Shadowing empathy

A study about the influence of empathy on speech shadowing

Lina van Otterdijk

s4767012

23-04-2020

Bachelor thesis

First supervisor: Dr. Linda Drijvers

Second supervisor: Dr. Judith Holler

Second reader: Dr. Asli Özyürek

Table of contents

Abstract

Human face-to-face communication entails the rapid integration of auditory as well as visual signals. For this thesis, we examined whether such a multimodal context offers a benefit during language processing and, in addition, whether empathy has an influence on this procedure of multimodal language processing. In employing a shadowing task, the participants were required to repeat (shadow) speech as he or she hears it (i.e. the shadower started to repeat the fragment before he or she has heard all of it). The stimuli consisted of natural dyadic conversations and were presented in three different conditions: audio only (AO), audiovisual with a blurred mouth (AB) and audiovisual (AV). Along with the shadowing experiment, participants were asked to fill out the Empathy Quotient. Results demonstrated that participants made less errors in the shadowing task as more visual context was present, implying that a multimodal advantage for language processing indeed can be found. We argue that a possible underlying cognitive mechanism for this multimodal advantage could be prediction. The error rates from the shadowing experiment compared to the empathy scores revealed an overall trend: the higher the empathy score, the lower the amount of errors. This correlation was significant for the AV condition, which means that listeners with a high empathy score seem to be more responsive to visual cues, and experience more of the multimodal benefit, than listeners with a low empathy score. However, no significant correlation could be found between empathy and the AB (or the AO) condition, for which we speculate that the blurring of the mouth may have been distracting and that the effect of empathy on multimodal language processing therefore may be diminished. To conclude, we believe that a multimodal context may facilitate predictive language processing and that someone's level of empathy can affect how much the listener benefits from this multimodal advantage.

Introduction

Language is a complex construct with many facets playing a role in comprehension and production in order to establish successful communication. The importance of the state in which language occurs, has become clear these past few years. It is overall accepted that most of the human face-to-face communication is multimodal (Chen & Rao, 1998; Holler & Levinson, 2019; Scollon & Levine, 2004): speech consists of more than just audio signals. We speak of multimodal communication when a combination of signals is received through multiple sensory channels. In human face-to-face conversation this entails mainly the sensory channels for auditory and visual signals. Speech is often accompanied by lip movements, other facial movements and bodily movement, such as gesturing. One might consider that this many cues can be distracting or confusing and therefore slow down one's process of language processing. Remarkably, the opposite seems to be true. Although there might be a lot of visual context to be taken in at once, all these cues combined, appear to be enhancing language processing (Drijvers & Özyürek, 2017; McGurk & MacDonald, 1976; Norris, 2004; Van Leeuwen, 2004). Empirical evidence of how visual cues from the facial area contribute to the procedure of language processing can be found in early as well as recent studies (Sumby & Pollack, 1954; Thompson, Garcia and Malloy, 2007). Venezia, Thurman, Matchin, George and Hickok (2016) found that visual speech signals often precede visual speech by 100-200 ms and suggested that visual input can aid in predicting upcoming speech. These studies however focus only on facial cues. Next to facial movements, bodily movements seem to play an important role in conveying a message; especially manual gestures appear to carry a great amount of meaning in concurrence with what is verbally being said (Gerwing & Allison, 2009; Kelly, Healy, Özyürek & Holler, 2015). Holler, Kendrick and Levinson (2018) provide some first insights about the influence of bodily movements on the psycholinguistic process of human communication. Their research focused on question-response sequences in conversation and they found that gestures accompanying the question ensure a faster response. This would support the hypothesis that visual information aids language processing. The current study aims to explore further insights in this subject and to establish whether visual context indeed aids the procedure of language processing.

Language comprehension and production

In order to understand how we expect visual input to be a beneficiary factor in language processing, it is important to first know how comprehension and production in language operate and coincide. Comprehension is about how people process and extract the overall meaning, how they deduce a lot of information that is not explicitly stated, and how they decide which details to leave out of their language analysis (Mandler, 1979). The speech-recognition system processes language mainly through word hypothesizing (i.e. word 'guessing'), to improve and speed up comprehension. From early on, it was believed that there were two ways in which people can manage this process: the bottom-up approach or the top-down approach (Smith & Erman, 1981). Top-down processing (conceptually driven) means that to understand language, people begin with the so-called 'higher-level' features, such as general knowledge, background information, context and semantics, and later proceed with 'lower-level' features, such as syntactic and phonological features. Bottom-up processing (data-driven) proceeds the other way around; starting with recognizing phonemes, which are combined to form syllables, syllables then form into words, multiple words together become clauses, and so on, to where the 'higher-level' features are used for analysis (Mandler, 1979). At first, it was thought that it was either the one route or the other; the top-

down approach being the most accurate and appropriate for use, placing language input immediately into context, but also being the slower procedure, going from a wide range of information to a smaller set (whereas bottom-down starts small and builds from there) (Smith & Erman, 1981). However, more recent research suggests that there is no such thing as the classical two-step model of either top-down or bottom-up, but rather that there is one concomitant process (Friederici, 2012; Hagoort & van Berkum, 2007). Since we are mainly interested in face-to-face communication (it is the way in which language occurs most often), it is good to look at the comprehension process during conversation specifically. When looking at discourse comprehension, it seems it is mainly making use of the 'higher-level' features; it requires many connections to be made between different parts of language, in order to comprehend it in its whole. There are two important points concerning these connections: one being that the knowledge used for making connections is mainly knowledge of the world, rather than knowledge of language itself; and two being that an indefinite amount of connections can be deduced and only specific selection can be used (Carpenter, Miyake & Just, 1995). Language comprehension is thus a multi-level and concomitant process (lexical, syntactic and thematic analyses have to be made) and it has to be coordinated properly as well in timing as in functionality.

Producing linguistic utterances is, just like comprehending them, a multicomponent multistage process. Word production knows roughly the following sequence of actions: conceptual preparation, lexical selection, morphological and phonological encoding, and phonetic encoding, which leads to articulation (Ferreira, 2010; Levelt, Roelofs & Meyer, 1999). Sentence production, however, becomes more complex already and consists of the word production process as well as function assignment and constituent assembly: words are to be given grammatical roles and relations, which have to be framed in hierarchical order by the syntax (Ferreira, 2010). Speakers employ both hierarchical and incremental planning (Lee, Brown-Smith & Watson, 2013): meaning that speakers prepare a chunk of an utterance before speech onset and prepare the next chunk of the utterance during speech (incremental), and that they plan syntactically dependent lexical items together (hierarchical). Another step further in the language production process is dialogue. For conversational language to be produced it is important to separate the knowledge both the speaker as the hearer know (common ground) from the knowledge only the speaker knows (privileged ground) (Ferreira, 2010). Besides the content of the information being transferred in a conversation, the produced language also shows a high level of alignment. Interlocutors usually use the same words and the same aspects of meaning (Garrod & Anderson, 1987), the same grammatical forms (Branigan, Pickering & Cleland, 2000) and also imitate accent and speech rate, and adopt increasingly similar phonetic realizations of repeated words (Pardo, 2006). The language production process therefore has different cognitive processes working together.

Both production and comprehension have to be coordinated carefully in order for people to use language successfully and communicate. These two processes can occur separately from each other in rare cases: such as understanding a certain dialect or a second language quite well, but being unable to produce it (comprehension alone); or reciting a part of the bible for example, without understanding what it entails. This suggests that the process for language production is indeed quite different from the process of language comprehension. However, there is also a certain level of coordination of what we can produce with what we can understand (Clark & Hecht, 1983); we attach meaning to a word on the basis of our own production, when hearing a word in a different dialect for instance. Christiansen and Chater (2016) and Chater, McCauley and Christiansen (2016) support the idea that language production and comprehension are combined into one skill (rather than being two separate skills) with the Now-or-Never Bottleneck theory: an important constraint on the language

system is the combination of our poor memory for auditory and visual signals together with the fast and fleeting nature of linguistic input. Incoming information needs to be processed rather immediately in order for it not to be overwritten with quickly following new information. This suggests indeed that comprehension and production have to go hand in hand. In exploring dual-tasking, Fairs (2019) proposed a model in which comprehension and production overlap in timing as well. According to her model, she suggested that lexical concept and lemma selection stages cannot occur in overlap in production and comprehension, but lexical input and output word forms and input and output phonemes can be selected in parallel. Furthermore, seeing and hearing speech appears to excite the motor system involved with speech production (Watkins, Strafella & Paus, 2003; Fadiga, Craighero, Buccino & Rizzolatti, 2002). Considering that incoming speech will have to be understood first, it can be assumed that there is a close relation between the comprehension process and the production process, since the motor system appears to become activated already when speech is just being comprehended. Hickok and Poeppel (2007) suggest a dual-stream neuroanatomical model in which the ventral stream processes speech signals for comprehension and the dorsal stream maps acoustic signals to frontal lobe articulatory networks, at the same time. Existing literature therefore suggests that, even though language comprehension and production can be considered as two different processes, they also show a lot of similarities and appear to be working in a coordinated structure as well.

Speech shadowing

An experimental setting which is often used to test (the concurrence of) comprehension and production during online language processing is called speech shadowing. With this experiment, the subject is required to repeat ('shadow') speech as he or she hears it, as accurately and quickly as possible. This means that the shadower is presented with a fragment of speech and will start repeating it before he or she has even heard the end of that fragment (Marslen-Wilson, 1973, 1985). There exists a delay between hearing something and being able to repeat it. This delay reflects the process of speech perception, and the shorter this latency, the more directly it represents this process (Marslen-Wilson, 1985). Participants who partake in shadowing also make errors in their shadowed speech, which can tell us something about language processing as well. The errors made during such a shadowing task reflect how well someone can coordinate perception and production; the lesser the amount of errors, the better both processes coincide. Marslen-Wilson (1973) showed that these errors usually remain suitable, as well syntactically as semantically, in that specific context. This would suggest that not only comprehension and production form a concomitant process, but that this procedure of language processing entails prediction of some kind, in 'guessing' what types or forms of words might follow. As suggested earlier, visual signals could provide a benefit in predicting upcoming speech (Venezia et al., 2016). Nonetheless, the existing studies that have employed the speech shadowing experiment have either focused on a unimodal setting, presenting only auditory signals (Bailly, 2003; Brouwer, Holger & Falk, 2010), made use of rather simple stimuli, such as single words or syllables (Beautemps, Cathiard & Borgne, 2003; Scarbel, Beautemps, Schwartz & Sato, 2014), or conducted the experiment both in unimodal setting and with short stimuli (Mitterer & Ernestus, 2008). Even more, when researchers did look into multimodality when testing the shadowing paradigm, the visual signals mainly existed of facial movements from the neck up. This seems to be somewhat insufficient, since suggestions arose from earlier work that bodily movements might have significant influence as well (Kelly et al., 2015). The current study utilizes the shadowing

experiment in order to explore the influence of the multimodal context (including bodily movements) on language processing. Moreover, the stimuli consist of longer fragments of speech taken from a conversational context, which resembles more to human face-to-face communication than the short stimuli and therefore makes the results more generalizable. We expect the participants to have a lower amount of errors in their performance of the shadowing task in multimodal setting, compared to the unimodal setting. In other words, we expect that multimodal context indeed helps in speeding up language processing.

One explanation for this expectation derives from Holler and Levinson (2019), who pose prediction as an essential factor in multimodal language processing. During comprehension, multimodal signals can provide a meticulous framework in searching for the correct lexical item (top-down); lip movement, for example, can indicate the start of a w-sound, limiting the lexical search to a selection of w-words. Moreover, due to frequent cooccurrence of multimodal signals, someone can predict what multimodal signals are coming up by assessing statistical association (bottom-up); a certain lip movement, for instance, can often go together with an eyebrow raise, so when seeing that specific lip movement, a quick prediction can be made that an eyebrow raise will follow, and with that, a question. Thus, prediction appears throughout the procedure of (multimodal) language processing on interacting bottom-up and top-down levels and could therefore be a solid explanation of the supposed multimodal advantage.

Predictive language processing

More research on the role of prediction during (multimodal) language processing can be found. Pickering and Garrod (2013) state how production and comprehension are interwoven and how this enables people to predict themselves and each other. We begin predicting language input at a very young age. Mani and Huettig (2012) employed a preferential looking paradigm and found that 2-year-olds, with a large productive vocabulary, already used information of the verb to predict an upcoming theme. The fact that children at such a young age already make use of the predictive language strategy, while still learning, is an important indicator of how important this technique is for language processing. Especially production seems to have an influence on prediction during comprehension (Hintz, Meyer & Huettig, 2016; Pickering & Garrod, 2007). Pickering and Garrod (2007) state that just like motor involvement during the perception of body movement can facilitate that process of perception, so involvement of the production system during comprehension can facilitate that process of comprehension. Hintz, Meyer and Huettig (2014) even found that prediction occurs to a stronger degree when a production task is used as to when only a comprehension task is used, which implies just how essential the production process is when it comes to predictive language processing. Predictions occur at different times and at various linguistic levels. They are core to the survival of species and the multimodal aspect of language, having different kinds of signals over different modalities, seems to facilitate this predictive language use in face-to-face communication (Holler & Levinson, 2019); especially in comparison to unimodal language processing, which takes relatively long (perhaps due to the lack of visual context providing meaning to the message, unimodal language processing has to make up for this missing information and consumes more time in doing so). Hintz, Meyer and Huettig (2017) used three eye-tracking experiments to test the verb-mediated anticipatory eye gaze and their

findings agreeingly showed that predictive language processing is a pluralistic approach: multiple mechanisms, factors and situational context interact.

There are reportedly two routes that lead towards prediction: prediction-by-association and prediction-by-simulation (Pickering & Garrod, 2013). The first drawing from previous experience with others' speech that you have heard, while the latter draws from your own speech use (what would you yourself do in a similar situation). Prediction-by-association seems to be especially of relevance when interlocutors don't have that much in common. These two routes support the idea that alignment is key in successful conversation (Garrod & Pickering, 2004; Pickering & Garrod, 2006, 2007), since alignment is in a way making use of associations and simulations. Especially in dialogue, people tend to imitate each other. Alignment at one linguistic level causes more alignment at another level and in this manner, complex language use (like face-to-face conversation) can successfully take place (Pickering & Garrod, 2013). So, people predict upcoming language and imitate the linguistic input that has just come in, at various levels. Especially covert imitation enables the production system to make predictions and, in this manner, help the comprehension system significantly in speeding up the process.

Empirical evidence of benefits during low-level audiovisual integration can be found in several studies. As said before, lip movements and facial articulation movements of the speaker are of significant influence in predicting sounds; next to previous mentioned literature, Van Wassenhove, Grant and Poeppel (2005), for example, show how visual speech speeds up the processing of auditory signals in the cortical regions early. Shams, Kamitani and Shimojo, (2000) looked at the effect of sounds and flashes. They found that the auditory input (short beeps) had a significant effect on how the flashes were perceived; more specifically the number of beeps affected the amount of flashes that were seen. Van der Smagt, van Engeland and Kemner (2007) performed the same type of experiment, but with participants with autism as well. They remarkably found no difference between this test group and a control group (of people without autism), which means that this integration of audiovisual information must take place in a higher-level processing stage. Additionally, Giard and Peronnet (1999) recorded event-related potentials for subjects who had to choose one of two subjects, which was presented with audio only, audiovisuals only or both. Their results confirm that multimodal integration indeed takes place at high-level processes and also occurs rather early in the sensory processing chain. These findings together suggest that low-level audiovisual signals indeed contribute positively to the procedure of language processing.

Nonetheless, it is good to keep in mind that prediction might not be the sole explanation for the potential processing advantages of the multimodal context; multimodal integration may be solely a facilitator in the process of perception, or multimodal context may cause an improvement with regards to attention (Clark, 2016) or motivation during processing. Previous work does seem to indicate that a listener indeed predicts language and that a multimodal context could possibly facilitate this prediction process, and therefore facilitate the language processing system. On the other hand, the role of prediction in language processing should perhaps be nuanced, since some other work in the unimodal domain shows there might be other factors that have an influence on this effect of prediction. Namely someone's propensity to employ prediction possibly depends on individual factors or on the context in which the language occurs. Huettig and Janse (2016), for example, showed how individual differences with regards to cognitive abilities, such as working memory or

processing speed, can cause a moderation on the effect of prediction. In addition, Huettig and Guerra (2019) found that contextual issues, such as the instructions of the experiment, the speech rate of the stimuli or the preview time of the visual context, can limit the effects of prediction on language processing. If individual differences and contextual factors affect the effects of prediction in a unimodal context, the same could go for a multimodal context.

Empathy

Since individual differences (in working memory or processing speed for instance) might affect the tendency of the listener to predict upcoming language, differences in cognitive measures may also have an effect. Various research suggests that it might be interesting to look at the effects of empathy in particular. Firstly, empathy appears to share functional mechanisms with language and imitation (Iacoboni, 2005). Thereby, cognitive style (i.e. the way individuals think, perceive and remember information) seems to influence one's (social) language processing (Van den Brink et al., 2012). Van den Brink and colleagues tested the influence of semantic information by using speaker characteristics which sometimes conflicted with the message (like a child saying he likes to smoke). They found that the empathy quotient score was the determinant of inter-individual variability in the pragmatic N400-effect. More specifically, high-empathizing skills would show a capability of rapidly integrating information about the speaker with the content of the message, as making use of voice-based inferences about the speaker to process language in a top-down manner. Low-empathizing skills, on the other hand, suggested more a bottom-up approach to processing social pragmatic utterances, rather than using information about social stereotypes in implicit speech comprehension. Another one of their findings was that women seem to be generally better at language processing than men. Together with other research that showed that women have higher empathy levels than men (Gault & Sabini, 2000; Touissant & Webb, 2005), it could be that empathy is the cognitive ability that makes all the difference in language processing.

Considering the multimodal context, studies again point towards a possible important role for empathy during language processing. Mandel, Helokunnas, Pihko, and Hari (2015), for instance, indicated that a higher empathy score reflected back on being more responsive to visual cues, like blinks. Next to these low-level visual cues, the entire visual context exists also of facial and bodily movements, and for these two, a correlation with empathy can also be found. Participants with high empathy scores proved to perform better at imitating visual signals from the facial area than participants with low empathy scores (Williams et al., 2013). On top of this, Chu, Meyer, Foulkes and Kita (2014) found a correlation between gestures and empathy: cognitive abilities and empathy levels are related to individual differences in gesture frequency and saliency. These studies therefore imply that empathy could be a cognitive aspect that plays a valuable role in language processing.

Current study

In summary, existing research already provides adequate insights into language processing via the use of the speech shadowing paradigm. However, these insights apply mainly to a unimodal context or to short phrases or syllables and fails to provide sufficient information about the multimodal context and more complex linguistic situations, such as human face-to-

face communication. Therefore, this study will employ the shadowing experiment in a multimodal context and make use of stimuli more consistent with face-to-face communication. Via this experimental setting, we hope to indeed demonstrate a multimodal advantage during language processing. As hypothesized earlier, prediction might underlie this potential multimodal advantage. The effects of prediction in turn, could be moderated by individual cognitive differences. So, in addition to attempting to show a possible multimodal advantage, this study will focus on the influence of empathy.

In conclusion, this thesis aims to explore whether a correlation can be found between empathy and language processing in a multimodal context. More accurately, does visual context positively contribute to language processing and does the level of empathy influence this process at all? The experiments used for this study are a shadowing experiment and the empathy test (Baron-Cohen & Wheelwright, 2004). As for the shadowing experiment, three conditions will be employed: audio only, audio with visual context and audio with visual context, but with lips blurred. The more visual context someone is presented with, the better the language processing will probably go, so we would expect the least amount of errors in the shadowing experiment for the audiovisual condition, followed by the audiovisual with blurred mouth condition, and the audio only condition. When comparing the empathy scores to the results of the shadowing experiment, we expect to see that the level of empathy does have an influence on the performance of the shadower. As outlined above, people with higher empathy scores tend to be more receptive to visual cues, so we would expect that the higher the level of empathy, the lower the amount of errors for mainly the multimodal conditions. More specifically, we expect to see a negative correlation between empathy and language processing, which is the strongest for the audiovisual condition, followed by the audiovisual with blurred mouth condition and lastly the audio only condition.

Besides the main shadowing experiment and the empathy quotient test, this study has also employed several other cognitive measures, to explore other individual differences that might explain the potential multimodal advantage during language processing. These were the Digit Span Test, to measure the capacity of working memory; the Animal Naming Task and the Letter Fluency task, to measure semantic and phonemic fluency respectively; and the Trail Making Test (A and B), to measure effects of inhibitory control on speech shadowing. These, however, won't be discussed in this paper, as that is outside of the scope of this thesis. Furthermore, this study will take the percentage of errors as only measure for the shadowing task, since assessing latency additionally would also be outside of the scope of this thesis.

Method

Subjects

In total, a group of 37 subjects (34 females and 3 males) participated in the experiment. All were native Dutch-speaking, had normal hearing, normal (corrected) sight and no further cognitive disabilities. They were between the ages of 19 and 32 (mean age = 23.8, SD = 2.97), were randomly selected from the participant database of the Max Planck Institute and were given a small compensation (€10,00) for their partaking in the study.

One participant was excluded from the analysis, because it turned out after the experiment that she was familiar with the stimuli; she had participated in the study from which the stimuli were taken. That makes that 36 participants (33 females and 3 males, mean age = 23.9, SD = 2.97) were included for analysis.

Materials

Stimuli main experiment

The shadowing experiment consisted of 36 video fragments in total, of which 30 were used as experimental stimuli and 6 were used as practice stimuli. These stimuli lasted for 30-40 seconds (mean length = 36.4 sec, SD = 5.96 sec), and were presented in three conditions: audio only (AO), audiovisual with blurred mouth (AB) and audiovisual (AV).

The video fragments used for the stimuli were retrieved from the CoAct corpus of the Max Planck Institute (ERC project #773079). In these videos, two friends were talking to each other face-to-face about various subjects and were filmed in doing so. Using ELAN, we selected sections in these videos in which people were talking for a longer period of time without getting interrupted by their conversational partner. The speaker in these videos was seated in a chair and was always visible from a frontal perspective, and from head to knee.

We extracted the audio from these video files, intensity-scaled the speech to 70 dB and de-noised the speech in *Praat* (Boersma & Weenink, 2009). All sound files were then recombined with their corresponding video files using Adobe Premiere Pro. The selected fragments underwent video editing to fit the testing conditions. For the AO condition, the participant saw only a black screen with a white dot in the middle as focusing point; for the AV condition, the participant saw the video in its whole; and for the AB condition, the participant saw the video, but the mouth of the speaker in the video was blurred. The blurring of the lips was done manually with the program Adobe Premiere Pro, and were fitted per frame per stimulus to ensure a dynamic presentation of the blurred area that fitted the speaker's mouth, even when the speaker would be moving around in the video.

Further equipment

Next to the main shadowing experiment, cognitive measures were tested prior (in the following order): The Digit Span Test, the Animal Naming Task, Trail Making Test A, the Letter Fluency Task and Trail Making Test B. For the digit span task, we used Presentation (Neurobehavioural Systems), the fluency tests were recorded with the program Audacity and the trail making tests were made with pen and paper. Each test was timed and for the fluency tests the amount of words were counted as well. These tests were mainly employed for further research and the results of these will therefore not be discussed in this thesis. After the shadowing experiment, the participants were also asked to fill out a questionnaire digitally in Excel, which, amongst a few general questions, contained the empathy test (Baron-Cohen & Wheelwright, 2004). This test consists of 60 questions, of which 20 are filler items. For each question the participant could choose one out of for answers: strongly disagree, slightly disagree, slightly agree or strongly agree. They were unaware of the fact that their empathy was being tested.

Procedure

First of all, the participants read through the instructions and signed the required consent forms. They were then asked to perform the cognitive measures before the main experiment. For the digit span task, the participants were shown a series of numbers (one-by-one) on the computer screen and they had to memorize these numbers in the right order. A series would become longer (and therefore harder) or a bit easier in proportion to how the participant performed. Presentation would stop automatically when three errors in a row were made. The time it took the participants to finish this task was recorded. The Animal Naming Task was

one minute in which the participant had to mention as many animals as he or she could possibly think of. The time was measured, as well as the number of mentioned animals. For the Letter Fluency Task, the procedure was the same, but instead of animals, the participant had to name as many words as he or she could think of that started with the letter 'p'. The trail making test came in two forms; the first was a series of numbers (1-2-3-etc. and the second included letters as well (1-a-2-b-etc.). The participant had to connect the numbers (and letters) in the correct order and could not remove the pen from the paper in doing so. If they made a mistake, they had to correct it. These were also timed.

After the cognitive measures were recorded, the participants commenced with the main experiment. Participants were tested in a soundproof booth and seated in front of the computer with headphones on. In the booth, we set up a webcam to monitor the participant's attention, as well as their speech output. This video footage came through on a separate laptop. We then repeated the instructions for the main experiment. The participant's goal was to repeat the person they heard as quickly and correctly as possible, whilst still listening to the video fragment. They were also instructed to look at the screen at all times and press the space bar at the end of a video to move on to the next trial.

All videos were presented by using Presentation (Neurobehavioural Systems) and displayed on a 24-inch monitor with 1920x1080-pixel resolution. The experiment consisted of three blocks, of which the order was different each time: AO-AV-AB, AV-AB-AO and AB-AO-AV. Every block contained ten videos and two practice videos beforehand, so that the participant could get familiar with the upcoming stimuli. The videos were ordered randomly for each participant as well, both in terms of presentation order, and in occurrence per condition. Every video was only played once, and no videos were repeated over conditions. After every block, the participants could take a self-paced break and could drink some water if they wanted to. During testing, comments on how the experiment went, were written down in a notebook.

When the shadowing experiment was concluded, the participant was brought back to the fore room and had to fill in the questionnaire and sign the last consent forms. The consent forms (the ones prior to conducting the experiment included) and the questionnaires were coded, so that the data is anonymized. In the end the participant got a debriefing and was asked what he or she thought of the experiment. After the experiment, an USB-stick was used to transfer and save the data carefully.

There were also some guidelines for any inconveniences, such as: if the participant is waiting (i.e. tries to memorize the speech and then repeats it at once), pause the experiment immediately. Open the door and tell them to talk as soon as possible, even if it is quite hard, because you're still listening. If the participant is looking away during the video, pause the experiment during the count down for the next video. Open the door and tell them to please always look at the screen. If the participant complains that he or she doesn't hear anything or that it's too difficult, pause the video immediately. Open the door, reassure them that whatever they are able to do, is fine and to please just try.

The total duration of the experiment is about an hour: 15 minutes for first consent forms and the cognitive measures, 30 minutes for the main shadowing experiment and 15 minutes for the final consent forms and questionnaire.

Design and analysis

Before starting with actually testing participants, the stimuli were transcribed with the program *Praat* (Boersma & Weenink, 2009). The videos were first transcribed by the automatic speech recognizer and then checked manually. Word boundaries, as well as which word was being said, were carefully transcribed. Also stutters, mispronunciations and hesitations for example were included. This was done in preparation for comparing the initial fragment to the (to be) recorded shadowed fragment. After the experiment was conducted, the obtained data had to be processed. A total of 1.080 audio files (36 participants, 30 stimuli per participant), in which the stimuli were shadowed, were transcribed in the same manner as the original stimuli files at the beginning (initial transcription by the automatic speech recognizer and manually checking after that).

For analysis of this thesis, the scores on the empathy test were used together with to the error rates of the shadowing experiment. The error rate was determined by comparing the annotated audio file of the participant to the matching original stimulus file. The following were determined to be an error: everything that the participant said which didn't appear in the original file (this includes stutters, mispronunciations, unfinished words, self-entered words), everything that was said for the second time (the first 'studie' (= study) was counted as correct, but the second 'studie' was counted as incorrect for e.g.) and missed words (everything that was in the original file, but wasn't said by the participant). The scores of the empathy test were determined by its guidelines (Baron-Cohen & Wheelwright, 2004). For every test item the participant scored 1 point if he or she chose the 'slightly' answer and 2 points if he or she chose the 'strongly' answer. Approximately half of the test questions tried to elicit a 'disagree' response and the other half an 'agree' response.

The program SPSS was then used to analyze both the empathy scores and error rate. For the error rate, a repeated measures design (within-subject) was used, so that the effect of the testing condition on the error rate would become clear. Then, a correlation test was performed on the error rates per condition in combination with the empathy scores, to see whether empathy has an effect on the error rate.

Results

Shadowing errors

Overall, we observed an error rate of 26.54% (SD = 7.99) for the AV condition, an error rate of 28.76% (SD = 8.51) for the AB condition, and an error rate of 29.72% (SD = 8.18) for the AO condition (figure 1). A Repeated Measures ANOVA showed that the condition in which the stimuli were presented had a significant effect on the percentage of shadowing errors, $F(1.96, 68.73) = 6.91$, $p = .002$, $\eta^2_p = .16$. Mauchley's test showed that the assumption of sphericity wasn't violated, $\chi^2(2) = .63$, $p = .729$.

Pairwise comparisons revealed that the error rate for the AV condition was lower than the error rate for the AB condition. This difference, -2.23, 95% CI [-4,40 - -.06], was statistically significant, $p = .042$. The error rate for the AV condition was also lower than the error rate for the AO condition. This difference, -3.19, 95% CI [-5.29 - -1.08], was statistically significant, $p = .002$. The error rate of the AB condition was not statistically different from the error rate of the AO condition, -.96, 95% CI [-3.31 – 1.40], $p = .940$.

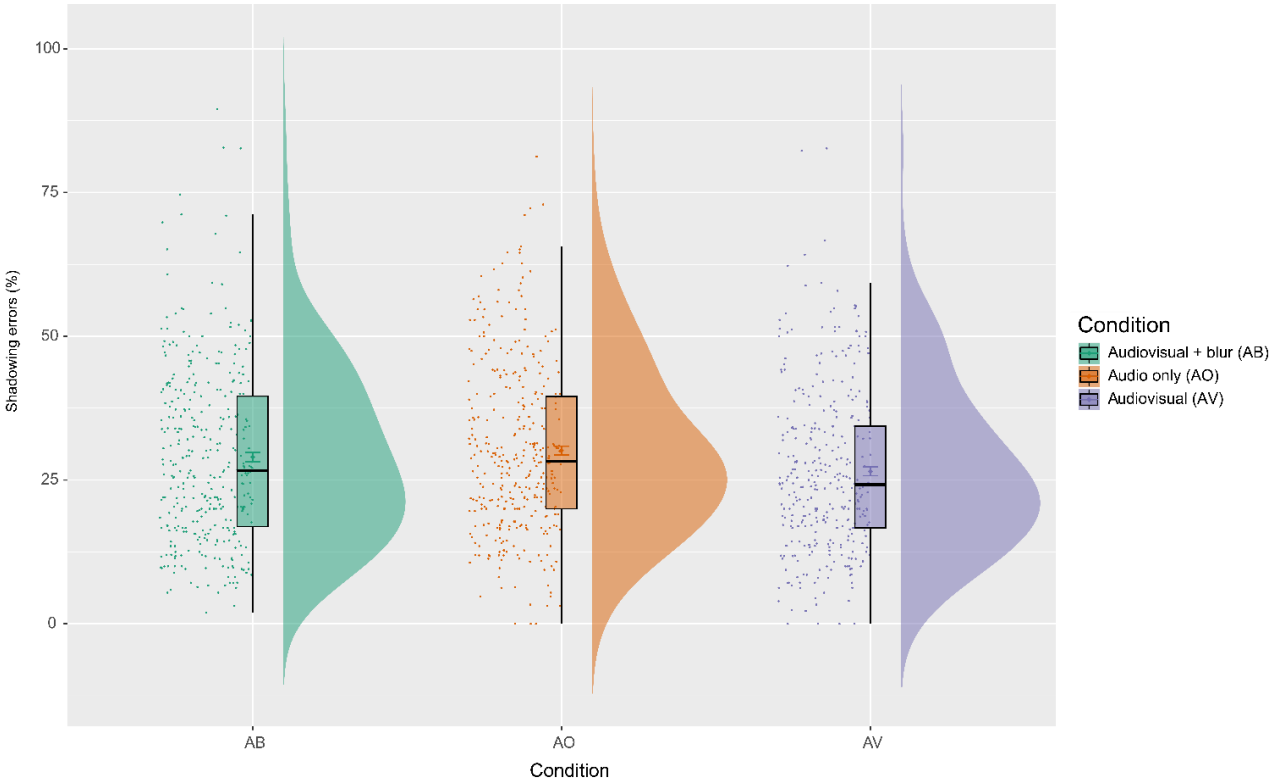

Figure 1: The percentage of errors displayed per condition; every dot is a shadowed video and the density plots show how the error rates are distributed.

Empathy quotient scores

Out of the empathy scores (mean score = 47, SD = 12.00), the most common score was 50 in a range of 16 being the lowest and 67 being the highest. The most participants scored between 50-59, which is considered a rather high level of empathy (table 1).

Table 1: Distribution of the empathy quotient scores over the participants and the average error rate in percentage per condition and per group.

EQ score	**Number of participants**	**AB (%)**	**AO (%)**	**AV (%)**
0-9	0	-	-	-
10-19	1	26,18	29,32	34,10
20-29	4	34,38	31,53	33,25
30-39	4	37,77	37,24	32,75
40-49	9	28,59	29,35	26,93
50-59	13	25,39	27,68	22,68
≥ 60	5	26,67	28,30	24,00

Correlation between empathy and shadowing errors

Pearson's correlations show that there is a significant strong negative correlation between empathy and the audiovisual condition, $r(35) = -.474$, $p = .004$. Empathy does not seem to have a significant correlation with the condition audio only, $r(35) = -.196$, $p = .253$, nor with the condition audiovisual and blur, $r(35) = -.299$, $p = .077$.

Discussion

Where existing studies portrayed language processing through the use of a unimodal model or short phrases, this study adds to the knowledge by looking at the multimodal aspect and by making use of more complex language use in employing the shadowing paradigm. After all, language processing during human face-to-face communication entails more than processing just auditory signals; listeners also need to process many visual signals. Additionally, this research explored the potential impact that individual differences may have, by including the level of empathy (a cognitive measure that possibly plays an important role in language processing) in the analysis. The first research question was whether a multimodal advantage can be found for language processing. Like we hypothesized, our results suggest that this is indeed the case; participants performed better in the shadowing task as more visual context was present. The second research question was whether the cognitive measure empathy has an influence on language processing in a multimodal context. The results show that this was true for the condition audiovisual, but not for the condition audiovisual with blurred mouth or for the condition audio only. So, irrespective of the level of empathy, the performance of the shadower seemed to be more or less the same for the conditions audiovisual with blurred mouth and audio only, but when observing the condition audiovisual, shadowers with a high empathy score appear to have a lower error rate than shadowers with a low level of empathy. Since the level of empathy only seems to become of significant relevance when the full multimodal context is present, inferences could be made that empathy has the most influence on visible cues, and less on auditory cues. Thus, participants with a high empathy score seem to be more receptive to especially the visual cues (than participants with a low empathy score) and therefore probably experience more of the multimodal benefit, resulting in a lower percentage of errors in their shadowed speech. The results will now be discussed more intricately per research question.

Multimodal advantage for language processing

The results showed an overall trend in the shadowed speech with the least amount of errors in the AV condition, followed by the AB condition, and then the AO condition. However, the difference between the AB and AO condition was not significant.

Based on these data, we can conclude that in general, a multimodal context contributes to language processing to a great extent in possibly facilitating incremental integration and aiding predictive language processing. Participants performed better as the condition in the shadowing experiment contained more visual cues. This is along the lines of our hypotheses, other than the observation that the error rate for the AB condition does not seem to differ from the error rate for the AO condition as much. There are two possible reasons that could explain this: one being that it is the combination of all visual cues that ensures a temporal advantage and, when some of the visual context is missing, this benefit diminishes; the other being that the blurred mouth creates confusion and distracted the listener. The latter can be reasoned quite simply by imagining how we do encounter an AV or an AO situation in daily life, like face-to-face communication and talking over the phone respectively, but we don't ever encounter an AB situation. For this reason, it could be that the participants were distracted or uncomfortable with seeing a blurred mouth and therefore produced about as many errors as the AO condition. The other possible explanation, that the blurred mouth takes away the temporal advantage, can be deduced from the fact that visual speech can speed up the processing of auditory signals (Van Wassenhove, Grant and Poeppel, 2005) and can aid in

predicting upcoming speech (Venezia et al., 2016). Visual speech therefore seems to secure a temporal advantage during language processing. Furthermore, this temporal advantage seems to be the most prominent when all visual cues are present, including facial and bodily movements (Gerwing & Allison, 2009; Kelly, Healy, Özyürek & Holler, 2015). This means that when one of these cues is missing, the temporal advantage may be reduced. This may happen when the mouth is blurred, but also in a situation where only the lip movements are presented to the listener (without showing any of the bodily movements). This would clarify why the results show no significant difference between the AB and AO condition. The main take-away from this is however, that the results from the shadowing paradigm show an overall trend which demonstrates that there is indeed a multimodal advantage for language processing.

Furthermore, it would be wise for further research to inspect the latencies from the shadowing task as well. In order to establish a more complete image of the temporal advantage brought by a multimodal context. Like Marslen-Wilson (1985) said, the latency of the shadowed speech reflects the process of speech perception directly. It could therefore be a sensitive and dynamic indicator of the predictability of the shadowed material. In other words, latency may very well be an efficient measure to investigate language processing in a multimodal context. The error rate on its own does reflect multimodal language processing, but by including latency as a variable, the results and conclusions can be presented with more certainty. Unfortunately, inspecting latencies would have been outside the scope of this particular thesis and therefore we only looked at the error rates for the different conditions.

Even though this thesis only used error rates for analyses, the results can still point towards the idea that prediction might play an important role in language processing. The manner in which visual speech can aid in predicting upcoming speech (Venezia et al., 2016) and the way in which we observe a significant lower amount of errors in the performance for the AV condition, in comparison to the AB and AO condition (with the only difference being the removal of the lip movements), may lead us to believe that prediction is involved in multimodal language processing. As other explanations for a multimodal benefit remain likely as well when considering the data of this research, the possibility of the multimodal advantage laying in predictive language processing will be discussed further below.

Prediction as underlying cognitive mechanism

As proposed at the beginning of this research, prediction holds great potential for being the underlying cognitive mechanism for multimodal language processing (Holler & Levinson 2019). In order to prove this theory with more certainty, the shadowed speech should be searched for words that are uttered before they are even heard in the speech from the stimuli. If prediction is indeed an important cognitive mechanism for the multimodal advantage, we would expect to find the most of such early predicted words in the conditions with multimodal context. Or, if latency is being looked at as well, as suggested earlier, these predicted words could occur in equal proportions in any condition (regardless of being uni- or multimodal), but faster for multimodal contexts for example. The latter would be in line with suggestions made by Holler and Levinson (2019), saying that the multimodal utterances, with its compositional and temporal architecture, facilitate predictive language processing. Thus, prediction as an underlying cognitive mechanism remains a possibility, but based solely on these data, further research is needed to validate this idea.

Before further research can validate (or discard) the interpretation of a multimodal context aiding predictive language processing, we ought to be cautious in interpreting our data and should not assume prediction as an underlying cognitive mechanism too rapidly. An alternative explanation is that the results do not show how a multimodal context contributes to predictive language processing, but rather how it involves a combination of multiple processes. A multimodal context might still support predictive language processing, but listeners might as well experience this multimodal advantage at other levels too. At the level of perception, for example, a multimodal context could facilitate incremental integration. Another alternative theory includes motivation or attention as potential factors; listeners possibly experience a higher level of motivation or are more actively engaged in a multimodal context (relative to a unimodal context). So, a multimodal context does not serve as an advantage at perceptual level or during predictive language processing, but rather increases the level of motivation or attention.

In general, prediction as an underlying cognitive mechanism remains a possible explanation for the multimodal advantage during language processing, even though alternative explanations cannot be ruled out based upon the findings in this thesis. Our finding that there seems to be no significant difference between AB and AO would perhaps lean more to such an alternative explanation. As we speculated before, the blurred mouth could potentially have distracted the listener, which may have led to a reduction in the level of motivation or attention. On the other hand, these alternative explanations cannot fully explain the other finding that the AV condition does differ significantly from the AB condition. Based on this, an explanation is more likely to be either the assumption of prediction as an underlying cognitive mechanism or the theory that listeners might experience a multimodal advantage at other levels, such as at the perceptual level for incremental integration, along with predictive language processing as well. Thus, the results from this research advocate different theories about the underlying cognitive mechanism for the multimodal benefit during language processing. Considering the main finding that there is a significant difference for the condition AV (with respect to the AB and AO condition), we propose that the explanation of a multimodal context that aids predictive language processing is nonetheless very likely. However, to exclude alternative explanations with more certainty, more research is needed on predicted words in the shadowed speech. This was unfortunately out of the scope of this thesis, so for now, we can only speculate.

Correlation between empathy and multimodal language processing

The results demonstrated that there is a strong correlation between empathy and the AV condition of the shadowing task. However, no correlation could be found for empathy related to the AB condition, nor for empathy related to the AO condition.

In general, we do see a trend in favor of empathy playing an important role during multimodal language processing. The lowest error rates can be found mainly for participants with a high empathy score. For the AV condition this correlation is the most significant, from which we can infer that empathy has the greatest effect on multimodal language processing (in comparison to unimodal language processing). Remarkably, no strong correlation was found for the AB condition, which is still a multimodal condition with only the lip movements missing. For this observation, we speculate again that the blurring of the mouth could be distracting for the listener, which could temper the effects of empathy on language processing.

Nonetheless, we did find a trend for the AB condition and the finding that empathy correlates strongly with the AV condition (having the most complete multimodal context possible) is the most important one and indicates that listeners with a higher empathy score are indeed more responsive to visual cues in comparison to listeners who have a lower empathy score (Mandel et al., 2015). In order to provide this second research question with a more complete image as well and illustrate the influence of empathy on multimodal language processing more thoroughly, latency should be taken on as an additional measure. As stated before, this was unfortunately outside of the scope for this thesis, so this is strongly recommended for further research. All in all, the main conclusion we can take from these data is that empathy indeed plays a significant role in language processing: the higher the level of empathy, the bigger the benefit from visual signals during language processing.

A note on these findings is that most of our participants were female and women seem to have higher empathy scores than men (Van den Brink et al., 2012). This could be the reason that most of the participants scored relatively high on the Empathy Quotient, whereas if more men would have been included, the distribution of empathy scores could have been somewhat more even in the group of subjects. Nevertheless, the data still contains a sufficient spread of variety qua empathy scores, so the data is still generalizable.

Conclusion

After conducting the shadowing paradigm in multimodal contexts in comparison to unimodal contexts, results show that there indeed exists a multimodal advantage for language processing. When taking empathy levels in account, results verify a correlation between empathy and multimodal language processing. Based on these results, prediction is possibly an important underlying cognitive mechanism for multimodal language processing. In conclusion, we suggest that a multimodal context may help in predictive language processing and that individual differences, such as the level of empathy, can affect the extent to which listeners experience this multimodal advantage.

For future research in language processing with the shadowing paradigm, we advise to include latency as a variable, to portray multimodal language processing in its entirety. In addition, to verify prediction as an underlying cognitive mechanism, we recommend filtering the words that are being said before listeners have heard them, to exclude other theories regarding multimodal language processing. Also, when upcoming studies are to investigate the influence of empathy, we suggest making use of a varied group of subjects with regards to gender, as it appears that women score higher levels of empathy. Irrespective of these recommendations, this study has succeeded in finding a multimodal advantage for language processing and a correlation between empathy and multimodal language processing.

References

Bailly, G. (2003). Close shadowing natural versus synthetic speech. *International Journal of Speech Technology*, *6*(1), 11-19.

Baron-Cohen, S., & Wheelwright, S. (2004). The Empathy Quotient: An Investigation of Adults with Asperger Syndrome or High Functioning Autism, and Normal Sex Differences. *Journal of Autism and Developmental Disorders, 34*(2), 163–175.

Beautemps, D., Cathiard, M. A., & Le Borgne, Y. (2003). Benefit of audiovisual presentation in close shadowing task. In *15th International Congress of Phonetic Sciences, volume* (Vol. 1, pp. 841-844).

Boersma, P., & Weenink, D. (2009). *Praat: doing phonetics by computer (Version 5.1.05) [Computer program]*. Retrieved May 1, 2009.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*(2), B13–B25.

Brouwer, S., Mitterer, H., & Huettig, F. (2010). Shadowing reduced speech and alignment. *The Journal of the Acoustical Society of America*, *128*(1), EL32–EL37.

Carpenter, P. A., Miyake, A., & Just, M. A. (1995). Language Comprehension: Sentence and Discourse Processing. *Annual review of psychology, 46*(1), 91-120.

Chater, N., McCauley, S. M., & Christiansen, M. H. (2016). Language as skill: Intertwining comprehension and production. *Journal of Memory and Language*, *89*, 244–254.

Chen, T., & Rao, R. R. (1998). Audio-visual integration in multimodal communication. *Proceedings of the IEEE*, *86*(5), 837–852.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences, 39.*

Chu, M., Meyer, A., Foulkes, L., & Kita, S. (2014). Individual differences in frequency and saliency of speech-accompanying gestures: The role of cognitive abilities and empathy. *Journal of Experimental Psychology: General*, *143*(2), 694–709.

Clark, A. (2016). Attention alters predictive processing. *Behavioral and Brain Sciences, 39.*

Clark, E. V., & Hecht, B. F. (1983). Comprehension, Production, and Language Acquisition. *Annual Review of Psychology*, *34*(1), 325–349.

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*(1), 212-222.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically

modulates the excitability of tongue muscles: A TMS study: Tongue involvement during speech listening. *European Journal of Neuroscience*, *15*(2), 399–402.

Fairs, A. (2019). *Linguistic dual-tasking: Understanding temporal overlap between production and comprehension* (Doctoral dissertation, Radboud University, Nijmegen, the Netherlands).

Ferreira, V. S. (2010). Language production. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 834–844.

Friederici, A. D. (2012). The cortical language circuit: From auditory perception to sentence comprehension. *Trends in Cognitive Sciences*, *16*(5), 262–268.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*(2), 181–218.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*(1), 8–11.

Gault, B. A., & Sabini, J. (2000). The roles of empathy, anger, and gender in predicting attitudes toward punitive, reparative, and preventative public policies. *Cognition & Emotion*, *14*(4), 495–520.

Gerwing, J., & Allison, M. (2009). The relationship between verbal and gestural contributions in conversation: A comparison of three methods. *Gesture*, *9*(3), 312–336.

Giard, M. H., & Peronnet, F. (1999). Auditory-Visual Integration during Multimodal Object Recognition in Humans: A Behavioral and Electrophysiological Study. *Journal of Cognitive Neuroscience*, *11*(5), 473–490.

Hagoort, P., & van Berkum, J. (2007). Beyond the sentence given. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 801–811.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402.

Hintz, F., Meyer, A. S., & Huettig, F. (2014). Prediction using production or production engaging prediction?. *In the 20th Architectures and Mechanisms for Language Procesing Conference (AMLAP 2014).*

Hintz, F., Meyer, A. S., & Huettig, F. (2016). Encouraging prediction during production facilitates subsequent comprehension: Evidence from interleaved object naming in sentence context and sentence reading. *Quarterly Journal of Experimental Psychology*, *69*(6), 1056–1063.

Hintz, F., Meyer, A. S., & Huettig, F. (2017). Predictors of verb-mediated anticipatory eye movements in the visual world. *Journal of Experimental Psychology: Learning,*

Memory, and Cognition, *43*(9), 1352–1374.

Holler, J., Kendrick, K. H., & Levinson, S. C. (2018). Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychonomic Bulletin & Review*, *25*(5), 1900–1908.

Holler, J., & Levinson, S. C. (2019). Multimodal Language Processing in Human Communication. *Trends in Cognitive Sciences*, *23*(8), 639–652.

Huettig, F., & Guerra, E. (2019). Effects of speech rate, preview time of visual context, and participant instructions reveal strong limits on prediction in language processing. *Brain Research*, *1706*, 196–208.

Huettig, F., & Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Language, Cognition and Neuroscience*, *31*(1), 80–93.

Iacoboni, M. (2005). Understanding others: Imitation, language, empathy. *Perspectives on imitation: From cognitive neuroscience to social science*, *1*, 77-99.

Kelly, S., Healey, M., Özyürek, A., & Holler, J. (2015). The processing of speech, gesture, and action during language comprehension. *Psychonomic Bulletin & Review*, *22*(2), 517–523.

Lee, E.-K., Brown-Schmidt, S., & Watson, D. G. (2013). Ways of looking ahead: Hierarchical planning in language production. *Cognition*, *129*(3), 544–562.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). Multiple perspectives on word production. *Behavioral and Brain Sciences*, *22*(1), 61–69.

Mandel, A., Helokunnas, S., Pihko, E., & Hari, R. (2015). Brain responds to another person's eye blinks in a natural setting-the more empathetic the viewer the stronger the responses. *European Journal of Neuroscience*, *42*(8), 2508–2514.

Mandler, J. M. (1979). Language Comprehension. *Science*, *203*(4377), 259–260.

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(4), 843–847.

Marslen-Wilson, W. (1973). Linguistic Structure and Speech Shadowing at Very Short Latencies. *Nature*, *244*(5417), 522–523.

Marslen-Wilson, W. D. (1985). Speech shadowing and speech comprehension. *Speech Communication*, *4*(1–3), 55–73.

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

Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition*, *109*(1), 168–173.

Norris, S. (2004). Multimodal discourse analysis: A conceptual framework. In *Discourse and technology: Multimodal discourse analysis*, 101-115.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382–2393.

Pickering, M. J., & Garrod, S. (2006). Alignment as the Basis for Successful Communication. *Research on Language and Computation*, *4*(2–3), 203–228.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*(3), 105–110.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347.

Scarbel, L., Beautemps, D., Schwartz, J.-L., & Sato, M. (2014). The shadow of a doubt? Evidence for perceptuo-motor linkage during auditory and audiovisual close-shadowing. *Frontiers in Psychology*, *5*.

Scollon, R., LeVine, P. (2004). Multimodal discourse analysis as the confluence of discourse and technology. In *Discourse and technology: Multimodal discourse analysis,* 1-6.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). What you see is what you hear. *Nature*, *408*(6814), 788.

Smith, A. R., & Lee D. Erman. (1981). Noah-A Bottom-Up Word Hypothesizer for Large-Vocabulary Speech Understanding Systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-3*(1), 41–51.

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.

Thompson, L., Garcia, E., & Malloy, D. (2007). Reliance on Visible Speech Cues During Multimodal Language Processing: Individual and Age Differences. *Experimental Aging Research*, *33*(4), 373–397.

Toussaint, L., & Webb, J. R. (2005). Gender Differences in the Relationship Between Empathy and Forgiveness. *The Journal of Social Psychology*, *145*(6), 673–685.

Van den Brink, D., Van Berkum, J. J. A., Bastiaansen, M. C. M., Tesink, C. M. J. Y., Kos, M., Buitelaar, J. K., & Hagoort, P. (2012). Empathy matters: ERP evidence for inter-individual differences in social language processing. *Social Cognitive and Affective Neuroscience*, *7*(2), 173–183.

Van der Smagt, M. J., van Engeland, H., & Kemner, C. (2007). Brief Report: Can You See What is Not There? Low-level Auditory–visual Integration in Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *37*(10), 2014–2019.

Van Leeuwen, T. (2004). Ten Reasons Why Linguists Should Pay Attention to Visual Communication. In *Discourse and technology: Multimodal discourse analysis*, 7-19.

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.

Venezia, J. H., Thurman, S. M., Matchin, W., George, S. E., & Hickok, G. (2016). Timing in audiovisual speech perception: A mini review and new psychophysical data. *Attention, Perception, & Psychophysics*, *78*(2), 583–601.

Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, *41*(8), 989–994.

Williams, J. H. G., Nicolson, A. T. A., Clephan, K. J., Grauw, H. de, & Perrett, D. I. (2013). A Novel Method Testing the Ability to Imitate Composite Emotional Expressions Reveals an Association with Empathy. *PLoS ONE*, *8*(4).