513195	0.9999999

P2T-P15D	-0.727662438	-14.012.205.430	-0.054104332	0.0166856
P3T-P15D	-0.594153684	-12.648.070.175	0.076499650	0.1834573
P4T-P15D	-0.735638243	-14.121.881.611	-0.059088326	0.0150642
P5T-P15D	-0.305277044	-0.9982435069	0.387689419	0.9990040
P6T-P15D	-0.923014812	-15.936.681.460	-0.252361479	0.0001031
P7D-P15D	-0.243272012	-0.9083618607	0.421817836	0.9999729
P7T-P15D	-0.771833072	-14.396.648.527	-0.104001292	0.0052849
P8T-P15D	1.893.790.899	12.231.375.653	2.564.444.232	0.0000000
P9T-P15D	-0.252138507	-0.9256966123	0.421419598	0.9999562
P1T-P15D	-0.567054412	-12.321.442.605	0.098035436	0.2514195
P4D-P1D	-0.284137363	-0.9657902554	0.397515528	0.9996359
P5D-P1D	0.203095035	-0.4814159531	0.887606023	0.9999998
P6D-P1D	0.150657344	-0.5338536442	0.835168332	10.000.000
P8D-P1D	0.106521162	-0.5839682017	0.797010525	10.000.000
P9D-P1D	-0.104201113	-0.7887121008	0.580309875	10.000.000
P2D-P1D	0.316042957	-0.3684680310	1.000.553.945	0.9977693
P3D-P1D	0.297707924	-0.3959100407	0.991325888	0.9993754
P10T-P1D	0.474882757	-0.2039943002	1.153.759.813	0.6945378
P11T-P1D	-0.014088926	-0.6902688410	0.662090989	10.000.000
P12T-P1D	0.304513604	-0.3829415307	0.991968739	0.9989032
P13T-P1D	0.018435087	-0.6660759010	0.702946075	10.000.000
P14T-P1D	0.329989456	-0.3574656785	1.017.444.591	0.9957984
P15T-P1D	0.307928998	-0.3709480583	0.986806055	0.9983467
P16T-P1D	0.769107084	0.0874541921	1.450.759.976	0.0081315
P2T-P1D	-0.152415215	-0.8369262032	0.532095772	10.000.000
P3T-P1D	-0.018906462	-0.7005593535	0.662746430	10.000.000
P4T-P1D	-0.160391021	-0.8478461559	0.527064114	10.000.000
P5T-P1D	0.269970178	-0.4336471112	0.973587468	0.9999275
P6T-P1D	-0.347767590	-10.294.204.821	0.333885302	0.9894004
P7D-P1D	0.331975210	-0.3442047052	1.008.155.125	0.9940266
P7T-P1D	-0.196585850	-0.8754629067	0.482291207	0.9999999
P8T-P1D	2.469.038.121	17.873.852.292	3.150.691.013	0.0000000
P9T-P1D	0.323108715	-0.3614022725	1.007.619.703	0.9967901
P1T-P1D	0.008192810	-0.6679871050	0.684372725	10.000.000
P5D-P4D	0.487232398	-0.1944204937	1.168.885.290	0.6486909
P6D-P4D	0.434794707	-0.2468581848	1.116.447.599	0.8522487
P8D-P4D	0.390658525	-0.2969975906	1.078.314.641	0.9560790
P9D-P4D	0.179936250	-0.5017166414	0.861589142	10.000.000
P2D-P4D	0.600180320	-0.0814725716	1.281.833.212	0.1935790
P3D-P4D	0.581845287	-0.1089522614	1.272.642.836	0.2753988
P10T-P4D	0.759020120	0.0830249790	1.435.015.261	0.0088472
P11T-P4D	0.270048437	-0.4032380172	0.943334892	0.9998254

P12T-P4D	0.588650968	-0.0959583626	1.273.260.298	0.2354239
P13T-P4D	0.302572450	-0.3790804416	0.984225342	0.9988624
P14T-P4D	0.614126820	-0.0704825104	1.298.736.150	0.1640271
P15T-P4D	0.592066362	-0.0839287792	1.268.061.503	0.2024835
P16T-P4D	1.053.244.447	0.3744616858	1.732.027.209	0.0000026
P2T-P4D	0.131722148	-0.5499307438	0.813375040	10.000.000
P3T-P4D	0.265230902	-0.4135518599	0.944013664	0.9998955
P4T-P4D	0.123746342	-0.5608629878	0.808355672	10.000.000
P5T-P4D	0.554107542	-0.1467295738	1.254.944.658	0.4142019
P6T-P4D	-0.063630227	-0.7424129884	0.615152535	10.000.000
P7D-P4D	0.616112573	-0.0571738813	1.289.399.028	0.1358149
P7T-P4D	0.087551513	-0.5884436276	0.763546655	10.000.000
P8T-P4D	2.753.175.485	20.743.927.229	3.431.958.246	0.0000000
P9T-P4D	0.607246079	-0.0744068131	1.288.898.971	0.1746759
P1T-P4D	0.292330173	-0.3809562811	0.965616628	0.9992309
P6D-P5D	-0.052437691	-0.7369486790	0.632073297	10.000.000
P8D-P5D	-0.096573873	-0.7870632365	0.593915490	10.000.000
P9D-P5D	-0.307296148	-0.9918071356	0.377214840	0.9986152
P2D-P5D	0.112947922	-0.5715630658	0.797458910	10.000.000
P3D-P5D	0.094612889	-0.5990050755	0.788230853	10.000.000
P10T-P5D	0.271787722	-0.4070893350	0.950664779	0.9998315
P11T-P5D	-0.217183961	-0.8933638758	0.458995954	0.9999984
P12T-P5D	0.101418569	-0.5860365655	0.788873704	10.000.000
P13T-P5D	-0.184659948	-0.8691709358	0.499851040	10.000.000
P14T-P5D	0.126894421	-0.5605607133	0.814349556	10.000.000
P15T-P5D	0.104833964	-0.5740430931	0.783711020	10.000.000
P16T-P5D	0.566012049	-0.1156408427	1.247.664.941	0.3047938
P2T-P5D	-0.355510250	-10.400.212.380	0.329000738	0.9863246
P3T-P5D	-0.222001496	-0.9036543883	0.459651395	0.9999979
P4T-P5D	-0.363486056	-10.509.411.907	0.323969079	0.9824727
P5T-P5D	0.066875144	-0.6367421460	0.770492433	10.000.000
P6T-P5D	-0.550862625	-12.325.155.168	0.130790267	0.3640230
P7D-P5D	0.128880175	-0.5472997399	0.805060090	10.000.000
P7T-P5D	-0.399680885	-10.785.579.415	0.279196172	0.9338490
P8T-P5D	2.265.943.086	15.842.901.944	2.947.595.978	0.0000000
P9T-P5D	0.120013681	-0.5644973073	0.804524668	10.000.000
P1T-P5D	-0.194902225	-0.8710821397	0.481277690	0.9999999
P8D-P6D	-0.044136182	-0.7346255454	0.646353181	10.000.000
P9D-P6D	-0.254858457	-0.9393694445	0.429652531	0.9999608
P2D-P6D	0.165385613	-0.5191253747	0.849896601	10.000.000
P3D-P6D	0.147050580	-0.5465673844	0.840668545	10.000.000
P10T-P6D	0.324225413	-0.3546516439	1.003.102.470	0.9961239

P11T-P6D	-0.164746270	-0.8409261847	0.511433645	10.000.000
P12T-P6D	0.153856260	-0.5335988743	0.841311395	10.000.000
P13T-P6D	-0.132222257	-0.8167332447	0.552288731	10.000.000
P14T-P6D	0.179332113	-0.5081230221	0.866787247	10.000.000
P15T-P6D	0.157271655	-0.5216054020	0.836148712	10.000.000
P16T-P6D	0.618449740	-0.0632031515	1.300.102.632	0.1475271
P2T-P6D	-0.303072559	-0.9875835469	0.381438429	0.9989118
P3T-P6D	-0.169563805	-0.8512166972	0.512089087	10.000.000
P4T-P6D	-0.311048365	-0.9985034995	0.376406770	0.9984152
P5T-P6D	0.119312835	-0.5843044548	0.822930124	10.000.000
P6T-P6D	-0.498424934	-11.800.778.257	0.183227958	0.5978484
P7D-P6D	0.181317866	-0.4948620488	0.857497781	10.000.000
P7T-P6D	-0.347243194	-10.261.202.504	0.331633863	0.9889988
P8T-P6D	2.318.380.777	16.367.278.856	3.000.033.669	0.0000000
P9T-P6D	0.172451372	-0.5120596162	0.856962360	10.000.000
P1T-P6D	-0.142464534	-0.8186444486	0.533715381	10.000.000
P9D-P8D	-0.210722275	-0.9012116378	0.479767089	0.9999995
P2D-P8D	0.209521795	-0.4809675680	0.900011159	0.9999996
P3D-P8D	0.191186762	-0.5083317500	0.890705274	10.000.000
P10T-P8D	0.368361595	-0.3165430162	1.053.266.206	0.9779561
P11T-P8D	-0.120610088	-0.8028413868	0.561621212	10.000.000
P12T-P8D	0.197992443	-0.4954156844	0.891400569	0.9999999
P13T-P8D	-0.088086075	-0.7785754379	0.602403289	10.000.000
P14T-P8D	0.223468295	-0.4699398322	0.916876422	0.9999983
P15T-P8D	0.201407837	-0.4834967743	0.886312448	0.9999998
P16T-P8D	0.662585923	-0.0250701931	1.350.242.038	0.0784593
P2T-P8D	-0.258936377	-0.9494257402	0.431552986	0.9999545
P3T-P8D	-0.125427623	-0.8130837388	0.562228492	10.000.000
P4T-P8D	-0.266912183	-0.9603203097	0.426495944	0.9999226
P5T-P8D	0.163449017	-0.5459856584	0.872883692	10.000.000
P6T-P8D	-0.454288752	-11.419.448.673	0.233367364	0.8003332
P7D-P8D	0.225454048	-0.4567772510	0.907685347	0.9999970
P7T-P8D	-0.303107011	-0.9880116227	0.381797600	0.9989206
P8T-P8D	2.362.516.960	16.748.608.440	3.050.173.075	0.0000000
P9T-P8D	0.216587554	-0.4739018094	0.907076917	0.9999991
P1T-P8D	-0.098328352	-0.7805596508	0.583902948	10.000.000
P2D-P9D	0.420244070	-0.2642669181	1.104.755.058	0.8972466
P3D-P9D	0.401909037	-0.2917089278	1.095.527.001	0.9446821
P10T-P9D	0.579083870	-0.0997931872	1.257.960.926	0.2505077
P11T-P9D	0.090112187	-0.5860677280	0.766292102	10.000.000
P12T-P9D	0.408714717	-0.2787404177	1.096.169.852	0.9263874
P13T-P9D	0.122636200	-0.5618747880	0.807147188	10.000.000

P14T-P9D	0.434190569	-0.2532645655	1.121.645.704	0.8649969
P15T-P9D	0.412130111	-0.2667469453	1.091.007.168	0.9082915
P16T-P9D	0.873308197	0.1916553051	1.554.961.089	0.0006214
P2T-P9D	-0.048214102	-0.7327250902	0.636296885	10.000.000
P3T-P9D	0.085294651	-0.5963582405	0.766947543	10.000.000
P4T-P9D	-0.056189908	-0.7436450429	0.631265227	10.000.000
P5T-P9D	0.374171291	-0.3294459982	1.077.788.581	0.9810522
P6T-P9D	-0.243566477	-0.9252193691	0.438086415	0.9999835
P7D-P9D	0.436176323	-0.2400035922	1.112.356.238	0.8367488
P7T-P9D	-0.092384737	-0.7712617937	0.586492320	10.000.000
P8T-P9D	2.573.239.234	18.915.863.422	3.254.892.126	0.0000000
P9T-P9D	0.427309828	-0.2572011595	1.111.820.816	0.8790605
P1T-P9D	0.112393923	-0.5637859920	0.788573838	10.000.000
P3D-P2D	-0.018335033	-0.7119529976	0.675282931	10.000.000
P10T-P2D	0.158839800	-0.5200372570	0.837716856	10.000.000
P11T-P2D	-0.330131883	-10.063.117.978	0.346048032	0.9945182
P12T-P2D	-0.011529353	-0.6989844875	0.675925782	10.000.000
P13T-P2D	-0.297607870	-0.9821188578	0.386903118	0.9992120
P14T-P2D	0.013946499	-0.6735086353	0.701401634	10.000.000
P15T-P2D	-0.008113958	-0.6869910151	0.670763098	10.000.000
P16T-P2D	0.453064127	-0.2285887647	1.134.717.019	0.7903151
P2T-P2D	-0.468458172	-11.529.691.600	0.216052816	0.7379348
P3T-P2D	-0.334949418	-10.166.023.103	0.346703473	0.9939474
P4T-P2D	-0.476433978	-11.638.891.127	0.211021157	0.7133764
P5T-P2D	-0.046072778	-0.7496900680	0.657544511	10.000.000
P6T-P2D	-0.663810547	-13.454.634.389	0.017842345	0.0693165
P7D-P2D	0.015932253	-0.6602476620	0.692112168	10.000.000
P7T-P2D	-0.512628807	-11.915.058.635	0.166248250	0.5227803
P8T-P2D	2.152.995.164	14.713.422.724	2.834.648.056	0.0000000
P9T-P2D	0.007065759	-0.6774452293	0.691576746	10.000.000
P1T-P2D	-0.307850147	-0.9840300618	0.368329768	0.9982385
P10T-P3D	0.177174833	-0.5108837740	0.865233440	10.000.000
P11T-P3D	-0.311796850	-0.9971944465	0.373600747	0.9982623
P12T-P3D	0.006805680	-0.6897179377	0.703329298	10.000.000
P13T-P3D	-0.279272837	-0.9728908014	0.414345128	0.9998119
P14T-P3D	0.032281532	-0.6642420855	0.728805150	10.000.000
P15T-P3D	0.010221075	-0.6778375321	0.698279681	10.000.000
P16T-P3D	0.471399160	-0.2193983884	1.162.196.709	0.7433968
P2T-P3D	-0.450123139	-11.437.411.036	0.243494825	0.8280195
P3T-P3D	-0.316614385	-10.074.119.340	0.374183163	0.9980276
P4T-P3D	-0.458098945	-11.546.225.629	0.238424673	0.8075026
P5T-P3D	-0.027737745	-0.7402178352	0.684742344	10.000.000

P6T-P3D	-0.645475514	-13.362.730.626	0.045322035	0.1099216
P7D-P3D	0.034267286	-0.6511303107	0.719664883	10.000.000
P7T-P3D	-0.494293774	-11.823.523.805	0.193764833	0.6376253
P8T-P3D	2.171.330.197	14.805.326.487	2.862.127.746	0.0000000
P9T-P3D	0.025400792	-0.6682171729	0.719018756	10.000.000
P1T-P3D	-0.289515114	-0.9749127105	0.395882483	0.9995340
P11T-P10T	-0.488971683	-11.594.476.651	0.181504300	0.6038168
P12T-P10T	-0.170369153	-0.8522146834	0.511476378	10.000.000
P13T-P10T	-0.456447670	-11.353.247.264	0.222429387	0.7704728
P14T-P10T	-0.144893300	-0.8267388312	0.536952231	10.000.000
P15T-P10T	-0.166953758	-0.8401497354	0.506242219	10.000.000
P16T-P10T	0.294224327	-0.3817708136	0.970219469	0.9991965
P2T-P10T	-0.627297972	-13.061.750.287	0.051579085	0.1232914
P3T-P10T	-0.493789218	-11.697.843.592	0.182205923	0.6001550
P4T-P10T	-0.635273778	-13.171.193.086	0.046571753	0.1132794
P5T-P10T	-0.204912578	-0.9030501404	0.493224984	0.9999998
P6T-P10T	-0.822650347	-14.986.454.878	-0.146655206	0.0019231
P7D-P10T	-0.142907547	-0.8133835292	0.527568436	10.000.000
P7T-P10T	-0.671468607	-13.446.645.838	0.001727371	0.0516633
P8T-P10T	1.994.155.365	13.181.602.235	2.670.150.506	0.0000000
P9T-P10T	-0.151774041	-0.8306510980	0.527103016	10.000.000
P1T-P10T	-0.466689947	-11.371.659.290	0.203786036	0.7046434
P12T-P11T	0.318602530	-0.3605576477	0.997762708	0.9970983
P13T-P11T	0.032524013	-0.6436559020	0.708703928	10.000.000
P14T-P11T	0.344078382	-0.3350817955	1.023.238.560	0.9904285
P15T-P11T	0.322017925	-0.3484580579	0.992493907	0.9957613
P16T-P11T	0.783196010	0.1099095555	1.456.482.465	0.0046851
P2T-P11T	-0.138326289	-0.8145062042	0.537853626	10.000.000
P3T-P11T	-0.004817536	-0.6781039901	0.668468919	10.000.000
P4T-P11T	-0.146302095	-0.8254622729	0.532858083	10.000.000
P5T-P11T	0.284059104	-0.4114560104	0.979574219	0.9997525
P6T-P11T	-0.333678664	-10.069.651.186	0.339607791	0.9931047
P7D-P11T	0.346064136	-0.3216807722	1.013.809.044	0.9867296
P7T-P11T	-0.182496924	-0.8529729063	0.487979059	10.000.000
P8T-P11T	2.483.127.047	18.098.405.926	3.156.413.502	0.0000000
P9T-P11T	0.337197641	-0.3389822735	1.013.377.556	0.9924285
P1T-P11T	0.022281736	-0.6454631720	0.690026644	10.000.000
P13T-P12T	-0.286078517	-0.9735336519	0.401376618	0.9996471
P14T-P12T	0.025475852	-0.6649108742	0.715862579	10.000.000
P15T-P12T	0.003415394	-0.6784301365	0.685260925	10.000.000
P16T-P12T	0.464593480	-0.2200158501	1.149.202.810	0.7537214
P2T-P12T	-0.456928819	-11.443.839.542	0.230526315	0.7902850

P3T-P12T	-0.323420066	-10.080.293.957	0.361189264	0.9967471
P4T-P12T	-0.464904625	-11.552.913.516	0.225482101	0.7677286
P5T-P12T	-0.034543426	-0.7410252442	0.671938393	10.000.000
P6T-P12T	-0.652281194	-13.368.905.243	0.032328136	0.0890800
P7D-P12T	0.027461606	-0.6516985721	0.706621784	10.000.000
P7T-P12T	-0.501099454	-11.829.449.849	0.180746077	0.5862114
P8T-P12T	2.164.524.517	14.799.151.870	2.849.133.847	0.0000000
P9T-P12T	0.018595111	-0.6688600234	0.706050246	10.000.000
P1T-P12T	-0.296320794	-0.9754809719	0.382839384	0.9991609
P14T-P13T	0.311554369	-0.3759007653	0.999009504	0.9983706
P15T-P13T	0.289493912	-0.3893831452	0.968370968	0.9994452
P16T-P13T	0.750671997	0.0690191053	1.432.324.889	0.0122624
P2T-P13T	-0.170850302	-0.8553612901	0.513660686	10.000.000
P3T-P13T	-0.037341548	-0.7189944404	0.644311343	10.000.000
P4T-P13T	-0.178826108	-0.8662812427	0.508629027	10.000.000
P5T-P13T	0.251535092	-0.4520821980	0.955152381	0.9999833
P6T-P13T	-0.366202677	-10.478.555.689	0.315450215	0.9782809
P7D-P13T	0.313540123	-0.3626397920	0.989720038	0.9976040
P7T-P13T	-0.215020937	-0.8938979936	0.463856120	0.9999989
P8T-P13T	2.450.603.034	17.689.501.424	3.132.255.926	0.0000000
P9T-P13T	0.304673628	-0.3798373594	0.989184616	0.9988067
P1T-P13T	-0.010242277	-0.6864221918	0.665937638	10.000.000
P15T-P14T	-0.022060458	-0.7039059887	0.659785073	10.000.000
P16T-P14T	0.439117628	-0.2454917023	1.123.726.958	0.8447014
P2T-P14T	-0.482404672	-11.698.598.063	0.205050463	0.6880065
P3T-P14T	-0.348895918	-10.335.052.479	0.335713412	0.9895665
P4T-P14T	-0.490380477	-11.807.672.038	0.200006249	0.6623278
P5T-P14T	-0.060019278	-0.7665010964	0.646462541	10.000.000
P6T-P14T	-0.677757046	-13.623.663.764	0.006852284	0.0567591
P7D-P14T	0.001985754	-0.6771744243	0.681145931	10.000.000
P7T-P14T	-0.526575306	-12.084.208.371	0.155270225	0.4694824
P8T-P14T	2.139.048.665	14.544.393.348	2.823.657.995	0.0000000
P9T-P14T	-0.006880741	-0.6943358756	0.680574394	10.000.000
P1T-P14T	-0.321796646	-10.009.568.241	0.357363532	0.9965876
P16T-P15T	0.461178086	-0.2148170555	1.137.173.227	0.7438863
P2T-P15T	-0.460344214	-11.392.212.706	0.218532843	0.7551551
P3T-P15T	-0.326835460	-10.028.306.011	0.349159681	0.9952923
P4T-P15T	-0.468320020	-11.501.655.505	0.213525511	0.7310168
P5T-P15T	-0.037958820	-0.7360963823	0.660178742	10.000.000
P6T-P15T	-0.655696589	-13.316.917.296	0.020298552	0.0726317
P7D-P15T	0.024046211	-0.6464297711	0.694522194	10.000.000
P7T-P15T	-0.504514848	-11.777.108.257	0.168681129	0.5406020

P8T-P15T	2.161.109.123	14.851.139.816	2.837.104.264	0.0000000
P9T-P15T	0.015179717	-0.6636973398	0.694056774	10.000.000
P1T-P15T	-0.299736188	-0.9702121709	0.370739794	0.9987122
P2T-P16T	-0.921522299	-16.031.751.913	-0.239869407	0.0001657
P3T-P16T	-0.788013546	-14.667.963.074	-0.109230784	0.0048634
P4T-P16T	-0.929498105	-16.141.074.353	-0.244888775	0.0001481
P5T-P16T	-0.499136906	-11.999.740.213	0.201700210	0.6565551
P6T-P16T	-1.116.874.674	-17.956.574.359	-0.438091912	0.0000003
P7D-P16T	-0.437131874	-11.104.183.288	0.236154580	0.8273330
P7T-P16T	-0.965692934	-16.416.880.751	-0.289697793	0.0000363
P8T-P16T	1.699.931.037	10.211.482.754	2.378.713.799	0.0000000
P9T-P16T	-0.445998369	-11.276.512.606	0.235654523	0.8156344
P1T-P16T	-0.760914274	-14.342.007.286	-0.087627819	0.0079019
P3T-P2T	0.133508754	-0.5481441382	0.815161646	10.000.000
P4T-P2T	-0.007975806	-0.6954309405	0.679479329	10.000.000
P5T-P2T	0.422385394	-0.2812318958	1.126.002.683	0.9183942
P6T-P2T	-0.195352375	-0.8770052667	0.486300517	0.9999999
P7D-P2T	0.484390425	-0.1917894898	1.160.570.340	0.6438471
P7T-P2T	-0.044170635	-0.7230476913	0.634706422	10.000.000
P8T-P2T	2.621.453.337	19.398.004.446	3.303.106.228	0.0000000
P9T-P2T	0.475523931	-0.2089870571	1.160.034.919	0.7086085
P1T-P2T	0.160608025	-0.5155718896	0.836787940	10.000.000
P4T-P3T	-0.141484560	-0.8260938896	0.543124771	10.000.000
P5T-P3T	0.288876640	-0.4119604757	0.989713756	0.9997052
P6T-P3T	-0.328861129	-10.076.438.902	0.349921633	0.9951375
P7D-P3T	0.350881671	-0.3224047832	1.024.168.126	0.9856503
P7T-P3T	-0.177679388	-0.8536745294	0.498315753	10.000.000
P8T-P3T	2.487.944.583	18.091.618.210	3.166.727.344	0.0000000
P9T-P3T	0.342015177	-0.3396377149	1.023.668.069	0.9917029
P1T-P3T	0.027099272	-0.6461871830	0.700385726	10.000.000
P5T-P4T	0.430361200	-0.2761206190	1.136.843.018	0.9050205
P6T-P4T	-0.187376569	-0.8719858990	0.497232761	10.000.000
P7D-P4T	0.492366231	-0.1867939469	1.171.526.409	0.6172942
P7T-P4T	-0.036194829	-0.7180403597	0.645650702	10.000.000
P8T-P4T	2.629.429.142	19.448.198.122	3.314.038.472	0.0000000
P9T-P4T	0.483499737	-0.2039553982	1.170.954.871	0.6832793
P1T-P4T	0.168583831	-0.5105763467	0.847744009	10.000.000
P6T-P5T	-0.617737769	-13.185.748.842	0.083099347	0.1917872
P7D-P5T	0.062005031	-0.6335100835	0.757520146	10.000.000
P7T-P5T	-0.466556028	-11.646.935.907	0.231581534	0.7810641
P8T-P5T	2.199.067.943	14.982.308.270	2.899.905.058	0.0000000
P9T-P5T	0.053138537	-0.6504787526	0.756755827	10.000.000

P1T-P5T	-0.261777368	-0.9572924833	0.433737746	0.9999510
P7D-P6T	0.679742800	0.0064563453	1.353.029.254	0.0441857
P7T-P6T	0.151181740	-0.5248134009	0.827176881	10.000.000
P8T-P6T	2.816.805.711	21.380.229.495	3.495.588.473	0.0000000
P9T-P6T	0.670876305	-0.0107765864	1.352.529.197	0.0610220
P1T-P6T	0.355960400	-0.3173260545	1.029.246.855	0.9824956
P7T-P7D	-0.528561060	-11.990.370.422	0.141914923	0.4209970
P8T-P7D	2.137.062.911	14.637.764.568	2.810.349.366	0.0000000
P9T-P7D	-0.008866494	-0.6850464093	0.667313420	10.000.000
P1T-P7D	-0.323782400	-0.9915273078	0.343962508	0.9950739
P8T-P7T	2.665.623.971	19.896.288.300	3.341.619.112	0.0000000
P9T-P7T	0.519694565	-0.1591824914	1.198.571.622	0.4903311
P1T-P7T	0.204778660	-0.4656973225	0.875254642	0.9999995
P9T-P8T	-2.145.929.406	-28.275.822.977	-1.464.276.514	0.0000000
P1T-P8T	-2.460.845.311	-31.341.317.657	-1.787.558.857	0.0000000
P1T-P9T	-0.314915905	-0.9910958203	0.361264010	0.9974239

7.4.2 Results of the analyses of variance for the noise pretest.

		Noise level (DF = 3)		**Cognate status (DF=1)**		**Response accuracy (DF = 1)**		**Trial (DF= 62)**		**Stimulus (DF=61)**	
time bin	DF of residuals	F	p	F	p	F	p	F	p	F	p
1	1565	1.162	> 0.1	0.092	> 0.1	6.600	< 0.05	0.727	> 0.1	1.084	> 0.1
2	1580	1.930	> 0.1	0.777	> 0.1	7.775	< 0.01	0.003	> 0.1	1.158	> 0.1
3	1593	2.124	< 0.1	0.816	> 0.1	7.666	< 0.01	0.117	> 0.1	1.295	< 0.1
4	1600	2.041	> 0.1	0.376	> 0.1	8.347	< 0.01	0.036	> 0.1	1.204	> 0.1
5	1606	2.424	< 0.1	0.134	> 0.1	4.997	< 0.05	0.019	> 0.1	1.279	< 0.1
6	1607	3.425	< 0.05	0.018	> 0.1	4.689	< 0.05	0.008	> 0.1	1.323	> 0.1
7	1613	3.190	< 0.05	0.036	> 0.1	3.511	< 0.1	0.014	> 0.1	1.296	< 0.1

8	1614	2.033	> 0.1	0.061	> 0.1	3.479	< 0.1	0.011	> 0.1	1.218	> 0.1
9	1661	1.076	> 0.1	0.016	> 0.1	3.886	< 0.05	0.113	> 0.1	0.908	> 0.1
10	2186	1.168	> 0.1	2.198	> 0.1	0.854	> 0.1	0.003	> 0.1	1.427	< 0.0
11	2130	1.204	> 0.1	0.888	> 0.1	1.040	> 0.1	0.001	> 0.1	1.320	< 0.1
12	1825	1.233	> 0.1	0.385	> 0.1	0.001	> 0.1	0.014	> 0.1	1.463	< 0.0
13	1717	1.358	> 0.1	0.221	> 0.1	0.552	> 0.1	0.000	> 0.1	1.269	< 0.1
14	1588	2.502	< 0.1	0.016	> 0.1	0.065	> 0.1	0.679	> 0.1	0.965	> 0.1
15	1593	4.879	< 0.05	0.051	> 0.1	1.000	> 0.1	0.243	> 0.1	1.038	> 0.1
16	1568	5.098	< 0.05	0.010	> 0.1	2.155	> 0.1	0.222	> 0.1	1.219	> 0.1
17	1575	6.274.	< 0.05	0.240	> 0.1	3.730	< 0.1	0.288	> 0.1	1.260	< 0.1
18	1568	5.437	< 0.05	0.790	> 0.1	6.270	< 0.05	1.573	> 0.1	1.205	> 0.1
19	1478	5.978	< 0.05	1.490	> 0.1	12.234	< 0.05	8.140	< 0.05	1.104	> 0.1
20	1428	6.833	< 0.05	3.452	< 0.1	14.783	< 0.05	2.885	< 0.1	1.000	> 0.1
21	1367	7.079	< 0.05	7.639	< 0.05	11.882	< 0.05	0.500	> 0.1	1.046	> 0.1
22	1284	5.952	< 0.05	4.157	< 0.05	7.467	< 0.05	0.066	> 0.1	0.854	> 0.1
23	1242	8.090	< 0.05	1.243	> 0.1	2.284	> 0.1	0.004	> 0.1	0.765	> 0.1
24	1193	9.353	< 0.05	0.135	> 0.1	0.266	> 0.1	0.783	> 0.1	0.989	> 0.1
25	1161	4.727	< 0.05	0.110	> 0.1	0.004	> 0.1	0.727	> 0.1	1.004	> 0.1
26	1182	3.607	< 0.05	0.429	> 0.1	0.869	> 0.1	0.264	> 0.1	1.172	> 0.1
27	1192	1.412	> 0.1	0.159	> 0.1	1.856	> 0.1	0.073	> 0.1	1.256	> 0.1
28	1217	0.504	> 0.1	0.228	> 0.1	2.045	> 0.1	2.355	> 0.1	1.017	> 0.1
29	1242	0.109	> 0.1	0.265	> 0.1	1.687	> 0.1	4.087	< 0.05	1.027	> 0.1

30	1383	1.420	> 0.1	0.516	> 0.1	2.111	> 0.1	2.070	> 0.1	1.023	> 0.1

Table 33: Results of the analyses of variance for the noise pretest. F-values and p-values for the predictors noise level, cognate status, correctness of response, trial and stimulus in each 100 millisecond time bin.

7.4.3 Model estimates for the speech duration estimations

	Estimate	SE	t-value
(Intercept)	245.00	100.26	2.444
no noise	-45.00	55.57	-0.810
noise:P1D	55.00	141.79	0.388
no noise:P1D	100.00	141.79	0.705
noise:P5D	30.00	141.79	0.212
no noise:P5D	30.00	141.79	0.212
noise:P9D	20.00	141.79	0.141
no noise:P9D	100.00	141.79	0.705
noise:P10D	-100.00	141.79	-0.705
no noise:P10D	-25.00	141.79	-0.176
noise:P14D	-90.00	141.79	-0.635
no noise:P14D	-40.00	141.79	-0.282
noise:P6D	-5.00	141.79	-0.035
no noise:P6D	-50.00	141.79	-0.353
noise:P2D	115.00	141.79	0.811
no noise:P2D	65.00	141.79	0.458
noise:P11D	5.00	141.79	0.035
no noise:P11D	35.00	141.79	0.247
noise:P15D	-105.00	141.79	-0.741
no noise:P15D	-5.00	141.79	-0.035
noise:P3D	-80.00	141.79	-0.564
no noise:P3D	-65.00	141.79	-0.458
noise:P12D	-145.00	141.79	-1.023
no noise:P12D	-105.00	141.79	-0.741
noise:P4D	-55.00	141.79	-0.388
no noise:P4D	25.00	141.79	0.176
noise:P8D	125.00	141.79	0.882
no noise:P8D	160.00	141.79	1.129
noise:P13T	-15.00	141.79	-0.106
no noise:P13T	90.00	141.79	0.635
noise:P5T	-65.00	141.79	-0.458
no noise:P5T	-65.00	141.79	-0.458
noise:P9T	120.00	141.79	0.846
no noise:P9T	130.00	141.79	0.917

noise:P1T	-65.00	141.79	-0.458
no noise:P1T	70.00	141.79	0.494
noise:P10T	-15.00	141.79	-0.106
no noise:P10T	20.00	141.79	0.141
noise:P14T	-75.00	141.79	-0.529
no noise:P14T	5.00	141.79	0.035
noise:P2T	-10.00	141.79	-0.070
no noise:P2T	35.00	141.79	0.247
noise:P6T	30.00	141.79	0.212
no noise:P6T	25.00	141.79	0.176
noise:P11T	210.00	141.79	1.481
no noise:P11T	95.00	141.79	0.670
noise:P15T	-65.00	141.79	-0.458
no noise:P15T	5.00	141.79	0.035
noise:P3T	-40.00	141.79	-0.282
no noise:P3T	40.00	147.13	0.272
noise:P17T	-85.00	141.79	-0.600
no noise:P17T	-15.00	141.79	-0.106
noise:P7T	-105.00	141.79	-0.741
no noise:P7T	-50.00	141.79	-0.353
noise:P12T	-60.00	141.79	-0.423
no noise:P12T	-10.00	141.79	-0.070
noise:P16T	15.00	141.79	0.106
no noise:P16T	15.00	141.79	0.106
noise:P4T	-55.00	141.79	-0.388
no noise:P4T	10.00	141.79	0.070
noise:P8T	55.00	141.79	0.388
no noise:P8T	10.00	141.79	0.070

Table 34: Model estimates (intercept for all participants in the noise and no noise condition), standard error and t-value for estimation of speech duration.

7.4.4 Cognate translations: Cognate pairs with high Levenshtein ratio

English source word	German cognate translation	Phonetic Levenshtein ratio
passenger	Passagier	0.22
seat	Sitz	0.25
offer	offerieren	0.20

guest	Gäste	0.33
focus	fokussieren	0.20
discuss	diskutieren	0.28
oil	Öl	0.33
help	helfen	0.25
tours	Touren	0.20
number	Nummer	0.33
stop	stoppen	0.25
pass (on)	passieren	0.28
policy	Politik	0.33
especially	speziell	0.30
Italy	Italien	0.20
sit	sitzen	0.67
linked	verlinkt	0.67
(be) willing	willens (sein)	0.25
cost	kosten	0.50
longer	länger	0.33
present	präsentieren	0.22
often	oft	0.20
learn	lernen	0.20
group	Gruppe	0.20
clearly	klar	0.28
olive	Olive	0.20

prefer	präferieren	0.33
(brand) label	Label	0.36
coast	Küste	0.60
place	Platz	0.40
(youth) hostel	Hostel	0.25
ruins	Ruinen	0.20
young	jung	0.20
diet	Diät	0.50
find	finden	0.25
fact	Fakt	0.25
crisis	Krise	0.26
burden	Bürde	0.33
older	älter	0.20
sums	Summen	0.25

Table 35: Cognate pairs with a ratio Levenshtein distance over 0.167