university of
groningen

faculty of science
and engineering

Resolving the Pronoun Problem: Investigating the Influence of Linguistic and Visual Contexts on Object Pronoun Processing

Bachelor's Project Thesis

Boris Marinov, s2957531, b.m.marinov@student.rug.nl,
Supervisors: Dr J.C. van Rij-Tange & Dr J.P. Borst

Abstract: The current study investigated whether and how linguistic and visual contexts affect the processing of object pronouns. Past research has provided support for two theories: the initial-filter account, stating that grammatical principles first filter out impossible referents, following which, the linguistic and visual scene play a role; and the competing-constraints account, which argues that grammar competes with these contexts at an earlier stage of processing. Some recent studies have provided support for the competing-constraints account by measures of the pupil dilation (van Rij, 2012). In a follow up study, which employed a blank-screen, the findings however were not replicated (Verhoeven, 2018). To further investigate the effects of context and verify the difference in results, this study employed a 2x2x2 within-subject design in which we manipulated the visual scene (picture with self and other-oriented action), linguistic context (introducing the actor earlier or later), and the type of referent used (object and reflexive pronouns). For both referent types, the visual context affected pronoun resolution when the agent of the sentence was introduced first. Early after the onset of the pronoun, pupil dilations showed a reliable difference between the conditions of picture type and introduction order. These results are in line with earlier research (van Rij, 2012), supporting the competing-constraints account.

1 Introduction

In the field of linguistics, the term **anaphor** refers to the use of an expression in a sentence, of which the interpretation depends on another expression. The expression of which the interpretation depends on could exist within the described context, or outside of the sentence scope itself. The name given to that preceding expression is an antecedent. The most common examples of anaphors are pronouns. For example, in the sentence *”Lotte arrived at the station on time, but still missed her train”*, the pronoun *”her”* is an anaphor which refers back to the antecedent *”Lotte”*. Pronouns themselves take on many forms, with the two most common being object and subject pronouns. The distinction between the two is important, mainly determined by their grammatical role within the sentence. Humans have learned to subconsciously resolve the difference between the two, linking the object and subject antecedents in the sentence to the appropriate pronoun.

Exactly how that is done is still a matter of discussion, and the exact reference rules of pronouns vary between languages. For example, in Dutch and English it is impossible for the object pronoun to refer to the subject. That is, in the sentence: *”The dog barked at him with anger”*, one naturally assumes the dog as being the subject of the sentence (it is the one performing the action) and the pronoun *”him”* referring to some unknown object in this context. This restriction, and therefore interpretation guiding, is known as the Principle B of Binding Theory (Chomsky, 1981).

Language complexity, however, scales up rapidly and often sentences can be made up of multiple characters. The listener then needs to select which character is the most likely referent of the pronoun. Arnold (1999) provides a review which outlines how preceding language in the sentence can be of use to the listener. The preceding language can provide cues in the form of grammatical role (same number and gender), the order of mention (characters mentioned first are preferred over those that come second), and recency of introduction (characters should be mentioned as early as possible). These guiding factors together form the linguistic discourse of the sentence.

An important idea emerges from this. Are the grammatical principles described by Chomsky (Chomsky, 1981) applied first, followed by the rest of the linguistic discourse restrictions, or do the two compete in the early stages of pronoun resolution. The **initial-filter account** (Chow et al., 2014; Clifton et al., 1997; Nicol and Swinney, 1989) proposes that a listener would only consider candidates for potential referents if they haven't violated the language's specific grammatical principles. In other words Principle B would first filter out impossible candidates, and only then the linguistic discourse would play its role.

Competing evidence has emerged in recent years. Results suggest that linguistic discourse competes alongside the grammatical rules in the early stages, and thus influences the online interpretation of pronouns (Clackson et al., 2011; Spenader et al., 2009). This forms the basis of the **competing-constraints account** (Badecker and Straub, 2002; Kennison, 2003) which outlines this competition between grammatical rules and other sources of information that the listener can draw upon (the rest of the linguistic discourse, such as the one described by Arnold (1999)).

A recent study by van Rij (2012) reports effects which support the competing-constraint account. That is, the study reported a combined effect of the visual and linguistic context on the interpretation of object pronouns. During the study participants were presented with an image of an actor and a patient. After seeing the cartoon image, a recorded sentence introduced the referents, either the actor first or second (thus manipulating the linguistic context), followed by a description of the image. The description either matched the image presented or not, thus altering the congruency of the visual scene (manipulating the visual context). The study used the dilation of the pupil as an indicator of cognitive load, due to the pupil's sensitivity to linguistic differences (Engelhardt et al., 2010; Zellin et al., 2011). The dilation of the pupil has been established as a strong predictor of cognitive load, with higher dilation indicating higher load (Beatty et al., 2000). Effects in van Rij's study were found on pupil dilation 500-1000ms after the onset of the object pronoun. The results suggest that when an agent is introduced first (so increasing the prominence), the visual context had an effect on the pronoun processing. That is, a **canonical** (actor-first) introduction order elicited a larger pupil dilation when the visual scene was incongruent with the subject's interpretation of the sentence, compared to when the interpretation was congruent. On top of this, a canonical introduction order with a congruent scene elicits a smaller pupil dilation than both the **non-canonical** (actor-second) introductions with congruent and incongruent scenes. These results provide evidence against the initial-filter account, as it seemed that the linguistic and visual context had an influence on the pronoun processing.

One could interpret these findings as follows. The order of mention of a referent has an effect on its perceived prominence from the listener's perspective. That is, referents mentioned earlier are interpreted as more likely antecedents of a subject pronoun in a follow up sentence, compared to referents mentioned later (Gernsbacher and Hargreaves, 1988; Gordon et al., 1993; Kaiser and Trueswell, 2008). When people hear some referent mentioned earlier, they build an expectation of that referent to be mentioned again as a subject of the sentence. Any later mentioned referents are therefore expected to be linked to a less prominent grammatical position, such as to an object pronoun. A person would therefore accept a correct interpretation when the subject of the second sentence does indeed refer to the first mentioned referent. When this isn't the case, there is conflicting information and the expectation is violated. This could cause listeners to not build these expectations until they hear the object of the sentence. Therefore, when referents are introduced in a non-canonical order, the listener is less surprised when the scene is incongruent and can directly judge whether it was congruent. However, with a canonical introduction, some expectation has already been build, therefore a much larger surprise when the scene turns out to be incongruent.

In her study, van Rij (2012) used well established paradigms of research. One of them being the visual world paradigm, as well as the truth-value judgment task. The visual world paradigm provides a visual scene with which aids the interpretation of the language input. A review by Huettig et al. (2011), outlines the suitability of the paradigm to study the interplay between linguistic and visual information processing. As for the truth-value judgment task, it has provided some of the more insightful methods of assessing children's linguistic competence (Gordon, 1998), by asking them to make a bipolar judgment about whether a statement accurately describes some scene.

A follow-up study was carried out, using EEG as a second measure due to its high temporal resolution (Verhoeven, 2018). The study applied similar pupil dilation measurements as van Rij's experiment, however pupil dilations have been established to usually be evoked around 1000ms after the presentation of a stimulus (Hoeks and Levelt, 1993), thus their reliability on a temporal scale should be backed with other measurements. As well as adding a second measurement, the study employed the use of a blank-screen (Altmann, 2004). This meant that the image is removed from the screen before the recording of the description is played. Altmann showed that having the image present on screen wasn't necessary, subjects would simply rely on a mental representation. Another reason for using the blank screen paradigm was to try and minimize the visual system in language processing. By removing distracting stimuli during the presentation of the sentence, the researchers hoped to get a cleaner EEG signal which would focus on the features of language processing. A second important change in methodology is that Verhoeven (2018)'s study presented the image for a longer period of time before the recordings began, that being 2000ms instead of the 500ms in van Rij's experiment.

Removing the image and simply relying on a mental representation might, however, minimise the effects of visual context in the competing resolution process. In fact, the findings from the follow-up study were not in line with the original experiment by van Rij (2012). While van Rij did find an effect of linguistic and visual context on object pronoun processing, these findings were not replicated in last year's experiment. Instead, only an interaction between the visual scene (congruency) and introduction order (linguistic discourse) was found for the reflexive pronouns. The differences could arise from several places, with the main one being the removal of the visual scene, however other factors, such as different timings of stimuli presentation, could have also influenced the results.

1.1 Research Question

The current study was conducted with two main purposes in mind. The first aiming to provide more evidence for the competing-constraints account, and the second to potentially discover where the difference in findings between the two studies could arise from. Thus, the following question was formulated: ***Do visual and linguistic contexts have an influence on object pronoun processing in a visual world verification task?***

Moreover, if an effect is found:

- For which type of pronoun (object and reflexive pronouns) is the effect present?
- Do the introduction order and visual scene affect both pronoun types in the same way?
- How early do these effects occur?
- Does removing the visual scene have an influence on the found effect?

In the current study the blank-screen method was not utilised and the image was displayed on the screen for a longer duration compared to van Rij (2012)'s study, with the idea that if the findings now reflected van Rij's results, the presence of the image and timing of presentation could indeed alter the effects of linguistic and visual contexts on pronoun resolution.

1.2 Hypothesis

We hypothesise that both the linguistic and visual contexts would set up some expectation for the interpretation of the pronoun. This effect would be present for both pronoun types (object and reflexive). When the interpretation is not in line with the visual scene (incongruent) we would expect to see a larger pupil dilation, compared to the congruent case. In the case of an incongruent scene, we also expect the **canonical order** (actor-first introduction) to elicit a higher pupil dilation compared to the **non-canonical order** (actor-second introduction). In the case of a congruent scene, we expect the canonical order to elicit a smaller dilation than the non-canonical order of introduction. We expect the effects to take place early (around 1000ms) after the onset of the pronoun. On top of this, we expect the results to be somewhat in line with van Rij (2012) findings, as the image was kept on the display. In a case where the results differ slightly, this might be an indication that the timing of presentation plays a role in resolution too. Finally, if no effects of the introduction order are present, this would suggest further evidence for the initial-filter account.

2 Methods

2.1 Participants

In total, 32 subjects took part in the study. All participants were native Dutch speakers, of whom 20 were male and 12 were female. The age of the subjects ranged from 18-29 years old, with a mean age of 22.2 years old. Participants were recruited via direct approach around the University of Groningen campus, as well as fliers and inquiries on a social media based *"Paid Research Participants"* group. All participants were presented with the same written instructions, after which they were asked to sign an informed consent form as an agreement for taking part in the experiment. On average the experiment took approximately an hour, for which participants were rewarded with a compensation of €10.

2.2 Design

The experiment was designed on the basis of a picture verification task, with ideas drawn from the visual-world paradigm. The paradigm allowed for an investigation into the influence of both visual and linguistic context on the object and reflexive pronoun processing (Huettig et al., 2011).

Collection of data was done in a 2x2x2 design, where the following were varied across conditions:

- Picture Type (self-oriented or other-oriented picture)
- Introduction Order (canonical and non-canonical order)
- Anaphor (pronoun) type (object pronoun or reflexive pronoun)

The experiment followed a within-subject design.

Picture Type consisted of two levels. Either an other-oriented picture (Figure 2.1), in which the the actor (some animal) performs an action on a patient (another animal); or a self-oriented picture (Figure 2.2), in which the actor (again, an animal) performs the action upon him/herself. In total there were 80 unique pictures, which form the 40 pairs of self-and-other variants of the same picture. The pictures were sourced from Verhoeven (2018)'s study, with van Rij (2012) also using a subset of these. The same images were used in order to keep the set-up constant.

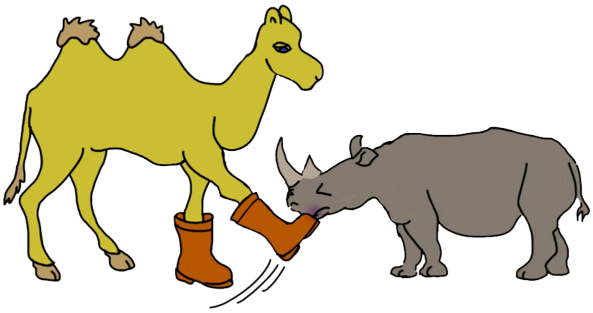

Figure 2.1: Example of an image illustrating an other-oriented action

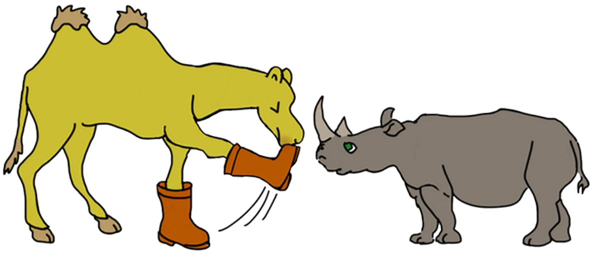

Figure 2.2: Example of an image illustrating a self-oriented action

The Introduction Order is related to the first sentence (the introduction sentence) which was presented after 2000ms after the image. The image was also kept on the screen during the presentation of the sentences. For this condition there were two levels: an A1 introduction (canonical) in which the actor is introduced first, followed by the patient (*"Zojuist zag je een muis en een eekhoorn."*, you just saw a mouse and a squirrel); or the A2 introduction (non-canonical) in which the agent is introduced second (*"Zojuist zag je een eekhoorn en een muis."*, you just saw a squirrel and a mouse).

Once the introduction sentence is played, it is followed by a test sentence. The test sentence contained the referring expression, or in other words the Anaphor Type. The anaphor used was either an object pronoun, **"hem"**, (*"De muis raakte hem aan met een lepel"*, the mouse touched him with a spoon), or a reflexive pronoun, **"zichzelf"** (*"De muis raakte zichzelf aan met een lepel"*, the mouse touched himself with a spoon).

The experiment constituted of one training block (made up of three training trials) and four experimental blocks (each made up of 40 test trials), resulting in 163 total trials. Participants were presented twice with each picture, with each image occurring in a different block and with two of the four conditions (Introduction Order x Anaphor Type). To make sure that no repetitions of the combinations occurred, as well as ensuring that no two combinations were shown after each other, unique lists were created per participant. The lists, just like the pictures were sourced from the previous two studies. Each list was segmented into four parts, corresponding to the four experimental blocks, with the order in each segment, for each participant, being randomized to avoid any further bias.

2.3 Materials/Stimuli

The entire procedure of the experiment, including the four separate blocks, ordering and presentation of the stimuli (pictures/sentences), and storing of the data was programmed in Experiment Builder (SR Research).

Pictures were presented against a light grey background (RGB: 153, 153, 153) with a width of 500 pixels. Images were centered, with the height of the background depending on the image ratio. Half of the pictures were randomly selected and mirrored.

For each picture, Verhoeven (2018) recorded two sentences. This study used the same recordings, that being the recordings for the introduction and test sentences. Sentences were recorded in the recording studio of the Faculty of Arts, University of Groningen. Afterwards, recordings were manipulated by means of splicing and normalising, using the PRAAT program (Boersma and Weenink, 2018). This was done to ensure that all sentences followed the same intonation to reduce the possible variable reaction of each participant to the recordings.

Both the introduction and test sentences were build in a similar style with artificial breaks. The introduction sentences were divided into two kinds based on the Introduction Order condition: A1 or A2 (canonincal and non-canonical order). The structure for both types was as follows: *"Zojuist zag je"* (you just saw) + 100ms silence + <referent> + 100ms silence + *"en"* (and) + <referent>. In the case of the A1 condition, the first referent would be the actor, and the second would be the patient. The opposite was the case for the A2 condition.

For the test sentences three variants were recorded: one being an object noun sentence, the other with an object pronoun and finally with a reflexive pronoun. The sentence which contained the object noun (e.g., *"De muis raakte de eekhoorn aan met een lepel"*, the mouse touched the squirrel with a spoon) was used as the basis for forming the other two conditions of the sentence. The object pronoun (*"hem"*) and the reflexive pronoun (*"zichzelf"*) were then spliced into the appropriate place, replacing the object noun. This method was chosen in order to keep the intonation of the rest of the sentence identical. The structure of the test sentences was therefore: <Actor> + 100ms silence + <verb> + 100ms silence + <anaphor> (object pronoun/reflexive pronoun) + 100ms silence + <prepositional phrase>. All test sentences ended with the propositional phrase. When presenting the sentences there was a fixed break of 200ms between the introduction and test sentence.

Once the sentence recordings were played to the participant, an answer screen would appear. The answer screen contained two boxes of rectangular shape: a green one with the word "correct" in it, and a red one with the word "incorrect" in it. The Ctrl-left button was linked to the answer on the left side of the screen, and the Ctrl-right button for indicating the answer on the right side. The order of the boxes was randomly swapped each trial, to minimize automatic motor responses of the participant as the experiment progressed into the later stages.

2.4 Apparatus

A computer screen was used to present the experiment to the participant. A Dell 2007FPB screen was used, measuring at 16.1 by 12.1 inches. The resolution was set to 1920 by 1080 pixels. The same screen was used for all subjects to ensure equal brightness levels. A 35mm lens was positioned approximately 70cm from a headrest, in which the participants positioned their head for the entirety of the experiment. The adjustable headrest ensured that that their head remained stationary throughout the trials, and provided some comfort for the subjects. Between the headrest and the lens, a keyboard was positioned for starting the experiment and giving the appropriate responses in each trial.

The eye-tracking device used was an EyeLink 1000 (SR Research) with a sampling rate of 500Hz. The pupil of the right eye was monitored continuously for each participant, with the measuring set to the diameter of the right eye.

The subject was seated in a non-adjustable chair, in the same room as the experimenters, with a shelf separating the experimental computer from the monitor displaying the eye-tracking data. A speaker was placed next to the screen and was used to play the recorded sentences to the subject.

2.5 Procedure

Before the participant arrived to the room, the set up of each session was checked against a pre-made checklist to ensure consistency for all subjects.

Subjects were warned that wearing mascara, glasses or hard-contact lenses was not allowed, since these are factors that all influence the precision of the eye-tracker. Wearing soft-contact lenses was permitted, however this was reported in the session log for future reference. Once the subject arrived, they were given written instructions about the experiment, and were asked to sign a consent form. The experimenter used a check-list to ensure that the subject received a unique list number. Before starting the experiment, the headrest was adjusted to meet the height of the participant, so that their head would sit comfortably to minimise movement throughout the experiment. Finally the keyboard was placed between the lens and the headrest, at a comfortable position for the subject.

Once set up, oral instructions were given to the participant, as well as similar written instructions given on the screen. The written instruction on the screen allowed the subjects to get familiar with the type of pictures they would see, as well as how they would need to answer the verification in each trial by pressing the appropriate keys.

After the subjects became clear with the task, a nine-point calibration was performed, followed by a validation. On average, a deviation of 0.5 was set as an aim for each calibration. A lower deviation value meant a more precise calibration and therefore more precise eye-tracking data. In case of a higher deviation (more than 0.5) the calibration was repeated.

Once the calibration was performed, the subjects were faced with three training trials. Each trial started with a fixation point, of up to 5000ms. An invisible square surrounded the fixation point, checking that the participant's gaze was present there for at least 100ms. Once the fixation check was cleared, an image would be presented for 2000ms. Following this, and while keeping the image on the screen, the two sentences (introduction and test sentence) associated with that specific image were played back to the participant. The test sentence was played 200ms after the introduction sentence. Afterwards, the two buttons (correct and incorrect) would appear on the screen for up to 5000ms, allowing the subject to press the corresponding key. This order constituted one trial in the experiment. A visualisation of a trial can be seen in Figure 2.3.

In a case where the eye-tracker recognized that the subject wasn't looking inside the invisible square during the fixation check for 100ms within the 5000ms, another calibration was performed and that specific trial was skipped.

The subjects started off with three practice trials, with pictures that weren't used during the actual experiment.

Once finished with the practice trials, another calibration was performed and the first block of the experiment would begin. A break was given to the participant between each block, as well as a new calibration before going onto the next block. During the experiment, both experimenters were present in the room. One was in charge of calibrations and ensuring the procedure was going smoothly, while the other kept a participant log with notes of the current session. Any comments, complaints or problems were noted down in the log, as this would be useful information later in the analysis.

At the end of the experiment the participants were thanked for the help, some demographic data

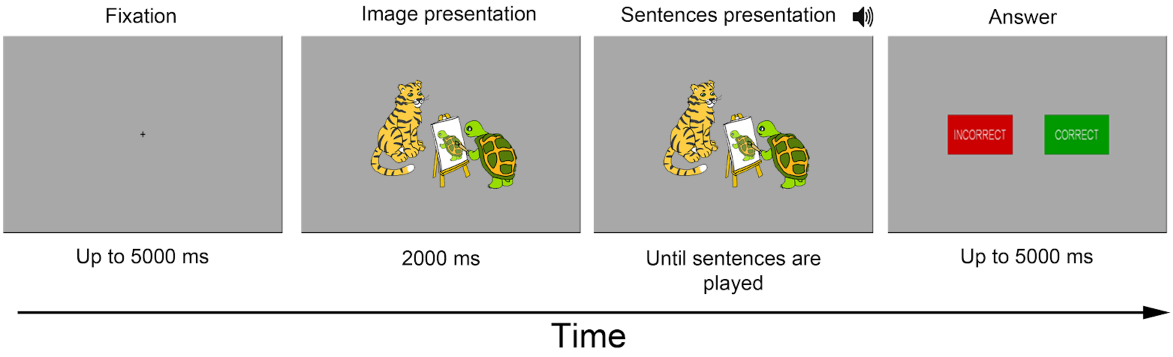

Figure 2.3: Visualisation of an entire trial

such as their age was recorded. The subjects were given a debriefing letter which described the aim of the experiment in more detail. They were also asked for their bank details, and compensated accordingly to the time it took to complete the experiment.

2.6 Data Preprocessing

The pupil dilation data was preprocessed with a script from van Rij, using R (R Core Team (2018)). Before preprocessing the data, an exploratory analysis revealed a large number of artifacts, especially in the later stages of the experiment. These artifacts included pupil occlusions, such as blinks, and random fluctuations in the size of the pupil due to saccades. Thus the script removed blinks with a padding of 200ms around the occlusion, while saccades were removed with a velocity threshold of 5. Once removed, the data was interpolated using a cubic spline method to fill in the missing data points (Mathôt, 2013).

The first 80 trials (the first two blocks) have been used for the analysis. After the first two trials participants had already seen each image at least once, and many of the participants reported fatigue in the second half of the experiment. Therefore the results from the first half would be more reliable, and less affected by artifacts and distractions on the participant's side.

The data was aligned with the onset of the pronoun, and baselined on 250-0ms before the onset of the test sentence.

2.7 Statistical analysis

Statistical analyses have been performed on the eye-tracking data, using linear mixed-effect (LME) models. Similar to the preprocessing, analysis was completed in R (R Core Team, 2018), using the ”lme4” package (Bates et al., 2015), and following a tutorial outlined by Winter (2013). Analysis was performed first on the most complex model, followed by split analyses on both sentence types (object pronouns and reflexive pronouns).

The chosen window for analysis was 750 to 1250ms after the onset of the pronoun. The reasoning behind this is that pupil dilation peaks around 1000ms after the stimulus onset which triggered the dilation Hoeks and Levelt (1993). For each subject, the median pupil size per trial within this window was computed and taken for analysis.

3 Results

This section starts off with an overview of the behavioural data, followed by a description of the pupil dilations for both Sentence Types. Analysis is first done on all conditions, followed by a split analysis on Sentence Type.

It is important to note that not all data was included in the subsequent sections. Firstly, the data from the second half of the experiment (block three and four) was discarded due to reasons mentioned in the previous section. The first three training trials were not included either. Following this, trials where the participant gave a wrong answer were also discarded from the results.

Before interpolating the data, an exploratory analysis was performed to check the percentages of NAs (missing data points) for each participant across the entire experiment. In cases where a participant had more than a quarter of their data missing (a 25% threshold), a decision was made to remove that participant from subsequent analysis. We believe that this was necessary as it was likely that such gaps in the data would prove to be unreliable for us to draw conclusions on. After applying the criteria, a total of 19 subjects were left for subsequent analysis.

Finally, many participants reported confusion with one of the images presented and its accompanying audio file, which contained an object many of them thought was named wrong (a spoon, instead of a stick). Therefore, all trials containing that image and its mirrored version were removed too.

3.1 Behavioural data

Accuracy and response times for all conditions were computed in order to gain insight into the overall performance during the experiment. Figure 3.1 illustrates the accuracy for the four conditions in the object pronoun trials, while Figure 3.2 shows the same for the reflexive pronoun trials. The other two figures (Figure 3.3 and Figure 3.4) show the response times (aligned to the pronoun onset) for each of the conditions.

’O’ and ’S’ refer to other or self-oriented Picture Type conditions respectively, while ’A1’ and ’A2’ refer to the Introduction Order. That is ’A1’ is the canonical introduction, and ’A2’ being the non-canonical introduction.

The first thing to observe from the plots is the high accuracy rate. The average accuracy across all conditions, for both Sentence Types, is 97.8%. The plots also seem to suggest no differences between the conditions, with all standard error bars overlapping. This suggests that the participants had no trouble understanding the experiment and the different conditions had no influence on how the participants performed.

The response time plots suggest something similar. The average response time across all conditions was 686.7ms. Large variability in the data (all conditions overlapping in their standard error bars) again suggests no difference between the conditions. Therefore, the different parts of the experiment seemed to have no influence on the response times of participants.

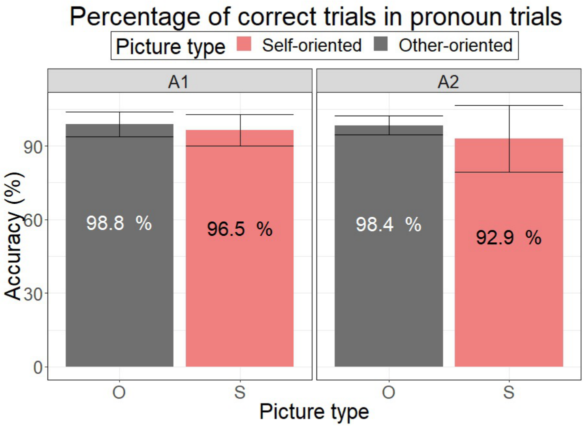

Figure 3.1: Accuracy for each condition for object pronoun trials

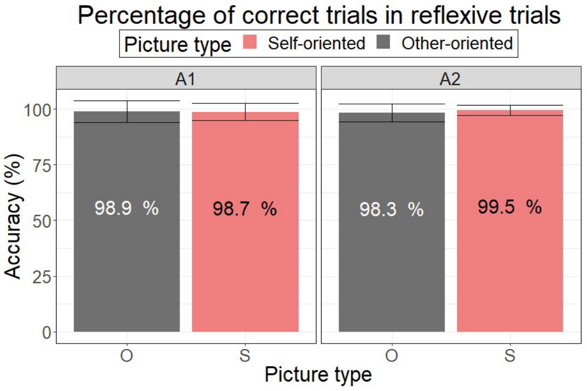

Figure 3.2: Accuracy for each condition for reflexive pronoun trials

3.2 Pupil dilation

Figure 3.5 shows the results of the pupil dilation for the object pronoun trials. The dashed line at 0 represents the onset of the object pronoun (in this case the beginning of the word *”hem”*). The x-axis represents Time in ms, while the y-axis indicates the baselined pupil dilation in arbitrary units, set by the eye-tracking software. The condition for Picture Type is represented with the two colours: the black lines represent other-oriented (’O’) pictures, while the red lines represent the self-oriented (’S’)

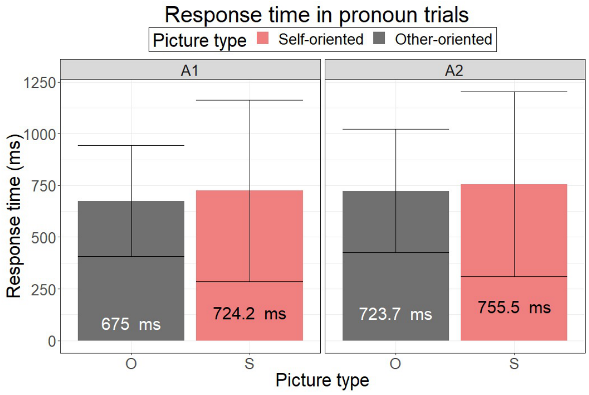

Figure 3.3: Response times for each condition for object pronoun trials

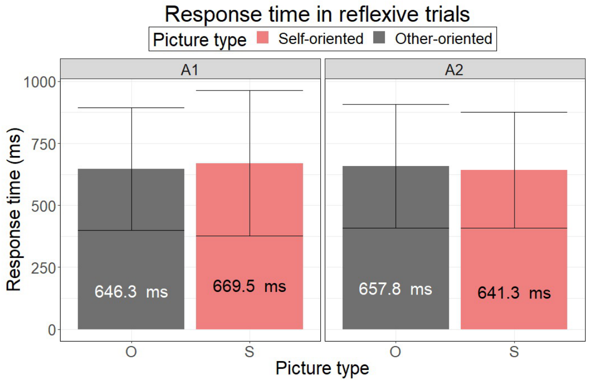

Figure 3.4: Response times for each condition for reflexive pronoun trials

pictures. In this case, for the object pronoun plot, the black lines ('O' condition) are congruent, while the red lines (S condition) are incongruent with the type of anaphor used in the sentence. The second condition, Introduction Order is represented by the style of the line: solid lines represent a canonical order (actor-first (A1)), while the dashed lines represent a non-canonical order (actor-second (A2)).

A first inspection of the plot indicates a difference between the conditions (Picture Type x Introduction Order). In the canonical order (A1) larger dilations are observed when the Picture Type is incongruent with the object pronoun (S), compared to the congruent case (O).

Figure 3.6 shows the results of the pupil dilation for the reflexive pronoun trials. Similar effects of Introduction Order x Picture Type can be observed from a first look. The other-oriented (O) condition for the canonical order (A1), which in this case was incongruent with the pronoun used, elicited a larger pupil dilation than the self-oriented (S) condition, which in this case was congruent.

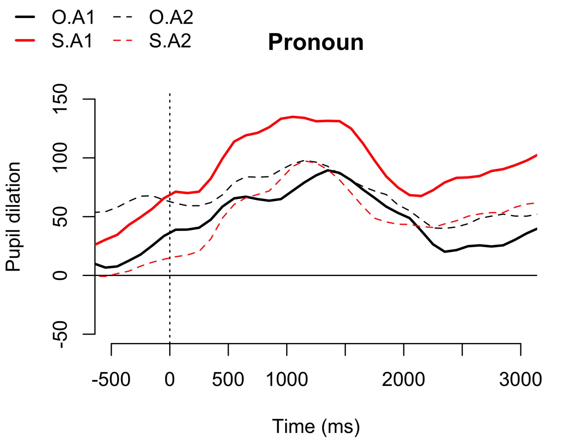

Figure 3.5: Pupil dilation for all participants in object pronoun condition

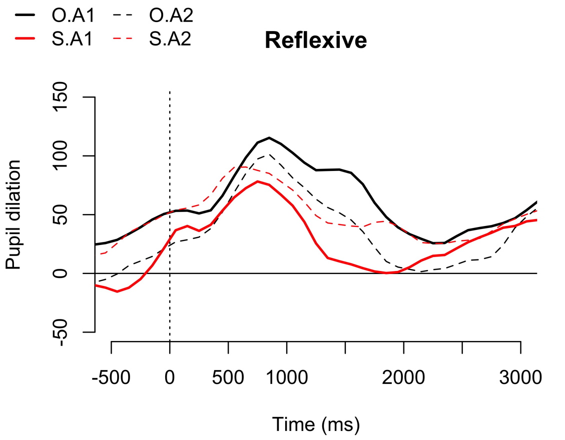

Figure 3.6: Pupil dilation for all participants in reflexive pronoun condition

A backward-fitting model comparison procedure (Winter, 2013) was performed on three separate linear mixed effect models: the simplest model containing only the main effects, a more complex model including the two-way interactions and then the most complex model which also includes the three-way interactions. As mixed effects, Subject and Image were included in the model, in order to account for variability between the subjects and possible different responses to each image presented. A Chi-squared test revealed that the model containing (1|Subject) as a random effect explained significantly more variance than the model without the random effect ($X^2(1) = 3413.2; p < 0.001$), and a similar significance was revealed for the model containing (1|image) ($X^2(1) = 1479.7; p < 0.001$). A Chi-squared test indicated that the three-way interaction model explains significantly more variance than the model containing only the two-way interactions ($X^2(1) = 62.17; p < 0.001$).

A summary of the most complex model can be seen in Table 3.1. The table summarises the significant differences in pupil size caused by each of the conditions. More importantly, it shows a significant difference caused by the three-way interaction of Picture Type x Introduction Order x Sentence Type (β = 99.587, SE = 12.624, t = 7.888). This significant difference suggests an influence of Sentence Type on the introduction of the other two conditions, therefore the analysis is now split into two categories: one for the object pronouns and one for the reflexive pronouns.

Table 3.1: Fixed effects on pupil dilation of the most complex model (three-way interaction)

Model: medianPupil ~(pictype + introtype + sentencetype)ˆ3 + (1|Subject) + (1|image)			
	Estimate	**Std. Error**	**t value**
(Intercept)	71.197	25.414	2.801
pictypeS	62.997	16.318	3.861
introtypeA2	20.806	6.283	3.312
sentencetypeR	35.576	6.233	5.708
pictypeS: introtypeA2	-71.729	8.971	-7.996
pictypeS: sentencetypeR	-107.804	8.879	-12.141
introtypeA2: sentencetypeR	-39.977	8.874	-4.505
pictypeS: introtypeA2: sentencetypeR	99.587	12.624	7.888

A similar backward-fitting model comparison process has been performed on each of the split data sets. Two models have been created for each sentence type: one with only the main effects, and one containing the two way interaction Picture Type x Introduction Order.

3.2.1 Pupil dilation: pronoun analysis

A Chi-squared test revealed that the more complex model, including the two-way interaction of Picture Type x Introduction Order, explains significantly more variance than the simpler model containing just the main effects ($X^2(1) = 76.17; p < 0.001$). Therefore the more two-way interaction model is chosen for the analysis. Table 3.2 shows a summary of the complex model.

Table 3.2: Fixed effects on pupil dilation of the model with two-way interaction for objective pronouns

Model: medianPupil ~(pictype + introtype)ˆ2 + (1|Subject) + (1|image)			
	Estimate	**Std. Error**	**t value**
(Intercept)	69.256	31.604	2.191
pictypeS	64.165	24.532	2.616
introtypeA2	26.620	6.164	4.318
pictypeS: introtypeA2	-76.897	8.800	-8.738

First, there is a significant difference caused by the Picture Type condition on the processing of object pronouns (β = 64.165, SE = 24.532, t = 2.616, for self-oriented images). This can be interpreted as follows: an actor-first introduction elicits a larger pupil dilation in self-oriented (incongruent) pictures than in the other-oriented (congruent) images. This can also be observed in Figure 3.5, the line for the self-oriented, actor-first condition is higher than the other-oriented, actor first condition.

Following this, there is also a significant difference caused by the Introduction Order condition (β = 26.620, SE = 6.164, t = 4.318, for an actor-second introduction). This means that an actor-second introduction elicits a larger pupil dilation than an actor-first introduction for the other-oriented pictures. Again, this can be seen in the same figure (Figure 3.5). The line for the other-oriented, actor-first condition is lower than the line for the other-oriented, actor-second condition.

Finally, the interaction of the conditions Picture Type x Introduction Order also causes a significant difference (β = -76.897, SE = 8.800, t = -8.738, for a self-oriented picture with an actor-second introduction). This suggests an interaction effect, which results in a stronger effect of introduction order for self-oriented images than for other-oriented images. Again, this can be observed in the graph (Figure 3.5). The lines for the self-oriented conditions are further apart than the lines for the other-oriented conditions.

3.2.2 Pupil dilation: reflexive analysis

The same backward-fitting procedure for model comparison is carried out for the reflexive pronouns. A Chi-squared test revealed that the more complex model, including the two-way interaction of Picture Type x Introduction Order, explains significantly more variance than the simpler model containing just the main effects ($X^2(1) = 7.65; p = 0.0056$). Therefore, the more complex model is chosen for further analysis. Table 3.3 shows the summary for the complex model.

Table 3.3: Fixed effects on pupil dilation of the model with two-way interaction for reflexive pronouns

<table>
<tr><th colspan="4">Model: medianPupil ∼(pictype + introtype)ˆ2 + (1|Subject) + (1|image)</th></tr>
<tr><th></th><th>Estimate</th><th>Std. Error</th><th>t value</th></tr>
<tr><td><b>(Intercept)</b></td><td>103.352</td><td>26.426</td><td>3.911</td></tr>
<tr><td><b>pictypeS</b></td><td>-40.494</td><td>21.327</td><td>-1.899</td></tr>
<tr><td><b>introtypeA2</b></td><td>-14.731</td><td>5.964</td><td>-2.470</td></tr>
<tr><td><b>pictypeS:
introtypeA2</b></td><td>23.418</td><td>8.465</td><td>2.766</td></tr>
</table>

First, no significant difference is caused by the Picture Type condition. It is however worth noting a borderline significant difference (β = -40.494, SE = 21.327, t = -1.899, for self-oriented images). This suggests that for an actor-first introduction, other-oriented (incongruent) images could elicit a higher pupil dilation than self-oriented (congruent) images. Figure 3.6 hints towards this difference, as we do observe some difference between the other-oriented, actor-first condition and the self-oriented, actor-first condition.

Moreover, there is a significant difference caused by the Introduction Order condition (β = -14.731, SE = 5.964, t = -2.470, for for an actor-second introduction). This means that an actor-second introduction elicits a smaller pupil dilation than the actor-first introduction, for other-oriented images. This can indeed be observed in the graph (Figure 3.6), as the line for the other-oriented, actor-first condition is higher than the line for the other-oriented, actor-second condition.

Finally, the interaction between the conditions of Picture Type x Introduction Order also causes a significant difference (β = 23.418, SE = 8.465, t = 2.766, for a self-oriented picture with an actor-second introduction). This suggests an interaction effect, which implies a stronger effect of introduction order for the other-oriented images than for the self-oriented images. This can also be observed in the figure (Figure 3.6), the lines for the other-oriented conditions are slightly further apart than those of the self-oriented conditions.

4 Discussion

The current study investigated the effects of visual and linguistic contexts on pupil dilation during object pronoun resolution. We hypothesized that indeed, in the early processing stages grammar is not the only criteria for possible antecedent candidates, rather linguistic and visual information compete with these grammatical principles. We found an influence of Sentence Type on the dilations of the pupil. In other words, the type of referent used in the sentence (objective pronoun or reflexive pronoun), had an effect on the pupil dilation, and thus on the processes which individuals use to resolve these referents.

During the split analysis on the objective pronouns, it was revealed that both the visual scene and the order of introduction, as well as the interaction between the two, cause significant differences in the pupil dilation. Similarly, for the reflexive pronouns, the order of introduction and the interaction between the two conditions caused significant differences, while the visual scene seemed to play less of a role with a borderline significant difference in pupil dilations.

This study was carried out with the purpose of investigating whether and how the visual and linguistic contexts influenced the processing of object pronouns. It also aimed to settle the different findings between van Rij's (2012) original experiment and the follow up study of Verhoeven (2018). More importantly it asked the following question: ***Do visual and linguistic contexts have an influence on object pronoun processing in a visual world verification task?***

Moreover, if an effect is found:

- For which type of pronoun (object and reflexive pronouns) is the effect present?
- Do the introduction order and visual scene affect both pronoun types?
- How early do these effects occur?
- Does removing the visual scene have an influence on the found effect?

The results provided above are in line with our expectations. That is, linguistic information (introduction order, grammatical role, recency etc) competes with visual information in the early stages of pronoun resolution. This competition was observed in both pronoun types. With these results in mind we accept our hypothesis, and thus conclude more evidence towards the competing-constraints account.

In a more general sense the results can be explained as follows. When the actor of the sentence is introduced first, the visual scene affects its resolution. In cases where the visual scene was congruent with the referent used, subjects were less surprised, compared to the incongruent visual scene. This reflects well in the plots of the object pronoun pupil dilations (Figure 3.5), as we observe the incongruent line ('S') eliciting a higher dilation than the congruent line ('O'), when the actor was introduced first. For the reflexive pronouns we can't conclude that this is indeed the case, due to the borderline significance of the Picture Type condition for a canonical introduction. However the plot (Figure 3.6) does show a higher pupil dilation for the incongruent trials, compared to the congruent ones, when an actor is introduced first.

As for the Introduction Order, when a subject was presented with a canonical order during a congruent scene, they were less surprised compared to when the order was non-canonical. Again, this can be observed in the object pronoun plot (Figure 3.5), as the other-oriented, actor-second condition elicits a higher pupil dilation than the other-oriented, actor-first condition. We also observe a similar effect in the reflexives pronoun plot (Figure 3.6), where the self-oriented, actor-second condition elicits a larger pupil dilation than the self-oriented, actor-first condition.

Finally the interaction between the two conditions of Picture Type x Introduction Order also had an effect on the object pronoun resolution. The interaction effect suggests that the Introduction Order played stronger effect during incongruent visual contexts, compared to the congruent visual context.

These findings are indeed in line with previous research (van Rij, 2012). The largest surprise (largest pupil dilation) can be observed in the cases where the actor was introduced first, but the visual context was incongruent with the description. This suggests that subjects indeed form some expectation. They expect the first mentioned referent to be mentioned as the subject of the next sentence. When this is the case, we observe the smallest pupil dilation (other-oriented, actor-first condition in Figure 3.5). When this expectation is violated, and the subject of the second sentence does not refer to the first mentioned referent (self-oriented, actor-first condition), we observe the largest pupil dilation. Hence, the surprise in the listener during resolution.

As for the Introduction Order, the results are also in line with previously mentioned ideas. As mentioned before, listeners might be more cautious building expectations in cases where the actor is introduced in a non-canonical order, and thus subjected to the prominence effects of language (Gernsbacher and Hargreaves, 1988; Gordon et al., 1993; Kaiser and Trueswell, 2008). The may wait for the object to be mentioned too before building a sentence representation. Thus, they would be better at resolving this in a congruent scene, and less surprised when the scene turns out to be incongruent. These effects can be observed from Figure 3.5 and Figure 3.6. A non-canonical introduction for the congruent visual scene elicits a slightly higher pupil dilation than the canonical introduction, as the listener is more cautious after hearing the non-canonical introduction. In the case of a incongruent scene, we observe a smaller dilation when in the non-canonical introduction. This again suggests the hesitance of the listener to form an expectation, and therefore they are less surprised when the actor was introduced second, even though the scene was incongruent.

Finally a note on where the possible difference in findings could have occurred between the previous two experiments. The results presented suggest that indeed the removal of the visual scene, and simply relying on a mental image, could influence the competing of the linguistic and visual contexts. In fact, Laeng and Sulutvedt (2014) provide some evidence for the luminance of mental pictures having an effect on pupil size during the retrieval of the scene. In Verhoeven (2018)'s experiment, when subjects were asked to press the corresponding key, and thus retrieve the mental image, it might have been the case that certain brightness features of the picture influenced the pupil size in ways the experimenters could not predict or control.

It is important to outline some potential problems with the design of the study. Firstly, the recorded sentences used in this version of the experiment were actually suited better for the blank-screen paradigm. All sentences told the listener that they *"just saw"* some agent and patient, as originally the visual scene was removed before playing the recordings. As we kept the picture on the screen during the playback of sentences, conflicting information was presented to the listener from the very beginning. This is something which was overlooked at the early stages of this study, and only noticed after some subjects reported confusion. However, because the conflict occurred in very early stages of the trial (at the start of the introduction sentence), it is unlikely that it had an influence on the pronoun effects, which occurred after the test sentence.

The length of the experiment itself could be a cause for some potential problems. Firstly, fatigue was reported by all subjects during the later stages of the experiment. This lead to more blinks, in some cases multiple calibrations needed to be redone (and thus trials skipped), as well as lack of focus and interest on the participant's side. We tried to overcome this issue by only looking at the first 80 trials as well as interpolating the data to fill in the missing data. While reliable for pupil time series (Mathôt, 2013), the interpolation could have introduced fluctuations into the data which were not present originally. Another downside to the length of the experiment is the opportunity of subjects to learn the task well and employ certain strategies. They could have not been paying attention to the introduction sentence, taking the time to maybe break or rest their eyes, and simply relied on the test sentence and the visual scene to give a response. While this is plausible, effects of Introduction Order were found for both referent types, showing that the introduction sentence did indeed have an effect on their understanding.

Another problem could be the external and ecological validity of the images. The images were originally designed for a children's experiment, and while their cartoon nature definitely did not reflect real life situations, adult subjects might find some of them comedic and too distracting. A few cases of ambiguity of the images/recordings were also noted. The most evident example of that being Image 05 which reported a spoon in the scene, while the recording did not match this description. As mentioned earlier, this image was removed from the analysis. A separate analysis with the image included can be seen in the Appendix A. Including the image in the analysis results in only the interaction Picture Type x Introduction Type to be significant in the reflexive pronoun analysis, thus indicating the effect this mistake had on the subjects. Further, it demonstrates how a single ambiguity could alter our findings. Several subjects reported confusion with the gender role of a recorded referent not being in line with the animal presented on screen. While it may not have been an influencing factor, past research has shown the importance of gender principles in pronoun resolution (Alves, 2016). Therefore this could be an important suggestion to fix in subsequent studies, as well as eliminating any possible ambiguities.

Currently, reproducibility might also be an issue in the experiment, mainly due to the many versions of the cartoon images. A future study might wish to present the same two animals in all of the trials. First, this would make the experiment easier to reproduce. Secondly, this would minimise the difference in individual reactions to specific animals, thus making the findings more credible.

The slight difference in findings between the objective and reflexive pronouns could be another basis for further investigations. Object pronouns are more ambiguous than their reflexive counterparts. An object pronoun could refer to two antecedents, whereas this is not the case for the reflexives as they can refer to only one antecedent. All images presented to the subjects contained one actor and one patient, in other words only one of the animal on screen was performing an action. When hearing a test sentence containing an object pronoun, while observing a scene where only one action is performed, the listener would need to hear our the entire test sentence to determine which of the two antecedents the pronoun could refer to. On top of this, all test sentences finished with a propositional phrase, which is another clue for a correct interpretation. In the reflexive trials, due to reflexive pronouns being able to refer to only one antecedent, subjects would be quicker in deciding what that antecedent was. A visual scene where only one of the animals is performing an action would also act as a really strong cue. A future study could adjust the nature of the images, so that both animals are performing some action. An example might be an image where a mouse is touching itself with a spoon, and a squirrel that is touching itself with a fork. When hearing the test sentence *"the mouse touched himself with a spoon"*, the subject might be slower in their decision, as they need to hear the entire sentence to determine exactly which of the two actions is being described. This change in ambiguity in the visual context might reveal stronger effects of the Picture Type during the reflexive pronoun processing.

Whether the suggested changes have a noticeable improvement on the findings presented here can only be answered through more research. We are however pleased with the presented results and support of the competing-constraints account.

5 Acknowledgements

Firstly, I would like to thank Federico Ferlito for his partnership during the conduction of the study. I would also like to thank Jacolien van Rij and Jelmer Borst for initiating the project in the first place and for their supervision, contribution and ideas during the study.

References

Altmann, G. T. (2004). Language-mediated eye movements in the absence of a visual world: The blank screen paradigm. *Cognition*, 93(2):B79–B87.

Alves, M. (2016). Gender features in pronoun resolution processing in brazilian portuguese. *Proceedings of the Linguistic Society of America*, 1:27–1.

Arnold, J. E. (1999). Reference form and discourse patterns.

Badecker, W. and Straub, K. (2002). The processing role of structural constraints on interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(4):748.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Beatty, J., Lucero-Wagoner, B., et al. (2000). The pupillary system. *Handbook of psychophysiology*, 2(142-162).

Boersma, P. and Weenink, D. (2018). Praat: doing phonetics by computer (version 6.0. 40).

Chomsky, N. (1981). Lectures on government and binding. dordrecht: Foris.-. 1986a. knowledge of language: Its nature, origin, and use.

Chow, W.-Y., Lewis, S., and Phillips, C. (2014). Immediate sensitivity to structural constraints in pronoun resolution. *Frontiers in Psychology*, 5:630.

Clackson, K., Felser, C., and Clahsen, H. (2011). Childrens processing of reflexives and pronouns in english: Evidence from eye-movements during listening. *Journal of Memory and Language*, 65(2):128–144.

Clifton, C., Kennison, S. M., and Albrecht, J. E. (1997). Reading the wordsher, his, him: implications for parsing principles based on frequency and on structure. *Journal of Memory and language*, 36(2):276–292.

Engelhardt, P. E., Ferreira, F., and Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken language comprehension. *The Quarterly Journal of Experimental Psychology*, 63(4):639–645.

Gernsbacher, M. A. and Hargreaves, D. J. (1988). Accessing sentence participants: The advantage of first mention. *Journal of memory and language*, 27(6):699.

Gordon, P. (1998). The truth-value judgment task. In *Methods for assessing childrens syntax*. Citeseer.

Gordon, P. C., Grosz, B. J., and Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive science*, 17(3):311–347.

Hoeks, B. and Levelt, W. J. (1993). Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research Methods, Instruments, & Computers*, 25(1):16–26.

Huettig, F., Rommers, J., and Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta psychologica*, 137(2):151–171.

Kaiser, E. and Trueswell, J. C. (2008). Interpreting pronouns and demonstratives in finnish: Evidence for a form-specific approach to reference resolution. *Language and Cognitive Processes*, 23(5):709–748.

Kennison, S. M. (2003). Comprehending the pronouns her, him, and his: Implications for theories of referential processing. *Journal of Memory and Language*, 49(3):335–352.

Laeng, B. and Sulutvedt, U. (2014). The eye pupil adjusts to imaginary light. *Psychological science*, 25(1):188–197.

Mathôt, S. (2013). A simple way to reconstruct pupil size during eye blinks. *FigShare*.

Nicol, J. and Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *Journal of psycholinguistic research*, 18(1):5–19.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Spenader, J., Smits, E.-J., and Hendriks, P. (2009). Coherent discourse solves the pronoun interpretation problem. *Journal of child language*, 36(1):23–52.

SR. Research experiment builder [computer software] (2017).

van Rij, J. (2012). *Pronoun processing: Computational, behavioral, and psychophysiological studies in children and adults*. Rijksuniversiteit Groningen.

Verhoeven, R. (2018). Influence of visual and linguistic context on object pronoun processing: Eeg and eye-tracking provide new information.

Winter, B. (2013). A very basic tutorial for performing linear mixed effects analyses. *arXiv preprint arXiv:1308.5499*.

Zellin, M., Pannekamp, A., Toepel, U., and van der Meer, E. (2011). In the eye of the listener: Pupil dilation elucidates discourse processing. *International Journal of Psychophysiology*, 81(3):133–141.

A Appendix

The following three tables were obtained after performing the same analysis, including the removed image from the experiment (Image 05). The first table shows the most complex model with the three-way interaction, followed by split tables for the most complex models (including two-way interactions) for both object and reflexive pronouns.

Table A.1: Fixed effects on pupil dilation of the most complex model (three-way interaction): Image 05 included

Model: medianPupil ~(pictype + introtype + sentencetype)ˆ3 + (1|Subject) + (1|image)			
	Estimate	**Std. Error**	**t value**
(Intercept)	71.471	25.206	2.835
pictypeS	59.392	16.066	3.697
introtypeA2	20.407	6.237	3.272
sentencetypeR	35.101	6.180	5.680
pictypeS: introtypeA2	-68.284	8.893	-7.679
pictypeS: sentencetypeR	-105.639	8.790	-12.018
introtypeA2: sentencetypeR	-35.339	8.789	-4.021
pictypeS: introtypeA2: sentencetypeR	91.753	12.521	7.328

Table A.2: Fixed effects on pupil dilation of the model with two-way interaction for objective pronouns: Image 05 included

Model: medianPupil ~(pictype + introtype)ˆ2 + (1|Subject) + (1|image)			
	Estimate	**Std. Error**	**t value**
(Intercept)	69.376	31.320	2.215
pictypeS	60.400	24.004	2.516
introtypeA2	25.099	6.109	4.108
pictypeS: introtypeA2	-72.074	8.711	-8.274

Table A.3: Fixed effects on pupil dilation of the model with two-way interaction for reflexive pronouns: Image 05 included

Model: medianPupil ~(pictype + introtype)ˆ2 + (1|Subject) + (1|image)			
	Estimate	**Std. Error**	**t value**
(Intercept)	102.901	26.115	3.940
pictypeS	-40.974	20.928	-1.958
introtypeA2	-10.042	5.910	-1.699
pictypeS: introtypeA2	18.771	8.443	2.223