L2 listeners benefit from audiovisual information in the processing of Swedish vowels

A speech comprehension experiment with L1 and L2 listeners

Helene Springer

Supervisor: Assoc. Prof. Victoria Johansson, Prof. Marianne Gullberg

Centre for Language and Literature, Lund University
MA in Language and Linguistics, General Linguistics
SPVR01 Language and Linguistics: Degree Project – Master's (Two Years) Thesis, 30 credits

November 2021

Abstract

This thesis investigates the effect of visible articulatory movements on speech comprehension in L1 and L2 listeners of Swedish. Previous research found faciliatory effects on speech perception when speech was presented audiovisually in various settings and under diverse listening conditions for both L1 and L2 listeners. However, it is still not investigated thoroughly, how L2 proficiency affects the way in which listeners benefit from visual information in the speech signal. Therefore, this study examined L1 and L2 listeners of Swedish in a phoneme recognition task including audio-only and audiovisual speech. Furthermore, in light of the Covid-19 pandemic, there has been an increasing interest in the effects of covering the mouth of a speaker on speech comprehension. Therefore, this study added a third condition in which the mouth of the speaker was blurred. The target phonemes were embedded in minimal pairs and inserted in decontextualized sentences in order to put the listeners in a challenging comprehension situation.

The results showed that audiovisual information facilitated the recognition of phonemes in clear speech, but only for L2 listeners. Although there was a floor effect with regards to the error rates, a temporal facilitation occurred in the form of decreased response times. L2 listeners benefit from visible speech in the recognition of Swedish vowels, especially when they perceive non-native sound contrasts. However, no correlation was found between the L2 proficiency and the processing of audiovisual speech. This study supports findings from previous research suggesting that audiovisual information is not an essential, but a facilitative aspect in speech perception. Thus, it has implications for speech perception of vulnerable listener populations such as L2 learners or hearing-impaired listeners, in educational, but also in health care contexts.

Keywords: *language processing, visual information, audiovisual speech perception, phoneme recognition, second-language sound contrasts*

Acknowledgements

First of all, I want to thank my supervisors Victoria Johansson and Marianne Gullberg for their support, encouragement, and inspiration throughout the process of this project, from turning my ideas and research interests into a study, arranging the use of the Humanities Lab, up to putting this final version into writing. Your knowledge and feedback made it possible to create this thesis. Tack så mycket!

I would also like to thank Peter Roslund for helping me with the recordings for the experiment, and the editing of the material. Thank you to Jordan Zlatev for organizing the vouchers for the participants, to Åsa Wikström for assisting in administrative matters throughout my studies, and to my teachers at Lund University for equipping me with the knowledge to form this thesis project.

A special thank you also belongs to my friends Annika, Michelle, Ana and Rocío, for helping me with the recordings, for all our linguistics-related discussions, fika breaks, and for sharing laughs and troubles along the way. We did not expect our Master studies to be like this, but I am more than happy to see what we made out of these two years. I also owe a thank you to every single one of my friends who encouraged me, and piloted my experiment until it was finally ready for the participants.

This brings me to all the people that took part in this study. Every single one of you was incredibly patient, even when technical issues got in the way of our zoom meetings. Tack!

Zu guter Letzt danke ich meiner Familie, meinen Eltern und Großeltern, Anni und Käthe, für eure grenzenlose Unterstützung. Danke, dass ihr immer nur einen Anruf entfernt seid und es in jeder Lebenslage schafft, mich aufzuheitern und zu motivieren, mein Bestes zu geben.

Table of contents

List of figures

List of tables

Abbreviations

A	audio-only
AoA	Age of Acquisition
AV	audiovisual
B	blurred
CG	Category Goodness assimilation
dB	Decibel
ERP	event-related potential
F0	fundamental frequency
fMRI	functional magnetic resonance imaging
H	hypothesis
IFG	inferior frontal gyrus
L1	first language
L2	second language
LHQ	Language History Questionnaire
M	mean
ms	milliseconds
PAM	Perceptual Assimilation Model
PoIE	principle of inverse effectiveness
(p)STS	(posterior) superior temporal sulcus
RQ	research question
RT	response time
SC	Single Category Assimilation
SLM	Speech Learning Model
STG	superior temporal gyrus

1. Introduction

The perception of spoken language in hearing individuals' face-to-face conversation involves visual information for the addressee, even though the dominant source of input is auditory. This visual information comprises manual gestures as well as facial movements. This is not to say that humans cannot communicate via the phone without any visual speech cues. Why is it the case then, that there are people who anecdotally report comprehension difficulties in telephone conversations, or conversations with a person wearing a face mask? Whether these reports actually give rise to an underlying speech perception phenomenon has been investigated by a number of studies, showing that gestures and lipreading ease language processing as well as production of speech for both first language (L1) and second language (L2) speakers (Alibali, Kita & Young, 2000; Drijvers & Özyürek, 2020; Gullberg, 2006; Hazan, Sennema & Faulkner, 2002; Sueyoshi & Hardison, 2005), in noise and in clear speech (Reisberg, McLean & Goldfield, 1987; Sumby & Pollack, 1954). Especially with regards to the recognition of phonemes, a significant amount of information can be retrieved from the face. Phonemes are representations of speech sounds and the smallest distinctive units in spoken language (Warren, 2013). Thus, if a phoneme is misperceived, the meaning of a word and consequently, the meaning of a whole utterance can be changed. This makes phoneme recognition an interesting domain to investigate speech perception benefits caused by visual cues. As Cutler et al. (2004) put it, "[p]honeme identification problems may be particularly important in that all later levels processing will be affected by the decisions made at the phonemic level" (p. 3676).

A lot of research carried out in the field of audiovisual (AV) speech perception addresses L1 speech perception, and the research on L2 acquisition gives a twofold picture of when in the course of L2 acquisition L2 listeners gain the most benefit from AV speech. This thesis aims to investigate the effects of visual speech input on speech comprehension in L1 users as well as L2 learners of Swedish, with a particular focus on lip movements and on the lowest level of speech comprehension, that is the recognition of phonemes. This is particularly interesting with respect to the recent pandemic situation in which people listen to speech without seeing the whole face

of the person they are interacting with, but also in health care, where face masks are used consistently.

The outbreak of the Covid-19 pandemic poses a challenge to a listener in face-to-face conversations because people are confronted with covered mouths in communication situations on a daily basis. Studies have investigated the effects of face covering on the recognition of emotions revealing that the mouth area can be an important cue for recognizing the interlocutor's state of mind (Mheidly et al., 2020). Studies on the effect of wearing face masks on speech recognition reveal not only that the speech signal is affected adversely on the acoustic level, but that hiding articulatory gestures can pose a challenge in speech comprehension (Bandaru et al., 2020).

There are a lot of factors that can affect the perception of spoken language adversely or beneficially. Among these factors are noise degradation by means of background noise or the use of face masks, the availability of other visual cues such as manual gestures, speaking style, complexity of the utterance, language proficiency and the L1 of the listener, hearing capacity, and individual lipreading abilities that may facilitate AV speech perception, and a major goal of research in this field has been, and still is, to find out to what extent and how visual cues improve speech perception.

Studies show that L1 listeners benefit more from visual input in the form of lipreading than L2 listeners (Drijvers & Özyürek, 2020), and that high-proficient L2 listeners use visual speech cues more efficiently than less-proficient L2 listeners (Hazan et al., 2002; Sueyoshi & Hardison, 2005). However, to date only a few studies have explicitly investigated if there is a systematic connection between the proficiency level of a listener and AV speech perception (Wang, Behne & Jiang, 2008b). This study attempts to address this gap by taking a closer look at the correlation between different proficiency levels and AV speech effects, and introduces the novel aspect of a continuous variable for L2 proficiency. Swedish L1 listeners and German L2 listeners of Swedish are examined on the level of phoneme recognition, not in isolation but embedded in minimal pairs in decontextualized sentences. More precisely, vocalic as well as consonantal minimal pairs are examined.

For this purpose, a speech perception experiment is carried out, in which spoken language is presented accompanied by a video showing the full face of the speaker, and a video in which the speaker's face is masked so that the lower part of the face is blurred, and thus, articulatory gestures are no longer visible. In a third condition, speech is presented without any visual information. Together with a questionnaire on the language background of the participants, potential correlations between proficiency levels and AV speech benefit are addressed in this study. This benefit is measured by means of performance in the perception task as well as response times (RTs) that can point to facilitating effects and processing load.

The findings may add knowledge to the field of visual cues in speech perception, which is of high interest for everyday communication, but also for educational and health care settings, not least in light of the ongoing pandemic.

1.1 Outline

The following study is organized into five further chapters. In Chapter 2, a background regarding AV speech perception in native speakers as well as second-language learners is provided, finished with section 2.4 stating the research questions and hypotheses of the current study. In Chapter 3, the methods are described, which includes information on the participants, the stimulus material and experiment design as well as the language background questionnaire and the experimental procedure. Chapter 4 presents the results of the questionnaire and the experiment. In Chapter 5, these results are discussed, finishing with a conclusion and outlook in Chapter 6.

2. Theoretical background

In order to elaborate on the current state of AV speech perception research and the resulting research niche for this study, this chapter provides a review of the relevant literature. First, section 2.1 focuses on the effects of visible speech and articulatory gestures on L1 listeners as well as L2 learners. Section 2.2 provides an overview about studies that have been carried out related to face masks and the resulting effects on the recognition of emotions and on spoken language. Section 2.3 provides complementary information on the scope of visual information available in speech processing. Finally, in section 2.4, the current study is described.

The term *first language (L1)* is used in this thesis for the first language learned as a child, whereas the *second language (L2)* is used for any other language learned after the L1 has been acquired.

2.1 Reading lips while listening

Phonemes are underlying representations of speech sounds that have a distinctive status in a language (Ladefoged, 1975). From the distinctive nature of the term phoneme, the visual pendant "visual phoneme", *viseme*, can be derived, which categorizes phonemes according to their visual distinctiveness (Fisher, 1968, p. 800). For example, phonemes with the same place of articulation, such as the bilabials [p], [b] and [m] can be described as belonging to the same visual category (Amcoff, 1970). However, as this example illustrates, these visemes are not sufficient to unambiguously specify a speech sound in spoken language. Concerning Swedish phonemes, Amcoff (1970) establishes seven viseme groups and orders them in terms of visual salience, starting with the labials [p], [b] and [m] being easily readable, followed by labiodentals

[f] and [v], while the viseme group containing [h], [s], [j], [ɕ], [ʂ], [r], [l], [g], [k] and the alveolar sounds [t], [d] and [n] are the most difficult to lipread[1] (p. 12).

2.1.1 Lipreading as a developing ability in L1 listeners

While it is well-known that deaf or hearing-impaired individuals rely strongly on the visual modality, the earlier mentioned studies point to the fact that even normal-hearing individuals make use of the visual modality in speech perception. This ability exists already in infancy. For instance, infants are sensitive to the correspondence of visual and auditory input (Kuhl & Meltzoff, 1982). Examples are the McGurk effect in infants (Burnham & Dodd, 2004), which is the phenomenon that conflicting visual and auditory information is fused or combined in the perceptual outcome (McGurk & MacDonald, 1976; see section 2.1.4.1), and infants' ability to distinguish their first language from another language in purely visual contexts with silent videos (Weikum et al., 2007).

Even though AV speech is already processed by infants, this is not to say humans are able to process AV speech highly efficiently from a young age. Instead, AV speech integration and concomitant benefits for perception can also be related to age and language experience in L1 acquisition. As discussed by Soto-Faraco et al. (2012), the degree to which AV information is integrated efficiently increases with age. This might be related to articulation experience – the more intact and trained the production ability and the more established the phonological representations, the higher the influence of visual information (Desjardins, Rogers & Werker, 1997; Siva et al., 1995). Lalonde and Werner (2019) show that infants cannot make use of visual cues in speech as effectively as adults can. Furthermore, infants' ability to distinguish even non-native phonemes decreases in the first months of life which poses a later challenge to L2 acquisition. The so-called *perceptual narrowing* describes the specialization of speech perception on the phoneme inventory of an infant's L1 (Soto-Faraco et al., 2012). Touching upon the effects of AV speech on L2 acquisition, Werker, Frost and McGurk (1992) show that

[1] Amcoff (1970) acknowledges that the placement of [t], [d] and [n] in relation to [l], [k] and [g] in this hierarchy of salience was rather unexpected, since the former are articulated further in the front of the vocal tract, assuming this was found due to the openness of [l], [k] and [g].

language experience is a crucial factor that facilitates L2 lipreading by asking French-English bilinguals to repeat audiovisually presented syllables.

2.1.2 Lip movements precede the acoustic signal in AV speech

A considerable part of AV speech perception research concerns the time course of articulatory movements and their corresponding auditory events. Articulatory movements precede the speech output that is audible for the listener. Thus, visual information is visible before the auditory speech is heard, but still both modalities are integrated to one percept. Munhall et al. (1996) show that this temporal shift can vary up to a certain extent without diminishing AV integration by evoking McGurk effects with visual information presented up to 180 ms before the auditory information. This is not the case when auditory information is presented before the visual speech input, suggesting that listeners are sensitive to the natural dynamics of articulation when listening to speech (Munhall et al., 1996). Dynamic visual information has a much stronger effect on perception than static information from a speaker's face (Rosenblum & Saldaña, 1996).

It is further suggested that the visual input preceding the acoustic speech signal allows predictions to be formed regarding the following auditory input during AV speech processing, facilitating the perception of speech when both sources of input provide matching information (van Wassenhove, Grant & Poeppel, 2005). Navarra et al. (2010) extend these findings by showing that L2 listeners with higher experience in the L2 also show this anticipatory effect.

2.1.3 A model for the facilitation of speech perception by AV information

The issue of temporal aspects in AV speech, which modality is perceived first and at what point the two modalities are integrated, are the basis for many theoretical frameworks on AV speech perception (for an overview, see also Peelle & Sommers, 2015). The *Fuzzy Logical Model of Perception* (Massaro, 1987) is a late integration model, which states that the auditory and visual modalities are first perceived separately, and integrated at a later point. Early integration models, on the other hand, emphasize that the auditory and visual signals interact already at an early

stage of processing, and that perception results from earlier production experience in a language (*Motor Theory* by Liberman & Mattingly, 1985; *Analysis-by-synthesis* by van Wassenhove et al., 2005). Combining late and early integration models, a hybrid approach was suggested by Peelle and Sommers (2015), which involves several integration stages.

The analysis-by-synthesis model is based on the approach of Halle and Stevens (1962) assuming that incoming speech input is actively analyzed by synthesizing the production process according to stored phonemic representations. The model suggested by van Wassenhove et al. (2005) is based on the natural dynamics of speech, meaning that the visible articulatory gesture precede the acoustic signal. AV speech information facilitates the perception process in terms of salience and redundancy (van Wassenhove et al., 2005). According to this model, the visual input builds up a prediction, which is then checked against the following acoustic signal. If the visual input is highly salient, it allows for a strong prediction regarding the upcoming acoustic content of speech. For example, the visual signal for the phoneme /b/ is expected to facilitate the phoneme perception process more than the visual signal for /k/, since the latter is articulated further back in the vocal tract and thus, builds up a weaker prediction (van Wassenhove et al., 2005). If the following acoustic content, in turn, is redundant and matches the prediction from the visual signal, speech processing is temporally facilitated. In other words, when the visual and auditory information in the speech signal is congruent [2] and within the time frame of tolerated desynchronization, as in this current study, visual input facilitates the recognition of phonemes in speech by accelerating the perception process.

2.1.4 Lipreading facilitates perception on different levels of speech

After some basic characteristics of the visual cues from articulatory gestures have been discussed, the following sections focus on the effects of visual cues from lipreading on the different levels of speech perception, from lower level effects regarding phoneme recognition up to higher level effects on the comprehension of sentences and narratives.

[2] If the visual and auditory information is incongruent, phenomena like the McGurk phenomenon can be observed. In this case, there is no redundancy and fusion/combinations may be the perceptual result (see 2.1.4.1).

2.1.4.1 Phonemes and syllables

McGurk and MacDonald (1976) have contributed highly influential evidence for the importance of visual speech input by showing that conflicting visual and auditory information can lead to fused or combined percepts, known as the McGurk effect. One of the most-cited examples is the presentation of a speaker saying *ba*, while showing the lip movements of *ga* simultaneously, which causes listeners to perceive *da* (McGurk & MacDonald, 1976). Beyond this, AV speech facilitates the perception of sound contrasts. For instance, Navarra and Soto-Faraco (2007) examine Spanish-Catalan bilinguals' discrimination of the vowel contrast /e/ and /ɛ/, which is a phonemic contrast in Catalan, but an allophonic contrast in Spanish. RTs reveal that in unimodal auditory speech, Spanish-dominant listeners do not perceive a difference, whereas the AV speech input allows for better discrimination of the sounds in both groups, even the Spanish-dominant group, suggesting that reading from the lips supports both L1 and L2 sound discrimination (Navarra & Soto-Faraco, 2007).

Neuroimaging research has established regions for the processing of acoustic signals, located in the temporal lobe, especially Heschl's gyrus, as well as visual processing areas, which lie in the occipital lobe (Campbell & MacSweeney, 2012). The posterior superior temporal sulcus (pSTS) is commonly associated with the integration of multimodal input by a wide range of studies that have investigated, for example, written language, AV speech, speechreading in deaf individuals, or sign language processing (see Campbell & MacSweeney, 2012). In line with this, there is evidence from a transcranial magnetic stimulation study that the stimulation of the pSTS leads to a disruption of the McGurk effect[3] (Beauchamp, Nath & Pasalar, 2010).

Phoneme processing, which is the objective of the current study, involves the STS/STG regions as well as the supramarginal gyrus, and the inferior frontal gyrus (IFG) (Peelle, 2019). The supramarginal gyrus, for instance, is involved in the detection of unfamiliar phonemes, and the IFG has been shown to be of importance in tasks that require a decision on the phoneme level

[3] Even though the McGurk effect can be interpreted as evidence for the integration of information from different modalities, Van Engen, Xie and Chandrasekaran (2017) demonstrate that this effect is a very specific phenomenon, and that AV processing of natural speech involves other factors that enable a listener to benefit from AV speech, even if this particular listener does not exhibit the McGurk effect.

with similar sounding words (Peelle, 2019). Moreover, the left premotor cortex, involved in the execution of articulatory movements to produce certain speech sounds, can support the integration of the multimodal input in AV speech perception (Peelle, 2019). For instance, motor-evoked potentials of tongue-related areas in the motor cortex have been found for the visual and/or auditory perception of syllables that involve the tongue as an articulator (e.g. *ga* or *da*, but not *ba*) (Sato et al., 2010). Moreover, an event-related potential (ERP) study shows that audiovisually presented syllables reduce the N1 and P2 amplitudes, which is explained in terms of the early integration of visual information in the perception process that allows for the prediction of upcoming auditory speech input (van Wassenhove et al., 2005). However, even though seeing a talking face activates the visual movement cortex (V5/MT) in the occipito-temporal lobe, watching someone talk activates more left-lateralized regions as well, which suggests that articulatory movements are processed like language-specific movements (Campbell & MacSweeney, 2012).

In order to extract beneficial visual cues, spatial resolution does not need to be very high (Munhall et al., 2004b) or that an actual face is visible. Rosenblum and Saldaña (1996) show that a face marked with single illuminated points at anchor positions of the lips and lower part of the face is already sufficient to facilitate speech perception, which suggests that the kinematic features of articulatory movements are processed similarly to real visible speech as it is read from a talking face.

AV speech facilitates processing even on the suprasegmental level, for example with regards to lexical tone processing. In the acoustic signal, lexical tones are systematic variations of the fundamental frequency (F0) that can change the meaning of a word (Smith & Burnham, 2012). Native as well as tone-naïve listeners can distinguish Mandarin Chinese lexical tones in clear and degraded speech more accurate when visible speech provides them with durational information of these tones (Smith & Burnham, 2012). This does not only underline the fact that AV speech improves speech perception on many different levels, but also suggests that visual speech benefit is not restricted to language-specific elements (Smith & Burnham, 2012) and thus, may play a greater role in L2 acquisition as well.

When investigating phonemes in phonological contexts, the aspect of coarticulation comes into play. As soon as phonemes occur adjacent to others, their execution is affected by the preceding and/or following phoneme on both the visual and acoustic level (Benguerel & Pichora-Fuller, 1982). Therefore, one and the same phoneme may look differently in different phonological contexts. This means that the preceding viseme can already serve as a cue to the upcoming viseme. Benguerel and Pichora-Fuller (1982) show that coarticulation can impede lipreading in a visual-only paradigm especially in contexts involving visually less salient articulatory gestures, whereas labial, labiodental and linguodental articulatory gestures remain relatively resistant to these adverse coarticulation effects, presumably due to their visual saliency. However, Redford et al. (2018) show that AV input improves subjects' ability to anticipate upcoming sounds. Finally, even though coarticulation affects phoneme realization, the minimal pairs examined in the current study are embedded in the same sentence context. Thus, the phonological context is held constant for both target phonemes so that coarticulation is no longer a potential confounding factor.

Moreover, AV facilitation on the phoneme level has consequences for lexical activation. Visual cues do not only facilitate phoneme recognition, but this visual speech benefit is also stronger in words than in pseudowords, which suggests that visual phonemic cues in speech support the access of corresponding lexical entries (Fort et al., 2010).

2.1.4.2 Words

Moving on to a higher level of speech, previous research shows that listeners benefit from reading lips in word recognition as well. Kim, Aubanel and Davis (2015) compare an audio-only condition with an AV condition and a third one in which the mouth is masked, and measured accuracy in identification of keywords. Speech perception performance for English L1 listeners is best in the AV condition, followed by the masked condition, while the audio-only condition yields the lowest perception scores (Kim et al., 2015). Drijvers and Özyürek (2017) examine how co-speech gestures and articulatory gestures facilitate verb recall in L1 listeners of Dutch. The results show that the more input is given visually, the more words are identified correctly,

which means that both articulatory gestures and co-speech gestures facilitate the word comprehension in L1 listeners (Drijvers & Özyürek, 2017).

Reading the articulatory gestures from the lips of the speaker can further reduce the lexical competition, which supports word recognition. Tye-Murray, Sommers and Spehar (2007) investigate the acoustic and visual neighborhoods of AV words, these are sets of similar sounding and similar looking words, showing that the densities of the neighborhoods from both modalities affect the recognition of AV speech. If speech is presented in both modalities, and the intersection of the two modality sets contains only a small number of words, it is more likely that the presented word is identified faster because there are fewer competitors (Tye-Murray et al., 2007).

In order to understand speech, the speech stream needs to be segmented into words. Cunillera et al. (2010) shows that AV input eases the segmentation when the auditory and visual information components are presented synchronously, albeit this synchronicity does not need to be perfect.

2.1.4.3 Sentences and narratives

As early as in the 1930s, studies have provided quantitative evidence that when hearing-impaired as well as normal-hearing individuals listen to spoken sentences, they are not doing this exclusively by means of hearing but also by using visible articulation (Cotton, 1935). Sueyoshi and Hardison (2005) test comprehension of a lecture in intermediate and highly proficient learners of English, using manual gestures and articulatory gestures separately and combined, and compare these conditions to an audio-only narrative. Intermediate learners perform best when both visual cues are present, whereas the high proficiency learner group perform best when the face of the speaker is visible, suggesting that manual gestures are more important in lower levels of proficiency and that lipreading information is used more efficiently by advanced L2 learners (Sueyoshi & Hardison, 2005). This can be explained in terms of phonological representations, which are more developed in more advanced L2 listeners (Sueyoshi & Hardison, 2005).

Reisberg et al. (1987) use the speech shadowing method to show that L2 speakers of French and German benefit from lipreading even in processing of clear L2 speech, since these L2 listeners

have been able to repeat more words when visual speech input is provided compared to purely auditory speech input. In two further experiments, Reisberg et al. (1987) show that AV speech benefits are even applicable to the comprehension of L1 speech with a foreign accent[4], and also to complex speech input such as philosophical narratives. However, these effects are somewhat smaller than the effects observed in L2 listeners, suggesting that lipreading plays a greater role in L2 comprehension (Reisberg et al., 1987). Arnold and Hill (2001) extend these findings regarding speechreading advantages beyond the method of shadowing and investigate AV speech input in syntactically and semantically complex speech, and stories presented in French as well as in English with a Glaswegian accent. All three experiments show significant improvement of comprehension when listeners can see the speaker (Arnold & Hill, 2001).

Neuroimaging evidence on the sentence level of speech is provided by Calvert, Campbell and Brammer (2000), whose functional magnetic resonance imaging (fMRI) study shows that the pSTS is activated for AV narrative speech input when auditory and visual information is congruent, whereas incongruent AV input leads to sub-additive response patterns (Calvert et al., 2000). Evidence from a more recent fMRI study supports this view of the pSTS as an integration area of multimodal input for L1 as well as L2 perception (Barrós-Loscertales et al., 2013).

Although this study has the recognition of phonemes as its objective, these phonemes are embedded in a sentence context. In contrast to many of the above-described studies, real life conversation seldomly confronts people with isolated words or syllables, but with speech in the form of sentences. The more sentences are strung together, the more context builds up that eases the processing of incoming speech[5]. L2 learners often have much more difficulties in processing degraded speech signals, for example in noisy surroundings or when higher-level context is missing (Cutler et al., 2007). However, even when higher-level context is given, studies show that L2 listeners cannot rely on this context information as efficiently as native speakers can (Shi, 2010). Semantic context, for example, is used by advanced L2 learners but not by beginners and

[4] The benefits from lipreading for comprehension of accented speech found in Reisberg et al.'s (1987) study vary to a large extent and can be interpreted as further evidence that AV advantages vary strongly between individuals (Arnold & Hill, 2001).

[5] That is not to say that word context is not sufficient to facilitate phoneme recognition. The so-called word superiority effect holds even for L2 listeners (Hommel, 2018; Inceoglu, 2021).

intermediates (Oliver et al., 2012). These studies suggest that providing a larger reference frame for the listener might lead to advantages for L1 listeners as well as for advanced L2 learners. These different degrees of context-sensitivity are a further reason for removing sentence context in this study and choosing a comprehension situation in which the listeners cannot make use of contextual information and have to rely on what they hear and see.

2.1.5 Lipreading affects AV speech perception in different acoustic conditions

Some studies emphasize the effects of visual speech input by demonstrating that visual access to articulatory movements allows for the discrimination of languages and syllable identification in foreign languages, even in purely visual paradigms (Davis & Kim, 2006; Ronquest & Hernandez, 2005; Ronquest, Levi & Pisoni, 2007; Soto-Faraco et al., 2007). However, AV speech perception research often includes several levels of signal degradation. One of the most influential studies by Sumby and Pollack (1954) shows that speech intelligibility under noisy conditions increases when the face of the speaker is visible compared to speech comprehension with purely auditory input. Indeed, the majority of studies find lipreading benefits for L1 and L2 listeners is greatest in moderate noise (Drijvers & Özyürek, 2017; Drijvers & Özyürek, 2020; Fitzpatrick & Kim, 2010; Ross et al., 2007). Xie et al. (2014) support this finding for English L1 listeners, but find greater facilitatory effects of AV speech at higher signal-to-noise ratios for L2 listeners, which suggest that L2 listeners benefit more from lipreading cues, but only in contexts with less noise. Interestingly, neuroimaging studies show that the pSTS, claimed to be the integration area of multimodal input in the brain, is especially active when speech is presented in noise, whereas clear AV speech activates more anterior parts of the STS (Ozker et al., 2017).

Another factor is the number of potential sources of disturbance and the degree to which the signal is degraded. The principle of inverse effectiveness (PoIE) in AV speech perception states that when the quality of unimodal input is decreasing, multimodal integration effects are increasing, as it was described first by Stein and Meredith (1993). The PoIE is supported by studies showing that visual speech cues become more effective for the listener when the speech signal is degraded, or when there is a larger range of possible utterances out of which the subject

has to choose a response (Sumby & Pollack, 1954). However, Blackburn et al. (2019) show that the more background talkers are present, the less listeners benefit from lipreading information, which suggests that there might be limitations to the PoIE. In clear speech, the benefits from seeing lip movements of the speaker remain unchanged regardless of how many background talkers are interfering (Blackburn et al., 2019).

Even though the effects of AV speech seem to emerge especially under noise for hearing individuals, effects are observable even in clear speech. Hardison (1999) presents evidence for significant improvement of sound perception for L2 listeners in clear speech input. Davis and Kim (2004) add that seeing the face of the speaker facilitates syllable monitoring compared to audio-only speech, but also enables a more reliable estimation regarding the duration of speech stimuli. Traunmüller and Öhrström (2007) show that AV benefits can be gained of L1 listeners of Swedish even in clear speech. As described in section 2.1.4.3, other studies provide additional evidence that speech comprehension is improved even in clear speech, for example when the information is complex (Arnold & Hill, 2001; Reisberg et al., 1987). The more demanding and complex a task, the more listeners rely on the AV benefit (Arnold & Hill, 2001). Nath & Beauchamp (2011) reinforce these findings by demonstrating that less degradation on both the auditory and visual levels of presented words and syllables lead to higher activation of the pSTS, the integration area of AV speech perception.

2.1.6 How L1 can affect speech comprehension in L2

2.1.6.1 L1 phoneme inventories affect lipreading in L2 speech perception

There is empirical evidence that the L1 phoneme inventory determines to some extent how L2 learners process visual input. Werker et al. (1992) show that the English interdental fricative /ð/ was confused with /d/ or /t/ by French native listeners, suggesting that L2 listeners adjust their percepts according to the phonology of their L1. Hazan et al. (2002) present sound contrasts that differed visually (e.g. /b/ and /v/) and those that do not differ visually (e.g. /b/ and /p/) to Spanish L2 learners of English, revealing that the visually overt sound contrast is perceived by some L2 learners, but also suggested that learners at earlier stages of L2 acquisition are not as sensitive to

visual cues as more advanced learners. They further state that the contrast of /b/ and /v/ has an allophonic status in the participants' L1 Spanish and can, therefore, not be discriminated as easily as a contrast of two sounds with phonemic status (Hazan et al., 2002). Ortega-Llebaria, Faulkner and Hazan (2001) show that Spanish L2 learners of English make fewer errors concerning the recognition of manner and place of articulation when visual information from lipreading is available. Overall, the perception of unfamiliar sounds in L2 speech is more difficult than perceiving sounds in the L2 that also exist in the L1, and AV speech input can facilitate this process (Cutler et al., 2004; Hommel, 2018; Wang, Behne & Jiang, 2008a; Wang, Behne & Jiang, 2009). However, there may be differences concerning particular sound contrasts, in the way that more frequent unfamiliar contrasts are acquired faster than the infrequent unfamiliar ones (Best & Tyler, 2007; Hommel, 2018). Moreover, when identifying fricatives, Spanish subjects show similar strategies as English L1 subjects, whereas Dutch subjects differ in this respect, suggesting that similarities in Spanish and English phoneme inventories enable Spanish L2 learners of English to use similar strategies in sound identification (Cutler et al., 2007).

Language-specific properties such as the overall number of phonemes and the role of suprasegmental features, such as lexical tone, can also affect the extent to which listeners rely on visual cues in speech perception. Since Japanese L1 listeners, for example, need to distinguish only a relatively small number of phonemes compared to English L1 listeners, they do not rely as much on lipreading as English L1 listeners do (Sekiyama, 1997). In the same vein, listeners of a tonal L1 rely more on auditory information because it contains richer information (Sekiyama, 1997). These listening habits[6] formed during L1 acquisition can thus, also affect the L2 speech perception later on and force the L2 listener to develop new listening strategies.

There are two highly influential approaches to the issue of L2 sound acquisition, the *Speech Learning Model* (SLM) by Flege (1995) and the *Perceptual Assimilation Model* (PAM) by Best

[6] Crosslinguistic AV speech perception research has been aware of cultural and language-specific factors that affect gaze behavior as well. However, even though studies show that humans prefer to look at faces already in infancy (e.g., Frank, Amso & Johnson, 2014), and it has been assumed that culturally conditioned avoidance of gazing at a speaker's face in Chinese listeners might affect AV language processing, a recent study shows that the likelihood of the McGurk effect is similar in Chinese and American subjects (Magnotti et al., 2015).

(1994). As humans acquire their first language, the perceptual narrowing can pose difficulties later in life during L2 acquisition because the perception system specializes on the phonemes of the L1 (Soto-Faraco et al., 2012). Earlier approaches to the field, as the *Contrastive Analysis* approach, note that deviating structures in L1 and L2 lead to learning difficulties, by means of contrasting the two phoneme inventories in question (Lado, 1957; Moulton, 1962). In contrast to this approach, which exclusively uses the phonological level to account for L2 sound acquisition, following approaches such as the SLM and PAM consider interaction on both the phonetic and phonological level (Best & Tyler, 2007; Flege, 1995; see also Escudero & Boersma (2004) for an optimality theory approach).

The PAM states that when L2 listeners encounter an unfamiliar speech sound, they assimilate it to a familiar exemplar from their L1 phoneme inventory (Best, 1994). Even when they assimilate L2 sounds to a L1 phoneme, they can still perceive differences in the realization of these phonemes in the phonetic domain (Best & Tyler, 2007). Thus, the sound can be either *categorized* as a more or less good exemplar of the L1 phoneme, remain *uncategorized* because it is too deviant from any L1 phoneme, or be of the *Non-Assimilable* type (Best & Tyler, 2007).

The *Two Category* assimilation predicts an unproblematic discrimination when two L2 sounds are assimilated to two different L1 phonemes, whereas the *Single Category* assimilation assumes that two L2 sounds are equally poor or good exemplars of the same L1 phoneme, which results in poor discrimination (Best, 1994). In *Category Goodness* differences, the two sounds are both assimilated to the same L1 phoneme, but one of them is a better fit than the other, which results in intermediate to good discrimination performance (Best, 1994). Finally, the *Non-Assimilable* types are discriminated very well because the listener does not recognize the L2 sounds as a part of speech but as a non-linguistic sound, and thus, a comparison of speech vs. not speech (Best, 1994). There are cases in which only one of two L2 sounds is similar to a L1 phoneme (*Uncategorized-Categorized* assimilation), and cases in which both L2 sounds fail to be categorized (*Uncategorized-uncategorized* assimilation) (Best & Tyler, 2007). The former type is expected to be discriminated well because it is a contrast of a familiar and an unfamiliar sound, whereas the discrimination performance for the latter type depends on the degree of deviance from the L1 phonemes (Best & Tyler, 2007).

The phoneme level is only existent for the L1 in this initial PAM version. However, this version is based on observations of naïve listeners, whereas the PAM-L2 accounts for the developing phonological system of L2 phonemes as soon as learners gather experience in their target language, which is in line with the SLM (Best & Tyler, 2007). Also in line with SLM is the assumption that L2 listeners refine their L1 sound categories during L2 acquisition, adjusting the phonological system so that it fits both sound inventories the listener is exposed to, since both L1 and L2 categories share the same phonological space (Best & Tyler, 2007; Flege, 1995). In line with the SLM, the PAM-L2 states that perceptual learning mechanisms are used throughout the listeners' whole life span in both L1 and L2 development, and by combining the SLM and PAM views on perceptual objects, the PAM-L2 states that there are phonetic as well as gestural perceptual objects, which are used depending on the listeners' perception goals and attention (Best & Tyler, 2007).

2.1.6.2 The impact of L1 orthography on AV speech perception

The written modality has effects on the perception of speech as well. Orthographic consistency describes the extent to which phonemes are represented transparently by graphemes (Goswami, 2005). Instances in which the novel phoneme is represented by a novel grapheme, or by a novel combination of graphemes, support the L2 learner in the perception and production of L2 phonemes (Hommel, 2018). For example, the novel sound [ɧ] for German L2 learners of Swedish can be represented by the novel grapheme combination <sj>, whereas languages with a low grapheme-phoneme correspondence can cause difficulties for learners (Hommel, 2018). An example for this case is *-ough* in English, which is pronounced differently in *though*, *through* and *tough* (Hommel, 2018, p. 66). Such inconsistent (opaque) grapheme-to-phoneme mappings affect the development of reading abilities in children (Goswami, 2005), but also hinder L2 sound perception (Hommel, 2018). Another example relevant to this study is the novel phoneme /ɕ/ in for German L2 learners of Swedish, which is written with the graphemes <tj>, as in *tjena* 'hello', <kj> as in *kjol* 'skirt', and <k>, as in *köra* 'drive' (Riad, 2014). The latter may pose a challenge to the German L2 learner of Swedish, since <k> in German is always pronounced [k] (O'Brien & Fagan, 2016).

Erdener and Burnham (2005) show that visual cues from AV speech facilitate the production of non-native speech, and that Turkish subjects with a transparent L1 orthography rely more on orthographic cues than on AV cues in foreign speech comprehension when both types of cues are available. Their performance worsens when this orthographic information stems from a language with an opaque orthography (Erdener & Burnham, 2005). In contrast, Australian subjects focus more on AV cues than on orthographic cues, being primed by the opaque orthography of their L1 English (Erdener & Burnham, 2005). Since this study includes a listening task with a following written comprehension question, it may be the case that differences in the graphemic coding of phonemes between German and Swedish affects the L2 subjects' performance, possibly diminishing the AV benefits.

2.1.7 AV benefit – a matter of proficiency or other individual factors?

Both L1 and L2 listeners benefit from additional visual cues from lipreading. It is suggested that just as L1 listeners, L2 listeners pass different developmental stages regarding the ability to integrate lipreading in AV speech. This gives reason to suspect that L2 proficiency is a crucial factor for gaining AV speech benefits.

There is evidence that the benefit of AV speech is greater for L1 listeners, for example in speech comprehension in noise (Drijvers & Özyürek, 2020; Fitzpatrick & Kim, 2010). Direct comparisons of L1 and L2 listeners show that L1 listeners benefit more from visual cues than L2 listeners (Drijvers & Özyürek, 2020). This suggests that the higher the proficiency, the more listeners are able to extract information from visual cues. This assumption is supported by various studies of AV speech comprehension (Hazan et al., 2002; Sueyoshi & Hardison, 2005) as well as by studies of purely visual discrimination of languages. In a study by Öhrström et al. (2009), for example, Swedish and Finnish are visually distinguished by L1 and L2 listeners, with Swedish L1 listeners showing the highest scores, and L2 listeners of Swedish showing a relatively high score which was correlated with their use of Swedish. Drijvers, Vaitonytė and Özyürek (2019) show that manual gestures and lipreading reduce RTs of L1 and L2 listeners of Dutch, especially in noise. L1 listeners' gaze is more directed to the lips, whereas L2 listeners

focus more on the manual gestures, suggesting that L2 listeners benefit more from visual information that is connected to the semantic level of speech and not to the phonological information that is conveyed by the lip movements (Drijvers et al., 2019). Even though they compare a homogeneous L2 listener group with L1 listeners, they point out to the fact that experience may be crucial to the way gaze is directed to different visual articulators[7] (Drijvers et al., 2019). Another study by Xie, Yi and Chandrasekaran (2014) reinforces this assumption by showing that the facilitatory effect of AV speech input in Korean L2 listeners of English is positively correlated with L2 proficiency, but only when an English L1 speaker is talking. However, a correlation between AV speech perception and proficiency is further suggested by Wang et al. (2008b), showing that higher L2 proficiency in Mandarin L1 listeners is related to better performance in a phoneme identification task. The less-proficient L2 listeners rely more on visual cues when perceiving unfamiliar L2 sounds (Wang et al., 2008b). However, these effects are found between different proficiency groups, and not in correlation to a continuous scale of L2 proficiency, as attempted by this study.

However, even though phoneme inventories of the native languages and experience with a certain language play a role in AV speech perception of L2, lipreading abilities vary widely across individuals and can only be improved to a small extent (Summerfield, 1992). However, the study of Hardison (2003) shows that AV speech training yields higher identification scores of the English /ɹ/ – /l/ contrast in native listeners of Korean and Japanese compared to audio-only training (see also Hardison, 1999 for a similar study). Wang et al. (2008a) show that in the initial phase of L2 acquisition unimodal training in the auditory or visual modality can be more effective than AV training.

Another factor examined by previous studies is musical experience. For example, musically trained individuals exhibit different brain responses than individuals without musical training

[7] With increasing age, children's gaze to a speaker's mouth increases as well (Irwin, Brancazio & Volpe, 2017). Drijvers & Özyürek (2020) suggest that L1 listeners benefit more from AV speech than L2 listeners. However, both native and non-native listeners focus on lip movements in difficult listening conditions, and especially monolinguals and highly proficient L2 listeners gaze to the speaker's mouth when non-native speech is presented (Barenholtz, Mavica & Lewkowicz, 2016; Birulés et al, 2020). However, this degree of attention to the speaker's mouth does not correlate with the L2 listener's proficiency (Birulés et al., 2020).

when confronted with AV speech, which may account for some of the observed individual variations in early processing of AV speech (Sorati & Behne, 2019). Inceoglu (2019) shows that subjects' performance in a phonological short-term memory task can predict the L2 perception performance, but that is not the case for working memory. Traunmüller and Öhrström (2007) discuss influences of gender, since females are known to look more at a speaker's face and are said to be better lipreaders. These factors may explain why there is often a large variation in the data between the participants in AV speech perception studies.

2.1.8 Are consonants and vowels affected differently by lipreading information?

As described in the previous sections, visual speech can support the recognition of spoken language. Consonants can be distinguished because they have different places and manners of articulation, and thus, they have distinguishable visemes (Ladefoged, 1975). Vowels differ phonetically in terms of roundedness of the lips, tongue height and backness (Robert-Ribes et al., 1998). Even though AV training can facilitate the acquisition of quantitative phoneme length contrasts as well (Hirata & Kelly, 2010), this study focuses on qualitative phoneme contrasts. Given these fundamentally different properties of consonants and vowels, there is good reason to assume that they might affect speech perception in different ways and/or to different extents.

Ortega-Llebaria et al. (2001) show that consonant recognition of English consonants is facilitated by 3.7% for L2 learners and 5.7% for L1 listeners, whereas vowel recognition is not enhanced considerably. Lip movements reduce the likelihood of perception errors regarding manner and place of consonants, but L2 learners do not use voice onset time information from the visual cues, which may be due to lacking language experience and training (Ortega-Llebaria, 2001). In line with these findings, Mártony (1974) presents AV input in the form of pseudowords to Swedish L1 listeners, showing that vowels are perceived without major difficulties, whereas consonants are more prone to misperceptions. Difficulties in the perception of consonants concerns especially fricatives, whereas laterals and nasals are more resistant to errors in L2 perception of English by Dutch and Spanish subjects (Cutler et al., 2007).

Roundedness is a dominant feature on the visual level, whereas tongue height is most dominant on the auditory channel (Amcoff, 1970; Robert-Ribes et al., 1998). However, all of these three features are transferred better when presented audiovisually, which results in the higher identification scores of isolated French vowels (Robert-Ribes et al., 1998). Similarly, Traunmüller and Öhrström (2007) show that openness of vowels is perceived more on the auditory level, whereas the visual input has a greater impact in terms of roundedness, which is interpreted as evidence for the *information reliability hypothesis* that assumes that, in case of incongruent information, the more reliable modality determines if a phonetic feature is perceived by means of the auditory or visual information[8]. Moreover, the roundedness feature is a major source of vowel confusions for L1 listeners in audio-only conditions (Traunmüller & Öhrström, 2007). In Swedish, for example, there is a considerable number of rounded vowels in the vowel inventory. Considering that the roundedness feature is highly salient, this suggests that visual information may be of higher importance for speech perception in Swedish than in other languages.

Inceoglu (2021) presents evidence for stronger AV enhancement in the recognition of French vowels for L1 than for L2 listeners, and discusses potential effects of visual salience that may has an impact on these findings and that some sounds, such as nasal vowels in French, can be more difficult for L2 learners to acquire than others. Phonemes that are easier to distinguish from others in the visual modality, because they belong to different viseme categories, are visually more salient. These phonemes were more affected by AV enhancement than less salient visemes (Inceoglu, 2019). Indeed, visual salience has been already discussed by Amcoff (1970), who shows that certain consonant visemes of Swedish are harder to lipread in a purely visual task, especially those that are articulated further back in the vocal tract. In the speech sounds involving labial features, there are fewer confusions for both vowels and consonants (Amcoff, 1970). In sum, empirical evidence shows that both the perception of consonants and vowels can be facilitated by visual speech. However, it is less clear whether removing the visual domain in speech perception affects one of them more than the other.

[8] However, fusions and combinations might still be possible. Indeed, Traunmüller and Öhrström (2007) found McGurk-like fusions for vowels (auditory *geg* presented with visual *gyg* caused the perception of *gøg*).

2.2 How wearing a face mask affects communication and lipreading

In the context of the Covid-19 pandemic, face-to-face conversation is hindered in different ways. First, the social distance in general leads to situations in that speakers involved in a conversation are facing each other with a greater distance than usual. Secondly, according to the World Health Organization, the risk of spreading or becoming infected with the virus is significantly lower when face masks are worn on a regular basis (World Health Organization, 2020), which has led to several laws and recommendations to wear these masks in everyday life situations to cover mouth and nose. For that reason, it is more and more common to communicate without having visual access to the speaker's articulatory gestures. Whereas these communication scenarios are rather unfamiliar and challenging for Western cultures, covering parts of the face is more common in other regions of the world, for religious or cultural reasons (Ong, 2020), but also for medical reasons (Yang, 2014). Thirdly, the presence of a mask attenuates the speech signal (Mendel, Gardino & Atcherson, 2008; Atcherson et al., 2017).

These alterations of the speech signal can impede speech perception. Especially higher frequencies are mitigated, since the alternation caused by a face mask functions similar to a low-pass filter on the speech signal (Bottalico et al., 2020; Corey, Jones & Singer, 2020; Rahne et al., 2021). When masks serve as a kind of low-pass filter, this has a considerable impact on the comprehension of sibilants and consonants, which leads to an overall higher processing and listening effort (Rahne et al., 2021). At which frequencies this attenuation emerges, and to what extent face masks dampen the intensity of speech, depends, for example, on the kind of mask that is used. Surgical masks and N95 are not affecting the speech signal drastically, while cotton masks and KN95 masks distort the speech signal more (Bottalico et al., 2020; Corey et al., 2020). Transparent face masks, which have the advantage of showing the mouth of the speaker, show the strongest adverse effects on the speech signal, yielding distortions up to 14 Decibel (dB) (Corey et al., 2020). Face masks lower the root mean square, which is an indicator of loudness of speech (Atcherson et al., 2017). Bandaru et al. (2020) show that when masks and face shields are used, perception threshold increases 12.4 dB on average, which is problematic in

communication between patients and medical staff. However, Magee et al. (2020) add that voice quality features are maintained to a great degree, regardless of the mask type.

Face masks hide information from the mouth of the speaker. Behavioral studies on AV speech perception show how speech is understood in different listening conditions and with different degrees of visual input. While some of these studies demonstrate that the hidden visual information seems to be a smaller hindrance for speech understanding in normal-hearing individuals, while background noise is more problematic (Atcherson et al., 2017; Mendel et al., 2008), other studies showed that even normal-hearing listeners benefit from visual information when speech is presented in noise (Giovanelli et al., 2021; Thibodeau et al., 2021). Intelligibility of words can be hampered when face masks are involved across all age, gender, and different occupation groups (Bandaru et al., 2020; Bottalico et al., 2020). Hampton et al. (2020) place speaker and listener in the same room at a two-meter distance, masked and unmasked, showing that noise impedes speech comprehension, and that face masks lead to decreasing word comprehension scores, while raising the voice in turn improved the scores[9] (Hampton et al., 2020). Truong and Weber (2021) show that intelligibility of adult as well child speakers is significantly reduced to the same extent when they wear face masks, which has adversely affected German L1 listeners' performance in a recall task, even though children's voices have different F0s and often use different pronunciation patterns than adults. Finally, Smiljanic et al. (2021) show that the speech of non-native speakers was significantly less intelligible especially when the speaker was wearing a mask. Post-experimental questionnaires and surveys have revealed that people are aware of these effects and perceive a higher effort in listening and understanding and feel less confident, even if they perform well in the perception tasks (Atcherson et al., 2017; Giovanelli et al., 2021; Saunders, Jackson & Visram, 2021). Moreover, face masks have been shown to affect the recognition of emotions and nonverbal communication as well (Calbi et al., 2021; Carbon, 2020; Grundmann, Epstude & Scheibe, 2021; Mheidly et al., 2020; Spitzer, 2020; Wegrzyn, 2017).

[9] Speakers often adapt their speaking style according to the situation or background noise. Cohn, Pycha and Zellou (2021) show that more accurately articulated words are recognized better. Performances for clear speech were even better in the masked condition, which suggest that speakers adapt their speech when wearing a face mask (Cohn et al., 2021).

To conclude, face masks do not render speech recognition impossible. However, considering that they can cause a reduced speech signal in terms of intensity and perceived loudness, that they hide lipreading information, and that natural conversation can involve a certain level of background noise, talking with a masked face can result in higher processing effort. On the other hand, listeners can make use of various compensation strategies when articulatory gestures are not accessible. Spitzer (2020) stresses for instance, that listeners can use prosody and gestures. Moreover, voices are often raised automatically in noise or under face masks (Cohn et al., 2021). It remains questionable, however, whether these compensation strategies can maintain efficacy.

2.3 Completing the picture: Processing other visual information

AV speech perception does not only involve phonological information gained from lipreading and articulatory movements (and, in a wider sense, associated jaw and cheek movements), but also visual information from other parts of the face and the body. Along with lipreading, co-speech gestures, cues from the upper half of the face, and posture can provide useful information for the listener on the visual level.

2.3.1 Co-speech gestures

Co-speech gestures are hand movements that accompany speech, carried out spontaneously and not according to any rules of well-formedness (Gullberg, 2006). These gestures can have representational functions, reflecting the content of speech in the form of *iconic* or *metaphoric* gestures, the former related to specific actions or objects and the latter to abstract concepts (Kendon, 1986; McNeill, 1992). *Deictic* gestures are used to indicate objects or actions by pointing (McNeill, 1992). Other non-representational functions, for example in *beat* gestures, are related to the rhythmic properties of speech (Kendon, 1986; McNeill).

Semantic information from iconic co-speech gestures improves the recall of action verbs in spoken language comprehension in L1 listeners, and even advanced L2 listeners can benefit from co-speech gestures, although this effect is not as strong as for L1 listeners (Drijvers & Özyürek,

2017; Drijvers & Özyürek, 2020). Effects of co-speech gestures also apply to the perception of longer parts of speech. Sueyoshi and Hardison (2005) show that English L2 learners' performance in a comprehension task after listening to a coherent lecture declines progressively when reducing the amount of visual input from gestures and the speaker's face to the speaker's face only to audio-only speech input. When perceiving AV speech, L1 and L2 listeners show different gaze patterns, however. L1 listeners' is was more directed to the lips, whereas L2 listeners focus more on the manual gestures, suggesting that L2 listeners benefit more from visual information that is connected to the semantic level of speech and not to the phonological information that is conveyed by the lip movements (Drijvers et al., 2019).

2.3.2 Extraoral face movements as visual cues from the upper part of the face

Visual speech information on a speaker's face is not limited to mouth movements only. Instead, the movement of the articulatory organs causes movement in cheeks and jaw. Beyond that, eyebrow and head movements accompany spoken language as well.

Thomas and Jordan (2004) examine the effects of different face regions on speech recognition by editing certain parts of the face so that these appear static, revealing that seeing the full face of a speaker facilitates speech perception as efficiently as seeing lip movements only. As long as mouth movements are visible, other parts of the face can be removed without significantly impairing perception, but this is not the case for extraoral movements, and these effects persist even when the face is presented upside-down (Thomas & Jordan, 2004). Even though this underlines the importance of lipreading as a facilitatory strategy, facilitating effects can be found for the visibility of the nose and the eyes as well, suggesting that the latter serve as anchor points to perceive other related movements of the face (Thomas & Jordan, 2004).

The upper part of a face and the outline of a speaker's head serve as cues for prosodic properties of speech (Cvejic, Kim & Davis, 2010). Similarly, other studies show that comprehension is better in AV conditions with visible head movements compared to unimodal auditory conditions, and that eyebrow raising can be a cue for word prominence (Granström, House & Lundeberg, 1999; Munhall et al., 2004a). As mentioned earlier, gaze can be directed to different articulators.

Thus, it can also be directed to different parts of the face. When prosodic information is needed, the upper side of the face is focused, whereas the lower part is focused when segmental decisions are required (Lansing & McConkie, 1999). Davis and Kim (2006) conclude that "[…] the effect of head movements might interact with that of the mouth and jaw to produce an amplified audio-visual advantage" (pp. B29-B30). Moreover, listeners do not need to fixate the mouth with their gaze in order to extract linguistic information from this region and, for example, exhibit the McGurk effect, but attention is required to benefit from visual cues (Paré et al., 2002). The current study, however, does not examine the prosodic level of speech. Therefore, it is more likely that the upper face cues are not sufficient to facilitate speech comprehension.

2.4 The current study

The vast literature allows for postulating hypotheses, that can be tested in a psycholinguistic experiment implementing a speech comprehension task. Listeners are presented with speech input through films in three conditions: The audiovisual (AV) condition shows the full face of the speaker, the blurred (B) condition hides the articulatory gestures from the lower part of the face, and the audio-only (A) condition presents speech without visual information. Participants' performance in the perception task is examined by measuring error rates of their answers to comprehension questions and the respective RTs.

The phoneme inventories of Swedish and German are to a large extent similar. However, German L2 learners of Swedish are confronted with the novel vowel phoneme /ʉ/ occurring for example in the word *nu* 'now', and with the novel consonantal phonemes /ɕ/ and /ʂ/[10], as in the words *köra,* [ɕœːra] 'drive' and *kurs*, [kʊʂː] 'course', respectively[11] (O'Brien & Fagan, 2016; Riad, 2014). Moreover, the Swedish [y] is characterized by a larger lip opening, which is a reason why both sounds sound different and the Swedish [y] sounds different than [y] in German

[10] Some approaches use the symbol /ʃj/ instead of /ʂ/ for this phoneme. However, Riad (2014) choose /ʂ/, also because it is the allophone that occurs in both quantities in Central Swedish (short and long).

[11] Differences occur also in the quantity of sounds (vowels), but since these quantitative differences are not an objective of this study, they were not described in detail here (for further elaborations on vowel quantity in German see O'Brien & Fagan, 2016; for vowel quantity in Swedish see Riad, 2014).

(Lindau, 1978). This makes L2 subjects with a native background in German an interesting participant group for this study.

Swedish and German are similar with regards to syllable structures, and both languages belong to the more transparent systems in comparison to other European languages, albeit the German orthography is classified in the group that is slightly more transparent with regards to the grapheme-phoneme mapping than Swedish (Seymour, Aro & Erskine, 2003). This suggests that German and Swedish listeners, conditioned by their L1 orthography, are relatively sensitive to orthographic influences, which causes trouble in the case of different phonetic realizations of the same grapheme, for example <g> in words like *göra* 'do' which is pronounced [j] before front vowels in Swedish, whereas <g> in a German word with a similar phonological context, such as *göttlich* 'divine', is realized as [g]. Preceding other vowels or consonants, <g> is realized as [g], for example in *gul* 'yellow', *glad* 'happy', or as [ɧ] as for example in *generad* 'embarrassed'. Differences like these can potentially pose an extra challenge to the L2 listener when they carry out the comprehension task.

In contrast to the majority of studies, this study examines AV speech effects in clear speech. Previous studies have shown that AV effects can be observed even without noise degradation (Arnold & Hill, 2001; Davis & Kim, 2004; Hardison, 1999; Reisberg et al., 1987; Traunmüller & Öhrström, 2007). Moreover, this variable of noise degradation is left out also for technical reasons due to the online design of the study. Although there is already a lot of evidence showing that visual speech contributes positively to speech perception, this question is taken up here again because it serves as a starting point for looking more deeply at the consequences of covering the mouth and nose, and also regarding language proficiency. Based on the empirical findings described in the previous sections, the research questions (RQs) 1 and 2 and the corresponding hypotheses (H) 1a, 1b, 2a and 2b are motivated by studies suggesting that visible speech facilitates speech comprehension in various settings and circumstances (Arnold & Hill, 2001; Cotton, 1935; Drijvers & Özyürek, 2017; Hardison, 1999; Kim et al., 2015; Navarra & Soto-Faraco, 2007; Reisberg et al., 1987; Sueyoshi & Hardison, 2005; Sumby & Pollack, 1954; Traunmüller & Öhrström, 2007; van Wassenhove et al., 2005).

RQ_1 Does the visibility of the speaker's face facilitate speech comprehension in Swedish L1 and L2 listeners?

➔ H1a: L1 and L2 listeners reach lower error rates in the AV condition than in the audio-only condition.

➔ H1b: L1 and L2 listeners show faster RTs in the AV condition than in the audio-only condition.

RQ_2 Does the visibility of the full face facilitate speech comprehension in Swedish L1 and L2 listeners to the same extent as the visibility of a speaker with a covered mouth and nose?

➔ H2a: L1 and L2 listeners reach lower error rates in the AV condition than in the blurred condition.

➔ H2b: L1 and L2 listeners show faster RTs in the AV condition than in the blurred condition.

RQ_3 Is there a difference between the effect of visual input on the recognition of consonants and on the recognition of vowels in AV speech perception of Swedish?

RQ_3 is treated as an explorative question in this study since the primary goal is to examine overall effects of visible speech and the material is not explicitly constructed for this purpose. Moreover, the literature suggests both advantages for vowels (Inceoglu, 2021; Robert-Ribes et al., 1998; Traunmüller & Öhrström, 2007) and consonants (Cutler et al., 2007; Mártony, 1974; Ortega-Llebaria et al., 2001). Therefore, no hypothesis is formulated regarding this question.

RQ_4 Does the proficiency level of the L2 listeners correlate with the effect of visible speech on speech comprehension?

RQ_4 examines potential correlations between the proficiency level of the L2 listeners on a continuous scale and the perception of speech in different modalities. Since language proficiency comprises many different aspects, the grouping of participants in different proficiency groups is rather complicated. Depending on the factors that are measured to determine proficiency, establishing groups of proficiency in a categorial way is at risk of skewing the reality, especially when an intermediate group is involved, because intermediate groups can comprise a large range

of actual proficiencies. For instance, abilities related to vocabulary and/or grammatical skills can be more or less developed in an intermediate stage. Moreover, while some previous studies show that especially beginners benefit from AV speech in L2 acquisition when they perceive unfamiliar sounds (Wang et al, 2008b), there are also studies showing that highly proficient listeners can make more effective use of AV speech (Drijvers & Özyürek, 2020; Hazan et al., 2002; Öhrström et al., 2007; Sueyoshi & Hardison, 2005; Wang et al., 2008b; Xie et al., 2014). Therefore, RQ_4 is treated exploratively in this study.

3. Method

This section focuses on the methods chosen to investigate the facilitatory effects of visual information on speech perception. In a psycholinguistic experiment, the amount of visual information that accompanies the speech signal can be manipulated but also controlled. The listeners were confronted with speech with or without the visible face of the speaker and answered comprehension questions afterwards. The performance in behavioural tasks can indicate the amount of processing effort. Therefore, the error rates of participants' answers to comprehension questions as well as the respective RTs were measured as the dependent variables. Due to the ongoing Covid-19 pandemic, this study required an online setup to suit the safety recommendations. The recordings of the stimulus material were done in the Lund University Humanities Lab following the social distance recommendations, and the data collection itself was carried out online.

3.1 Participants

Thirty-two users of Swedish living in Sweden participated in the study (8 males, mean age *M*= 40.2, *SD*=10.88, see Table 1 for an overview of the participants). Twenty-two participants were L2 speakers of Swedish with German as their L1 (3 males, *M*=41.9, *SD*=10.35)[12]. Ten participants were native speakers of Swedish (5 males, *M*=36.4, *SD*=12.4). Thus, there was a between-subject variable, namely the status of Swedish as either the L1 or L2 of the participant. The reasons for choosing Swedish listeners were practical, due to accessibility of these participants. In order to keep the L2 listeners homogenous, only L1 speakers of German were chosen. This specific target group was also easily accessible in Sweden, which facilitated the recruitment. Moreover, German L1 speakers are said to acquire Swedish relatively quickly because of the typological proximity, which is why they can reach higher levels of proficiency in

[12] Data of four further L2 participants were excluded due to technical problems during the experiment procedure. They are not included in the numbers indicated here.

a shorter time compared to L2 learners from typologically distant backgrounds. This increased the chance that they can master the comprehension task.

	L1	*L2*
sum	10[13]	22
gender/ female	5	19
male	5	3
mean age	36.4	41.9
mean proficiency score		0.863
proficiency minimum		0.71
maximum		1.00

Table 1. Participant data

The participants were recruited using the snowball sampling technique. A couple of target groups was addressed directly on social media platforms by contacting acquaintances, language-related groups, and groups with potentially high numbers of German L2 learners of Swedish. To gather a sufficient number of participants, some of the previously participating subjects forwarded the recruitment text (Appendix A) to other potential participants. The recruitment text was addressed to Swedish native speakers between eighteen and sixty years of age, and to German native speakers with good to very good knowledge of Swedish being between eighteen and sixty years of age. The age limit was chosen to avoid potential interfering effects due to age-related decrease of hearing capacity (Huang & Tang, 2010).

3.2 Background tests

3.2.1 Language history questionnaire

To assess the L2 participants' proficiency in Swedish and to address the question of correlation between proficiency and the benefits of visual cues in speech comprehension tested in the behavioral task, the Language History Questionnaire 3.0 (LHQ; URL:

[13] One participant from the L1 group grew up trilingual.

https://lhq3.herokuapp.com/) was used (Li, Zhang, Yu & Zhao, 2020). It allowed for quantification of language experience based on an algorithm that aggregates scores for language dominance, proficiency, and immersion (Li et al., 2020). The questionnaire was generated itemized, so that less relevant aspects and questions were skipped, and the duration of the questionnaire was reduced to approximately ten to fifteen minutes (see Appendix B for the complete questionnaire). Participants provided personal information and information about their native language(s) and other languages. This included personal information such as age, gender, educational background, handedness, country of origin and residence. Furthermore, the age of acquisition (AoA) for the respective languages concerning reading, writing, listening, and speaking were factored in, as well as considerable stays abroad and the language use during this period. Moreover, participants were asked to provide information on how they learned their non-native languages, via classroom instruction, self-learning and/or immersion[14]. They also provided information on specific environments in which they use the respective languages (at home, at work, or online etc.). They rated their ability to acquire new languages and finally, rated their current ability in the respective languages concerning reading, writing, listening, and speaking on a scale from 1 to 7. For this study, the function of aggregated scores for language proficiency was used.

A score between 0 and 1 was generated, 1 signaling the highest proficiency. The self-assessment of the participants in terms of reading, writing, speaking, and listening was weighted with 25 percent for each ability[15]. This score served as a continuous variable for the correlation analysis.

[14] Immersion, as used by Li et al. (2020), is language learning by immigrating to a country or region where the target language is spoken.

[15] Involved in the task were listening and reading only, which is why a higher weighting for these abilities was considered. However, since most participants assessed themselves as being better at perception than at production of language, and therefore indicate higher scores for reading and listening, ranking these abilities higher would have led to a more compact proficiency score range, which would have been of disadvantage for this study. Thus, each modality (listening, reading, writing, speaking) was weighted with 25%, so that a wider range of proficiency scores enabled a more informative correlation analysis.

3.2.2 Word test

In this study, L2 learners of Swedish were asked to answer comprehension questions. For that reason, it needed to be ensured that the L2 listeners are familiar with the target words. Otherwise, answering the question incorrectly might not have been due to manipulation of the independent variable, that is the amount of visual information, but rather reflected a general lack of vocabulary. Therefore, a word test was administered after the L2 participants had completed the experiment. The isolated target words were presented to them without sentence context, and they were asked to translate these words into their L1, German.

3.3 Experimental stimulus materials

3.3.1 Minimal pairs and target words

Minimal pairs are two words that differ in a single phoneme only, leading to a change in meaning (Warren, 2013). The difference in the phoneme can be characterized by a difference in a consonant, an example being the pair *lever* 'lives' – *leder* 'leads' in Swedish. Minimal pairs can also involve quantitative phoneme contrasts, for example in the Swedish pair *kall* 'cold' – *kal* 'bald'. An example for a qualitative vowel contrast is the pair *dag* 'day' – *deg* 'dough'. Minimal pairs were chosen as the testing domain because they reflect rather small changes in the speech signal that can change the meaning of the whole utterance. This posed a comprehension challenge to the listener[16]. However, even though visual information can be helpful to distinguish quantitative as well as qualitative contrasts concerning sound length (see Hirata & Kelly, 2010), this study focused on qualitative contrasts only. Choosing qualitative contrasts had the advantage that the visual signal differed in the way this sound looked on the speaker's mouth. In

[16] It has to be noted here that these minimal pairs may differ in more than one grapheme on the orthographic level, for example in the minimal pair *hästen* 'the horse' – *festen* 'the party'.

accordance with the viseme groups found in the study of Amcoff (1970), minimal pairs were constructed so that the two sounds belong to different viseme groups[17].

As the minimal pairs, eighteen pairs of three word categories (verbs, nouns, and adjectives/adverbs/pronouns[18]) were chosen. In order to replace one word of the minimal pair with its counterpart within the same sentence, the words needed to be held consistent in terms of grammatical aspects, including word category, number, grammatical gender, or inflectional ending. To ensure that the L2 listeners are familiar with the target words, word frequency was checked in the Swedish Parole Corpus using SketchEngine (Kilgarriff et al., 2014)[19]. According to Brysbaert, Mandera and Keuleers (2018), words with a relative frequency below five hits per million are considered to be low-frequent. Therefore, only target words with a frequency above five hits per million were chosen. Moreover, in a pilot study these words were checked by two L2 learners of Swedish and replaced by more common target words if one of them did not know the word.

3.3.2 Sentence structure

In order to investigate the effect of visual information on speech perception and the recognition of phonemes, specific stimulus sentences were constructed. By controlling linguistic parameters such as tense, mood, aspect, or sentence structure, these stimulus sentences ensured that the observed response to the stimulus was due to the independent variable instead of being a result of varying tenses or differences in the complexity of sentence structures. A total number of thirty-six stimulus sentences was constructed in order to create a comprehension situation for the listeners that required them to perceive the speech up to the lowest level, namely single

[17] This applies to the consonants, whereas for vowels it was to find pairs that involve roundedness to achieve a certain degree of visual salience. Exceptions from this were the pairs *reta* 'annoy' – *rita* 'draw' and *dag* 'day' – *deg* 'dough'.

[18] The majority of the target words in the third group were adjectives, except for *nu* 'now', *bort* 'away' (adverbs) and *ingen* 'none' (pronoun)

[19] In this corpus, the relative frequency per one million words was checked by typing in the word in the simple concordance search field, exactly in the inflected/uninflected form in which they were used in the stimulus sentences. For words like *låt* 'song', which was used as a noun but is also part of the verb *låta* 'to let', additional specifications were made to check the noun frequency only.

phonemes. For this reason, eighteen minimal pairs were chosen as the target words in the stimulus sentences.

To make the stimulus sentences comparable, the sentences were all active sentences without subordinated clauses, and they are between 11 and 13 words in length. Eighteen of the thirty-six sentences were in the present tense[20], the other eighteen sentences in the simple past (preterite). Using a consistent sentence structure in the form of *There is a X …* was avoided for the most part to vary the positions of the target word more and make its position less predictable for the listeners. Only four of the thirty-six stimulus sentences had a structure of this type.

It was ensured that the target word was neither the first nor last word of the sentence. This way, it was avoided that the target phoneme stands out because it occurs as the onset or offset of the stimulus sentence. A sentence context was chosen so that the target word can be plausibly replaced by the other word from the minimal pair. Since the sentences were manipulated in a way that makes both words from the minimal pair plausible, the listeners could not make use of semantic context to answer the comprehension questions. An example can be found in (1).

(1) Mannen *lever* med en positiv inställning.
'The man *lives* with a positive attitude.'
Mannen *leder* med en positiv inställning.
'The man *leads* with a positive attitude.'

3.3.3 Comprehension questions

The comprehension questions for the stimulus sentences had three answer options each and were formulated so that the correct answer was the target word *lever* 'lives', as in example (2). The second option was the minimal pair associate which was phonologically related and highly similar to the target word *leder* 'leads'. The third answer option was a word that is a semantically plausible answer but clearly distinct in sound and phonological aspects, in this case *jobbar*

[20] Eight of the present tense sentences involved a modal verb + infinitive construction, the other ten present tense sentences involve a simple present form.

‘works’. The position of the correct answer as the first, second or third option was varied throughout the experiment. The complete list of all stimulus sentences and the corresponding comprehension questions can be found in Appendix C.

(2) Vad gör mannen? ‘What does the man do?’

1. lever med en positiv inställning ’lives with a positive attitude’
2. leder med en positiv inställning ’leads with a positive attitude’
3. jobbar med en positiv inställning ’works with a positive attitude’

3.3.4 Fillers

Along with the eighteen stimulus sentence pairs, eighteen filler sentence pairs were created to draw the participants’ attention away from the minimal pair distinctions. The comprehension questions for these sentences had either two or three answer options, some of them were yes-or-no and right-or-wrong questions. These questions were not aiming for minimal pair contrasts but were rather related to the understanding of prepositions, tense, or word order. Example (3) is related to prepositions (see Appendix D for the complete list of filler items).

(3) Marie dricker kaffe utan mjölk. ‘Marie drinks coffee without milk.’
Finns det mjölk i Maries kaffekopp? ‘Is there milk in Marie’s cup?’
1. Ja ‘yes’ 2. Nej ‘no’

3.3.5 Recordings

The stimulus and filler sentences were recorded in the LARM studio in the Lund University Humanities Lab. The speaker was a female native speaker of Central Swedish. She was instructed to speak in a moderate tempo, to keep a neutral facial expression, and to use a neutral declarative intonation without stressing a certain word or phrase in the sentence. The speaker sat in front of a black background. Seventy-two sentences were recorded in 4K by a Sony PXW-Z100 camera and edited in full HD, with a Sennheiser MKE600 microphone.

These video recordings served as the AV video stimuli. For the audio-only condition, the sound was extracted from these videos[21]. The blurred videos were created by editing the video so that mouth and nose of the speaker as well as parts of the jaw and cheek were blurred to a degree of five hundred percent. The idea was that listeners were able to see a difference between the initial and contrasting phoneme in the target words *hästen* 'the horse' and *festen* 'the party' (Figure 1) in the AV condition, whereas this visual input was hidden in the blurred condition (Figure 2).

Figure 1. Target phonemes /h/ in the word *hästen* 'the horse' (left) and /f/ in the word *festen* 'the party' (right)

[21] A visual-only condition, as it can be found in other studies, was not included here, because the nature of the task is not compatible with a visual-only condition, and also because visual-only input is a rather peculiar input for hearing individuals.

Figure 2. Blurred condition stimulus

3.4 Design and procedure

Participants listened to and looked at the video and audio files, respectively, on a screen, followed by a forced-choice comprehension question that forced them to choose what they perceived in the stimulus sentence.

3.4.1 Apparatus

The recorded sound files and videos were inserted as the stimuli in the experiment using the PsychoPy software (v2021.1.4) (Peirce et al., 2019). The builder interface was used to create the experiment routines. The experiment was then uploaded to the Pavlovia website (URL: https://pavlovia.org/) which allowed for a link to the experiment to be forwarded to the participants who then ran the experiment online on their own computers. The experiment opened in their web browser and after entering their participant IDs, a start screen with the most important summarized instructions appeared (see Appendix E for figures of the screen display). The PsychoPy software was programmed to play the stimuli in a loop that was randomized. The participants were instructed to answer the comprehension question following each stimulus by pressing 1, 2 or 3 on their keyboards. The audio and video files were played only once, and the following question appeared immediately afterwards. This question was visible until the moment the subject pressed one of the answer keys 1, 2 or 3 (Figure 3). Subsequently the next stimulus

appeared after a short break of one second. After half of the stimuli, a screen showed up signaling the break.

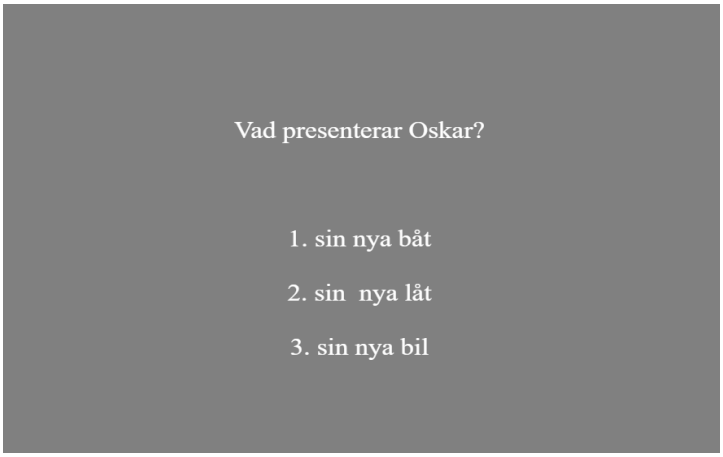


Figure 3. Question screen

Studies about AV speech integration often use background noise to create more realistic and more difficult listening conditions. Since the participants went through the experimental sessions at home and not under laboratory conditions, there was no control for the volume, earphones, or potential background noise. Therefore, the noise factor was excluded from this study. However, the sentence constructions described above confronted the listeners with a perceptual challenge even in clear speech by withdrawing higher-level context information. Other factors that could not be controlled for included participants' distance from or the size of their screen.

3.4.2 Experiment design

Altogether, 216 sentences constituted the material for the behavioral task. This material was split up in two parts for two main reasons: Firstly, the task load as well as the experiment duration for each participant was halved. Secondly, running the sessions online in a web browser worked smoothly up to a certain number of media files. Especially video files increased the browser load which is why reducing the stimulus material for the experiment session prevented delays, asynchrony of audio-track and videos, and malfunctions.

For this purpose, two versions of the experiment were compiled (see Appendix F). Both versions involved all three levels of the independent variable of modality, that is audio-only as well as the blurred and non-edited AV stimuli. Moreover, both versions were compiled so that each participant encountered both target words from the minimal pair for all minimal pairs. Therefore, effects of modality on speech comprehension could be examined. This also implied that both versions contained the same number of vowel and consonant pairs, respectively, which was of importance for the investigation of RQ_3. Each version contained eighteen audio stimuli, eighteen AV stimuli and eighteen blurred stimuli, leading to fifty-four stimulus sentences. Filler sentences were also divided in a way that both versions had the same number of audio-only (A), audiovisual (AV) and blurred (B) video stimuli, respectively. Each version of the experiment consisted of 108 sentences in total, that were presented randomized for each participant.

According to this division, participants were assigned to either version 1 or version 2 in alternation. To counterbalance, it was ensured that the same number of participants completed the experiment in version 1 and version 2. Each experimental trial consisted of a stimulus sentence in one of the three conditions (A, AV, and B) followed by a comprehension question with two or three answer options.

Moreover, each version was divided into two halves. Splitting up the experiment into two halves (and thus, two weblinks) minimized the memory load on the browser and reduced technical issues when loading the video files. The participants went through both halves within one experiment meeting, with a break in between.

3.4.3 Procedure

The link to the LHQ was sent to the participants and they filled in this questionnaire prior to the experiment meeting, which took approximately ten to fifteen minutes. Experiment meetings were organized via Zoom, in which participants first gave oral consent to participate in the study. Afterwards, the procedure was explained, and questions were answered. Once participants were ready to start, they received the link to the first part of the perception experiment. The break screen instructed the participants to go back to the Zoom meeting in order to receive the link to

the second half of the experiment. All participants completed this part of the experiment in about twenty to twenty-five minutes, except for one participant who encountered some file loading difficulties. Stimuli were presented synchronously and complete, albeit with some delays between the single items. For the L2 listeners, a word test followed after the experiment to ensure that they are familiar with the target words. The word test took about three to five minutes. Finally, after the procedure all participants were debriefed about the exact purpose of the study.

3.4.4 Ethics

Participants were informed that their participation in the study is voluntary, that they have the right to withdraw their participation at any time, and that they are guaranteed anonymity. To avoid participant bias, they did not get a detailed description of the study's aspects beforehand. During the recruitment phase they were merely informed that they participate in a speech perception experiment regarding the comprehension of spoken Swedish language (see Appendix A). They gave informed oral consent which was recorded during the experimental session and received compensation for their participation in the form of a discount code. After they performed the experimental task, they were informed in more detail about which sentences are analyzed and were debriefed about the study's exact purpose. Informed consent was also provided by the speaker in the recordings who agreed that the recordings may be used for the experiment and as illustrations in reports and presentations of this project.

3.5 Data treatment, coding, and analysis

3.5.1 PsychoPy data, error rates and response times

In order to investigate the effect of the modality variable on speech perception, the incorrect answers of the participants as well as the RTs were measured as the dependent variables. An erroneous answer to the comprehension question reflected the incorrect perception of the crucial phoneme and hence, the target word. RTs can indicate the amount of processing load that the participant had to manage in order to react to a certain stimulus, longer response time therefore indicating difficulties in processing (Kaiser, 2013).

In order to derive the error rates, it was measured for each participant how many of the eighteen questions in each condition (A, AV, B) were answered incorrectly. This number was then converted into a percentage. Each participant was presented with 18 stimuli in each of the three conditions. Hence, if 2 out of 18 questions were answered incorrectly in condition A, the corresponding error rate was calculated as 11,1%. Error rates were chosen instead of accuracy rates because of the ceiling effects observed in the participants' performance.

As described above, each participant was shown 18 stimuli in each of the three conditions. In order to investigate RQ_3 the dataset was divided by vocalic and consonantal minimal pairs. Hence, there are 9 vowel and 9 consonant stimuli per condition. If 2 out of 9 questions were answered incorrectly in condition A, this yielded a corresponding error rate of 22,2%.

The output files generated by PsychoPy listed the comprehension questions and the respective answer options together with the selected answer of the participant (Figure 4). For instance, in line 237 the question *Vad stämmer in på cirkusen?* 'What applies to the circus?' was listed. V to the left indicated that the target word involved a vowel contrast in the minimal pair, and A indicated the audio-only condition. The number in quotation marks showed the selected answer, which was 1. Below the selected answer, in line 241, it was indicated that *2* was the correct answer to this question. Therefore, the number *0* to the right side of the selected answer showed

that the answer was incorrect[22]. The RT was measured from the offset of the stimulus and the display of the question until the moment the participant pressed one of the answer keys. The RTs were presented in seconds, for the question in Figure 4, for example, it was 6,699 seconds[23] (6699 ms). For each participant, average RTs for the three conditions were calculated.

237			"1"	0	6,699	0	51	51	21	1	0256.mp4	adjective	V	A	Vad stämmer in på cirkusen?			
238																		
239	1. är ny i staden																	
240	2. är nu i staden																	
241	3. är här i staden	2	namf8	1	2021-05-1	test3	2021,1,4	MacIntel	14,92537									

Figure 4. PsychoPy output file

3.5.2 Word test data

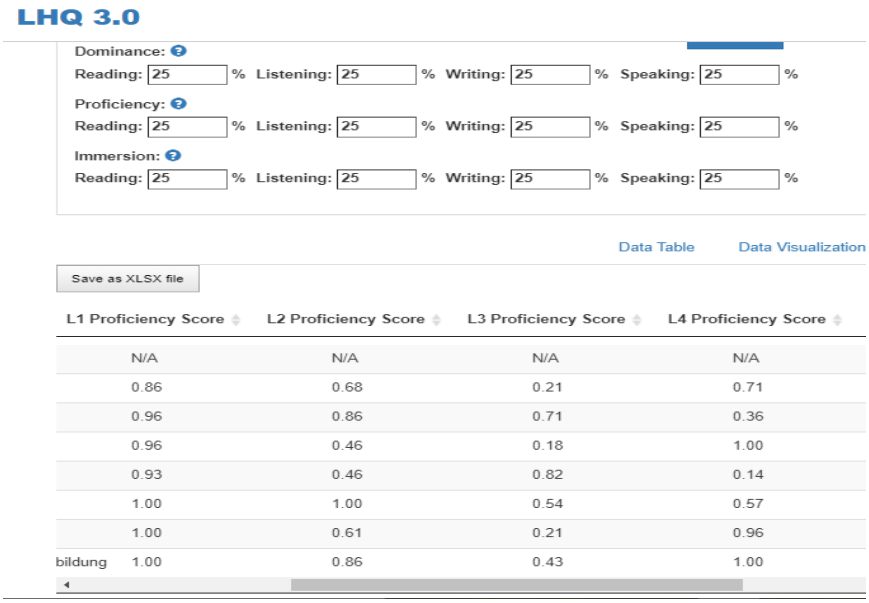
The target words that were not correctly translated by the L2 speakers in the word test were noted and subsequently excluded from the analysis for the particular participants. For example, if the word *reta* 'annoy' was not named correctly, the minimal pair *reta* 'annoy' – *rita* 'draw' was removed from the analysis for this participant, reducing the total number of stimuli per condition from 18 to 17 (since they were presented with only one version of the minimal pair in each condition)[24].

3.5.3 Language History Questionnaire data

In order to compile the proficiency scores that were used for the correlation analysis for RQ_4, question 15 was used by the proficiency score algorithm of the LHQ, by adding the values from 1 to 7 with regards to speaking, writing, listening, and reading (Li et al., 2020). The proficiency

[22] 1 was displayed here when the answer is correct.
[23] Outliers, for example caused by technical issues, were excluded from the analysis.
[24] This approach was chosen because the purpose of the analysis was to investigate whether the sound contrasts can be perceived with or without AV information. Therefore, even when a participant was acquainted with one word from the minimal pair, assertions regarding this sound contrast could not be made any longer, which is why both stimulus words were excluded.

score was calculated automatically and could be retrieved under the *aggregate score* interface page (Figure 5).



5. Proficiency scores in the LHQ interface

3.5.4 Statistics

Since there was a within-subject variable (modality) and a between-subject variable (group L1 or L2), the data were analyzed by performing two separate within-subject analyses for L1 and L2 listeners, respectively. Descriptive statistics were carried out to identify the basic characteristics of the dataset related to error rates and RTs. Inferential statistics were carried out in order to test the experiment data for significance. For this purpose, non-parametric repeated measure analyses of variance (Friedman tests) were carried out for the error rate and RT data of L1 and L2 listeners, separately. Eventually, the analyses of L1 and L2 listeners were compared with one another, but only descriptively. The software used for the descriptive and inferential statistics was Jamovi version 1.8.2 (The Jamovi project, 2021, https://www.jamovi.org/)

The calculated error rates and average RTs were further divided into vowel and consonant data in order to investigate potential differences between this factor in this study. Based on this

divided dataset, the non-parametric repeated Friedman tests were carried out once again in order to examine potentially significant differences between the perception of vowels and consonants. These results were compared descriptively for L1 and L2 listeners as well.

In order to examine if there was a connection between language proficiency and AV speech perception in L2 listeners, the data from the LHQ were linked to the error rates and RTs that were calculated before, by performing a correlation analysis.

4. Results

In this chapter the results are presented, starting with the effect of the modality variable for RQ_1 and RQ_2, and followed by RQ_3 that separately considered the vowel and consonant data. In the course of these analyses, the L1 subgroup was analysed first, followed by the L2 listener data, and a descriptive comparison of both groups. Finally, a correlation analysis for the L2 listeners examined RQ_4.

4.1 The effect of the modality variable on error rates and RTs

4.1.1 L1 listeners

On average, the ten L1 subjects made more mistakes in the audio-only (A) condition (*M*=2.2%) than in the audiovisual (AV) (*M*=1.1%) and the blurred (B) condition (*M*=1.1%) (Figure 6). However, the maximum error rates for each participant never exceeded 11.1%, and the median value was 0 in all three conditions, which emphasizes the floor effect.

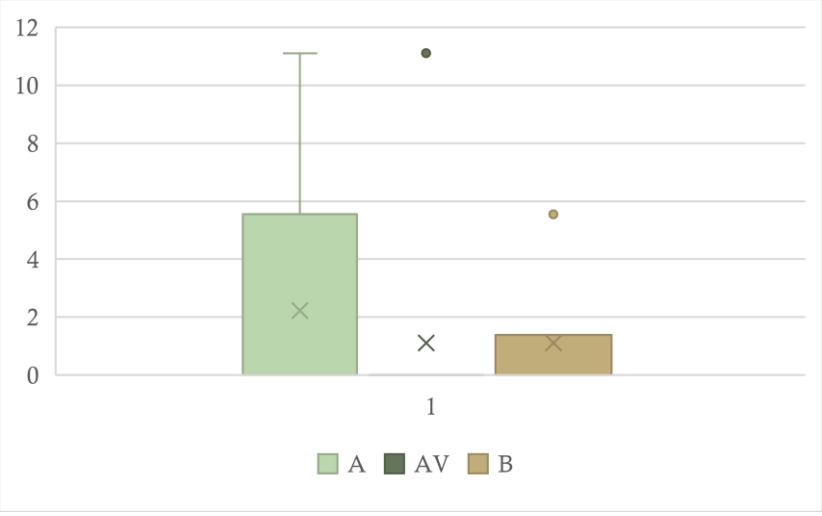


Figure 6. Boxplot of the mean error rates in % by condition (A=audio-only, AV=audiovisual, B=blurred) in the L1 listener group

The mean RTs were highest in the B condition (*M*=3519 ms), whereas the mean RTs in the AV (*M*=3318 ms) and A (*M*=3219 ms) condition were lower in comparison (Figure 7). The mean values for error rates and RTs, displayed in Figure 6 and Figure 7, respectively, therefore present a mirror image of one another. As indicated by Figure 7, there were considerable individual differences regarding the mean RTs in the three conditions.

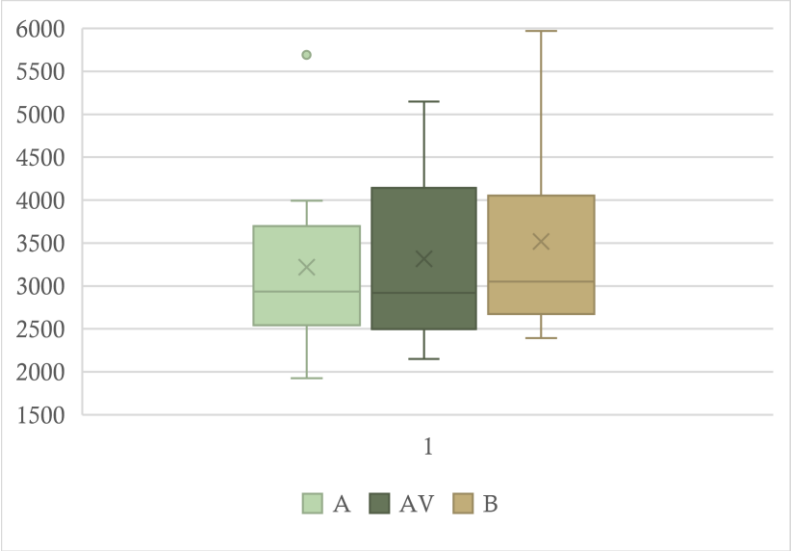


Figure 7. Boxplot of the mean RTs in ms by condition (A=audio-only, AV=audiovisual, B=blurred) in the L1 listener group

Sphericity checks revealed that data were not normally distributed. Thus, a non-parametric repeated measures Friedman test was conducted. There was no significant effect of the modality variable on the error rates of the L1 listeners, $\chi^2(2)=1.4$, $p=0.497$, and no significant effect of modality on the RTs either, $\chi^2(2)=0.8$, $p=0.67$.

4.1.2 L2 listeners

The twenty-two subjects made more mistakes in the A (*M*=2.0%) and B (*M*=2.0%) condition compared to the AV condition (*M*=1.0%) (Figure 8). However, the maximum error rates per participant in the A and B condition only reached 11.1%, compared to 5.9% in the AV condition.

The median value of 0 in all three conditions indicates that there was a floor effect in terms of error rates.

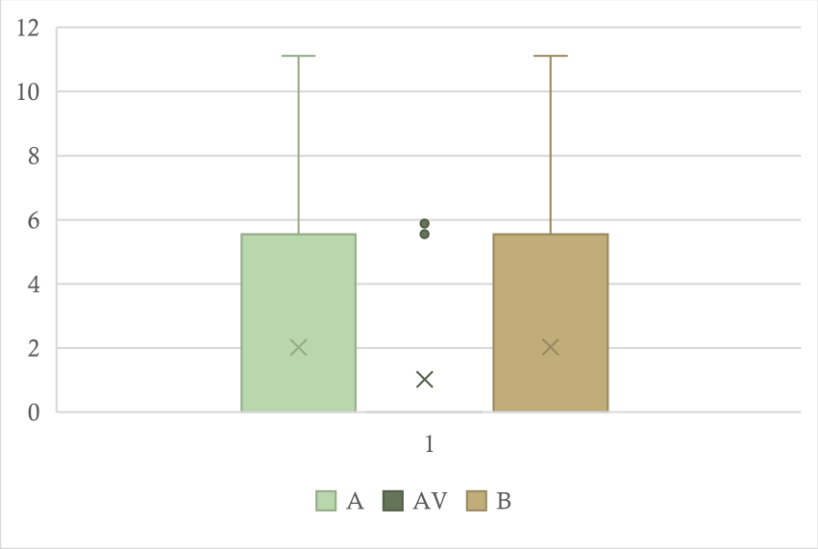


Figure 8. Boxplot of the mean error rates in % by condition (A=audio-only, AV=audiovisual, B=blurred) in the L2 listener group

The mean RTs were highest in the B condition (*M*=3634 ms), followed by the A condition (*M*=3437 ms), and the AV condition (*M*=3249 ms) (Figure 9). Thus, the means of error rates were parallel to the means of the RTs, both were lowest in the AV condition. Considerable individual variances were observable regarding the mean RTs in the three conditions (Figure 9).

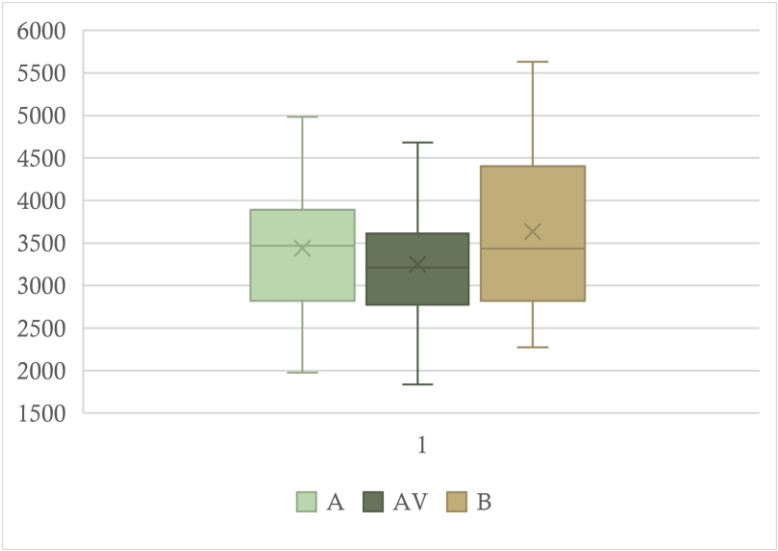


Figure 9. Boxplot of the mean RTs in ms by condition (A=audio-only, AV=audiovisual, B=blurred) in the L2 listener group

The sphericity checks revealed that the L2 listener data were not normally distributed. The Friedman test showed no significant effect of the modality variable on error rates for the L2 listener group, $\chi^2(2)=1.51$, $p=0.469$. However, a significant difference was found with regards to the RTs, $\chi^2(2)=9.09$, $p=0.011$. Pairwise Durbin-Conover comparisons revealed that the difference was significant between condition AV and B ($p=0.002$).

4.1.3 Comparison of the modality effect on L1 and L2 listeners

The descriptive comparison of the results for the L1 and the L2 listener groups is displayed in Figure 10, showing that the only difference regarding the error rates was found in the B condition, where error rates are higher for L2 than for L1 listeners. Overall, L2 listeners made more errors than L1 listeners.

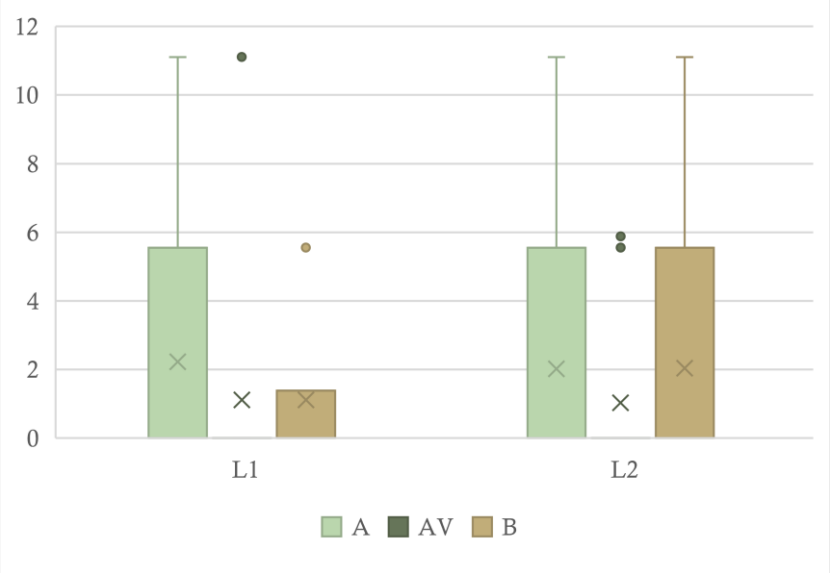


Figure 10. Boxplot of the mean error rates in % by condition (A=audio-only, AV=audiovisual, B=blurred) for L1 and L2 listeners

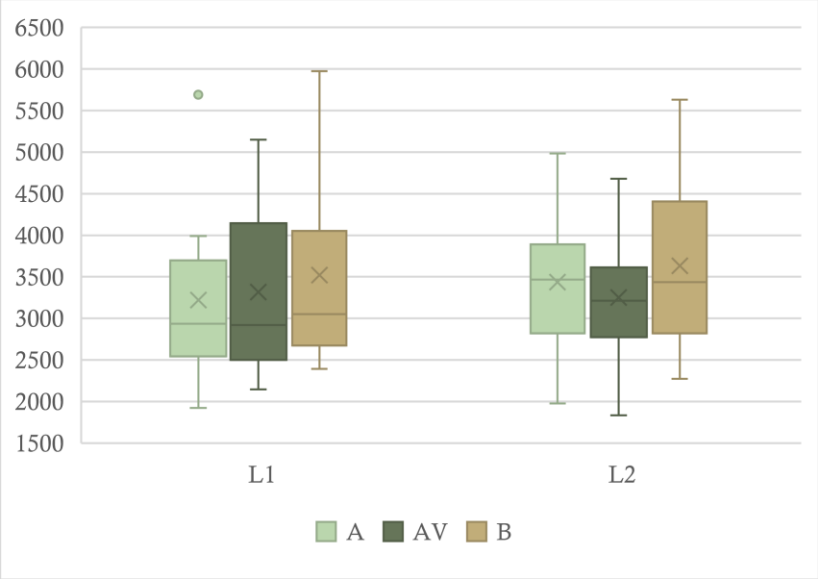


Figure 11. Boxplot of the mean RTs in ms by condition (A=audio-only, AV=audiovisual, B=blurred) for L1 and L2 listeners

As displayed in Figure 11, L1 responses were fastest in the A condition, followed by the AV condition and the slowest responses were found in the B condition, whereas L2 responses were fastest in the AV condition and comparatively slower in the A and B condition. Thus, L1 and L2 listeners showed different patterns regarding the RTs in the three conditions.

4.2 The effect of the modality variable on error rates and RTs of vocalic and consonantal minimal pairs

Figure 12 presents error rates in L1 and L2 groups by items. For both L1 and L2 listeners, more errors occurred in the consonant targets than in the vowel targets. However, a further inspection of the results showed that this was mainly due to the minimal pair *lever* 'lives' – *leder* 'leads', which caused 59.2% of the errors overall, and 76.2% of the errors in the consonant data[25] (Figure

[25]This might be due to the position of the target phonemes in the words, which was within the word and not at its beginning, as in the other target pairs. The position of the critical phonemes within the target words might make them more prone to coarticulation. In other words, the other consonantal minimal pairs might elicit fewer errors because word-initial phonemes are more salient and were perhaps articulated more precisely.

12). Leaving aside this target pair, the total sums of errors were clearly more balanced, and even slightly higher for the vowels (∑=6) than for the consonants (∑=5).

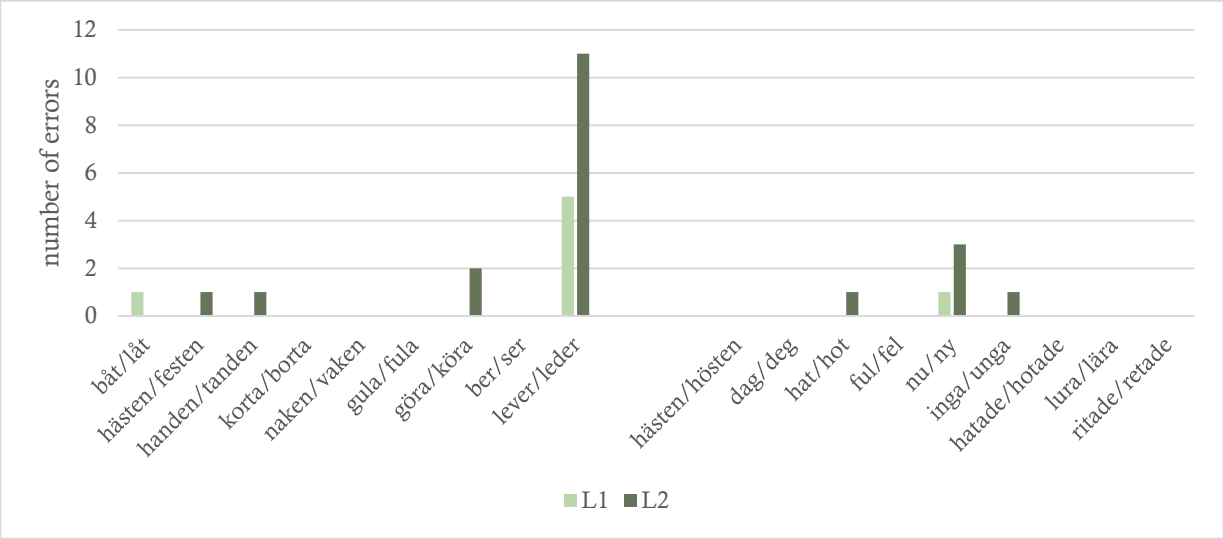


Figure 12. Errors grouped by vocalic and consonantal minimal pairs and by L1/L2 group collapsed across conditions

4.2.1 L1 listeners

Errors in the comprehension of vowel targets occurred only in the A condition (*M*=2.2%) (Figure 13), whereas the same error rate of 2.2% was found in all three modality conditions of the consonantal targets (Figure 14).

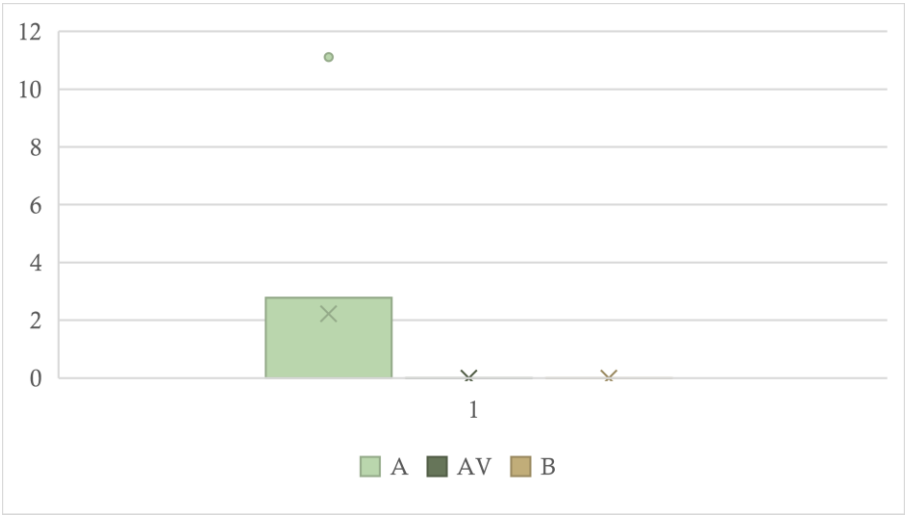


Figure 13. Boxplot of the mean error rates in % by condition (A=audio-only, AV=audiovisual, B=blurred) in the L1 listeners (vowel targets)

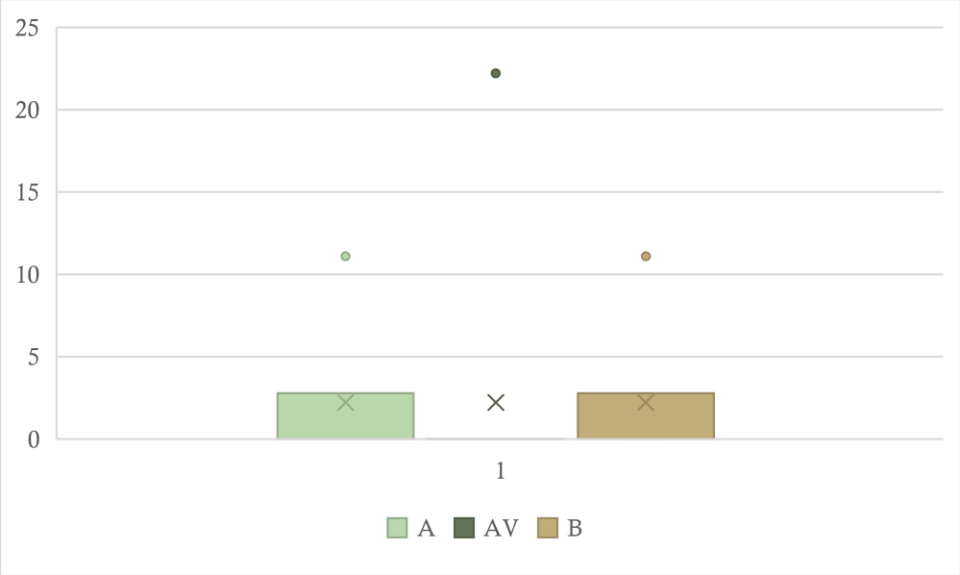


Figure 14. Boxplot of the mean error rates in % by condition (A=audio-only, AV=audiovisual, B=blurred) in the L1 listeners (consonant targets)

The RTs were faster for the consonantal targets in all three conditions compared to the vocalic targets. In the vowel group, RTs were fastest in the A condition (*M*=3334 ms), followed by the AV condition (*M*=3346 ms), with a greater distance to the B condition (*M*=3456 ms) (Figure 15). The RTs in the consonant group followed the same pattern, being fastest in the A condition (*M*=3105 ms), followed by the AV condition (*M*=3324 ms) and the B condition showing the slowest responses (*M*=3401 ms) (Figure 16).

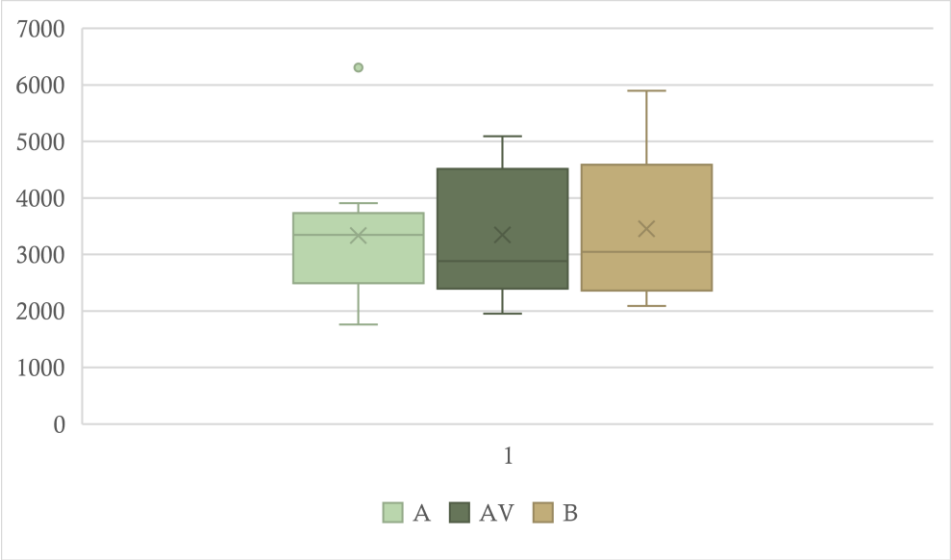


Figure 15. Boxplot of the mean RTs in ms by condition (A=audio-only, AV=audiovisual, B=blurred) in the L1 listeners (vowel targets)

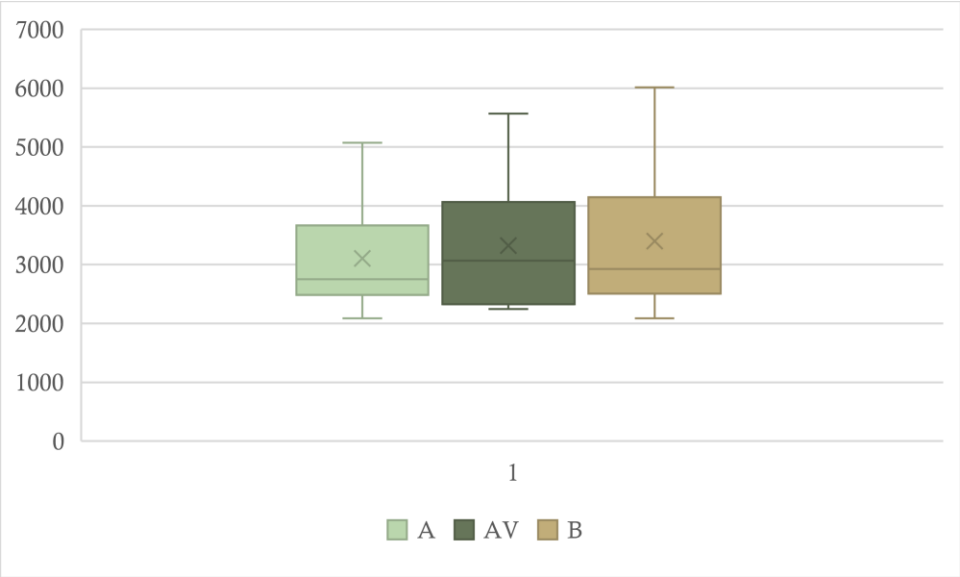


Figure 16. Boxplot of the mean RTs in ms by condition (A=audio-only, AV=audiovisual, B=blurred) in the L1 listeners (consonant targets)

Within the vowel group, the Friedman test did not show any significant differences between the modality conditions in terms of error rates, $\chi^2(2)=4.0$, $p=0.135$, nor in terms of RTs, $\chi^2(2)=0.2$, $p=0.905$. Within the consonant group, the modality variable yielded no significant effect, neither on the error rates ($\chi^2(2)=0.2$, $p=0.905$) nor on the RTs ($\chi^2(2)=0.2$, $p=0.905$).

4.2.2 L2 listeners

In the vowel group, most errors occurred AV condition (M=1.1%) and the B condition (M=1.0%), followed by the A condition (M=0.5%) (Figure 17).[26] In the consonant targets, most errors were found in the A condition (M=3.5%), followed by the B condition (M=3.0%) and, with a greater distance, the AV condition, which showed the lowest error rates (M=1.0%) (Figure 18).

[26] While this might seem odd, the median was 0 for all three conditions, showing a strong floor effect. Moreover, the descriptive findings reached no statistical significance.

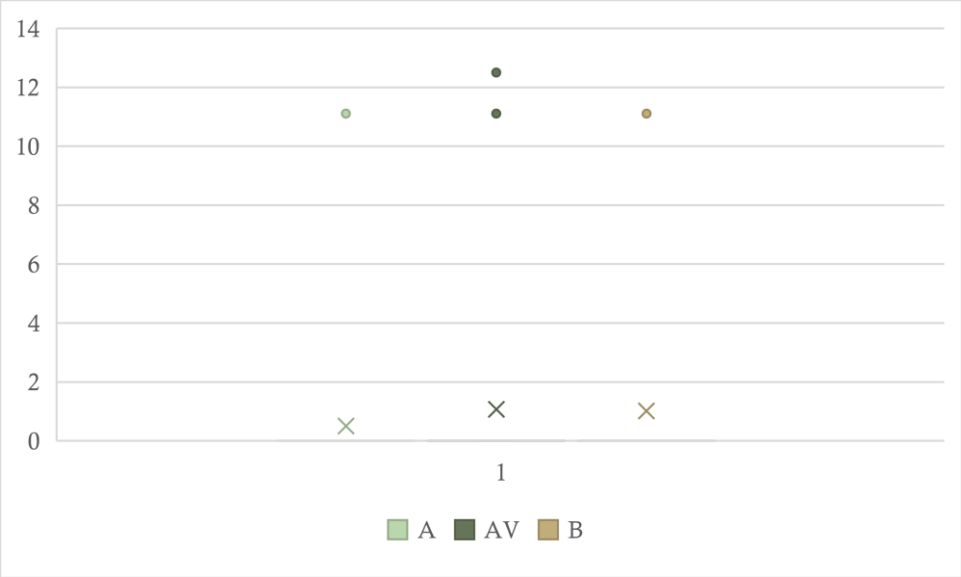


Figure 17. Boxplot of the mean error rates in % by condition (A=audio-only, AV=audiovisual, B=blurred) in the L2 listeners (vowel targets)

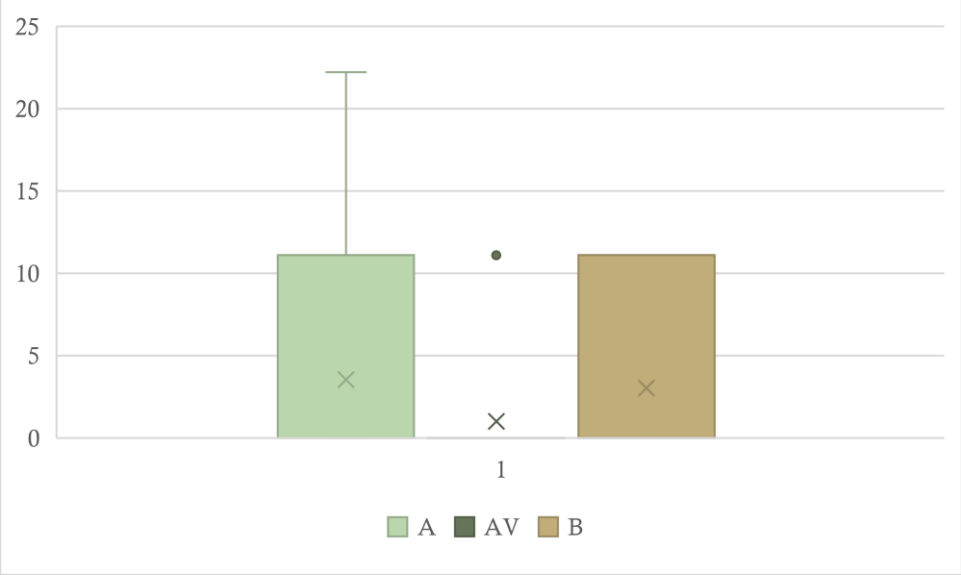


Figure 18. Boxplot of the mean error rates in % by condition (A=audio-only, AV=audiovisual, B=blurred) in the L2 listeners (consonant targets)

Overall, the RTs were faster for the consonants than for the vowels. Within the vowel group, RTs were fastest in the AV condition (*M*=3342 ms), followed by the A condition (*M*=3495 ms), and the B condition yielding the slowest responses (*M*=3709 ms) (Figure 19). The RTs in the consonant group followed this pattern as well, with condition AV yielding the fastest responses (*M*=3158 ms), followed by condition A (*M*=3372 ms) and condition B (*M*=3479 ms) (Figure 20).

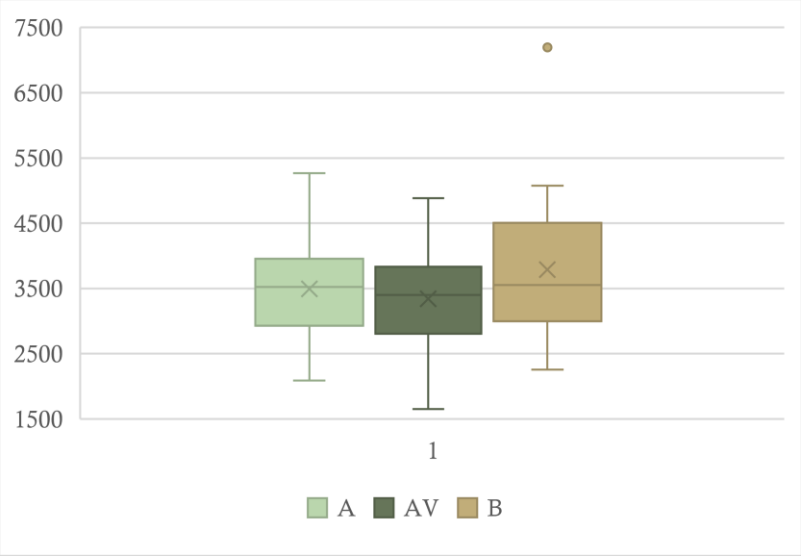


Figure 19. Boxplot of the mean RTs in ms by condition (A=audio-only, AV=audiovisual, B=blurred) in the L2 listeners (vowel targets)

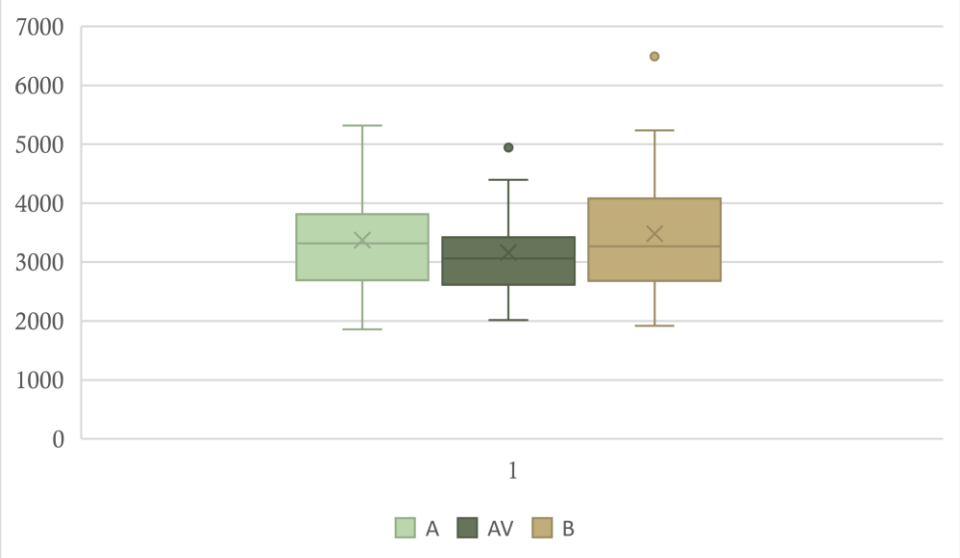


Figure 20. Boxplot of the mean RTs in ms by condition (A=audio-only, AV=audiovisual, B=blurred) in the L2 listeners (consonant targets)

Within the consonant group, the modality variable reached no significant effect for the L2 listeners, neither in terms of error rates ($\chi^2(2)=4.2$, $p=0.122$) nor in terms of RTs ($\chi^2(2)=2.82$, $p=0.244$). For the error rates of the vowel data, the Friedman test showed no significant differences between the modality conditions, $\chi^2(2)=0.4$, $p=0.819$. As the overall analysis of the modality variable showed in section 4.1, there were significant differences between the RTs in the AV and B condition for L2 listeners. Since the Friedman test showed no significant effect for the RTs of the consonant targets, it seemed likely that the observed effect in the overall analysis in section 4.1 is due to effects in the vowel targets. Indeed, there was a significant difference in the RTs between the conditions, $\chi^2(2)=8.82$, $p=0.012$. Post-hoc Durbin-Conover comparisons showed that this difference lies between the condition AV and B ($p=0.003$) and between the conditions A and B ($p=0.026$).

Figure 12 shows that especially the minimal pair *ny* 'new' – *nu* 'now' was prone to errors. This suggests that these target phonemes required more time to be recognized in the perception task. A follow-up inspection of the RTs for this minimal pair revealed that the RTs indeed followed the pattern that the AV condition yielded the quickest responses ($M=3578$ ms), whereas RTs were clearly slower in the A condition ($M=5033$ ms) and in the B condition ($M=6221$ ms). Unsurprisingly, these means were above the mean RTs from all vowel data and above the means from the overall RTs. Moreover, a Friedman test applied to the RTs of these target phonemes showed a significant effect of modality, $\chi^2(2)=6.63$, $p=0.036$. The post-hoc Durbin-Conover comparison revealed that a significant difference exists between the conditions A and AV ($p=0.044$) as well as between the conditions AV and B ($p=0.013$).

4.2.3 Comparison of the vowel and consonant analysis of L1 and L2 listeners

The descriptive comparison of the two subject groups revealed that in terms of error rates, there were no statistically significant differences for neither consonant nor vowel targets. Considering the insignificant results for the overall error rates in section 4.1, this result only supports the strong floor effects of the error rates, which did not reach 5% for neither of the conditions in vowels nor consonant targets, for neither L1 nor L2 listeners. The cross symbols in Figure 21 indicate these mean error rates.

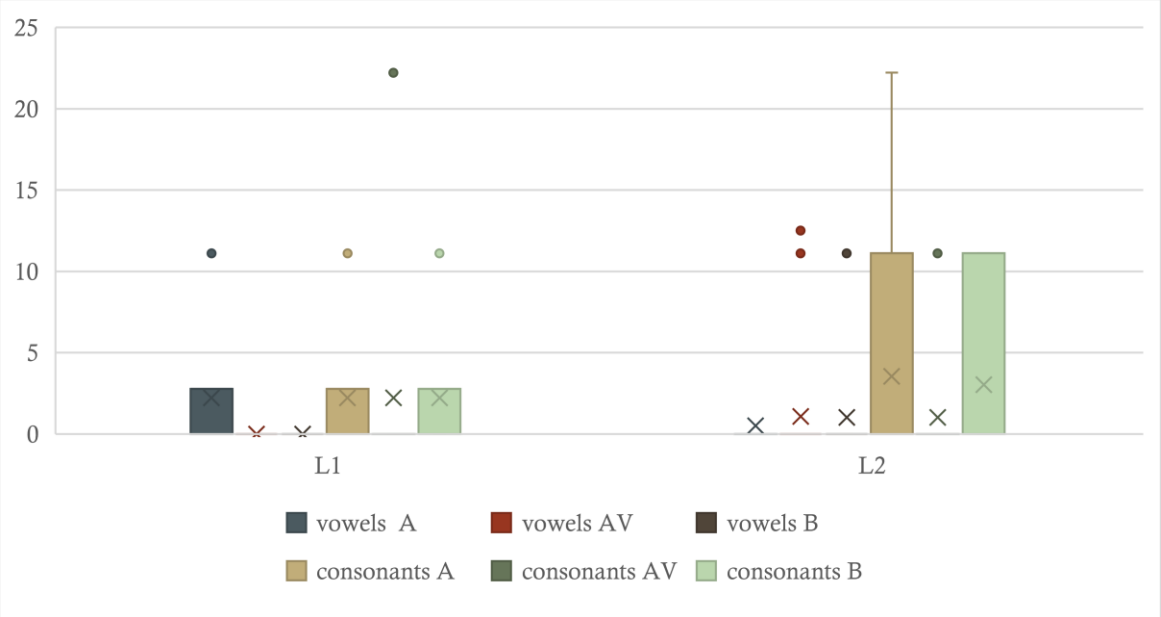


Figure 21. Mean error rates in % grouped by vowel and consonant targets, by condition (A=audio-only, AV=audiovisual, B=blurred), and by L1/L2 group

In terms of RTs, there was no significant effect in neither the consonant nor vowel targets for L1 listeners. However, for L2 listeners the visual speech input had an effect on the RTs in vowel recognition. Figure 22 displays the RT data side by side for the L1 and L2 listeners.

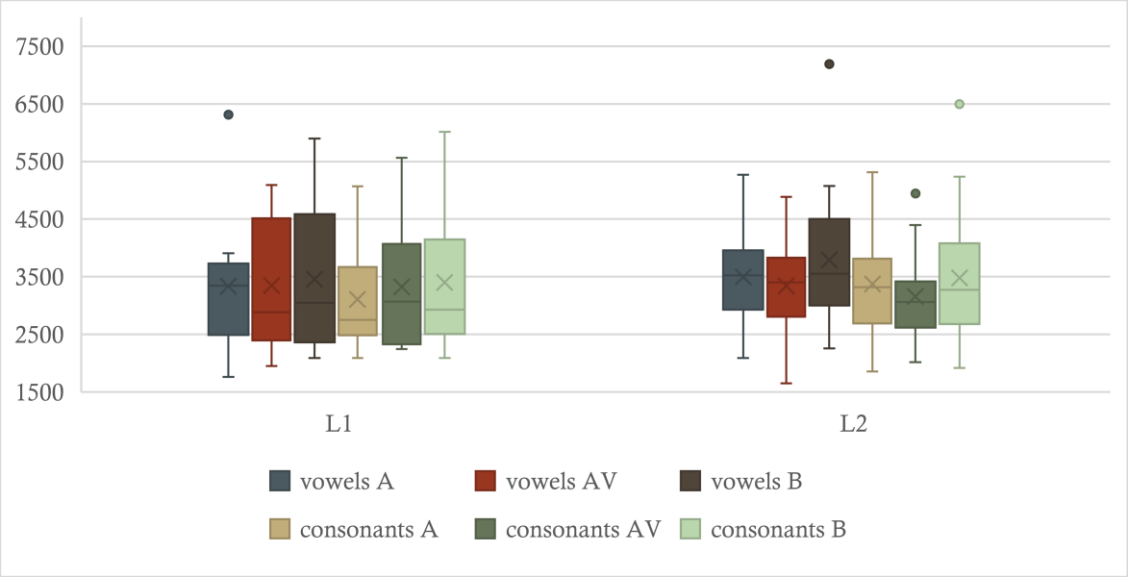


Figure 22. Mean RTs in ms grouped by vowel and consonant targets, by condition (A=audio-only, AV=audiovisual, B=blurred), and by L1/L2 group

4.3 Correlation between proficiency and the effect of the modality

For the correlation analysis, a potential connection between L2 proficiency and error rates was examined. The L2 proficiency was neither significantly correlated with the error rates in the A condition (r=-0.016, p=0.945), nor with the error rates in the AV condition (r=-0.164, p=0.465), nor the error rates in the B condition (r=-0.036, p=0.874).

The correlation analysis for the RTs showed no significant correlation between L2 proficiency and the RTs in the A condition (r=-0.183, p=0.415), AV condition (r=-0.075, p=0.74), nor the B condition (r=-0.136, p=0.547). However, the analysis in section 4.2 gave rise to the speculation that a correlation might be found between L2 proficiency and the RTs for the vowel targets. To this end, a follow-up correlation analysis was conducted with RT data of the vowel targets of L2 listeners. However, there was no significant correlation for these data either, neither for the A

condition (r=-0.052, p=0.818), nor for the AV condition (r=-0.054, p=0.811) and B condition (r=-0.179, p=0.424).

A further follow-up analysis considering the AoA showed no significant correlations with neither error rates in the conditions A (r=0.374, p=0.086), AV (r=0.235, p=0.293) and B (r=0.295, p=0.182), nor RTs in all three conditions A (r=0.183, p=0.414), AV (r=0.086, p=0.704) and B (r=0.103, p=0.647).

5. Discussion

This study investigated the effect of visual information from articulatory gestures on speech comprehension of L1 and L2 listeners of Swedish. AV information facilitated the recognition of phonemes in clear speech only for L2 listeners. This facilitation was not reflected in decreased error rates, but in the form of a temporal facilitation, as reflected in the RTs. The L2 listeners were supported by visual information in the recognition of vowel phonemes, and the results pointed to a considerable importance for non-native sound contrasts. Moreover, the correlation analysis did not yield significant correlations between the L2 proficiency and the processing of AV speech. This chapter elaborates on these findings, finishing with some considerations concerning the method of the study.

5.1 AV information facilitated comprehension in clear speech for L2 listeners

The following hypotheses were tested to examine differences between the conditions with audio-only (A), audiovisual (AV) and blurred (B) stimuli:

H1a: L1 and L2 listeners reach lower error rates in the AV condition than in the audio-only condition.

H1b: L1 and L2 listeners show faster RTs in the AV condition than in the audio-only condition.

H2a: L1 and L2 listeners reach lower error rates in the AV condition than in the blurred condition.

H2b: L1 and L2 listeners show faster RTs in the AV condition than in the blurred condition.

5.1.1 AV information from the speaker's face did not facilitate speech comprehension in L1 listeners

The results obtained in this experiment showed that the visual information in the form of the speaker's face did not significantly facilitate speech comprehension in L1 listeners. There were no differences in the error rates between the AV and A condition, and neither between the AV and B condition. Thus, H1a and H2a were not supported by the data. The most natural explanation for this finding might be that the task was not difficult enough for the L1 listeners to elicit perception errors, especially in the absence of noise. This finding is in line with studies showing considerable differences caused by the accessibility of lip movements when speech was presented in noise (Drijvers & Özyürek, 2017; Ross et al., 2007; Sumby & Pollack, 1954).

H1b and H2b assumed faster RTs in the AV condition compared to the A and B condition, respectively. This was not confirmed by the data for the L1 listeners either. Interestingly, this RT pattern was the mirror image to the pattern found in the error rates. This was rather unexpected, since conditions with higher error rates were expected to elicit longer RTs as well, as a sign of higher processing effort. In the data obtained in this experiment, the opposite was the case. The low number of participants might be a reason for this unexpected finding, since deviances in very few cases could already cause greater variances. Indeed, within the small group of L1 listeners (N=10), two participants showed higher error rates compared to the others, which might be the reason for this rather unexpected finding.

5.1.2 AV information from the speaker's face decreased response times in L2 listeners

The data obtained in the experiment showed that the AV condition did not yield significantly lower error rates compared to the A and B condition, and thus, H1a and H2a were not corroborated by the data, respectively. All in all, this means that as for the L1 listeners, the L2 listeners did not benefit from the AV input in the form of significantly increased comprehension scores. This is in line with studies suggesting that L2 listeners rely more on manual gestures than visual cues from the lips (Drijvers et al., 2019; Drijvers & Özyürek, 2020) and with studies showing that L2 listeners also benefit more from AV information when speech was presented in

noise (Fitzpatrick & Kim, 2010; Xie et al., 2014). However, it contradicts findings that more advanced listeners benefit more from visible speech than intermediate learners (Sueyoshi & Hardison, 2005). All L2 participants had a proficiency score of ≥ 0.71 and thus, were relatively advanced L2 listeners, which explains why the perception of the phonemes in clear speech was an easy task for them. Again, this might be explained by the low difficulty level, which failed to provoke perception errors, not least because of the missing auditory degradation (see section 5.5).

H1b assumed significantly faster RTs in the AV condition compared to the A condition, which was not corroborated by the data. However, significantly faster RTs between the AV and B condition were found, which corroborates H2b for the L2 listener group. Moreover, the error rate pattern was parallel to the pattern of RTs, meaning that the condition with the highest number of errors also had the highest mean RTs. What caused the most errors took the subjects on average the longest to process. This temporal facilitation could be explained in terms of the analysis-by-synthesis model, assuming that the visible articulatory gestures in the AV condition allowed for the prediction of upcoming auditory speech in the listener (van Wassenhove et al., 2005). Depending on the degree of salience, these predictions were more or less strong, but eventually accelerated the phoneme recognition process (see also section 5.3 for a more detailed analysis that considers salience of a specific vowel). Regardless of the degree of salience, a temporal facilitation was likely to occur, which can explain the enhanced RTs for the L2 listeners.

5.2 Only L2 listeners benefited from AV speech cues

The data obtained in this study showed that the visual information in AV speech reduced processing effort in L2 listeners. That the AV input did not yield any significant effects otherwise might be due to the fact that the subjects were proficient enough to comprehend the phoneme contrasts in clear speech without major difficulties. As Reisberg et al. (1987) pointed out, speech content in clear speech can be "[…] easy to hear but difficult to understand" (p. 99). In this study, this was attempted by removing higher-level context. However, this did not provoke frequent misperceptions. A reason for this was the native listener status of the L1

subjects and the overall very high proficiency of the L2 subjects. The presence of noise, for example, might have put both listener groups in a more challenging listening situation and yielded higher error rates (see section 5.5).

Surprisingly, L2 subjects showed faster RTs in the AV conditions than L1 subjects in the descriptive comparison. This is interesting in the sense that L1 listeners were expected to outperform L2 listeners in every condition, due to the high automaticity of language processing in L1. It was striking that the AV input brought L2 listeners on the processing level of a native listeners. Soto-Faraco al. (2007) pointed out that AV enhancement might occur stronger in L2 subjects because the input gained from the unimodal auditory and visual levels is in general lower than for L1 subjects, which in turn increases the multisensory benefit, in line with the PoIE (Stein & Meredith, 1993). A potential addition for a follow-up study might be the consideration of neurophysiological measurements, in order to examine whether ERP data confirm these findings in the form of reduced N1 and P2 amplitudes, similar to the findings of van Wassenhove et al. (2005). Moreover, considering neuroimaging research, it would be interesting to examine if the forced decision on the phonemic level led to an increased activation of the IFG (Peelle, 2019). Whether this comparison between L1 and L2 listeners was of statistical significance was not examined in this study, but might be an interesting addition to further studies. In the current study, a statistical comparison would have required more similar L1 and L2 groups, for example in terms of sample sizes.

It can be concluded that AV speech input facilitated comprehension for L2 listeners in the form of faster processing when compared to a masked face, whereas the overall performance in the comprehension experiment remained largely unaffected for both L1 and L2 listeners. However, the comparison of the AV and A condition showed no significant difference, which means that the AV facilitation might be based on overall expectations from the input in general. The audio-only condition put the listeners in a phone conversation mode, leaving all their cognitive resources for the listening process, whereas the blurred condition presented visual information in the form of a face while withholding the crucial part of this information, the lip movements. When listening to speakers wearing face masks, processing might be slowed down because the listeners have to divide their cognitive resources to a) scan the visual information to check

whether there is rich information in it, and b) detect that there is no viseme present in order to build a prediction, which means that they have to shift their focus to the auditory information instead. Thus, attention might play a role as well. Finally, the comprehension situation in the current study was the recognition of phonemes in minimal pairs in decontextualized speech. Therefore, the higher processing loads might have been due to the modality variable, but also, in parts, due to this unusual listening situation.

5.3 Vowels were more difficult to process for L2 listeners than consonants

The separate inspection of the data for vowels and consonants yielded one main result, namely that significant effects of AV speech was found only for vowel perception of L2 listeners. This effect concerned the RTs and not the error rates. This effect lied between conditions AV and B, reflecting a beneficial effect of the additional lipreading information for the recognition of vowels. However, the difference found between condition A and B gave rise to further questions. One possible explanation might be that this reflected a general confusion about the unusual input of a half-blurred face, and the extra load on cognitive resources, as discussed in section 5.2. L2 listeners of Swedish relied strongly on lipreading information, especially when they had to make decisions on vowel identity. Listening to audio-only speech did not raise the expectation of visual input, which is why the cognitive resources could focus on the auditory input, whereas the blurred condition initially suggested visual information, but did not allow for the prediction of an upcoming sound.

Since RTs showed that L2 listener made use of visual information in vowel recognition, a further evaluation of the error instances revealed that misperception happened most frequently for the target pair *ny* ‘new’ and *nu* ‘now’. In fact, a lot of the L2 subjects reported in the post-experimental word test, that this contrast was especially difficult to discriminate. Indeed, this minimal pair confronted the L2 listeners with a sound contrast that is non-existent in their native language German. In this particular minimal pair, significantly faster responses were observed in the AV condition. These results are in line with studies showing that L2 listeners benefited

especially from lipreading in AV speech when they had to deal with non-native speech sounds (Cutler et al., 2004; Hommel, 2018; Wang et al., 2008a; Wang et al., 2009).

The minimal pair *ny* 'new' and *nu* 'now' was a particular case in which the salience aspect of the analysis-by-synthesis model came into play. The two target phonemes involved the roundedness feature, which is a high salient feature (Amcoff, 1970; Robert-Ribes et al., 1998; Traunmüller & Öhrström, 2007). According to the analysis-by-synthesis model, these high salient visual features allowed for building strong predictions, which led to a temporal facilitation (van Wassenhove et al., 2005). This was corroborated by the RT results for the target pair, showing significantly faster responses in the AV condition compared to the other conditions without the information from lip movements, which was in line with the study of Navarra and Soto-Faraco (2007) who found enhanced RTs in an AV condition. Further studies might draw on this observation and examine high and low salient targets separately, in order to investigate whether RTs are significantly affected by the salience of the phonemes. However, the results showed that sound contrasts like these might be contexts in which the visual information of lip movements is essential for successful speech perception for some L2 listeners, but presumably also for other vulnerable populations, such as hearing-impaired individuals. It also suggested that visual information from lipreading might be more important in languages that have many rounded vowels, which is in line with Robert-Ribes et al. (1998) showing that roundedness is an important visual feature in French as well.

A further reason for the slower RT results for *ny* 'new' – *nu* 'now' might be that the L2 listeners were confronted with a sound contrast that does not exist in their L1. According to the PAM, the front rounded vowel [y] in *ny* 'new' and the central rounded vowel [ʉ] in *nu* 'now' were perceptually assimilated to the German category /y/, but the Swedish [ʉ] was perceived closer (and thus, a better match) to the category than the Swedish [y]. Indeed, the Swedish [y] is characterized by a larger lip opening than the German [y], which is why the former is perceived as less good match to the German /y/ (Lindau, 1978). Thus, this contrast resembled a Category Goodness (CG) assimilation, and the predicted discrimination performance for this contrast was indeed intermediate, compared to other target pairs (Best, 1994; Best & Tyler, 2007). Cutler (2012) further suggested that even though this is a property of CG contrasts, there might be slight

differences in terms of goodness between the L2 sounds in Single Category (SC) assimilations as well, and that L2 learners can shift such a contrast to a CG contrast after sufficient exposure made them more sensitive to these differences. In this case, beginning L2 learners of Swedish considered both [ʉ] and [y] as equally good or poor exemplars of the German /y/, and then shifted this contrast to a CG assimilation when noticing that one sound is a better match to the native category than the other (Best & Tyler, 2007). However, this remains open to speculation, not least because there were no beginner L2 listeners in this study.

The L2 listeners were in the higher range of proficiencies, and thus, L2 categories might have been developed already. Listeners with sufficient exposure to the sounds in question could master the discrimination successfully without making comprehension errors. That the AV condition supported L2 speech perception is in line with previous studies suggesting that with increasing proficiency, the phonological representations of L2 sounds are strengthened, and thus, their visemic representation are a stronger facilitator for perception than for less proficient L2 listeners (Drijvers et al., 2019; Drijvers & Özyürek, 2020). However, the significantly faster RTs in the AV condition suggested that these contrasts were still cognitively demanding.

It seems likely that as a result of L2 acquisition and exposure to a certain L2 sound contrast, German learners of Swedish stored [ʉ] as a Swedish realization of the already existing category /y/, whereas for [y] a new phoneme category was developed. This category formation can proceed faster when it is a contrast involved in a minimal pair of two high-frequent words (Best & Tyler, 2007). This might be the case for *ny* ‘new’ and *nu* ‘now’ as well, since both words occur frequently in Swedish and are easier to distinguish for L2 listeners than other minimal pairs. The PAM-L2 assumes that listeners use phonetic as well as gestural perceptual objects (Best & Tyler, 2007). This is reflected in the findings of this study that listeners mostly relied on phonetic cues in clear speech, but used the visual information from articulatory gestures increasingly when it was regarded as helpful for the discrimination of particular sound contrasts. Considering a neuroimaging approach, it may be interesting to investigate if the supramarginal gyrus was increasingly active when L2 listeners perceived these more demanding sound contrasts, as suggested by previous research (Peelle, 2019).

The comprehension question required participants to read target words as well, after hearing them. Studies showed that mismatches in the grapheme-phoneme correspondence have an impact on how effective listeners can use AV information in speech processing (Erdener & Burnham, 2005). Swedish and German are both languages that have a relatively consistent relation of the mapping of phonemes in speech to the corresponding graphemes in written language (Seymour et al., 2003). However, some graphemes are used in both Swedish and German, but with a different phonetic realization. The grapheme <y> is pronounced as [y] (*süß* 'sweet') or [ʏ] (*Mütze* 'cap')[27] in German, but can also be realized as [i], as in *Handy* 'mobile phone'. In Swedish, <y> is phonetically realized as [y] (*ny* 'new'), but as mentioned earlier, this realization differs perceptually from the German [y] (Lindau, 1978; Riad, 2014). Moreover, the grapheme <u> is pronounced [u] (*Suche* 'search') or [ʊ] (*Mutter* 'mother') in German, whereas it is realized as [ʉ] in Swedish, as in *nu* 'now' (O'Brien & Fagan, 2016; Riad, 2014).

Another case is the grapheme <g>, which is pronounced as [ɡ] in German, whereas in Swedish, word-initially the pronunciation ranges from [j] before front vowels (*göra* 'do') to [ɡ] before other vowels and consonants (*gul* 'yellow', *glad* 'happy'), or [ɧ] in some other cases, for example *generad* 'embarrassed' (O'Brien & Fagan, 2016; Riad, 2014). Moreover, the grapheme <k> in *köra* 'drive' is pronounced [ɕ], which is not only a novel sound for the German L2 listeners, but also exhibits another orthographic difference between German and Swedish. The grapheme <k> is realized as [k] when it is not occurring before a front vowel (*kaka* 'cake'), whereas in German, the spelling <k> always reflects the sound [k] (O'Brien & Fagan, 2016; Riad, 2014). These orthographic dissimilarities could have caused additional processing effort for the L2 listeners. Thus, the more discrepancies in terms of orthographic and phonemic dissimilarities occur in both languages, the more difficult speech perception might become in such a task. Consequently, the word pairs discussed above are examples for such discrepancies and were therefore, prone to errors and elicited longer RTs.

Furthermore, the missing benefits of AV speech on consonant recognition might be due to the absence of noise. Since the acoustic structure of consonants is inherently less intense compared

[27] Note that variations in vowel length can occur.

to vowels, they are expected to benefit more from AV information when speech is presented in noise, since noise degrades acoustic features of consonants more than those of vowels (Summerfield, 1987). The absence of acoustic degradation in this study might, therefore, explain why there were no significant effects of AV information.

Now that certain effects on L2 speech perception was attributed to orthographic influences from L1, further L2s could affect the performance in such an experimental task as well. The subjects here were not controlled in terms of further L2s, which might be an important consideration for a further study. Alternatively, a follow-up study could choose a different methodological approach that does not require reading in order to examine AV impacts.

To answer RQ_3, the data obtained in this study showed that the visual information in AV speech reduced processing effort in German L2 listeners of Swedish, especially when they were asked to recognize vowels that do not exist in their L1. Thus, the perception of non-native sound contrast was a situation that required listeners to obtain information from the visual domain. It is possible that not only L1 phoneme inventories but even orthographic properties of the L1 or other L2s play a role in this process.

5.4 No correlation between L2 proficiency and AV benefit

The correlation analysis did not reveal any significant connections between the effect of the modality and the L2 proficiency, neither in terms of error rates nor RTs. The missing significance in this respect might be due to the floor effects in general. It might also be explained in terms of the small subject sample size. Moreover, all the L2 subjects were within the upper 30% of the continuous proficiency scale by the LHQ. This is interesting insofar as the addition of less proficient subjects might provide more informative insights. However, the results were compatible with studies showing that L2 listeners focus the speaker's mouth regardless of L2 proficiency (Birulés et al., 2020). Even though some previous studies suggested that there is an increasing AV benefit with increasing proficiency (Hazan et al., 2002; Sueyoshi & Hardison, 2005; Wang et al., 2008b; Xie et al., 2014), it is not without risk to compare these studies to one

another, since the issue of proficiency was often treated differently in the respective studies. Often some other pre-existing language tests were used to assess proficiency, or other benchmarks such as length of residency or age of acquisition. However, it seems likely that there is something to these studies and that the missing correlation is due to methodological issues.

The LHQ is a very useful tool in order to retrieve information about a subject's language background. However, the algorithm to determine a proficiency score is based only on the self-assessment of the subject regarding the current level in speaking, writing, listening, and reading in the target language (Li et al., 2020). Further studies might look at this by considering not only self-assessment of subjects, which is at risk of being imprecise, but also length of residency, age of acquisition (AoA) or quality/quantity of exposure to the target language. In this study, the AoA was also not correlated to the AV effect. However, the AoA by itself does not give insights about the quality and quantity of exposure. Combined, these characteristics could provide a more precise classification of the proficiency level.

With respect to RQ_4, this means that no correlation between the effects of visible speech and L2 proficiency of the language in question was observed. However, since findings of previous studies with regards to an AV faciliatory effect for L2 listeners were corroborated, a different methodological approach to the issue of proficiency might be a fruitful source for further research.

5.5 Discussion of the method

As discussed above, studies of AV speech perception often include several levels of noise, since the AV benefit is greatest under noise degradation (Drijvers & Özyürek, 2017; Drijvers & Özyürek, 2020; Fitzpatrick & Kim, 2010; Ross et al., 2007; Sumby & Pollack, 1954; Xie et al., 2014). In this study, this factor was not considered due to the online procedure of the experiment, which did not allow for the control of volume or other background noises. However, as described similarly in the study of Giovanelli et al. (2020), the acoustic signal quality in the current study was the same across all conditions, which is why observed effects could be attributed to the

modality variable and not, for example, to the modality in an interplay with noise, or exclusively to the noise because speech was distorted, for example, by recording it with a face mask. Due to different technical devices, or volumes used for the experiment, a residual risk remained that the auditory level varied slightly between the participants.

However, the online setup had an advantage compared to in-laboratory experiments. The latter often relies on students as participants, whereas the online setup allowed for leaving the university demographics and finding participants with a wider range of occupations, which resembles the actual population more.

As mentioned before in this chapter, a further limitation was that the L2 listener sample was not controlled for other second languages. Experience with other L2s that involve similar, or the same non-native phonemes might have had an effect on the performance in the experiment. Furthermore, choosing L2 listeners from a different L1 background could have yielded different results, depending on the respective phoneme inventory as compared to Swedish, which might be a starting point for further research. Moreover, the investigation of how Swedish-German bilingual children and adults deal with these phoneme recognition tasks, is another possible domain worth investigating.

6. Conclusion

This study provides empirical evidence for the effect of AV information from visible speech gestures on speech comprehension in L1 and L2 listeners of Swedish. The main finding is that L2 listeners benefit from AV information when they perceive Swedish vowels, which is reflected in faster response times, presumably due to the reduction of processing load. Therefore, this study emphasizes that visual information is not only used in difficult listening situations, but that listeners even use it while listening to clear speech. This contributes to the generalizability of previous findings in the research area of AV speech perception. However, the study did not find correlations between the L2 proficiency and AV speech benefit.

Decontextualized sentences have been created in order to examine the level of phoneme recognition in spoken language. However, natural conversation provides the listener with much more context that allows for efficient interpretation and comprehension of speech, not least also by means of co-speech gestures, and extraoral face movements. The visual information from speech gestures may not be essential to comprehend spoken language – undoubtedly people are still able to communicate via telephones, and also when they wear masks, and this is reflected well in the performances at ceiling levels. However, reading articulatory gestures from the interlocutor's face can ease speech perception and take away extra cognitive load for the L2 listeners, especially when they are confronted with unfamiliar speech sounds deviating from the phoneme categories established in their L1 system. This might explain why listeners sometimes have comprehension difficulties in everyday communication when visual information is missing. Beyond that, it emphasizes that vulnerable populations of listeners gain even stronger AV effects, and it is in line with various studies stating that a covered mouth affects speech perception even more adversely for hearing-impaired individuals, not to mention users of cued-speech language, in which the visibility of the mouth is of much higher importance than in spoken or signed language (see also Irwin & DiBlasi, 2017). Thus, in educational settings, in L2 classrooms, but also in medical care contexts, it may be worth considering how listeners can be provided with this extra visual input to support their speech comprehension, not least in light of

the ongoing pandemic situation (Erdener, 2016; Hardison, 1999). Future studies may further explore these findings on AV effects by investigating specific phonemes, L2 learners with a typologically unrelated L1, or the factor of language proficiency and benefit from AV speech.

References

Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes*, *15*(6), 593-613.

Amcoff, S. (1970). *Visuell perception av talljud och avläsestöd för hörselskadade* (Report No. 7). Lärarhögskolan i Uppsala, Pedagogiska Institutionen.

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, *92*, 339-355.

Atcherson, S. R., Mendel, L. L., Baltimore, W. J., Patro, C., Lee, S., Pousson, M., & Spann, M. J. (2017). The effect of conventional and transparent surgical masks on speech understanding in individuals with and without hearing loss. *Journal of the American Academy of Audiology*, *28*(1), 58-67.

Bandaru, S. V., Augustine, A. M., Lepcha, A., Sebastian, S., Gowri, M., Philip, A., & Mammen, M. D. (2020). The effects of N95 mask and face shield on speech perception among healthcare workers in the coronavirus disease 2019 pandemic scenario. *The Journal of Laryngology & Otology*, *134*(10), 895-898.

Barenholtz, E., Mavica, L., & Lewkowicz, D. J. (2016). Language familiarity modulates relative attention to the eyes and mouth of a talker. *Cognition*, *147*, 100-105.

Barrós-Loscertales, A., Ventura-Campos, N., Visser, M., Alsius, A., Pallier, C., Rivera, C. Á., & Soto-Faraco, S. (2013). Neural correlates of audiovisual speech processing in a second language. *Brain and Language*, *126*(3), 253-262.

Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). fMRI-Guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *Journal of Neuroscience*, *30*(7), 2414-2417.

Benguerel, A. P., & Pichora-Fuller, M. K. (1982). Coarticulation effects in lipreading. *Journal of Speech, Language, and Hearing Research*, *25*(4), 600-607.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In Goodman, J. C. & Nusbaum, N. C. (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167-224). MIT Press.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementaries. In Bohn, O. & Munro, M. J. (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13-34). John Benjamins Publishing Company.

Birulés, J., Bosch, L., Pons, F., & Lewkowicz, D. J. (2020). Highly proficient L2 speakers still need to attend to a talker's mouth when processing L2 speech. *Language, Cognition and Neuroscience*, *35*(10), 1314-1325.

Blackburn, C. L., Kitterick, P. T., Jones, G., Sumner, C. J., & Stacey, P. C. (2019). Visual speech benefit in clear and degraded speech depends on the auditory intelligibility of the talker and the number of background talkers. *Trends in Hearing*, *23*, 1-14.

Bottalico, P., Murgia, S., Puglisi, G. E., Astolfi, A., & Kirk, K. I. (2020). Effect of masks on speech intelligibility in auralized classrooms. *The Journal of the Acoustical Society of America*, *148*(5), 2878-2884.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*(1), 45-50.

Burnham, D., & Dodd, B. (2004). Auditory–visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, *45*(4), 204-220.

Calbi, M., Langiulli, N., Ferroni, F., Montalti, M., Kolesnikov, A., Gallese, V., & Umiltà, M. A. (2021). The consequences of COVID-19 on social interactions: an online study on face covering. *Scientific Reports*, *11*(2601), 1-10.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, *10*(11), 649-657.

Campbell, R., & MacSweeney, M. (2012). Brain bases for seeing speech: fMRI studies of speechreading. In Bailly, G., Perrier, P. & Vatikiotis-Bateson, E. (Eds.), *Audiovisual speech processing* (pp. 76-103). Cambridge University Press.

Carbon, C. C. (2020). Wearing face masks strongly confuses counterparts in reading emotions. *Frontiers in Psychology*, *11*, 1-8.

Cohn, M., Pycha, A., & Zellou, G. (2021). Intelligibility of face-masked speech depends on speaking style: Comparing casual, clear, and emotional speech. *Cognition*, *210*(104579), 1-5.

Corey, R. M., Jones, U., & Singer, A. C. (2020). Acoustic effects of medical, cloth, and transparent face masks on speech signals. *The Journal of the Acoustical Society of America*, *148*(4), 2371-2375.

Cotton, J. C. (1935). Normal "visual hearing". *Science*, *82*(2138), 592-593.

Cunillera, T., Camara, E., Laine, M., & Rodríguez-Fornells, A. (2010). Speech segmentation is facilitated by visual cues. *The Quarterly Journal of Experimental Psychology*, *63*(2), 260-274.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. MIT Press.

Cutler, A., Cooke, M., Lecumberri, M. L. G., & Pasveer, D. (2007, August 27-31). *L2 consonant identification in noise: Cross-language comparisons* [paper presentation]. Interspeech 2007, Antwerp, Belgium.

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America*, *116*(6), 3668-3678.

Cvejic, E., Kim, J., & Davis, C. (2010). Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Communication*, *52*(6), 555-564.

Davis, C., & Kim, J. (2004). Audio-visual interactions with intact clearly audible speech. *The Quarterly Journal of Experimental Psychology*, *57*(6), 1103-1121.

Davis, C., & Kim, J. (2006). Audio-visual speech perception off the top of the head. *Cognition*, *100*(3), B21-B31.

Desjardins, R. N., Rogers, J., & Werker, J. F. (1997). An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks. *Journal of Experimental Child Psychology, 66*, 85-110.

Drijvers, L., & Özyürek, A. (2017). Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language, and Hearing Research, 60*, 212-222.

Drijvers, L., & Özyürek, A. (2020). Non-native listeners benefit less from gestures and visible speech than native listeners during degraded speech comprehension. *Language and speech, 63*(2), 209-220.

Drijvers, L., Vaitonytė, J., & Özyürek, A. (2019). Degree of language experience modulates visual attention to visible speech and iconic gestures during clear and degraded speech comprehension. *Cognitive Science, 43*, 1-25.

Erdener, D. (2016). Basic to applied research: the benefits of audio-visual speech perception research in teaching foreign languages. *The Language Learning Journal, 4*(1), 124-132.

Erdener, V. D., & Burnham, D. K. (2005). The role of audiovisual speech and orthographic information in nonnative speech production. *Language Learning*, *55*(2), 191-228.

Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition, 26*(4), 551-585.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, *11*(4), 796-804.

Fitzpatrick, M., & Kim, J. (2010, August 23-27). *Audio-visual speech perception in noise by first and second language listeners* [paper presentation]. 20th International Congress on Acoustics, Sydney, Australia.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In Strange, W. (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp.233-277). York Press.

Fort, M., Spinelli, E., Savariaux, C., & Kandel, S. (2010). The word superiority effect in audiovisual speech perception. *Speech Communication*, *52*(6), 525-532.

Frank, M. C., Amso, D., & Johnson, S. P. (2014). Visual search and attention to faces during early infancy. *Journal of Experimental Child Psychology*, *118*, 13-26.

Giovanelli, E., Valzolgher, C., Gessa, E., Todeschini, M., & Pavani, F. (2021). Unmasking the difficulty of listening to talkers with masks: lessons from the COVID-19 pandemic. *i-Perception*, *12*(2), 1-1-11.

Goswami, U. (2005). Synthetic phonics and learning to read: A cross-language perspective. *Educational Psychology in Practice, 21*(4), 273-282.

Granström, B., House, D., & Lundeberg, M. (1999, August 1-7). Prosodic cues in multimodal speech perception [paper presentation]. International Congress of Phonetic Sciences. San Francisco, California, USA.

Grundmann, F., Epstude, K., & Scheibe, S. (2021). Face masks reduce emotion-recognition accuracy and perceived closeness. *PloS ONE, 16*(4). 1-18,

Gullberg, M. (2006). Some reasons for studying gesture and second language acquisition (Hommage à Adam Kendon). *International Review of Applied Linguistics, 44*, 103-124.

Gullberg, M. (2008). Gestures and second language acquisition. In Robinson, P. & Ellis, N. C. (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 276-305). Routledge.

Halle, M., & Stevens, K. (1962). Speech recognition: A model and a program for research. *IRE Transactions on Information Theory, 8*(2), 155-159.

Hampton, T., Crunkhorn, R., Lowe, N., Bhat, J., Hogg, E., Afifi, W., De, S., Street, I., Sharma, R., Krishnan, M., Clarke, R., Dasgupta, S., Ratnayake, S., & Sharma, S. (2020). The negative impact of wearing personal protective equipment on communication during coronavirus disease 2019. *The Journal of Laryngology & Otology, 134*(7), 577-581.

Hardison, D. M. (1999). Bimodal speech perception by native and nonnative speakers of English: Factors influencing the McGurk effect. *Language Learning, 49*, 213-283.

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics, 24*(4), 495-522.

Hazan, V., Sennema, A., & Faulkner, A. (2002, September 16-20). *Audiovisual perception in L2 learners* [paper presentation]. Seventh International Conference on Spoken Language Processing, Denver, Colorado, USA.

Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research, 53*, 298-310.

Hommel, M. (2018). The role of orthography and phoneme inventory in Dutch students' speech perception in the EFL classroom. *Philologia, 16*(1), 65-75.

Huang, Q., & Tang, J. (2010). Age-related hearing loss or presbycusis. *Eur Arch Otorhinolaryngol, 267*, 1179-1191.

Inceoglu, S. (2019). Individual differences in L2 speech perception: The role of phonological memory and lipreading ability. *The Modern Language Journal, 103*(4), 782-799.

Inceoglu, S. (2021). Language experience and subjective word familiarity on the multimodal perception of non-native speakers' vowels. *Language and Speech, 00*(0) 1-20.

Irwin, J., Brancazio, L., & Volpe, N. (2017). The development of gaze to a speaking face. *The Journal of the Acoustical Society of America, 141*(5), 3145-3150.

Irwin, J., & DiBlasi, L. (2017). Audiovisual speech perception: A new approach and implications for clinical populations. *Language and Linguistics Compass, 11*(3), 77-91.

Kaiser, E. (2013). Experimental paradigms in psycholinguistics. In Podesva, R. J. & Sharma, D. (Eds.), *Research methods in linguistics* (pp. 135-168). Cambridge University Press.

Kendon, A. (1986). Some reasons for studying gesture. *Semiotica 62*(1/2), 3-28.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1: 7-36. http://www.sketchengine.eu

Kim, J., Aubanel, V., & Davis, C. (2015, August 10-14). *The effect of auditory and visual signal availability on speech perception*. [paper presentation]. International Congress of Phonetic Science, Glasgow, Scotland, UK.

Kuhl, K. P., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science 218*(4577), 1138-1141.

Ladefoged, P. (1975). *A course in phonetics*. Harcourt Brace Jovanovich.

Lado, R. (1957). Sentence structure. *College Composition and Communication*, *8*(1), 12-16.

Lalonde, K., & Werner, L. A. (2019). Infants and adults use visual cues to improve detection and discrimination of speech in noise. *Journal of Speech, Language, and Hearing Research*, *62*(10), 3860-3875.

Lansing, C. R., & McConkie, G. W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech, Language, and Hearing Research*, *42*(3), 526-539.

Li, P., Zhang, F., Yu, A., & Zhao, X. (2020). Language History Questionnaire (LHQ3): An enhanced tool for assessing multilingual experience. *Bilingualism: Language and Cognition*, *23*(5), 938-944.

Lindau, M. (1978). Vowel features. *Language*, *54*(3), 541-563.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1-36.

Magnotti, J. F., Mallick, D. B., Feng, G., Zhou, B., Zhou, W., & Beauchamp, M. S. (2015). Similar frequency of the McGurk effect in large samples of native Mandarin Chinese and American English speakers. *Experimental brain research*, *233*(9), 2581-2586.

Mártony, J. (1974). On speechreading of Swedish consonants and vowels. *KTH STL-Quarterly Progress and Status Report, 15*(2-3), 11-33.

Massaro, D. W. (1987). Speech perception by ear and eye. In Dodd, B. & Campbell, R. (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 53-83). Lawrence Erlbaum Associates Ltd.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746-748.

McNeill, D. (1992). *Hand and mind. What the hands reveal about thought.* Chicago University Press.

Mendel, L. L., Gardino, J. A., & Atcherson, S. R. (2008). Speech understanding using surgical masks: a problem in health care?. *Journal of the American Academy of Audiology*, *19*(9), 686-695.

Mheidly, N., Fares, M. Y., Zalzale, H., & Fares, J. (2020). Effect of face masks on interpersonal communication during the COVID-19 pandemic. *Frontiers in Public Health*, *8*, 1-6.

Moulton, W. G. (1962). *The sounds of English and German: [a systematic analysis of the contrasts between the sound systems]*. Chicago University Press.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, *58*(3), 351-362.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004a). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, *15*(2), 133-137.

Munhall, K. G., Kroos, C., Jozan, G., & Vatikiotis-Bateson, E. (2004b). Spatial frequency requirements for audiovisual speech perception. *Perception & Psychophysics*, *66*(4), 574-583.

Nath, A. R., & Beauchamp, M. S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *Journal of Neuroscience, 31*(5), 1704-1714.

Navarra, J., Alsius, A., Velasco, I., Soto-Faraco, S., & Spence, C. (2010). Perception of audiovisual speech synchrony for native and non-native language. *Brain Research*, *1323*, 84-93.

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychological Research*, *71*, 4-12.

O'Brien, M. G., & Fagan, S. M. (2016). *German phonetics and phonology: Theory and practice*. Yale University Press.

Öhrström, N., Bulukin Wilén, F., Eklöf, A., & Gustafsson, J. (2009, June 10-12). *Visual discrimination between Swedish and Finnish among L2-learners of Swedish* [paper presentation]. FONETIK 2009, Stockholm, Sweden.

Oliver, G., Gullberg, M., Hellwig, F., Mitterer, H., & Indefrey, P. (2012). Acquiring L2 sentence comprehension: A longitudinal study of word monitoring in noise. *Bilingualism: Language and Cognition*, *15*, 841-857.

Ong, S. (2020, June 9). *How face masks affect our communication*. BBC Future. https://www.bbc.com/future/article/20200609-how-face-masks-affect-our-communication [accessed August 17, 2021]

Ortega-Llebaria, M., Faulkner, A., & Hazan, V. (2001, September 7-9). *Auditory-visual L2 speech perception: Effects of visual cues and acoustic-phonetic context for Spanish learners of English* [paper presentation]. International Conference on Auditory-Visual Speech Processing, Aalborg, Denmark.

Ozker, M., Schepers, I. M., Magnotti, J. F., Yoshor, D., & Beauchamp, M. S. (2017). A double dissociation between anterior and posterior superior temporal gyrus for processing audiovisual speech demonstrated by electrocorticography. *Journal of Cognitive Neuroscience*, *29*(6), 1044-1060.

Paré, M., Richler, R. C., ten Hove, M., & Munhall, K. G. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. *Perception & Psychophysics*, *65*(4), 553-567.

Peelle, J. E. (2019). The neural basis for auditory and audiovisual speech perception. In Katz, W. F. & Assmann, P. F. (Eds.), *The Routledge handbook of phonetics* (pp. 193-216). Routledge.

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169-181.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods, 51,* 193-205.

Rahne, T., Fröhlich, L., Plontke, S., & Wagner, L. (2021). Influence of surgical and N95 face masks on speech perception and listening effort in noise. *Plos ONE*, *16*(7), 1-11.

Redford, M. A., Kallay, J. E., Bogdanov, S. V., & Vatikiotis-Bateson, E. (2018). Leveraging audiovisual speech perception to measure anticipatory coarticulation. *The Journal of the Acoustical Society of America*, *144*(4), 2447-2461.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli. In Dodd, B. & Campbell, R. (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97-113). Lawrence Erlbaum Associates Ltd.

Riad, T. (2014). *The phonology of Swedish*. Oxford University Press.

Robert-Ribes, J., Schwartz, J. L., Lallouache, T., & Escudier, P. (1998). Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise. *The Journal of the Acoustical Society of America*, *103*(6), 3677-3689.

Ronquest, R. E., & Hernandez, L. (2005). *Lip-reading skills in bilinguals: Some effects of L1 on visual-only language identification* (Report No. 27). Speech Research Laboratory, Department of Psychological and Brain Sciences, Indiana University.

Ronquest, R. E., Levi, S. V., & Pisoni, D. B. (2007). *Language identification from visual-only speech* (Report No. 28). Speech Research Laboratory, Department of Psychological and Brain Sciences, Indiana University.

Rosenblum, L. D., & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(2), 318-331.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*(5), 1147-1153.

Sato, M., Buccino, G., Gentilucci, M., & Cattaneo, L. (2010). On the tip of the tongue: Modulation of the primary motor cortex during audiovisual speech perception. *Speech Communication*, *52*(6), 533-541.

Saunders, G. H., Jackson, I. R., & Visram, A. S. (2021). Impacts of face coverings on communication: an indirect impact of COVID-19. *International Journal of Audiology*, *60*(7), 495-506.

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics*, *59*(1), 73-80.

Seymour, P. H., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, *94*(2), 143-174.

Shi, L.-F. (2010). Perception of acoustically degraded sentences in bilingual listeners who differ in age of English acquisition. *Journal of Speech, Language, and Hearing Research, 53*, 821–835.

Siva, N., Stevens, E. B., Kuhl, P. K., & Meltzoff, A. N. (1995). A comparison between cerebral-palsied and normal adults in the perception of auditory-visual illusions. *Journal of the Acoustical Society of America*, *98*(5), 2983.

Smiljanic, R., Keerstock, S., Meemann, K., & Ransom, S. M. (2021). Effects of face masks and speaking style on audio-visual speech perception and memory. *The Journal of the Acoustical Society of America*, *149*(6), 4013-4023.

Smith, D., & Burnham, D. (2012). Facilitation of Mandarin tone perception by visual speech in clear and degraded audio: Implications for cochlear implants. *The Journal of the Acoustical Society of America, 131*(2), 1480-1489.

Sorati, M., & Behne, D. M. (2019). Musical expertise affects audiovisual speech perception: Findings from event-related potentials and inter-trial phase coherence. *Frontiers in Psychology*, *10* (2562), 1-19.

Soto-Faraco, S., Navarra, J., Weikum, W. M., Vouloumanos, A., Sebastián-Gallés, N., & Werker, J. F. (2007). Discriminating languages by speech-reading. *Perception & Psychophysics, 69*(2), 218-231.

Soto-Faraco, S., Calabresi, M., Navarra, J., Werker, J., & Lewkowicz, D. J. (2012). The development of audiovisual speech perception. *Multisensory Development*, 207-228.

Spitzer, M. (2020). Masked education? The benefits and burdens of wearing face masks in schools during the current Corona pandemic. *Trends in Neuroscience and Education, 20*(100138), 1-8.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. MIT Press.

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*(4), 661-699.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*(2), 212-215.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In Dodd, B., & Campbell, R. (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3-51). Lawrence Erlbaum Associates Ltd.

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *335*, 71-78.

The jamovi project (2021). *jamovi* (Version 1.6) [Computer Software]. Sydney, Australia. Retrieved from https://www.jamovi.org

Thibodeau, L. M., Thibodeau-Nielsen, R. B., Tran, C. M. Q., & de Souza Jacob, R. T. (2021). Communicating during COVID-19: the effect of transparent masks for speech recognition in noise. *Ear and Hearing*, *42*(4), 772-781.

Thomas, S. M., & Jordan, T. R. (2004). Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(5), 873-888.

Traunmüller H. & Öhrström N. (2007) Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics*, 35, 244-258.

Truong, T. L., & Weber, A. (2021). Intelligibility and recall of sentences spoken by adult and child talkers wearing face masks. *The Journal of the Acoustical Society of America*, *150*(3), 1674-1681.

Tye-Murray, N., Sommers, M., & Spehar, B. (2007). Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends in Amplification*, *11*(4), 233-241.

Van Engen, K. J., Xie, Z., & Chandrasekaran, B. (2017). Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. *Attention, Perception, & Psychophysics*, *79*(2), 396-403.

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, *102*(4), 1181-1186.

Wang, Y., Behne, D. M, & Jiang, H. (2008a). Effects of training modality on audio-visual perception of nonnative speech contrasts. *Canadian Acoustics*, *36*(3), 120-121.

Wang, Y., Behne, D. M., & Jiang, H. (2008b). Linguistic experience and audio-visual perception of non-native fricatives. *The Journal of the Acoustical Society of America*, *124*(3), 1716-1726.

Wang, Y., Behne, D. M., & Jiang, H. (2009). Influence of native language phonetic system on audio-visual speech perception. *Journal of Phonetics*, *37*(3), 344-356.

Warren, P. (2013). *Introducing psycholinguistics*. Cambridge University Press.

Wegrzyn, M., Vogt, M., Kireclioglu, B., Schneider, J., & Kissler, J. (2017). Mapping the emotional face. How individual face parts contribute to successful emotion recognition. *PloS ONE*, *12*(5), 1-15.

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. F. (2007). Visual language discrimination in infancy. *Science*, *316* (5828), 1159-1159.

Werker, J. F., Frost, P. E., & McGurk, H. (1992). La langue et les lèvres: Cross-language influences on bimodal speech perception. *Canadian Journal of Psychology*, *46*(4), 551-568.

World Health Organization. (2020, July 9). *Transmission of SARS-CoV-2: implications for infection prevention precautions*. https://www.who.int/news-room/commentaries/detail/transmission-of-sars-cov-2-implications-for-infection-prevention-precautions [accessed 17.08.2021]

Xie, Z., Yi, H. G., & Chandrasekaran, B. (2014). Nonnative audiovisual speech perception in noise: Dissociable effects of the speaker and listener. *PloS ONE*, *9*(12), 1-17.

Yang, J. (2014, November 19). A quick history of why Asians wear surgical masks in public. Quartz. https://qz.com/299003/a-quick-history-of-why-asians-wear-surgical-masks-in-public/ [accessed August 17, 2021]

Appendix

A. Recruitment letters

A1. German version

Hej!

Ich suche Teilnehmer*innen an einem Experiment zum Verständnis gesprochener schwedischer Sprache, das Teil meiner Masterarbeit im Fach Linguistik an der Universität Lund ist. Jede*r Teilnehmer*in erhält als Dankeschön einen 49kr Rabattcode für die Internetseite SF Anytime, auf der ihr euch Filme und Serien ausleihen könnt.

Um teilzunehmen solltest du

*18-60 Jahre alt sein,

*deutsche*r Muttersprachler*in sein und gute bis sehr gute Schwedischkenntnisse haben

Das Experiment findet im Mai statt und wird aufgrund der Corona-Situation online ablaufen. Zuerst wirst du gebeten, einen Fragebogen zu deinem sprachlichen Hintergrund auszufüllen (auf Deutsch, dauert ca. 10 Minuten). Das Experiment wird dann online durchgeführt (via Zoom) und ca. 20-30 Minuten dauern. In dem Experiment wirst du kurze Videos und Tonaufnahmen ansehen bzw. anhören und danach Fragen dazu beantworten, indem du aus 2-3 Antwortmöglichkeiten eine Antwort auswählst. Du bist als Teilnehmer*in selbstverständlich anonym.

Es ist also auch eine gute Möglichkeit, dein Schwedisch auf die Probe zu stellen. :)

Wenn du Fragen hast, Lust hast teilzunehmen oder jemanden kennst, der teilnehmen würde, melde dich gerne bei mir (direkt hier bei Facebook oder unter der E-Mail-Adresse)!

E-Mail: he4137sp-s@student.lu.se

Link zum Streamingdienst: Hyr eller köp film online - Streama direkt på SF Anytime

A2. Swedish version

Hej!

Jag söker deltagare till mitt experiment om förståelse av svenskt talspråk. Experimentet är en del av min masteruppsats i Allmän språkvetenskap. Datainsamlingen kommer att pågå i maj och är helt online. För att delta i undersökningen ska du vara mellan 18 och 60 år och ha svenska som modersmål.

Du kommer först att bli ombedd att fylla i ett frågeformulär om din språkbakgrund (formuläret är på engelska, det tar cirka 10 minuter) och efter det deltar du i experimentet (i Zoom, cirka 20-25 minuter). I experimentet kommer du få lyssna på korta ljudinspelningar och titta på korta videoklipp. Efteråt får du frågor som du besvara genom att välja ett av 2-3 svarsalternativ. Du garanteras anonymitet förstås.

Varje deltagare får en 49kr rabattkod för SF Anytime för att hyra filmer eller serier.

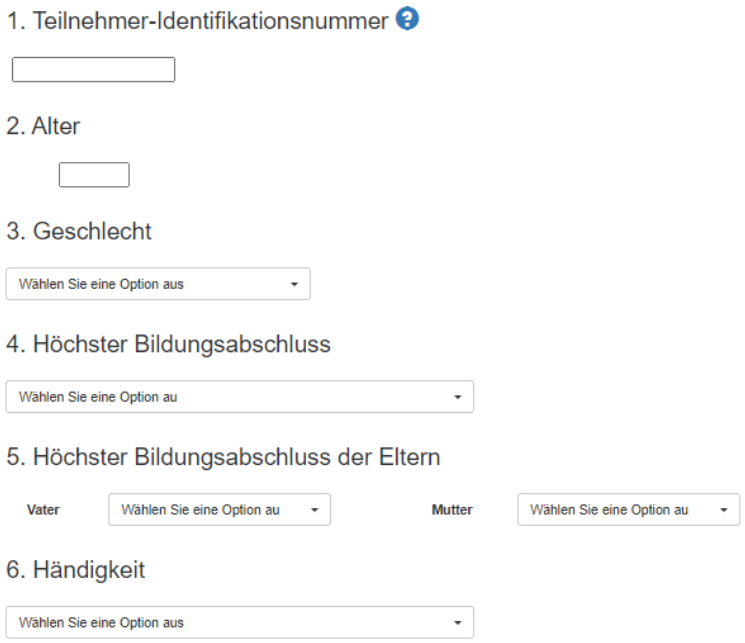
Om du har frågor eller är intresserad av att delta - hör gärna av dig (här eller via nedanstående epost)! Dela gärna inlägget om du vet någon som kan vara intresserad!

Epost: he4137sp-s@student.lu.se

SFanytime webbplats: Hyr eller köp film online - Streama direkt på SF Anytime

B. Language history questionnaire

B1. German version



7. Geben Sie Ihre Muttersprache(n) und alle anderen Sprachen an, die Sie erlernt haben, sowie das Alter, in welchem Sie begonnen haben, die jeweilige Sprache zu verstehen, zu sprechen, zu lesen und zu schreiben. Geben Sie auch die Anzahl der Jahre an, die Sie diese Sprache schon verwendet haben.

Sprache	Verstehen	Sprechen	Lesen	Schreiben	Verwendungsjahre
Wähle eine Optio					
Wähle eine Optio					
Wähle eine Optio					
Wähle eine Optio					

*Anmerkung: Sie können eine Sprache erlernt, sie eine Zeit lang nicht mehr genutzt und sie anschließend wieder verwendet haben. Bitte geben Sie die Gesamtzahl der Verwendungsjahre an.

8. Herkunftsland

Wähle eine Option

9. Land des derzeitigen Wohnsitzes

Wähle eine Option

10. Wenn Sie länger als drei Monate in Ländern gelebt haben oder gereist sind, die nicht dem Land Ihres Wohnsitzes entsprechen, nennen Sie den Namen des Landes, die Länge Ihres Aufenthalts (in Monaten), die Sprache, die Sie dort verwendet haben, und die Häufigkeit der Sprachverwendung für jedes Land.

Land	Länge des Aufenthalts (in Monaten)	Sprache	Häufigkeit der Verwendung
Wähle eine Option		Wähle eine Option	Wählen Sie eine Option au
Wähle eine Option		Wähle eine Option	Wählen Sie eine Option au
Wähle eine Option		Wähle eine Option	Wählen Sie eine Option au
Wähle eine Option		Wähle eine Option	Wählen Sie eine Option au

* Hinweis Sie waren möglicherweise mehrmals im Land, jeweils für einen anderen Zeitraum. Addiere alle Reisen zusammen

11. Geben Sie an, wie Sie Ihre Nicht-Muttersprache (n) gelernt oder erworben haben. Kreuzen Sie ein oder mehrere zutreffende Kästchen an.

Nicht-Muttersprache	Eintauchen in die Sprachumgebung*	Sprachunterricht	Selbststudium
Wähle eine Option	☐	☐	☐
Wähle eine Option	☐	☐	☐
Wähle eine Option	☐	☐	☐
Wähle eine Option	☐	☐	☐

* z. B. Einwanderung in ein anderes Land, in dem sich die vorherrschende Sprache von Ihrer Muttersprache unterscheidet, sodass Sie diese Sprache durch Eintauchen in die Sprachumgebung gelernt haben

12. Geben Sie für jede Sprache, die Sie erlernt haben (einschließlich Ihrer Muttersprache), an, in welchem Alter Sie begonnen haben, diese in den folgenden Situationen zu verwenden.

Sprache	Zu Hause	Mit Freund/innen	In der Schule	Im Beruf	Sprachlernsoftware	Online-Spiele
Wähle eine Option						
Wähle eine Option						
Wähle eine Option						
Wähle eine Option						

13. Stufen Sie Ihre Sprachlernfähigkeit ein. Mit anderen Worten, wie gut, denken Sie, sind Sie darin, neue Sprachen zu lernen, im Vergleich zu Ihren Freund/innen oder anderen Menschen, die Sie kennen?

Sehr schlecht

14. Stufen Sie für jede Sprache, die Sie erlernt haben (einschließlich Ihrer Muttersprache), Ihre derzeitigen Fähigkeiten in den Bereichen Verstehen, Sprechen, Lesen und Schreiben ein.

Sprache	Verstehen	Sprechen	Lesen	Schreiben
Wähle eine Option	Sehr schlecht	Sehr schlecht	Sehr schlecht	Sehr schlecht
Wähle eine Option	Sehr schlecht	Sehr schlecht	Sehr schlecht	Sehr schlecht
Wähle eine Option	Sehr schlecht	Sehr schlecht	Sehr schlecht	Sehr schlecht
Wähle eine Option	Sehr schlecht	Sehr schlecht	Sehr schlecht	Sehr schlecht

15. Falls Sie uns noch weitere Informationen bezüglich Ihres Sprachhintergrunds oder Sprachgebrauchs mitteilen möchten, schreiben Sie diese bitte in die folgende Kommentarbox.

Einreichen

B2. Swedish version

1. Deltagarens ID-nummer

2. Ålder

3. Kön

Välj ett alternativ

4. Utbildning

Välj ett alternativ

5. Föräldrarnas utbildning

Far Välj ett alternativ Mor Välj ett alternativ

6. Handedness

Välj ett alternativ

7. Ange ditt modersmål och alla andra språk du har studerat eller lärt dig, i vilken ålder du började använda varje språk när det gäller att lyssna, tala, läsa och skriva och det totala antalet år du har använt på varje språk .

Språk	Lyssnande	Tala	Läsning	Skrivning	År av användning *
Välj ett alternativ					
Välj ett alternativ					
Välj ett alternativ					
Välj ett alternativ					

* Anmärkningar: Under "Årsanvändning" kan du ha lärt dig ett språk, slutat använda det och sedan börjat använda det igen. Vänligen ange det totala antalet år.

8. Ursprungsland

Välj ett alternativ

9. Bostadsland

Välj ett alternativ

10. Om du har bott eller rest i andra länder än ditt bosättningsland i tre månader eller mer, ange sedan landets namn, vistelsens längd (i månader), språket du använde och hur ofta du använde språk för varje land.

Land:	Vistelsetid (i månader) *:	Språk:	Användningsfrekvens:
Välj ett alternativ		Välj ett alternativ	Välj ett alternativ
Välj ett alternativ		Välj ett alternativ	Välj ett alternativ
Välj ett alternativ		Välj ett alternativ	Välj ett alternativ
Välj ett alternativ		Välj ett alternativ	Välj ett alternativ

*** Du kan ha varit i landet vid flera tillfällen, var och en under olika tid. Lägg till alla resor tillsammans.**

11. Ange hur du lärde dig eller förvärvade ditt språk. Markera en eller flera rutor som gäller.

Icke-modersmål	Nedsänkning*	Klassrumsinstruktion	Självlärande
Välj ett alternativ	☐	☐	☐
Välj ett alternativ	☐	☐	☐
Välj ett alternativ	☐	☐	☐
Välj ett alternativ	☐	☐	☐

*** t.ex. att immigrera till ett annat land där det dominerande språket skiljer sig från ditt modersmål så att du lär dig det här språket genom nedsänkning i språkmiljön.**

12. Ange i vilken ålder du började använda vart och ett av de språk du har studerat eller lärt dig i följande miljöer (inklusive modersmål).

Språk:	Hemma:	Med vänner:	I skolan:	På jobbet:	Språkprogramvara:	Onlinespel:
Välj ett alternativ						
Välj ett alternativ						
Välj ett alternativ						
Välj ett alternativ						

13. Betygsätt din språkinlärningsförmåga. Med andra ord, hur bra känner du att du lär dig nya språk i förhållande till dina vänner eller andra människor du känner?

Välj ett alternativ

14. Betygsätt din nuvarande förmåga när det gäller att lyssna, tala, läsa och skriva på vart och ett av de språk du har studerat eller lärt dig (inklusive modersmål).

Språk:	Lyssnande:	Tala:	Läsning:	Skrivning:
Välj ett alternativ	Välj ett alternativ	Välj ett alternativ	Välj ett alternativ	Välj ett alternativ
Välj ett alternativ	Välj ett alternativ	Välj ett alternativ	Välj ett alternativ	Välj ett alternativ
Välj ett alternativ	Välj ett alternativ	Välj ett alternativ	Välj ett alternativ	Välj ett alternativ
Välj ett alternativ	Välj ett alternativ	Välj ett alternativ	Välj ett alternativ	Välj ett alternativ

15. Använd kommentarfältet nedan för att ge annan information om din språkbakgrund eller din användning.

Submit

C. List of stimuli with translations

-target words in **bold**

1	Flickan gillade **hästen** väldigt mycket. *The girl liked the horse very much.*	Vad gillade flickan väldigt mycket? *What did the girl like very much?*	**-hästen** -vännen *(friend)* -hösten	Noun, vowel ɛ/œ 11 syllables
	Flickan gillade **hösten** väldigt mycket. *The girl liked the autumn very much.*	Vad gillade flickan väldigt mycket? *What did the girl like very much?*	-hästen -vännen **-hösten**	
2	Köksmästaren förberedde sin **dag** i morse. *The chef prepared his day this morning.*	Vad förberedde köksmästaren? *What did the chef prepare?*	Han förberedde -sin deg -sin mat *(food)* **-sin dag**	Noun, vowel a/e 13 syllables
	Köksmästaren förberedde sin **deg** i morse. *The chef prepared his dough this morning*	Vad förberedde köksmästaren? *What did the chef prepare?*	Han förberedde **-sin deg** -sin mat -sin dag	

3	Människor måste ta våld och **hat** på allvar. *People need to take violence and hatred seriously.*	Vad måste människor ta på allvar? *What do people have to take seriously?*	-**våld och hat** -våld och lag *(law)* -våld och hot	Noun, vowel a/u 12 syllables
	Människor måste ta våld och **hot** på allvar. *People need to take violence and threat seriously.*	Vad måste människor ta på allvar? *What do people have to take seriously?*	-våld och hat -våld och lag **-våld och hot**	
4	Oskar presenterar sin nya **båt** på festen. *Oskar is presenting his new boat at the party.*	Vad presenterar Oskar? *What does Oskar present?*	-**sin nya båt** -sin nya låt -sin nya bil *(car)*	Noun, consonant b/l 13 syllables
	Oskar presenterar sin nya **låt** på festen. *Oskar is presenting his new song at the party.*	Vad presenterar Oskar? *What does Oskar present?*	-sin nya båt -**sin nya låt** -sin nya bil	
5	Det fanns en artikel om **hästen** i tidningen. *There was an article about the horse in the newspaper.*	Vad handlade artikeln om? *What was the article about?*	-kungen *(the king)* -festen -**hästen**	Noun, consonant h/f 13 syllables
	Det fanns en artikel om **festen** i tidningen. *There was an article about the party in the newspaper*	Vad handlade artikeln om? *What was the article about?*	- kungen -**festen** -hästen	

6	Ont i **handen** kan vara plågsamt och farligt. *Pain in the hand can be distressful and dangerous.*	Vad kan vara plågsamt och farligt? *What can be distressful and dangerous?*	-ont i tanden -ont i ryggen *(pain in the back)* **-ont i handen**	Noun, consonant h/t 12 syllables
	Ont i **tanden** kan vara plågsamt och farligt. *Pain in the tooth can be distressful and dangerous.*	Vad kan vara plågsamt och farligt? *What can be distressful and dangerous?*	**-ont i tanden** -ont i ryggen -ont i handen	
7	Skådespelarens kostym var **fel** igår. *The actor's costume was wrong yesterday.*	Vad stämmer in på kostymen? *What applies to the costume?*	-kostymen var ny *(new)* **-kostymen var fel** -kostymen var ful	Adverb, vowel e/~~u~~ 11 syllables
	Skådespelarens kostym var **ful** igår. *The actor's costume was ugly yesterday.*	Vad stämmer in på kostymen? *What applies to the costume?*	-kostymen var ny -kostymen var fel **-kostymen var ful**	
8	Den berömda cirkusen är **nu** i staden. *The famous circus is now in town.*	Vad stämmer in på cirkusen? *What applies to the circus?*	-är ny i staden **-är nu i staden** -är här i staden *(here in town)*	Adjective, vowel, ~~u~~/y 12 syllables

	Den berömda cirkusen är **ny** i staden. *The famous circus is new in town.*	Vad stämmer in på cirkusen? *What applies to the circus?*	**-är ny i staden** -är nu i staden -är här i staden	
9	På morgonen var **inga** hundar i parken. *There were no dogs in the park in the morning.*	Vem var i parken på morgonen? *Who was in the park in the morning?*	-många hundar *(many dogs)* -**inga hundar** -unga hundar	Adjective, vowel ɪ/ɵ 12 syllables
	På morgonen var **unga** hundar i parken. *There were young dogs in the park in the morning.*	Vem var i parken på morgonen? *Who was in the park in the morning?*	-många hundar -inga hundar -**unga hundar**	
10	Mamma tyckte inte om den **fula** dörren. *Mum didn't like the ugly door.*	Vad tyckte mamma inte om? *What didn't mum like?*	**-den fula dörren** -den gröna dörren *(the green door)* -den gula dörren	Adjective, consonant f/g 12 syllables
	Mamma tyckte inte om den **gula** dörren. *Mum didn't like the yellow door.*	Vad tyckte mamma inte om? *What didn't mum like?*	-den fula dörren -den gröna dörren -**den gula dörren**	

11	Barnen var **korta** och behövde hjälp. *The children were short and needed help.*	Vad stämmer in på barnen? *What applies to the children?*	De var -borta **-korta** -nära *(near)*	Adjective, consonant k/b 11 syllables
	Barnen var **borta** och behövde hjälp. *The children were gone and needed help.*	Vad stämmer in på barnen? *What applies to the children?*	De var **-borta** -korta -nära	
12	Det fanns en **naken** person i sovrummet. *There was a naked person in the bedroom.*	Vad stämmer in på personen? *What applies to the person?*	Personen var -hungrig *(hungry)* -vaken **-naken**	Adjective, consonant n/v 11 syllables
	Det fanns en **vaken** person i sovrummet. *There was an awakened person in the bedroom.*	Vad stämmer in på personen? *What applies to the person?*	Personen var -hungrig **-vaken** -naken	
13	Demonstranterna **hatade** politikerna. *The demonstrators hated the politicians.*	Vad gjorde demonstranterna? *What did the demonstrators do?*	**-de hatade politikerna** -de hotade politikerna -de gillade politikerna *(they liked the politcians)*	Verb, vowel a/u 12 syllables

	Demonstranterna **hotade** politikerna. *The demonstrators threatened the politicians.*	Vad gjorde demonstranterna? *What did the demonstrators do?*	-de hatade politikerna **-de hotade politikerna** -de gillade politikerna	
14	Mannen vill gärna **lura** mig ordentligt. *The man wants to trick me properly.*	Vad vill mannen göra? *What does the man want to do?*	-lära mig -visa mig *(show me)* **-lura mig**	Verb, vowel ʉ/ɛ 11 syllables
	Mannen vill gärna **lära** mig ordentligt. *The man wants to teach me properly.*	Vad vill mannen göra? *What does the man want to do?*	**-lära mig** -visa mig -lura mig	
15	Anna **ritade** sin familj efter skolan *Anna drew her family after school.*	Vad gjorde Anna efter skolan? *What did Anna do after school?*	-hon frågade sin familj *(she asked her family)* -hon retade sin familj **-hon ritade sin familj**	Verb, vowel i/e 12 syllables
	Anna **retade** sin familj efter skolan. *Anna annoyed her family after school.*	Vad gjorde Anna efter skolan? *What did Anna do after school?*	-hon frågade sin familj **-hon retade sin familj** -hon ritade sin familj	

16	Man får inte **göra** fel under körkortsprovet. *One may not make a mistake during the driving test.*	Vad får man inte göra? *What may one not do?*	**-göra fel** -räkna fel *(miscount)* -köra fel	Verb, consonant j/ɕ 13 syllables
	Man får inte **köra** fel under körkortsprovet. *One may not get lost during the driving test.*	Vad får man inte göra? *What may one not do?*	-göra fel -räkna fel **-köra fel**	
17	Föräldrarna **ber** henne komma till mötet. *The parents ask her to come to the meeting.*	Vad gör föräldrarna? *What do the parents do?*	-de ser henne **-de ber henne** -de hör henne *(they hear her)*	Verb, consonant b/s 12 syllables
	Föräldrarna **ser** henne komma till mötet. *The parents see her coming to the meeting.*	Vad gör föräldrarna? *What do the parents do?*	**-de ser henne** -de ber henne -de hör henne	
18	Mannen **lever** med en positiv inställning. *The man is living with a positive attitude.*	Vad gör mannen? *What does the man do?*	-leder med en positiv inställning -jobba med en positiv inställning *(works with a positive attitude)* **-lever med en positiv inställning**	Verb, consonant v/d 12 syllables

	Mannen **leder** med en positiv inställning. *The man is leading with a positive attitude.*	Vad gör mannen? *What does the man do?*	**-leder med en positiv inställning** -jobba med en positiv inställning -lever med en positiv inställning

D. List of filler sentences with translations

1a	Mammas hatt ligger på skåpet i sovrummet. *Mum's hat lies on the wardrobe in the bedroom.*	Ligger mammas hatt i skåpet? *Does mum's hat lie in the wardrobe?*	Ja/**nej** *(yes/no)*
1b	Mammas hatt ligger under skåpet i sovrummet. *Mum's hat lies under the wardrobe in the bedroom.*	Ligger mammas hatt under skåpet? *Does mum's hat lie under the wardrobe?*	**Ja**/nej
2a	Lisa fick brevet från Anna igår. *Lisa received the letter from Anna.*	Anna skrev brevet. *Anna wrote the letter.*	**Rätt**/fel *(right/wrong)*
2b	Brevet till Anna fanns i brevlådan. *The letter to Anna was in the mailbox.*	Anna skrev brevet. *Anna wrote the letter.*	Rätt/**fel**

3a	Pojkarna ville gömma sina kompisars väskor. *The boys wanted to hide their friends' bags.*	Ville pojkarna gömma sina kompisars flaskor? *Did the boys want to hide their friends' bottles?*	Ja/**nej**
3b	Pojkarna ville gömma sina kompisars flaskor. *The boys wanted to hide their friends' bottles.*	Ville pojkarna gömma sina kompisars flaskor? *Did the boys want to hide their friends' bottles?*	**Ja**/nej
4a	Pappa var borta, men Lisa visste inte det. *Dad was gone but Lisa didn't know that.*	Visste Lisa att pappa var borta? *Did Lisa know that dad was gone?*	Ja/**nej**
4b	Pappa var borta och Lisa visste det. *Dad was gone, and Lisa knew that.*	Visste Lisa att pappa var borta? *Did Lisa know that dad was were gone?*	**Ja**/nej
5a	Festen ägde rum igår och alla tyckte det var trevligt. *The party took place yesterday and everyone thought it was nice.*	Ska festen äga rum imorgon? *Is the party taking place tomorrow?*	Ja/**nej**
5b	Festen ska äga rum imorgon och alla hoppas det blir trevligt. *The party takes place tomorrow and everyone hopes it will be nice.*	Ska festen äga rum imorgon? *Is the party taking place tomorrow?*	**Ja**/nej
6a	Frågorna på provet är svåra. *The questions in the test were difficult.*	Frågorna är lätta. *The questions are easy.*	Rätt/**fel**

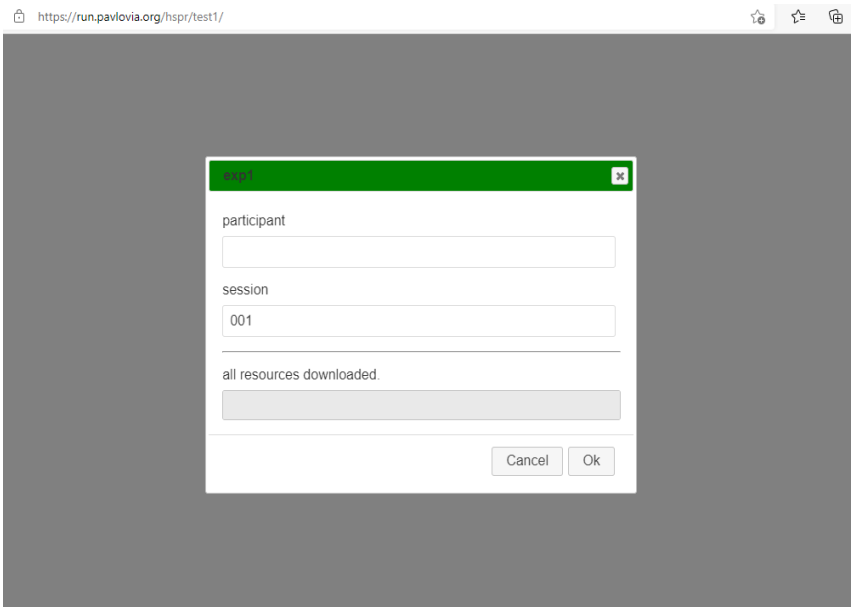
6b	Frågorna på provet är lätta. *The questions in the test were easy.*	Frågorna är inte svåra. *The questions are not difficult.*	**Rätt**/fel
7a	En olycka hände på gatan utanför staden. *An accident happened on the street out of town.*	Var hände olyckan? *Where did the accident happen?*	I staden**, På gatan**, framför rådhuset *(in the town, on the street, in front of the town hall)*
7b	En olycka hände på gatan mitt i staden. *An accident happened on the street in the middle of the town.*	Var hände olyckan? *Where did the accident happen?*	**I staden**, framför skolan, på trottoaren *(in the town, in front of the school, on the sidewalk)*
8a	Alla ser fram emot morgondagens fotbollsmatch. *Everyone looks forwards to tomorrow's football match.*	När är fotbollsmatchen? *When is the football match?*	i eftermiddag, om en vecka, **imorgon** *(in the afternoon, in one week, tomorrow)*
8b	Alla såg fram emot gårdagens fotbollsmatch. *Everyone has looked forward to yesterday's football match.*	När var fotbollsmatchen? *When was the football match?*	I förrgår, **igår**, förra veckan *(the day before yesterday, yesterday, last week)*
9a	Mannen kan vinna tävlingen imorgon. *The man can win the contest tomorrow.*	Har mannen redan vunnit tävlingen? *Has the man won the contest already?*	Ja/**nej**
9b	Mannen vann gårdagens tävling. *The man won yesterday's contest.*	Har mannen redan vunnit tävlingen? *Has the man won the contest already?*	**Ja**/nej

10a	Lisa och Anna letar efter Lukas, men de hittar honom inte. *Lisa and Anna looked for Lukas, but they didn't find him.*	Vem letar de efter? *Who are they looking for?*	**Lukas**, Anna, Lisa
10b	Lisa och Lukas letar efter Anna, men de hittar henne inte. *Lisa and Lukas looked for Anna, but they didn't find her.*	Vem letar de efter? *Who are they looking for?*	Lukas, **Anna**, Lisa
11a	Lunchen kostar 20 kronor. *The lunch costs 20 kronor.*	Hur mycket kostar lunchen? *How much does the lunch cost?*	70, **20**, 50 kronor
11b	Lunchen kostar 70 kronor. *The lunch costs 70 kronor.*	Hur mycket kostar lunchen? *How much does the lunch cost?*	**70**, 20, 50 kronor
12a	Alla, särskilt Tom, gillar matteläraren. *Everyone, especially Tom, like the math teacher.*	Gillar Tom matteläraren? *Does Tom like the math teacher?*	**Ja**/nej
12b	Alla utom Tom gillar matteläraren. *Everyone except Tom like the math teacher.*	Gillar Tom matteläraren? *Does Tom like the math teacher?*	Ja/**nej**
13a	Marie dricker kaffe utan mjölk. *Marie drinks coffee without milk.*	Finns det mjölk i Maries kaffekopp? *Is there milk in Marie's coffee cup?*	Ja/**nej**

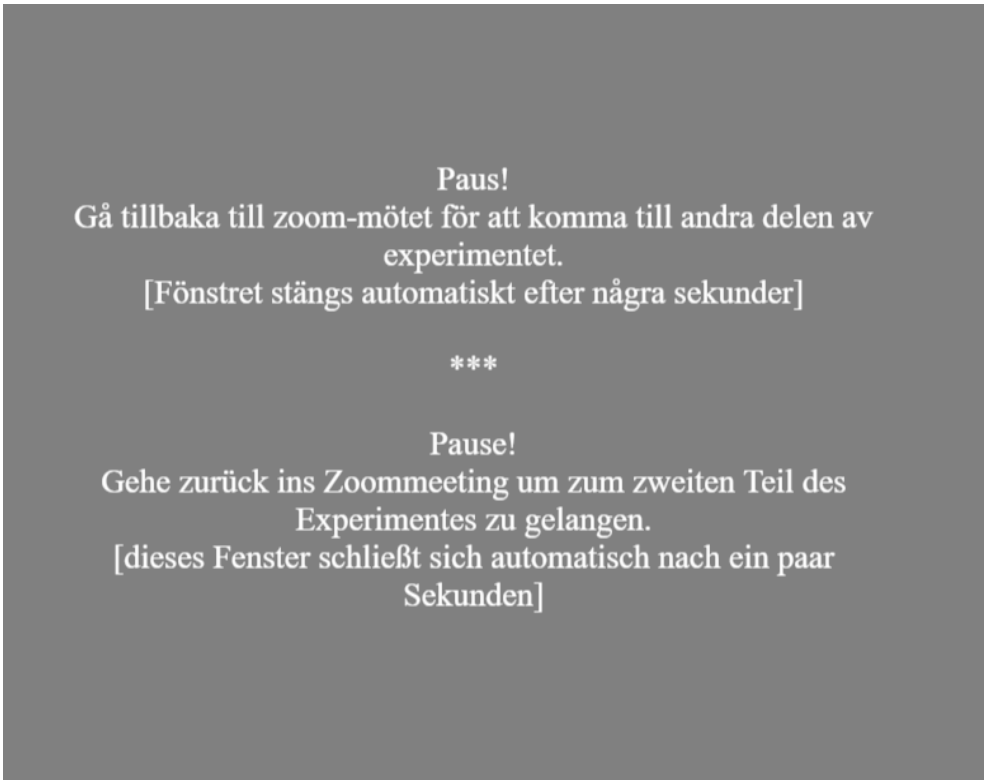
13b	Marie dricker kaffe med mjölk. *Marie drinks coffee with milk.*	Finns det mjölk i Maries kaffekopp? *Is there milk in Marie's coffee cup?*	**Ja**/nej
14a	Barnen vill inte äta grönsaker till middag. *The children don't want to eat vegetables for dinner.*	Barnen vill inte ha köttbullar till middag. *The children don't want to have meatballs for dinner.*	Rätt/**fel**
14b	Barnen vill inte äta köttbullar till middag. *The children don't want to eat meatballs for dinner.*	Barnen vill inte ha köttbullar till lunchen. *The children don't want to have meatballs for lunch.*	Rätt/**fel**
15a	En bil kör runt mannen på gatan. *A car is driving round the man on the street.*	Vad gör bilen? *What does the car do?*	Kör framför mannen, **kör runt mannen**, kör över mannen *(drives in front of the man, drives around the man, runs over the man)*
15b	En bil kör över mannen på gatan. *A car is knocking over the man on the street.*	Vad gör bilen? *What does the car do?*	Kör framför mannen, kör runt mannen, **kör över mannen**
16a	Nycklarna ligger i skålen på bordet. *The keys lie in the bowl on the table.*	Ligger nycklarna under bordet? *Are the keys lying under the table?*	Ja/**nej**

16b	Nycklarna ligger bredvid skålen på bordet. *The keys lie next to the bowl on the table.*	Ligger nycklarna på bordet? *Are the keys lying on the table?*	**Ja**/nej
17a	Barnen är inte hemma utan fortfarande i skolan. *The children are not home but still at school.*	Är barnen i skolan? *Is the child at school?*	**Ja**/nej
17b	Barnen är inte hemma utan redan i skolan. *The children are not home but already at school.*	Är barnen i skolan? *Is the child at school?*	**Ja**/nej
18a	Kursen är på tisdagar. *The course is on Tuesdays.*	Kursen är på torsdagar. *The course is on Thursdays.*	Rätt/**fel**
18b	Kursen är på torsdagar. *The course is on Thursdays.*	Kursen är på tisdagar. *The course is on Tuesdays.*	Rätt/**fel**

E. Apparatus


https://run.pavlovia.org/hspr/test1/
exp1
participant
session
001
all resources downloaded.
Cancel
Ok


https://run.pavlovia.org/hspr/test1/
Nu börjar experimentet.
Du kommer se varje klipp bara en gång. För att svara på frågorna tryck 1, 2 eller 3 på datorns tangentbord. Det handlar inte om hastigheten utan om att ditt svar är korrekt.
Om experimentet hänger sig eller om du vill avbryta, tryck på Escape-knappen för att stänga fönstret.
Tryck på mellanslagstangenten när du är redo att starta experimentet!
***
Das Experiment beginnt jetzt.
Du siehst jede Aufnahme nur einmal. Um die Fragen zu beantworten, drücke 1, 2 oder 3 auf deiner Computertastatur. Es geht nicht um die Schnelligkeit sondern die Korrektheit deiner Antwort.
Sollte das Experiment stehen bleiben oder solltest du abbrechen wollen, drücke die Escape-Taste um das Fenster zu schließen.
Drücke jetzt die Leertaste, wenn du bereit bist!


Paus!
Gå tillbaka till zoom-mötet för att komma till andra delen av experimentet.
[Fönstret stängs automatiskt efter några sekunder]
***
Pause!
Gehe zurück ins Zoommeeting um zum zweiten Teil des Experimentes zu gelangen.
[dieses Fenster schließt sich automatisch nach ein paar Sekunden]

F. Experimental design overview (stimulus items & filler)

In the following table, A stands for the audio-only conditions, whereas AV and B represent the audiovisual conditions with the full face(AV) and blurred face (B). One group was presented with all the green stimuli (randomized), the other group with the white stimuli (also randomized).

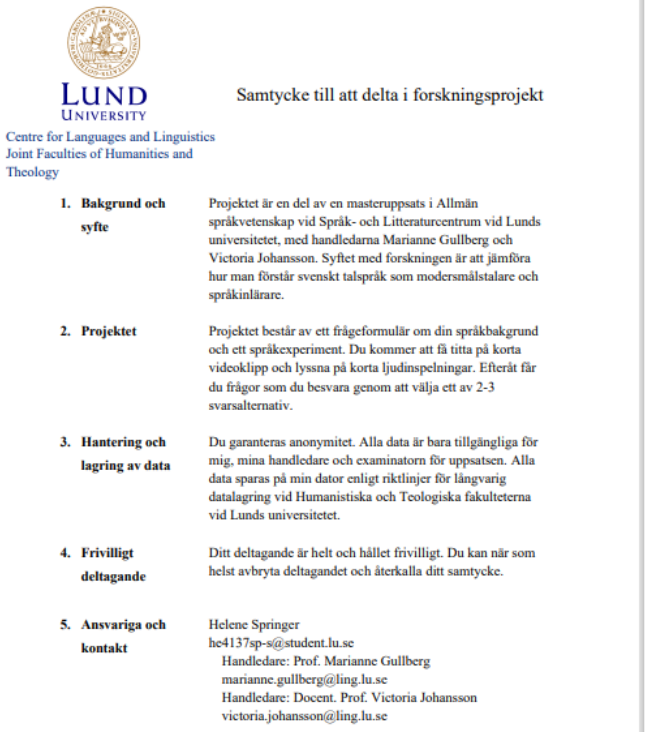
hästen A	hösten A	dag A	deg A
hästen AV	hösten AV	dag AV	deg AV
hästen B	hösten B	dag B	deg B
hat A	hot A	båt A	låt A
hat AV	hot AV	båt AV	låt AV
hat B	hot B	båt B	låt B
hästen/artikel A	festen A	handen A	tanden A
hästen/artikel AV	festen AV	handen AV	tanden AV
hästen/artikel B	festen B	handen B	tanden B
fel A	ful A	nu A	ny A
fel AV	ful AV	nu AV	ny AV
fel B	ful B	nu B	ny B
inga A	unga A	fula A	gula A
inga AV	unga AV	fula AV	gula AV
inga B	unga B	fula B	gula B
korta A	borta A	naken A	vaken A
korta AV	borta AV	naken AV	vaken AV
korta B	borta B	naken B	vaken B
hatade A	hotade A	lura A	lära A
hatade AV	hotade AV	lura AV	lära AV
hatade B	hotade B	lura B	lära B
rita A	reta A	göra A	köra A
rita AV	reta AV	göra AV	köra AV
rita B	reta B	göra B	köra B
ber A	ser A	lever A	leder A
ber AV	ser AV	lever AV	leder AV
ber B	ser B	lever B	leder B

Filler

1a A	1b A	2a A	2b A
1a AV	1b AV	2a AV	2b AV
1a B	1b B	2a B	2b B
3a A	3b A	4a A	4b A
3a AV	3b AV	4a AV	4b AV
3a B	3b B	4a B	4b B
5a A	5b A	6a A	6b A
5a AV	5b AV	6a AV	6b AV
5a B	5b B	6a B	6b B
7a A	7b A	8a A	8b A
7a AV	7b AV	8a AV	8b AV
7a B	7b B	8a B	8b B
9a A	9b A	10a A	10b A
9a AV	9b AV	10a AV	10b AV
9a B	9b B	10a B	10b B
11a A	11b A	12a A	12b A
11a AV	11b AV	12a AV	12b AV
11a B	11b B	12a B	12b B
13a A	13b A	14a A	14b A
13a AV	13b AV	14a AV	14b AV
13a B	13b B	14a B	14b B
15a A	15b A	16a A	16b A
15a AV	15b AV	16a AV	16b AV
15a B	15b B	16a B	16b B
17a A	17b A	18a A	18b A
17a AV	17b AV	18a AV	18b AV
17a B	17b B	18a B	18b B

G. Consent forms

G1. Swedish version for Swedish L1 participants



LUND UNIVERSITY

Centre for Languages and Linguistics
Joint Faculties of Humanities and Theology

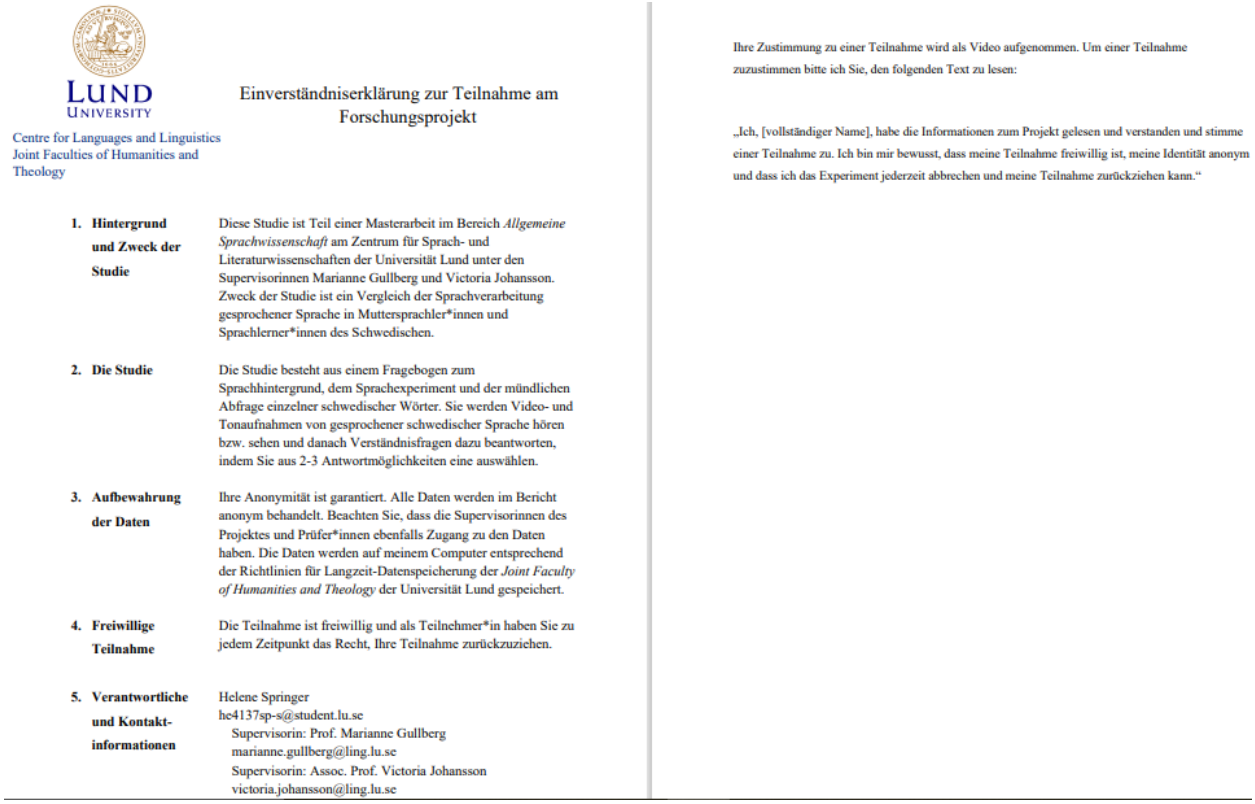
Samtycke till att delta i forskningsprojekt

1. Bakgrund och syfte	Projektet är en del av en masteruppsats i Allmän språkvetenskap vid Språk- och Litteraturcentrum vid Lunds universitetet, med handledarna Marianne Gullberg och Victoria Johansson. Syftet med forskningen är att jämföra hur man förstår svenskt talspråk som modersmålstalare och språkinlärare.
2. Projektet	Projektet består av ett frågeformulär om din språkbakgrund och ett språkexperiment. Du kommer att få titta på korta videoklipp och lyssna på korta ljudinspelningar. Efteråt får du frågor som du besvara genom att välja ett av 2-3 svarsalternativ.
3. Hantering och lagring av data	Du garanteras anonymitet. Alla data är bara tillgängliga för mig, mina handledare och examinatorn för uppsatsen. Alla data sparas på min dator enligt riktlinjer för långvarig datalagring vid Humanistiska och Teologiska fakulteterna vid Lunds universitetet.
4. Frivilligt deltagande	Ditt deltagande är helt och hållet frivilligt. Du kan när som helst avbryta deltagandet och återkalla ditt samtycke.
5. Ansvariga och kontakt	Helene Springer he4137sp-s@student.lu.se Handledare: Prof. Marianne Gullberg marianne.gullberg@ling.lu.se Handledare: Docent. Prof. Victoria Johansson victoria.johansson@ling.lu.se

Ditt samtycke spelas in. För att samtycka ber jag dig läsa nedanstående text:

"Jag, [fullständigt namn], här läst och förstått informationen ovan och samtycker till att delta i forskningsprojektet. Jag är medveten om att mitt deltagande är frivilligt, min identitet är anonym och att jag kan avbryta experimentet när som helst och återkalla samtycket."

G2. German version for German L2 participants



LUND UNIVERSITY

Centre for Languages and Linguistics
Joint Faculties of Humanities and Theology

Einverständniserklärung zur Teilnahme am Forschungsprojekt

1. Hintergrund und Zweck der Studie	Diese Studie ist Teil einer Masterarbeit im Bereich *Allgemeine Sprachwissenschaft* am Zentrum für Sprach- und Literaturwissenschaften der Universität Lund unter den Supervisorinnen Marianne Gullberg und Victoria Johansson. Zweck der Studie ist ein Vergleich der Sprachverarbeitung gesprochener Sprache in Muttersprachler*innen und Sprachlerner*innen des Schwedischen.
2. Die Studie	Die Studie besteht aus einem Fragebogen zum Sprachhintergrund, dem Sprachexperiment und der mündlichen Abfrage einzelner schwedischer Wörter. Sie werden Video- und Tonaufnahmen von gesprochener schwedischer Sprache hören bzw. sehen und danach Verständnisfragen dazu beantworten, indem Sie aus 2-3 Antwortmöglichkeiten eine auswählen.
3. Aufbewahrung der Daten	Ihre Anonymität ist garantiert. Alle Daten werden im Bericht anonym behandelt. Beachten Sie, dass die Supervisorinnen des Projektes und Prüfer*innen ebenfalls Zugang zu den Daten haben. Die Daten werden auf meinem Computer entsprechend der Richtlinien für Langzeit-Datenspeicherung der *Joint Faculty of Humanities and Theology* der Universität Lund gespeichert.
4. Freiwillige Teilnahme	Die Teilnahme ist freiwillig und als Teilnehmer*in haben Sie zu jedem Zeitpunkt das Recht, Ihre Teilnahme zurückzuziehen.
5. Verantwortliche und Kontakt-informationen	Helene Springer he4137sp-s@student.lu.se Supervisorin: Prof. Marianne Gullberg marianne.gullberg@ling.lu.se Supervisorin: Assoc. Prof. Victoria Johansson victoria.johansson@ling.lu.se

Ihre Zustimmung zu einer Teilnahme wird als Video aufgenommen. Um einer Teilnahme zuzustimmen bitte ich Sie, den folgenden Text zu lesen:

„Ich, [vollständiger Name], habe die Informationen zum Projekt gelesen und verstanden und stimme einer Teilnahme zu. Ich bin mir bewusst, dass meine Teilnahme freiwillig ist, meine Identität anonym und dass ich das Experiment jederzeit abbrechen und meine Teilnahme zurückziehen kann."

G3. Consent form recordings

LUND UNIVERSITY
Lund University Humanities Lab

CENTRE FOR LANGUAGES & LITERATURE
LUND UNIVERSITY
PO BOX 201, 211 00 Lund, Sweden

Consent Form

I hereby give my permission to Helene Springer (Student at Centre for Languages and Literature, Lund University, Sweden), to use today's recordings (audio and video) for the following purposes:

(Please tick the appropriate box, "☐", if you give your permission.)

☐ 1. analyses for scientific research

2. as illustrations of the above scientific research in professional seminars, lectures, conferences, and in scientific publications;

☐ as still photographs;

☐ as video clips.

My anonymity is guaranteed. Under no circumstances will my personal identity be revealed to anybody other than the above-mentioned researchers (e.g. no names will be used in presentations of the recording).

Name Signature Date

_____