Do coherence relations give an accurate indicator for ease of pronoun interpretation?

Douwe Schelvis
August 2013

Master Thesis
Artificial Intelligence / Human Machine Communication
Dept of Artificial Intelligence
University of Groningen, The Netherlands

Internal Supervisor:
Jennifer Spenader (Artificial Intelligence, University of Groningen)

External Supervisor
Jacolien van Rij (Quantitative Linguistics, Eberhard Karls Universität Tübingen)

Do coherence relations give an accurate indicator for ease of pronoun interpretation?

Douwe Schelvis
University of Groningen
Groningen, The Netherlands

Abstract

Research on pronoun interpretation has produced several theories to explain human behaviour, in particular the theories of Subject Preference, Parallel Function Preference and Coherence based Preference. We performed a pair of self paced reading experiments to test the predictions of these theories. Dutch participants were tested using sentences that modified the antecedent role (subject or object), grammatical match of the pronoun and antecedent (both subject or object, or different) and the coherence relation between sentence parts (cause-effect or resemblance). The second experiment goes into more depth and removes possible biases. Our findings do not match the predictions of any of the three leading theories, leading to the conclusion that online and unconscious behaviour differs from conscious judgement tasks.

Introduction

Pronoun interpretation is the process of deciding which is the best possible antecedent for a pronoun. Most people interpret pronouns so effortlessly they don't even realize they are doing it. However, how they do this and what features are guiding their interpretation is still being investigated. Please note the following two sentences:

(1) June had been to lunch with Sebastian, so she wasn't that hungry.

(2) Nicholas distracted Arnold at the last minute, so he didn't notice Sarah waving.

In sentence (1) the pronoun 'she' is interpreted as June. 'Sebastian' is a male name, and is therefore not considered a valid antecedent. This preference to have a match between the gender of the pronoun and its antecedent is one

of the strongest preferences playing a role in pronoun interpretation (together with number). Not all sentence constructions are this straight-forward though, and often preferences for antecedents is a lot weaker. For example, in sentence (2) 'he' can refer both to Nicholas and Arnold. Even though there is no strict rule saying which of the two the pronoun should refer to it is generally still read easily, with 'he' interpreted to mean Arnold. The mechanics underlying this antecedent decision process have been the topic of a lot of research which we'll cover in this paper. As we'll show, virtually all of the prior research has approached this topic using offline experiments like sentence judgement tasks. We have performed two experiments that will compare existing theories using the reading times of the participants to see which theories hold up.

In this paper we'll describe and test three theories that attempt to give a valid account of the way pronouns are interpreted. Our focus here is on a Coherence based account (Kehler, Kertz, Rohde, & Elman, 2008) but we will also consider the two preceding theories: The Subject Preference and then the Parellel Function Preference. The Subject preference theory states that a pronoun interpretation in which the antecedent was the subject of a previous sentence or sentence segment will get (some degree of) preference. The Parallel function preference refines this approach by saying that an antecedent gets preference when it fulfills the same grammatical role as the pronoun. The Coherence theory agrees with the Parallel Function Preference, but adds that it only holds when there is a resemblance relation between the sentences containing the pronoun and the antecedent.

These theories have been compared and verified quite extensively already, and I will cover a relevant segment of previous research in the next section. My hope is to offer new insights by looking at the online performance of participants rather than offline decisions. This is an approach that has not been covered as much when testing for interpretation preferences, and might yield new insights into the validity of the various theories and which makes the most accurate prediction for human pronoun interpretation. The problem with offline experiments is that the results only show effects that are large enough to influence the final interpretation of participants, which is only part of the impact a preference can have. Note the sentences below (3). In both cases the pronoun is interpreted to have Jonathan as its antecedent, however it could be argued that (3b) reads more naturally than (3a) since Erika comforting Jonathan is intuitively a result of him getting rejected.

(3) a. Markus rejected Jonathan, and Erika comforted him.
    b. Markus rejected Jonathan, so Erika comforted him.

In an offline judgement task, for example to assign an antecedent to the pronoun, there would be no difference in results between these two sentences.

Intuitively there is though, and the Coherence model in fact predicts a different level of preference for these two sentences. Though this might be possible to measure in an offline task, for example by asking participant for an acceptability rating of each sentence, the effects should become apparent far more easily in an online task where the reading times can be analyzed. Preferred forms may display shorter reading times, even if the final interpretation is the same. Another significant benefit of using an online task is that they provide an insight into the time course of the interpretation process, further advancing our understanding of the underlying mechanics.

In the next section we will introduce the three major competing theories of pronoun interpretation and give an account of the research behind each theory. As you will observe, virtually all of the experiments on which these theories have been based as well as the experiments with which they were tested have been offline tasks.

Background

Subject Preference Theory

The Subject Preference theory states that the referent for a pronoun is preferably the grammatical subject of either the same clause or a preceding sentence. So for example in sentence (4) 'he' is generally resolved to John.

(4) John serenaded David, then he ran off.

The intuitive explanation for this is that a sentence is about the agent or performer of an action, which is generally the subject, so it's more likely that the second clause is also about him or her. One of the first works to support this was that of Hobbs (1976). Hobbs was developing sentence parsing algorithms that also identify the correct antecedent of pronouns. He noticed that a large proportion of the found antecedents were in subject position. This has been raised as strong support for the Subject Preference theory, but it is only an indirect way of looking at what people actually do. As Hobbs (1976) pointed out himself, the way in which pronouns are used in published text isn't automatically a good representation of 'natural' English. As the test material for his parsing algorithm he used varied source material, ranging from a magazine to a novel, but it is possible that he would have found different results if he had looked at a series of letters or the transcripts of conversations.

Later research by Frederiksen (1981) takes a different approach to that of Hobbs (1976). Frederiksen (1981) looks at the reading times of participants in an online experiment. Though this study looked at anaphora in general and not just at personal pronouns (including abstract pronouns like 'it'), it still resulted in strong results to support the subject preference theory. In their

experiment they recorded and compared reading times of sentences based on the complexity of the context, including the number of possible referents for each anaphor. Their experimental items looked like (5) below.

(5) Education is, above all, supposed to produce a well-trained mind.
It should concern itself with developing the high ability to read, learn, and understand what men of intelligence have said about this world.

Their goal was to compile a set of rules or specifications that describes correct anaphor resolution. So not just for personal pronouns but for all anaphora. The sentences in their experiment were presented using an unchunked self-paced reading methodology, where the entire sentences were presented one at a time and the participants pressed a button to switch to the next sentence. After the sentence was read and the participant had pressed continue a line was displayed under the pronoun that meant that the participant had to explain what the pronoun referred to. Important to note is that the participants in the experiments were minors (grades 10-12).

Construction of the materials was done in such a way that there were different variants of each set of sentences where the complexity was modified by adding additional sentences, lengthening the existing sentences or both. To illustrate this, the sentence set (5) in another condition looked like (6) instead. This version has more sentences inserted to pad out the story and offers a wide variety of possible antecedents, some of which are even repeated within the text for additional difficulty.

(6) Education is, above all, supposed to produce a well-trained mind.
A well-trained mind possesses more than the ability to turn on a TV knob, fly an airplpane or make a good living.
Too often, the emphasis in our schools has been immediate practical goals, such as personal succes or wealth.
It should concern itself with developing the high ability to read, learn, and understand what men of intelligence have said about this world.

This is an example with an abstract antecedent for the pronoun, but there were also items where the underlined word in the final sentence was a personal pronoun that had to be interpreted. The data from these items is what I want to focus on, and the results of this are displayed below in Figure 1. The results show a clear increase in reading speed when the antecedent has a subject role in the preceding sentence when compared to otherwise similar material where the antecedent is in the object position.

There were three large drawbacks to this experiment which could undermine the validity and/or generalizability of their findings. The first

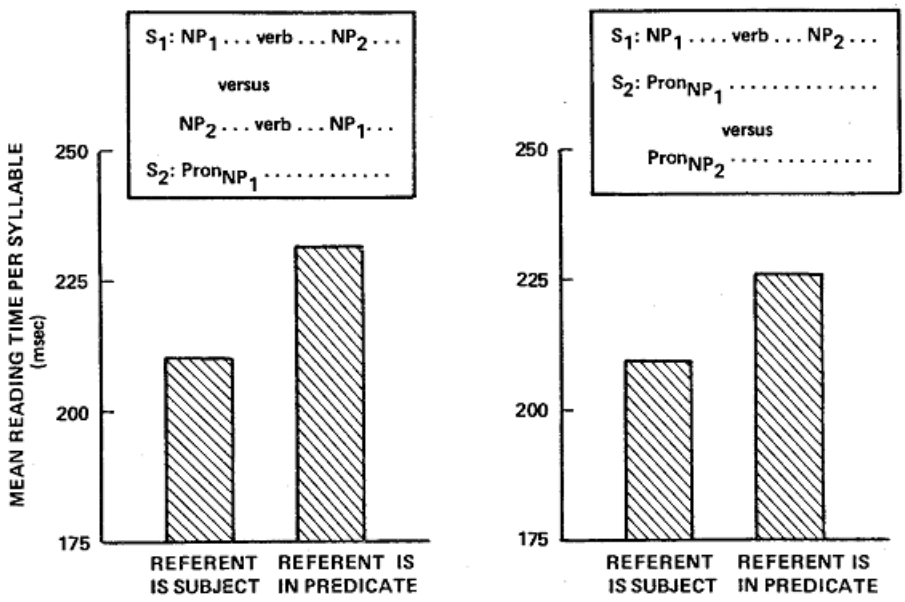

Figure 1. Reading time results from Frederiksen (1981) displaying an increase in reading speed when the antecedent is in subject position compared to it being in the object position.

problem is that the participants in this study were children with varying degrees of reading skill. This was taken into account in all of the analyses and showed that the effects were strongest in children with a lower skill level. This could imply that subject preference is not actually a strategy used by readers with an adult skill level (or at least that the prominence of the preference would be smaller in readers of a higher skill level). The second problem is that the interaction they looked at transcended sentence boundaries, which could also influence the results in unwanted ways. They did look at the influence of the amount of interjecting sentences and found no significant interaction as long as the content of the sentences was neutral. Additional antecedents did slow down readers though. The increase in speed might be because non-subject elements are harder to recall when several sentences have been read inbetween. The third problem lies in the setup of the experiment. Participants were asked to read entire sentences at a time. This makes the task more natural, but it also means that the reading times are averages over the entire sentence rather than just the relevant chunks. An analysis on a per-word basis might very well have shown very different results.

Crawley, Stevenson, and Kleinman (1990) also saw the drawbacks of the approaches in these two previous works (as well as other research on the topic)

and performed two experiments to investigate pronoun interpretation behaviour by looking at the judgements of pronoun interpretation problems as well as the speed of their resolution. Their goal was to investigate whether the Subject Preference Theory or the competing theory, of a Parallel Function Preference, proved to be more accurate to explain human behaviour. Parallel function preference states that there is no general preference for the pronoun to be interpreted as referring to a subject, but rather to an antecedent fulfilling the same grammatical role as the pronoun. We will cover this theory in more detail in the next subsection.

Similar to Frederiksen (1981) the stimuli used in their experiments consisted of multiple sentences per item, with sentence sets such as (7). Sentences were still presented completely and participants had to progress to the next sentence manually, but unlike the experiment from Frederiksen (1981) the preceding sentences (that still belonged to that item) remained on the screen for potential reference.

(7) Tommy and Kevin were twins and were trying to find out what their sister had bought them for their birthdays.
Samantha had bought them bikes and she hid them in case they started looking for them.
Tommy told Kevin to find the new bikes and Samantha shouted at him.

In the first experiment from Crawley et al. (1990), the pronoun in the final sentence of each item would either be ambiguous or only have a single correct interpretation determined by the gender of the characters. Each ambigous item would be followed by a question to ask the participant what named character the pronoun referred to. For the second experiment they only continued with the ambiguous sentences and did not ask direct questions between items. Instead they had people respond whether the pronoun refers to the first or second mentioned character in the text (where the first character was balanced equally between a subject and object role across items). Rather than asking people what antecedent the pronoun referred to they had two buttons to continue to the next stimulus. One signalling that it refers to the first mentioned character, and the other to the second. This way there could be no effect from one name being more common (either in general or that participant in particular), leading to more reliable data.

The results of both of the experiments showed significant preference towards antecedents in subject roles. The first experiment showed an increase in reading speed when the antecedent was interpreted as the subject, and also a preference towards replying that the antecedent was in subject position. This double effect of more subject antecedent replies as well as faster reading times for the critical sentence in those cases make for strong support.

Important to note because of a detail in our own experiment, Crawley et al. reported that there was no sign of an effect of the subject preference when the sentences were unambiguous, arguing that the pronoun interpretation was only being handled by the gender preference. The second experiment also showed an increased reading speed for subject assigned sentences as well as an increased rate of assignment for subject interpretations. These reading times that were compared were still based on the full sentence though, which is an imperfect indicator of the actual effect around the pronoun, which was presented at the end of the last item.

Further research into Subject Preference Theory has mostly used an offline methodology which is less relevant when considering the form of our own experiments.

Parallel Function Preference

The second theory that tries to give an account of pronoun resolution is the parallel function theory. The preference here is more sophisticated than a pure preference for the subject. Instead the claim is that an NP is a more suitable antecedent if its grammatical role is the same as that of the anaphor. Sentence (8) illustrates this, the object pronoun '*him*' is predicted to prefer the object Daniel over Adrian as its antecedent. Inversely, in (9) the pronoun 'she' is in a subject position so the predicted assignment would be to have Samantha as the antecedent instead of Lisa.

(8) Adrian deceived Daniel, and Edward ridiculed him.

(9) Samantha ridiculed Lisa, and she mocked Xavier.

If the pronoun is in a subject position this theory predicts the same preferences as the subject preference, but predicts an inversed preference if the pronoun is in object position. This theory was first introduced by Sheldon (1974) as a way to explain the behaviour she found in a toy-manipulation comprehension experiment with children.

Her research focused on the way children interpret relative clauses with the hopes of disproving a contemporary theory that stated that interruptions in the flow of a sentence increase the difficulty of correct interpretation. She performed a sentence interpretation experiment where experimental sentences like (10) would be read out twice to participants of ages 3 to 5 who then had to portray the events of the sentence using toy animals.

(10) The dog stands on the horse that the giraffe jumps over.

(11) The pig bumps into the horse that jumps over the giraffe.

In example (10) the expected actions from the child would be to place the dog on top of the horse for a second, then make the giraffe jump over the horse. In this example the horse is the object of both actions, so according to the proposed theory this sentence should be easier to interpret than an equivalent sentence with a nonparallel function (11). She recorded the acted out responses of the children and compared the number of errors for each category. Her findings supported the theory of a parallel function preference, with a significantly lower amount of errors in sentences where the recurring animal was fulfilling the same grammatical role. She also found no effect of sentence interruptions as she had hoped. This is, again, a purely action-driven analysis that does not take into account the processing speed, only the (accuracy of) the final response.

Sheldon (1974)'s research served as the basis for the parallel function preference, but she already pointed out herself that there is no guarantee that the results found in small children would generalize to adults. The form of her sentences also does not directly include pronoun interpretation but instead relies on relative clauses.

Crawley et al. (1990) attempted to compare the parallel function theory to the subject preference theory and found evidence to support the subject preference theory. The experiment performed was an offline judgement task where participants had to assign the correct antecedent to a pronoun over a series of filler sentences. (See previous section for examples and a more in-depth explanation)

The research by Crawley et al. (1990) has been challenged since, primarily by Smyth (1994). Smyth (1994) analyzed the findings of Crawley et al. (1990) and even performed an experiment using the last lines of the critical items in the experiments Crawley et al. (1990) performed. So for the example we listed earlier from the Crawley et al. (1990) paper, (7), Smyth (1994) only presented a stimulus like (12). Participants were then asked to fill in the blank with the correct antecedent.

(12) Tommy told Kevin to find the new bikes and Samantha shouted at him. Samantha shouted at _____

Smyth (1994) found that most items from Crawley et al. (1990) did not actually follow a strict parallel structure, claiming that a mere 4 out of the original 40 sentences were properly parallel. The difference in performance between these specific 4 sentences and the rest of the material was significant, only the truly parallel sentences followed the predictions of a parallel function preference strongly.

He then continued to propose an Extended Feature Match hypothesis to explain both his own findings as well as those of Crawley et al. (1990). His

theory states that assigning the correct antecedent is based on a search process that looks at the suitability of every prior NP based on its own properties like gender and number as well as the role within the sentence. As an example he gave the two sentences listed below (13).

(13) a. The carpenter gave the plumber an invoice, and the electrician gave him a cheque.

b. The carpenter invoiced the plumber, and the electrician gave him a cheque.

Here there are three possible targets for the pronoun to attach to. The carpenter, the plumber and the electrician. In both sentences the electrician can be excluded easily due to the form of the pronoun. If the electrician had been meant then it should have read '*himself*'. This leaves the carpenter and the plumber. According to Smyth (1994) only (13a) has a clear solution because the two clauses have a parallel form. Here '*him*' refers back to the plumber since he also received something in the first clause. For (13b) he makes no prediction.

Smyth (1994)'s goal of finding a way to explain both Crawley et al. (1990) and his own data in one unifying theory was accomplished using this. As he explained, there is a degree of parallelism between the two clauses that determines to which extent a parallel function is sought out by the reader of a sentence. If there are no, or very few, similarities between clauses, as was the case in Crawley et al. (1990)'s material, there will be no preference for an antecedent in a similar position and a subject resolution might be preferable. However, as the degree of similitude increases the reader is drawn to compare the two sentences more closely and an antecedent in the same grammatical role as the pronoun becomes preferable.

The theory by Smyth (1994) is not without its drawbacks though. The tested sentences followed a very strict form to allow for their specified degrees of parallelism to apply to the clauses, and it's not clear whether more general sentence interpretation would follow the same tendencies. It is a significant step up from having a preference purely on subject interpretations though, and is likely more similar to natural pronoun assignment. His experiment also only focused on the antecedent that was chosen though, and not on the ease with which an antecedent was found.

Furthermore, the theory does not actually explain everything and some researchers had already hinted that pronoun interpretation prferences might be even more complex.

Coherence

An alternate theory of pronoun interpretation is based on the coherence relations between sentences or sentence segments. One of the first papers to

formally discuss discourse coherence and pronoun interpretation was Hobbs (1979), who introduced it as a system to formalize sentences and the relations between them. It attempted to explain the goals of the speaker / writer, how they were being communicated and the influence of world knowledge on the way information is presented. His model was very theoretical and was originally not accompanied by experimental research to test his representation. However, he did define several types of coherence relations between sentences that were later tested and adopted into the model of coherence that we are employing here. These types are listed below, the listed examples are taken from Hobbs (1979):

Elaboration The second sentence is an explanation, supplement or further clarification of the first sentence. It always contains the same core information with an additional element that is not present in the first sentence.

(14) *Example*: "Go down Washington Street. Just follow Washington Street three blocks to Adams Street."

Parallel The second sentence has the same kind of structure and information the first sentence. Or, to put it in the words of Hobbs (1979), the propositions that follow from sentence one and two have identical predicates and similar arguments.

(15) *Example*: "Set the stack pointer to zero, and set link variable P to ROOT."

Contrast The second sentence is similar to the first sentence, in the same manner as the Parallel relation, with a single exception where the information is reversed. Or to rephrase it again, the propositions that follow from sentence one and two have similar elements except for one pair of elements that are contraries.

(16) *Example*: "You are not likely to hit the bull's eye, but you're more likely to hit the bull's eye than any other equal area."

This model is adapted almost 30 years later by, among others, Kertz, Kehler, and Elman (2006) as the basis of a model to specificlly explain pronoun interpretation. Hobbs (1979) does touch upon pronoun interpretation but sees it as an effect of discourse understanding. Kertz et al. (2006) distinguish two relevant coherence relations, the earlier mentioned Parallel relation and the Result relation, which describes a relation where the second sentence follows logically from the first sentence. (Example from Kertz et al. (2006): (17)) The experiment they performed contained items like (18) and they measured the response the participants gave to the questions.

(17) Dennis narrowly defeated Isaac, and Lilly congratulated him.

(18) Samuel threatened Justin with a knife, and he blindfolded Erin with a scarf.
Who blindfolded Erin?

They looked at the assignment rate of the antecedents split up by their grammatical role, starting out with all pronouns grouped together. In this case, the Subject Preference theory would predict a greater number of assignments to subject antecedents, but the data shows only a very limited effect, with only 52% of all pronouns being assigned to a subject antecedent. They showed a similar lack of effect for a split on the pronoun role, which the Parallel Function Preference would have predicted to have a greater number of assignments to antecedents fulfilling the same role. For subject pronouns the assignment to subject antecedents was a mere 51%, and for object pronouns the results were even contradicting the predictions from the parallel function preference, with only 48% of the assignments going to object antecedents. It was not until the data was split on the coherence between sentences that an effect emerged, with a preference of at least 90% in the predicted direction. Their precise findings are listed below in Figure 2.

These are impressive findings, and they definitely make a strong case for the coherence based interpretation of pronouns. However, there was already an extensive catalogue of research in favour of, among others, subject preference and parallel function preference. In order to strengthen the claim of coherence as an explanation a series of verification experiments was performed in a follow-up paper by Kehler et al. (2008). The first experiment addressed the earlier research by Smyth (1994) (see previous section) which found a strong impact of a parallel function preference, and Kehler et al. (2008) claimed that the observed results were caused by the modifications of Smyth (1994)'s sentences leading not just to the adjustment of the parallel relation, but also to changes in the coherence relation (from Occassion to Parallel). The experiment was constructed in such a way that coherence and syntactic parallelism were controlled for separately so that it could be determined what was the deciding factor in pronoun assignment. The experiment otherwise closely follows the design of experiment 3 in Smyth (1994) to allow for a direct comparison. This means that it was an offline pronoun assignment task with ambiguous sentences.

The analysis showed the predicted effects for coherence while showing no effect of the parallel function preference (or its expanded form, the qualified parallel structure preference, which is not covered further in this paper).

The second experiment investigated the effect of transfer verbs, like '*gave*' in sentence (19) which always include a 'source' argument which transfers the

	Antecedent Subj	Obj	n
Subject Preference			
all pronouns	0.52	0.48	512
Qualified Subject Preference			
non-biasing context	0.54	0.46	256
Parallel Structure Preference			
subject pronouns	0.51	0.49	256
object pronouns	0.52	0.48	256
Qualified Parallel Preference			
subject pronouns: fully parallel structure	0.52	0.48	128
object pronouns: fully parallel structure	0.50	0.50	128
Coherence Hypothesis			
subject pronouns: PARALLEL coherence	**0.98**	0.02	128
subject pronouns: RESULT coherence	0.05	**0.95**	128
object pronouns: PARALLEL coherence	0.10	**0.90**	128
object pronouns: RESULT coherence	**0.94**	0.06	128

Figure 2. Schematic overview of the findings by Kertz et al. (2006) showing that only the predictions for the coherence based model where observed.

object and a 'goal' argument which receives the object (Monika and Katie respectively in the example (19)). It featured a sentence completion task where participants were asked to finish sentences based on what they had read so far. The generated sentences were judged on what the intended antecedent for the provided pronoun was and what the coherence relation between the two sentences was. This judgement also included a third option, '*ambiguous*', for the situations where a pronoun was not sufficiently clear in referring to one of the two arguments. This is to test the theory that the end-state of a sentence (Katie having the pen in sentence (19)) is the deciding factor in how a pronoun referring back to it is interpreted since it is the goal argument. Previous research had shown a preference for interpretations where the goal argument returned in the new sentence. In a coherence-based explanation there will be a goal preference only in Occasion type coherence relations, and in a generally goal driven preference there would be no influence of the coherence between the two sentences. Looking at the data Kehler et al. (2008) found a strong influence of coherence, and no signs of a general goal driven preference, meaning the coherence theory again emerged as the more descriptive explanation.

(19) Monika gave a pen to Katie. She _____

The third experiment goes into the effects of implicit causality, but does not deal with pronouns directly so it is not discussed further. Online effects such as reading times are briefly touched upon and praised for being a potential source of further evidence, but they are not employed in any of the experiments.

Reading time / Online experiments

As we tried to show in the coverage of the previous research, the construction and verification of pronoun interpretation models has so far been explained almost exclusively using experimental designs measuring offline components. So a sentence completion task, or a judgement task, but not a reading time task. However, there is a lot of new information to be gained from an online experiment that measures the reading time of participants, information that would not be apparent in an offline design. For example, if a participant were to hesitate in a hypothetical offline pronoun judgement task then this hesitation would not be visible in the results (as long as the result of the judgement matched the predicted behaviour). A reading time measurement would clearly show any difference in speed for interpreting a sentence or judgement task.

The connection between reading speed and ease of processing is one that is covered in a lot of research, and the assumption is that faster reading of a particular segment can be interpreted as it being easier to comprehend or that it is more natural to the reader. Most papers do not further go into how or why this is the case. Witzel, Witzel, and Forster (2012) did go into this though, they compared various ways of measuring reading times in a series of experiments of which the predicted results were uncontroversial. This way they could verify to what extent each method would match the predicted results and based on that establish which method should be applied in order to answer a specific research question. They compared results from an eye tracker, a self paced reading task and two methods of their own design called G-Maze (for Grammaticality Maze) and L-Maze (for Lexicality Maze), an example can be found in figure 3. The two maze methods are similar to self paced reading tasks but offer the participant a binary choice at each read item where they have to determine which is the correct continuation of the sentence. In the L-Maze the choice is between a word and a non-word, and in the case of the G-Maze the participant has to choose between two words, but only one of them is a grammatical continuation of the sentence so far.

The methods were scored in two ways, the degree of natural reading that was still allowed and the accuracy with which the methods showed the predictions to the experiments. The experiments all dealt with ambiguity,

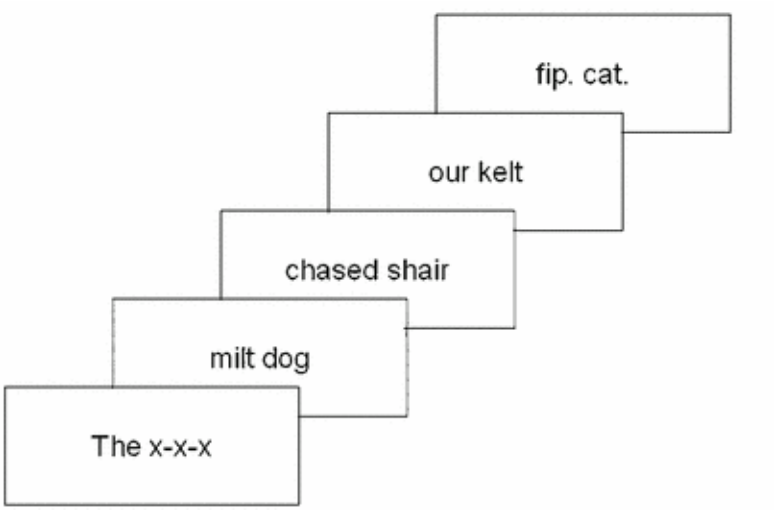

Figure 3. Example of an L-Maze taken from Witzel et al. (2012).

see (20) for an example. This is an example taken directly from their first experiment, which covered RC attachment ambiguity. In the first variant the person to shoot him- / herself is the actress, which is referred to as a low attachment (since it's the closer referent). In the second variant it is the son of the actress instead, which is considered a high attachment. The low attachment, so 'herself' in the example, should be easier to read and did indeed lead to lower reading times across methods.

(20) a. The son of the actress who shot *herself* on the set was under investigation.

b. The son of the actress who shot *himself* on the set was under investigation.

Their complete findings can be summarized as follows:

Eye-tracker Allows for the most natural reading, with no artificial restrictions placed on the users. It showed a strong effect (in the predicted direction) on each of the ambiguity experiments as well. The main problem with eye tracking experiments is that they allow for a lot of freedom in participant behaviour, so it can be hard to generalize between readings / participants since the amount of strategies that can be employed is so large.

Self-paced Reading Puts quite a few restrictions on natural reading, especially when compared to eye tracking experiments. Most pressing is that

participants can not jump forward or backward within a sentence, but instead have to progress through it linearly. This however comes with the added benefit that during the analysis each participant will have gone through the sentence in the same order, which makes for easier generalizations and conclusions. There is still the matter of different strategies being employed by different participants, just not the same extent as in eye-tracker experiments. The most relevant is that of pressing the button 'automatically', before the sentence so far has been fully comprehended. This leads to a spill-over effect or a hold-over effect. A spill-over effect shows a difference in reading time not just on the critical item but also on several items following it. A hold-over effect makes things even harder to analyze, with the effect being exclusively visible on items after the critical item. Witzel et al. (2012) have attempted to reduce the impact of this by removing participants with very monotonous reaction times, but found no effect on the size or placement of the effects. This would suggest that it is not necessarily just a strategy that can be employed by a participant but also something that will happen subconsciously over the course of an experiment.

G- and L-Mazes Here there are very strict restrictions put on the participant, they have to actively analyze the presented items and make a decision based on its form (though only based on whether or not it is a word in case of the L-Maze). This means that this forces the participants into a very defined procedure, making for easily generalizable data. It is, however, also the least like natural reading. The participant is constantly interrupted to perform this secondary judgement task and this severely limits the applicability of such Maze-type setups to experimental designs.

We ended up performing two self paced reading experiments. Based on Witzel et al. (2012) this should result in reasonably interpretable data (better than eye-tracking) while also still being sufficiently similar to natural reading (unlike the maze type experiments).

There are still general drawbacks to using reading times though. In a standard judgement task there is a very definitive set of possible results you can get. Participants pick from a set of possible options, and the proportions between these options are analyzed for the conditions and their interactions. For a reading time design this is a lot more complicated, since the conditions that are modified are spaced around the sentence, meaning that different modifications will lead to effects in different parts of the sentence. They also might not show an immediate response, instead needing several hundred milliseconds before they actively 'register' with a participant which could push the effect further down the sentence than the expected position through a spill-over or hold-over effect

Antecedent Role	Subject				Object			
Pronoun Role	Subject		Object		Subject		Object	
Coherence Relation	Result	Parallel	Res.	Par.	Res.	Par.	Res.	Par.
Subject Pref.	+	+	+	+				
Parallel Pref.	+	+					+	+
Coherence Theory		+						+

Table 1: Overview of three modifications that can be made in a sentence and the predictions each pronoun theory would make based on this. A '+' denotes a preference, but can also be interpreted as an increase in reading speed.

(explained in more detail above).

When we look at the existing material on each of the previously introduced pronoun interpretation theories we find the following predictions for each theory:

Subject Preference states that there is a preference for subject role antecedents, so the pronoun (or area directly following it) would be read faster if the antecedent is in subject position, regardless of the grammatical role of the pronoun itself.

Parallel Function Preference on the other has reading times around the pronoun decrease when the grammatical role of the pronoun is the same as that of the antecedent, since it claims this is the preferred form. This means that according to their predictions a subject pronoun is read faster when it is referring to a subject antecedent and an object pronoun is read faster when it is referring to an object antecedent.

Coherence Theory predicts an increase in speed when the grammatical role of pronoun and antecedent are the same *and* the coherence relation between the two sentence segments is Parallel. For a Result relation there should be an opposite reaction (mismatch = read faster) or no significant reaction.

Table 1 shows these predictions schematically, and attempts to also illustrate the gradual increase in complexity in the theories.

Now that we've covered these three theories and looked at what they would predict in an online measurement of reading speed we can look at a way to test for the validity of these theories, preferably in a single experiment for all three theories combined. We choose to focus on Coherence theory as most likely candidate of giving an accurate representation since the results produced by Kertz et al. (2006) were very convincing. We hope to find results supporting the predictions of Coherence theory, but will also look at the other theories and test their validity. We will attempt to follow the conditions from Table 1

in the construction of the experiments so that we can compare the results to these predictions. Relevant to note here is that when we mention a difference in match-mismatch we mean that the pronoun and its antecedent are either both fulfilling subject or object roles (a match) or in differing grammatical roles (a mismatch).

Experiment 1

Method

In this self-paced reading experiment, participants were presented with sentences of the form listed below in sentence (21). The experiment had a 2x2x2 design based on the different modifications presented in Table 1, and is illustrated with an example in (21a) to (21d) as well as one of the two possible starting segments. This means that each item had a total of 8 versions.

The self paced reading paradigm was designed by Just, Carpenter, Woolley, et al. (1982), who determined that it provided results comparable to natural reading speeds (which can for example be measured in eye tracking experiments). Because of this, and the discrete presentation of each single word which allows for easy data collection, the paradigm is widely used and accepted. The original paradigm showed a masked version of the complete sentence, with a single word becoming unmasked as the reader progresses through the sentence. This has later been named the moving window self paced reading task. In both of our experiments we will be using a more simple variant where a single word will be presented in the middle of the screen until the reader presses a button.

The different factors modify the following three aspects of each sentence. The level of resemblance between the two sentence segments is adjusted by the form of the first segment, which can be either in an active form, which is the same form as the second segment, or a passive form which is a mismatch with the second segment. The position of the pronoun of the segment also varies, differing between the subject and object position ((21a) & (21c) vs. (21b) & (21d)). The last modification is on the connective between sentence segments, which can be either 'en' ('*and*') or 'dus' ('*so*'), this effectively modifies the coherence relation between the sentences from parallel to result, respectively.

(21) Billy werd aangevallen door Ted,
Billy was attacked by Ted,
or
Ted viel Billy aan,
Ted attacked Billy,
a. dus hij verdedigde Sonya dapper.
so he defended Sonya bravely.

b. dus Sonya verdedigde hem dapper.
so Sonya defended him bravely.

c. en hij vloog Sonya aan in de avond.
and he went at Sonya in the evening.

d. en Sonya vloog hem aan in de avond.
and Sonya went at him in the evening.

Using these modifications we hope to see a significant impact of the various conditions and their interactions which would shed light on how well coherence holds up in a reading time task compared to the two alternative explanations we are considering (subject preference and parallel function preference). The goal of the experiment is to see whether a coherence based explanation is an accurate description of human pronoun interpretation. This means that the coherence relation between the clauses should be a critical element in the performance of the participants.

What we're expecting to see, if coherence is right, is a large difference in reading time in the Parallel coherence relation (connective: *en*'), where a match in the grammatical role of pronoun and antecedent (e.g. both are the object of their sentence) is read faster than a mismatch (e.g. subject antecedent, object pronoun).

Participants. The participants of the study were 32 native Dutch speakers (21 female, mean age = 30, range = 18 to 59) who had not participated in a similar study before and who did not receive any compensation for their participation.

Materials. The set contained 32 critical items with a corresponding question, each having 8 possible forms distributed fairly across 8 lists. Half the lists were in reverse order. Reverse ordered lists were used to prevent order effects. Each critical item was followed by a two-choice question asking the participant what they thought the correct antecedent was. In addition to the critical items there were also 28 filler items and 5 test items, all with questions as well, for a total of 65 items in each list.

The 28 filler items were sentences with a similar two-clause structure but without the same pronoun construction. The reason the filler items also had questions at the end of each item was to make them appear similar to the critical items. The 5 test items were non-ambiguous sentences, again of a similar structure, that served to verify whether each participant was paying sufficient attention to the task. The lack of ambiguity was caused by a forced decision due to the gender of the actors in the sentence. An example test sentence is listed below (22). The goal of this was to see if the participants were paying attention, and all test sentences were answered using the preferred antecedent

interpretation. Participants who made more than one error on test questions were excluded from the analysis.

(22) Linda ontving een brief van Harry, en Arend e-mailde haar een foto.
Linda got a letter from Harry, and Arend e-mailed her a picture.

Procedure. Participants were led to a quiet room and asked to turn off their phones and any other devices that could present a distraction. They were then given a brief explanation regarding the experiment that explained to them how a self paced reading task works, and that they will get a question about each sentence. They were not informed of the goal of the study until after the experiment was performed. Before the informed consent form was signed the experimenter pointed out specifically to the participants that their data would only be used anonymously and that they could stop at any time. The experiment was then started, and a brief explanation window appeared, showing how to progress the sentence and answer the question after each sentence. The space bar was used to progress the sentence, and the number 1 and number 2 keys to answer the question after each sentence. Since there was no correct or incorrect answer (for the critical items, which were ambiguous), no feedback was given to the participants after the questions.

After the 65 sentences and their corresponding questions had been presented, the experiment ended and the participant was thanked for his / her help and. When requested it was also explained what the study was about, and what parts of the participant's performance we'd be looking at.

Results

The average reading times per condition are portrayed below in Table 2. These are the reading times directly following the presentation of the pronoun, since this is where we are expecting to see a result. Effects can be observed, slower pronoun reading times are encountered when the antecedent is in subject position, when the connective is '*and*', and when the first segment of the sentence is passive. Relevant to our predictions, there appears to be a decrease in speed in the combined condition 'Match + En' and 'Match + Dus' (compared to 'Mismatch + En' and 'Mismatch + Dus' respectively).

Looking at this data using linear mixed effects models (Bates, 2007) we see a very different result though. Using a linear mixed effect model we can see which factors influence the behaviour of a participant. The upside to a linear mixed effect model compared to traditional approaches, like the ANOVA, is that a linear mixed effect model can take the random effects of both the participant and the critical item into account. This is something for which the ANOVA requires two separate analyses. A full comparison of mixed effect

models vs. ANOVA or t-test approaches goes beyond the scope of this paper, please refer to Baayen, Davidson, and Bates (2008) for an in-depth breakdown of the differences.

In constructing our model we started out with a complete set of all possible factors and their interactions. Then, using χ^2-analysis, we compared this model to a slightly simpler model to see if the modification had a significant impact on the model's predictions. Using this we systematically tested reductions of the original model to find the minimal model that still explained the data. This way there is the smallest number of degrees of freedom left, while still accurately portraying the behaviour of the original model. In Table 3 we show the original model chosen at the start of the analysis as well as the final model to which it was reduced. The factors that were considered for the model were the antecedent role, the pronoun role, the activeness of the first clause and the connective.

For the Coherence Theory it is important that the antecedent role that matches the pronoun role in a parallel coherence relation, which is not observed in the data ($\chi^2(4) = 1.31$; $p > 0.1$). When looking at the predicted outcomes for the Subject Preference (Subject antecedent = faster, $\chi^2(1) < 1$; $p > 0.1$) and Parallel Function Preference (same role antecedent & pronoun, $\chi^2(2) < 1$; $p > 0.1$) we also find no effect. The summary of pronoun reading times in Figure 4 further illustrates this. The only effects found were on the role of the pronoun itself, rather than its antecedent, as well as on the activeness of the first segment and coherence relation between segments. The other theories we covered (Subject Preference & Parallel Function Preference) both also predict an effect of the grammatical role of the antecedent.

Discussion

Looking at the reading time results in Table 2, as well as the outcome of the linear mixed effect model, it's apparent that the predictions made by the Coherence theory are not visible. A Match (subject antecedent with a subject pronoun or an object antecedent with an object pronoun) is read slower than a Mismatch not just in a Result coherence relation (which matches the prediction), but also in the Parallel condition (which explicitly contradicts it). It could be that Coherence theory is wrong, but when we look at the Subject Preference and Parallel Function Preference we see that these predictions are also contradicted (Subject antecedents and Match respectively read slower than their counterparts).

This means that either all research leading up to the three theories we cover is wrong, or something has muddled our data. We assumed that it is the latter, and went through both our stimuli and our results from the experiment extensively, looking for what might have caused these weird results.

	reading time	n
All pronouns	517.4 ms	1184
Pronoun Role		
Subject Position	599.6 ms	512
Object Position	454.8 ms	672
Antecedent Role		
Subject Position	530.8 ms	512
Object Position	504.4 ms	672
Connective		
En (*'and'*)	605.6 ms	512
Dus (*'so'*)	450.2 ms	672
Activeness		
Active first segment	482.0 ms	672
Passive first segment	563.8 ms	512
Activeness + Connective		
En + Active	386.1 ms	416
Dus + Active	637.9 ms	256
En + Passive	554.3 ms	256
Dus + Passive	573.3 ms	256
Grammatical Role Match - Pronoun & Antecedent		
Match	545.1 ms	572
Mismatch	494.2 ms	594
Role match + Connective		
Match + En	469.4 ms	314
Mismatch + En	435.6 ms	346
Match + Dus	637.2 ms	258
Mismatch + Dus	575.8 ms	248

Table 2: Overview of the average standard reading times per condition or combination of conditions.

Starting model	logRTs ∼ PronounRole * Connective * Activeness * AntecedentRole + (1|Item.) + (1|Participant.)
Final model	logRTs ∼ PronounRole * Connective * Activeness + (1|Item.) + (1|Participant.)

Table 3: The two linear mixed effect models showing the analysis performed on the data. The starting model includes all factors measured during the experiment, the final model only those aspects that offered a significant benefit to the model.

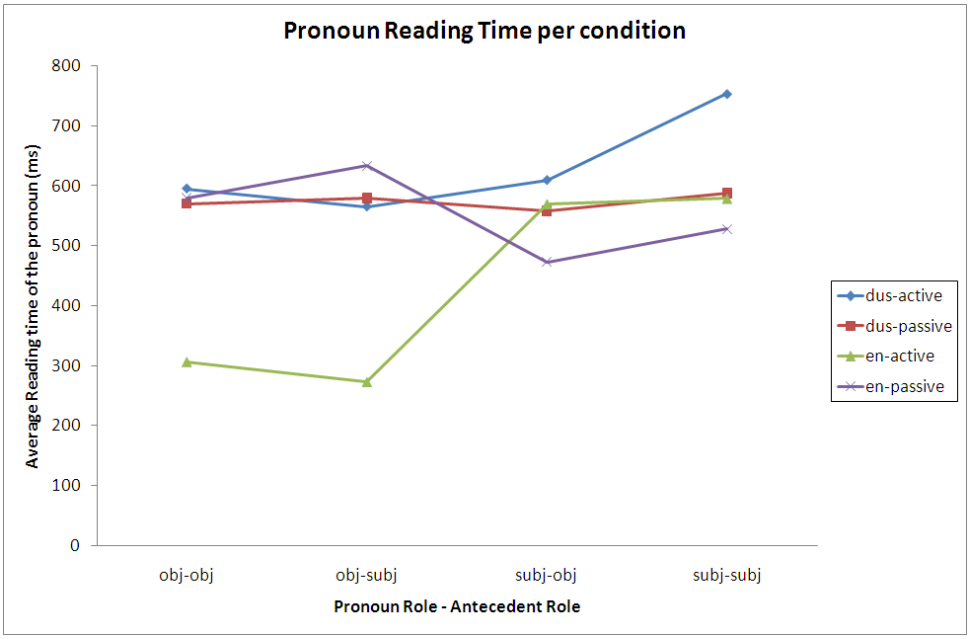

Figure 4. Average reading time of the pronoun per participant, split up on the grammatical role of the pronoun, that of the antecedent (determined by the reply of the participant, since sentences were ambiguous), the activeness of the sentence and the type of connective.

In doing so we discovered several aspects of the materials were such that they could have biased the results. First, consider the passive voice in the first segment of half the items. We didn't consider the reading times of the first segment at all in the analysis since the effects we'd be interested in would only show up around and following the position of the pronoun. What we did not account for is that the passive voice in the first segment would spill over considerably into the remainder of the sentence. The choice of a passive for the first segment was made in order to swith the subject and object roles, a modification that frequently appears in offline studies as well, but its impact on the reading times far exceeded what we anticipated, being consistently slower when the starting segment was passive. Another problem becomes apparent when you look more specifically at the subject preference theory. In the majority of existing research related to subject preference the antecedent can be either a subject or an object. However quite a few of the sentences in Experiment 1 were transfer of possession verbs similar to (23). For a full list of the critical items used in Experiment 1, including an overview of which sentences were problematic in this manner, please refer to the appendix.

(23) Margory threw a ball to Eve, and she...

Even though '*ball*' is not a valid antecedent for the pronoun it still has to be processed and considered on some level. The subject preference theory would still predict Margory to be the antecedent for the pronoun. The change of posession of the ball could for example also shift the attention of the users from Margory to Eve. This is something that has been shown to be the case in offline studies (Kehler et al., 2008) so it likely interfered with the results as well.

Based on these concerns we constructed a new item set. The new material we created always had a subject and object (and no further possible antecedents or objects that could interfere) in the first segment to make it as comparable as possible to the type of material earlier research used. The passive form was also cut from the design, which did mean that we needed a different way to lower the resemblance of the two sentence segments to properly investigate parallel function preference. To accomplish this we decided to manipulate the gender of the pronoun instead. With a differing gender for the subject and object of the first segment a manipulation of the pronoun's gender would enforce either a subject or object binding, thus altering the degree of similarity so the parallel function preference would predict different results depending on gender. The original modification to investigate coherence, based on altering the connective, was investigated as well but showed no problems.

Experiment 2

Method

For the second experiment a completely new self-paced reading task was constructed that was similar to that in Experiment 1 but incorporating the lessons we took from examining the shortcomings of Experiment 1 as well as using a more controlled approach for constructing the materials. An example sentence in all its forms can be seen below at (24). Similar to Experiment 1 there are 3 binary conditions resulting in 2x2x2 design with a total of 8 different versions for each item. Unlike the first experiment, however, all of the differences lie beyond (or at) the connective. This ensures that regardless of the form in which the participant is presented with the sentence he / she always gets the same first sentence segment.

The first factor adjusts the role of the pronoun, from the object to the subject of the second segment ((24a-24d) vs. (24e-24h) respectively). The second factor is the gender of the pronoun, which directly influences which prior character is its correct antecedent from an object antecedent (24a, 24b, 24e & 24f) to a subject antecedent (24c, 24d, 24g & 24h). The third factor modifies the coherence relation between the two sentences through adjustment of the

connective. From a Parallel coherence type (24a, 24c, 24e & 24g) to a Result coherence type (24b, 24d, 24f & 24f).

(24) Julia besloop Kevin,
Julia snuck up on Kevin,

a. en Jane bespioneerde hem stiekem vanaf de derde verdieping.
and Jane spied on him from the third floor.

b. dus Jane bespioneerde hem stiekem vanaf de derde verdieping.
so Jane spied on him from the third floor.

c. en Jane bespioneerde haar stiekem vanaf de derde verdieping.
and Jane spied on her from the third floor.

d. dus Jane bespioneerde haar stiekem vanaf de derde verdieping.
so Jane spied on her from the third floor.

e. en hij bespioneerde Jane stiekem vanaf de derde verdieping.
and he spied on Jane from the third floor.

f. dus hij bespioneerde Jane stiekem vanaf de derde verdieping.
so he spied on Jane from the third floor.

g. en zij bespioneerde Jane stiekem vanaf de derde verdieping.
and she spied on Jane from the third floor.

h. dus zij bespioneerde Jane stiekem vanaf de derde verdieping.
so she spied on Jane from the third floor.

The goal of the experiment is to use online reading time data to compare a Coherence based pronoun interpretation theory to its two main contenders, the Subject Preference and the Parallel Function Preference. By incorporating a condition for each of the pronoun interpretation theories that modifies a critical element we should be able to determine which theory (or theories) is supported by the reading time data.

As we already illustrated earlier in Table 1, the predicted outcome of such an experiment would be, per theory:

Subject Preference Faster reading of pronouns with a subject antecedent.

Parallel Function Preference Faster reading of pronouns that have the same grammatical role as their antecedent.

Coherence Faster reading of pronouns that have the same grammatical role as their antecedent when the coherence relation is Parallel, or a different role when it is Result.

Participants. The participants of the study were 28 native Dutch speakers (8 female, mean age = 25, range = 20 to 58) who had not participated in a similar study before. 27 participants were students and employees of the University of Groningen and one additional participant was recruited through an online forum and paid 5 €.

Materials. We created 32 critical items, along with 32 filler items for a total of 64 items. A list of names was constructed to serve as a basis for the critical items, each critical item would need a male and female name for the first sentence segment as well as a third name of arbitrary gender to fill the non-pronoun role in the second segment. For the gender-specific names it was critical that they were unambiguously male or female and clearly recognizable as Dutch first names. To ensure this all names were matched up with census data to determine both how common and how gender-directional each name was. Of the final items used all gender directional names in the critical items were at least 99.9% gender-biased. The length of the names was also controlled for, with each name requiring to be exactly two syllables in length, as well as a roughly equally spread number of names per starting letter. Additionally the final pairs were randomly distributed but ensured that the male and female names never had the same starting letter. For the third name in the sentence the gender was irrelevant since this character would be interacting with the pronoun so it could not be a valid antecedent. However, we still tried to balance these names between genders as well. This was made more difficult by the length restriction we set for ourselves here though, so the final ratio was roughly 2 male names for every female name in the third position. As a length restriction for the third name we chose to only use single-syllable names to rule out any effects that the length difference the pronoun might have compared to the name. All names used were also screened manually by several native speakers to determine whether they were valid and clear names as well as to test whether the gender was clear.

Using these names the critical sentences were constructed using the form explained below in (25). The additional padding of the sentence after the second verb phrase is to ensure that the entirety of any effect can be observed which generally does not (just) happen on the critical chunk but one or several chunks later, resulting in a spill-over effect that needs to be accounted for. The two verbs were chosen to generally have a categorical relationship to each other to increase the resemblance between the two segments. For example in (24) the verbs '*to spy*' and '*to sneak*' share a similar theme.

(25) [Name 1] [verb 1] [Name 2], [connective] [pronoun/Name 3] [verb 2] [Name 3/pronoun] [5 word long tail]

The filler items were constructed with the goal in mind to have sentences that would be as dissimilar as possible while having roughly the same complexity and exactly the same number of words as the critical items. To do this we constructed sentences with two to three actors, all with names that did not yet appear in the critical items, in various constructions and forms. An example is given below in (26).

(26) Lars is boos op Karin, maar Miranda had zijn avondeten stiekem opgegeten.
Lars is angry with Karin, but Miranda was secretly the one that had eaten his dinner.

Since there was no longer any ambiguity in the critical items there was nothing interesting to be gained from asking a comprehension question after each item. However, with no questions there would have been no incentive for the participants to keep reading the sentences properly. To ensure that participants could not rush through the sentences we added questions after 10 of the filler items. To illustrate, for the example filler item above (26) the question was whether or not Michael was angry with Lars.

Procedure. Participants were explained the basic premise of the experiment and asked to sign a consent form. At the start of the experiment a screen giving an explanation appeared. Users were given an example of sentence masking, and were explained how to answer the questions. After this explanation a set of three practice items followed, followed by further explanation to stress that every sentence had to be read carefully. Then the experiment would start. For the items that had questions participants were given feedback on whether their answer was correct. In case of a wrong answer they were also asked to slow down and read more carefully. A short break was added halfway through the items to limit the impact of fatigue and demotivation on the later items.

Results

Reading time data was analyzed using R (R Development Core Team, 2008) with linear mixed effects models (Bates, 2007). The three theories that are being discussed were analyzed to create complete models with all factors relevant to that theory included that were then reduced by removing parts and performing a χ^2-analysis on the model with and without each part to see if the removed part made a relevant contribution to the model. The final models are presented in Table 4.

Before we started to construct these models we first cleaned up the data. This was done in two steps. For the first step we looked at the replies to

Subject Preference	
Final model	logRTs ~ PronounType + Antecedent + logRT.prev + Trial + (1|Item.ID) + (1|Participant.)

General Explanation	
Final model	logRTs ~ PronounType + Antecedent + logRT.prev + Trial + logRT.prev:Trial + (1|Item.ID) + (1|Participant.)

Table 4: Overview of the two different linear mixed effect models per theory. There were two starting points, one for the Subject Preference theory (which did not care about connective, etc.) and a general theory which takes all factors into account. The final models shown were reached by reducing non-critical elements from the full starting models using Chi-square analyses.

the control questions on the filler items. Only participants with less than 4 errors were included in the final data set, which resulted in the exclusion of two participants. The second step looked for outliers in the responses to the critical items. Any response time that was more than 2 standard deviations faster or slower than the average would have been excluded (both standard deviation and average were calculated per participant), however all response times fell within this range.

It can be hard to interpret the meaning of linear mixed effect models, so in order to illustrate the difference between conditions and their directions the reading times are presented in Table 5. For a general overview showing these effects I also included reading time graphs for the complete sentence split up per word and the relevant conditions per factor (5, 6 & 7). When we do compare these models, using another χ^2-analysis, we find that the second model, which took into account all factors rather than just the ones related to subject preference, is the best fit for the data. Both models were also compared to null-models to ensure general significance, which both had. The details of the winning model are summarized in Table 6, showing an impact of the pronoun type (slower when subject) and the reading time of the previous word (slower when longer). There is also a small impact of the trial index (faster when later in the experiment).

Discussion

From the reading time table we can initially see that a Match in grammatical role pronoun and antecedent speeds up readers when the two sentence segments have a Parallel coherence relation and slows down when the segments have a Result coherence relation. This exactly fits the predictions made by coherence theory. The other two theories performed less favourably

	reading time	n
All pronouns	578.4 ms	768
Subject Preference		
Subject Antecedents	613.7 ms	384
Object Antecedents	543.2 ms	384
Parallel Function Preference		
Subject Ant. + Subject Pronouns	507.9 ms	192
Subject Ant. + Object Pronouns	719.4 ms	192
Object Ant. + Subject Pronouns	448.1 ms	192
Object Ant. + Object Pronouns	638.4 ms	192
Match*	573.1 ms	384
Mis-match*	583.8 ms	384
Coherence Based Preference		
S.A. + S.P. + Parallel	465.8 ms	96
S.A. + S.P. + Causal	550.0 ms	96
S.A. + O.P. + Parallel	785.4 ms	96
S.A. + O.P. + Causal	653.5 ms	96
O.A. + S.P. + Parallel	432.5 ms	96
O.A. + S.P. + Causal	463.7 ms	96
O.A. + O.P. + Parallel	588.5 ms	96
O.A. + O.P. + Causal	688.3 ms	96
Match* Parallel	527.2 ms	192
Match* Causal	619.2 ms	192
Mis-match* Parallel	609.0 ms	192
Mis-match* Causal	558.6 ms	192

Table 5: Overview of the average standard reading times per condition or combination of conditions. * Match and Mis-match are defined as the Antecedent and Pronoun role being the same (object-object / subject-subject) or different respectively.

Effect	Log RT Impact	Std. Error	t value
(Intercept)	5.111	0.224	22.78
PronounType (*Object: 1, Subject: -1*)	0.101	0.015	6.81
logRT.prev	0.171	0.035	4.93
Trial	-0.003	0.0008	-4.52

Table 6: Impact of the model's various effects on the logarithmic reading times of the winning linear effects model. (See Table 4 for the formula.)

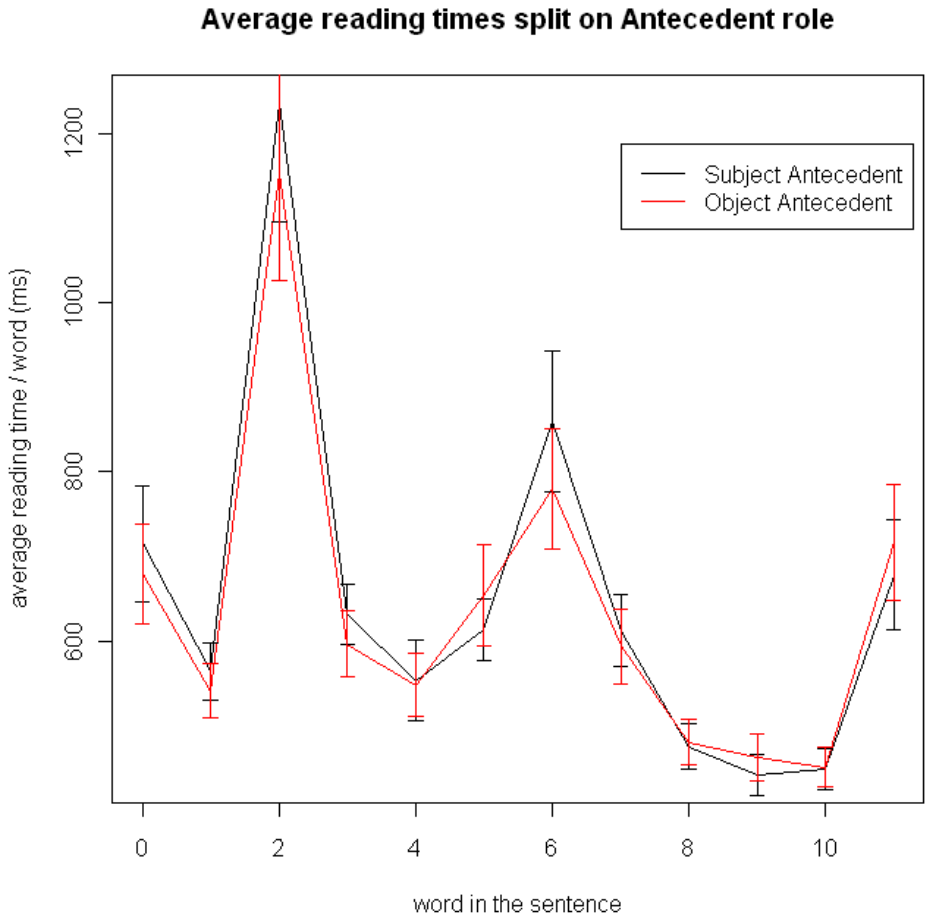

Figure 5. Average reading time over the sentence per participant, split up on the grammatical role of the antecedent. Subject preference predicts faster reading times for subject antecedents.

in this first look, with object antecedents having faster read pronouns than subject antecedents, contradicting the predictions from Subject preference, and Matches generally being ready only marginally faster than Mismatch, which Parallel Function predicts would be a strong effect.

However, when looking at the results of the Linear Mixed Effect Models we find that the picture isn't quite as simple (see Table 6). None of the three theories appear to be represented by the data. Pronouns referring to a subject antecedent are read slower than those referring to an object antecedent ($\chi^2(1) =$

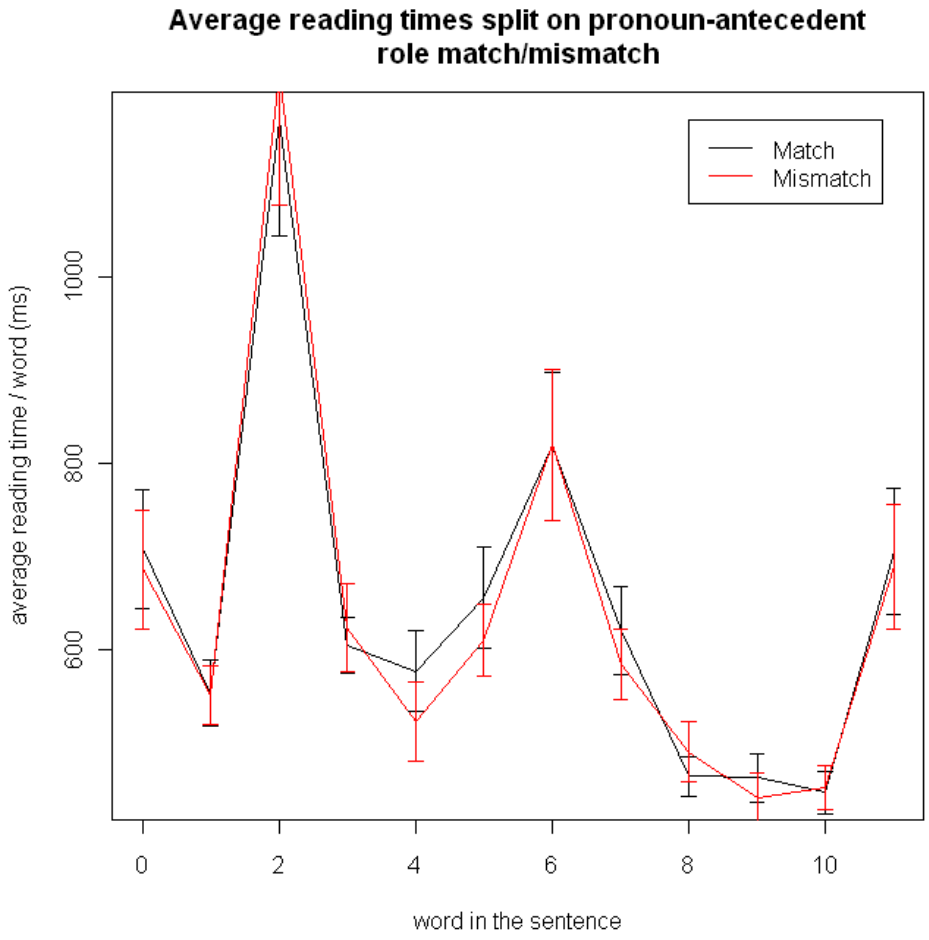

Figure 6. Average reading time over the sentence per participant, split up on whether the grammatical role of the pronoun and the antecedent are the same (match) or different (mismatch). Parallel function preference predicts faster reading times when there is a match.

7.29; $p < 0.01$), which goes against the Subject Preference Theory. The Parallel Function Preference Theory predicts an interaction between the grammatical role of pronoun and its antecedent, but there is a complete absence of such an effect ($\chi^2(1) < 1$; $p > 0.1$). This leaves the Coherence based explanation, but this relies on a critical role of the Coherence relation, here modified by the connective, but the connective was eliminated from the linear mixed effect model as a non-critical element, so this also discounts that theory. Again, none

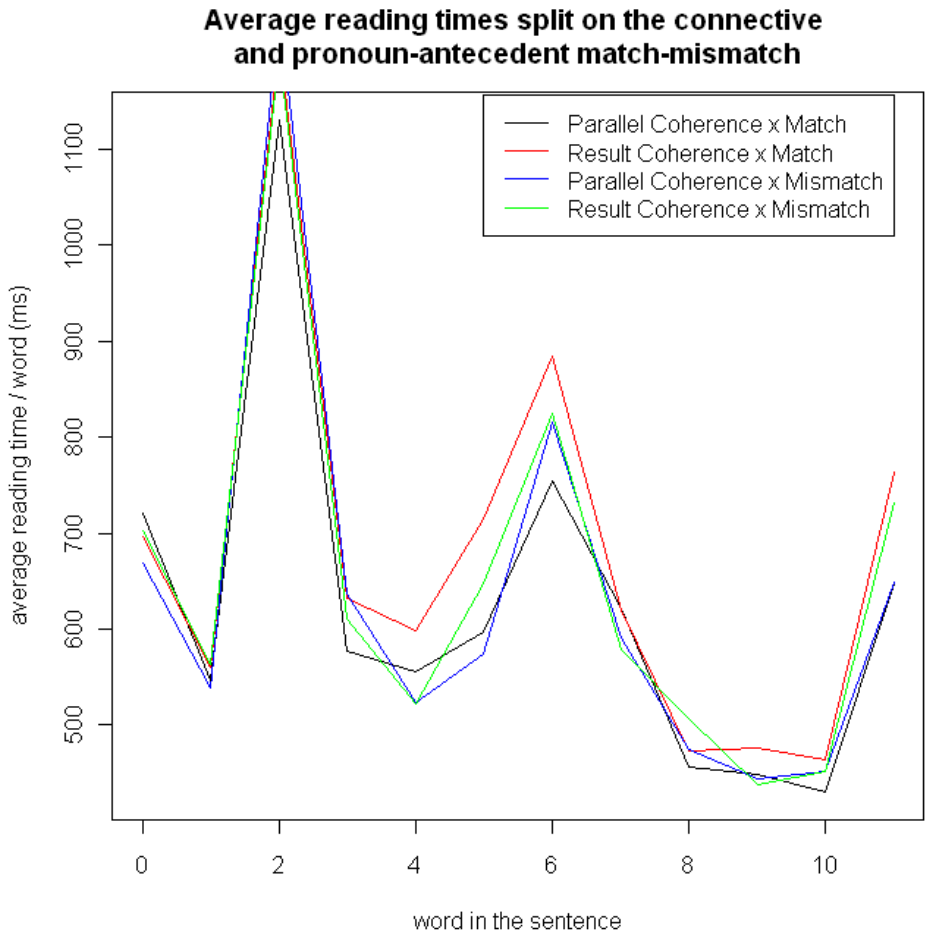

Figure 7. Average reading time over the sentence per participant, split up on the type of coherence and the match or mismatch of the grammatical role of pronoun and antecedent. Coherence theory predicts match to be read faster than mismatch in the parallel coherence relationship.

of the predictions of any of the three theories matched the observed results.

General Discussion

Neither of our two experiments produced the results we expected, which would have been matched predictions for the coherence based preference or at the very least the Subject Preference or Parallel Function Preference. It's important to look at how this can be, so we'll retrace our steps in getting to our

data and then propose possible explanations for these peculiar results.

Pronoun interpretation has been researched extensively over the past decades, and there are multiple theories attempting to explain which of the possible referents is considered 'best' by a human reader. Subject Preference (Hobbs, 1976) is based originally on findings in corpora, and states that an antecedent in subject role is the prefered referent for pronouns. Parallel function preference (Sheldon, 1974) proposes a more complex preference and was originally formalized to explain the behaviour of child participants in a comprehension experiment. It describes a referent preference for antecedents that fulfill the same grammatical role as the pronoun. So a subject pronoun would preferably be interpreted to refer to a subject antecedent, and an object pronoun to an object antecedent. Coherence theory (Hobbs, 1979) is a system that tries to formalize the type of relation between sentences. The different coherence relations specified in it were used decades later to explain different behaviour in pronoun interpretation (Kertz et al., 2006). Specifically, they showed that a Parallel coherence relation correlated with a preference for grammatical role symmetry between pronoun and antecedent and a Result coherence relation correlated with a mismatch of grammatical role.

Most of the research we covered has been done using offline judgement tasks like sentence completion or the answering of a question about each sentence. We felt that this does not give all information that might be relevant, and proposed an online experiment that looks at the reading times of participants when faced with a pronoun that they need to interpret.

An online experiment is not as easy to interpret as an offline task, since the amount of data produced by each item is considerably higher and the method of selecting meaningful measures from such a data set can already be grounds for debate. In an offline judgement task you get a small set of results, usually one data point per critical item per participant, but in a reading time experiment you get a set of reading time data and how to go about analyzing it is no trivial matter. Reading times themselves can be kept as-is, be normalized to eliminate the per-participant variants or be taken the logarithm of. Deciding which areas to look at is also tricky, since the entire sentence as a whole is unlikely to produce any significant effects it is important to look at the effects that are expecting and especially where you expect these effects to occur.

Earlier in the paper we already briefly touched on various types of experiments that measure reading times (Witzel et al., 2012). The fact that there is so much variety in methodology can be problematic when comparing different experiments, since each approach puts different restrictions on the participants which could in itself already influence behaviour. For example, one of the main differences in participant behaviour between eye tracker studies and self paced reading tasks is that the self paced reading task does not permit

backtracking (or jumping forward) within the sentence. When comparing two experiments, each using a different method, it would not be trivial to see what might be the cause of potential differences. By comparison, you have studies like those of Crawley et al. (1990) & Smyth (1994) that have research questions that challenge existing results and determining which pronoun interpretation theory is a better fit. These experiments closely resemble their precursors and attempt to overthrow the results by stricter control of the conditions, much like what we did from experiment 1 to experiment 2. Comparing data from, for example, Kertz et al. (2006) to our own data is a lot harder since the measures dimensions aren't the same.

It is possible that the prevalence of offline experiments is in part because of this comparability it offers to existing research. This view runs the risk of conforming too much to existing ideas, which I feel is slowing down the acquisition of understanding in the field. When looking at the progression of pronoun interpretation theories (subject preference, then parallel function preference, then coherence theory) you can see that each theory takes the core assumptions of the previous theory and refines it in an attempt to give a better interpretation of human preference. Each of these following theories was originally conceived due to information that did not follow directly in the wake of the research done on its preceding theory. The foundations of the parallel function preference were compiled based on a toy manipulation task in children (Sheldon, 1974), and coherence theory has existed long before it was applied directly to pronoun interpretation (Hobbs, 1979). Perhaps what is needed in order to take the next step in understanding is to look at sentence processing and pronoun interpretation behaviour in a more detailed manner, such as by using a self paced reading task.

In our first experiment we closely followed the design of Kertz et al. (2006). They als had ambiguous sentences consisting of two segments joined by a connective where participants had to determine which antecedent was the correct fit. This was to preserve the comparability between our results and those of Kertz et al. (2006). In hindsight our hope that the two results would be comparable, despite the different types of measurement may have been naive. When we looked at the reading times around the pronoun using linear mixed effects models we found no confirmation on any of our predictions, neither those of the coherence theory nor those of either of its predecessors.

Writing this off as a faulty experimental design we set up a second experiment with a very strictly constructed set of materials. It no longer featured ambiguity, passive sentence segments or questions and proved to paint a far more descriptive picture of online pronoun interpretation behaviour. None of the three factors predicted to have an impact on pronoun interpretation showed a significant effect.

A possible explanation for the lack of supporting evidence for the Coherence theory is that our sentence modification that was supposed to alter the coherence relationship (connective, 'en' ('*and*') or 'dus' ('*so*')) did not correctly change the sentence. This is a phenomenon that was also observed by Frazier and Clifton (2006). In their first two experiments they did not find results in line with predictions based on the Coherence theory. They decided they had probably not found anything because the changes to their critical items meant to modify the coherence relation were not strong enough. A similar problem might have occurred in our sentences, where the parallel connective 'en' ('*and*') might be interpreted in other ways (with differing coherence relations). This is because of the way the stimuli were created. They were designed to be easily switchable between coherence relations, just by switching out the connective. This could have resulted in peculiar or unnatural sentences that would have led to effects not directly related to the coherence relation. It's interesting to note that these were also experiments involving an online reading time component, and it might be possible that the lack of data matching the predictions from their experiment has the same underlying reason as that of our experiments.

Even if the above drawbacks of our experiments were resolved and a follow-up experiment would be performed I feel that the results probably still would not match those of the original experiments. There of course cases imaginable where the results will clearly be in the same direction. For example, in our first experiment we found that passive sentences were read consistently slower than active sentences. For example, if you were to take experimental results from experiments with eye trackers or even offline judgement tasks that asked to rate sentences based on how "easy" or "nice" they are the active sentences would win to the passive sentences. This has been shown in earlier research as well.

Pronoun interpretation is not a clear-cut problem though. Everybody that has done research into it so far, including myself, will admit that all the theories so far are merely there to indicate a preference. But even looking at elements on which all three theories would agree, for example a subject pronoun with a subject antecedent where the segments are in a coherence parallel relation being easier to interpret than a subject pronoun with an object antecedent in the same relation, it is still possible to acquire data that does not support this. Table 5 shows an average of 465.8 ms for this specific form (parallel - subject pronoun) with a subject antecedent compared to 432.5 ms for the same form with an object antecedent, meaning the observed difference is in the opposite direction of what any of the existing theories would predict.

Future Research

Our findings do not support the covered theories for pronoun interpretation. However, we do not mean to propose that this implies none of these theories are correct. Our experimental design allowed for a more controlled and objective measure of human interpretation, and we had hoped to find resuts more in line with the existing material. Especially the results from Kertz et al. (2006) and Kehler et al. (2008) made it surprising to observe the absence of a similar effect.

There are multiple possible reasons why the Coherence theory did not predict the behaviour as accurately as we expected. It could be that the definitions of the various coherence relations need to be formalized in such a way that our material would not have qualified as these two particular coherence relations. Another option could be that reading behaviour for Dutch is relevantly different to that for English which might have made our results harder to compare as well. Lastly, and I feel, one of the most important reasons and one that has so far been neglected, the difference in methodologies. In an offline study the only thing that is measured is the participant's final, conscious response. An online study, by comparison, gives a slew of data that can be interpreted in a variety of ways. We currently don't know if fluctuations in reading time are making an impact on the final opinion of the participant, so a difference in effects (even if the direction is completely opposite) might not mean that a theory is wrong, but rather that we do not yet understand how to compare the results from two different methodologies. Online tasks give a detailed look at the subconscious processing of a sentence, and there is definitely an impact to be seen here. I feel that further experiments looking at the timing, size and direction of reading time effects would be a huge help in figuring out the factors involved in correctly modelling how we decide on an antecedent.

In order to continue testing existing theories, and maybe even come up with an extension or alternative, I feel that the experiments will have to diversify. Within offline methodologies there is still a lot of research on the competing theories that has not had an equivalent study done for Coherence to see if it has a superior model for the observed behaviour. For online tasks I feel it would be beneficial if more methodologies would be explored and if we try to find a way to compare the data found in such experiments to those using offline methodologies. Witzel et al. (2012) has already shown us that different online methodologies can produce large differences in results, so it is fully possible that even more differences appear when comparing online and offline tasks. I am not sure whether pronoun interpretation is a good topic to focus such studies on, Witzel et al. (2012) intentionally performed their analysis on an easy subject so that the results of each test would not be controversial and they could focus

on the shape of the result. Regardless of whether or not pronoun interpretation would serve as the subject of such trials, I feel that it could benefit from such trials greatly. Right now our knowledge in cross-methodology comparisons is so limited, but with more research we would be able to determine which types of tests are sensitive to which kinds of modifications and effects. Of the three competing theories we have covered in our paper we feel that coherence theory is definitely the strongest contender. The lack of support for it in an online methodology was surpring, but shows us that there is so much more that we can investigate.

References

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390 - 412. Available from http://www.sciencedirect.com/science/article/pii/S0749596X07001398 (<ce:title>Special Issue: Emerging Data Analysis</ce:title>)

Bates, D. (2007). lme4: Linear mixed-effects models using s4 classes [Computer software manual]. (R package version 0.99875-9)

Crawley, R. A., Stevenson, R. J., & Kleinman, D. (1990). The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research*, *19*, 245-264.

Frazier, L., & Clifton, J., Charles. (2006). Ellipsis and discourse coherence. *Linguistics and Philosophy*, *29*(3), 315-346. Available from http://dx.doi.org/10.1007/s10988-006-0002-3

Frederiksen, J. R. (1981). Understanding anaphora: Rules used by readers in assigning pronominal referents. *Discourse Processes*, 323–347.

Hobbs, J. R. (1976). *Pronoun resolution* (Tech. Rep. No. 76-1). City University of New York, Department of Computer Sciences, City College.

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science*, 67–90.

Just, M. A., Carpenter, P. A., Woolley, J. D., et al. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228–238.

Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2008). Coherence and coreference revisited. *Journal of Semantics*, *25*, 1–44.

Kertz, L., Kehler, A., & Elman, J. L. (2006). Grammatical and coherence-based factors in pronoun interpretation. In *Proceedings of the 28th annual conference of the cognitive science society* (pp. 1605–1610).

R Development Core Team. (2008). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from http://www.R-project.org (ISBN 3-900051-07-0)

Sheldon, A. (1974). The role of parallel function in the acquisition of relative clauses in english. *Journal of Verbal Learning and Verbal Behavior*, *13*(3), 272 - 281.

Smyth, R. (1994). Grammatical determinants of ambiguous pronoun resolution. *Journal of Psycholinguistic Research*, *23*, 197-229.

Witzel, N., Witzel, J., & Forster, K. (2012). Comparisons of online reading paradigms: Eye tracking, moving-window, and maze. *Journal of Psycholinguistic Research*, *41*(2), 105-128. Available from http://dx.doi.org/10.1007/s10936-011-9179-x

Appendix

Critical Items Experiment 1

Underlined verb phrases indicate a transfer of posession context or otherwise non-standard subject-object relationship, which might have distorted the resulting reading times.

1. Hilbrand gooide de bal naar Anne en zij moedigde Eric vrolijk aan.
2. Thomas meldde Fred aan en Laura schreef hem in op de ledenlijst.
3. Bart werd geslagen door Wouter dus hij filmde Charlotte nauwkeurig.
4. Jaap wond Thijs op dus hij kalmeerde Petra in de kamer.
5. Emma kreeg een naambordje van Lars en zij heette Berend welkom.
6. Vincent bedreigde Sven en hij intimideerde Eliska in het zwembad.
7. Linda ontving een brief van Harry en Arend e-mailde haar een foto.
8. Daan bekladde Steven en hij bevuilde Lisa met modder.
9. Andre stalkte Ted dus Iris beboette hem met een werkstraf.
10. Ruud werd geassisteerd door Joris en Matilde ondersteunde hem bij het experiment.
11. Piet werd misleid door Ruud en hij verraadde Lisa op school.
12. Moniek haalde een bos bloemen bij Johan en zij versierde Kevin subtiel.
13. Michiel ontmoedigde Andre en hij demotiveerde Iris de hele dag.
14. Jaap werd geweigerd door Sven en hij negeerde Fatima de volgende dag.
15. Michiel verleidde Bart dus hij weerstond Sonya keer op keer.
16. Piet werd gepakt door Bart dus hij bevrijdde Charlotte een tijdje later.
17. Willem bedroog Henk dus Jasmijn verraadde hem in het geheim.
18. Fred werd afgekraakt door Ruud en Matilde haalde hem neer met een schop.
19. Bart verliet Piet en Lucy wees hem af om persoonlijke redenen.
20. Daan duwde Thomas en hij tackelde Jasmijn meerdere keren.
21. Jack huldigde Piet dus hij erkende Diana als winnaar.
22. Willem nodigde Wouter uit en Petra vroeg hem mee mee naar huis.
23. Tim trapte Daan dus hij stopte Annabel meerdere keren.
24. Jasper betaalde salaris aan Marthe en Tobias kookte voor haar een heerlijk diner.
25. Guus mishandelde Billy dus Eliska verzorgde hem in de keuken.
26. Willem werd geholpen door Steven dus Ingrid stoorde hem door te kloppen.
27. Henk werd berispt door Guus en Marjolein vermaande hem in de stilteruimte.
28. Vincent werd getest door Thomas dus Laura keurde hem af wegens

astma.

29. Fred werd gefraudeerd door Jack dus hij ontsloeg Suzanne meteen vandaag.
30. Jaap werd gemist door Vincent en hij negeerde Diana de hele dag.
31. Jack liet Henk achter dus Fatima strafte hem heel hard.
32. Guus werd gearresteerd door Joris dus Annabel liet hem vrij uit de cel.
33. Max werd verdacht door Fred dus Lucy gaf hem aan bij de politie.
34. Tim ondervroeg Andre en Lydia onderzocht hem op eventuele ziektes.
35. Billy werd vastgebonden door Thijs dus hij maakte Amanda los van de stoel.
36. Max werd verslagen door Wouter en hij overwon Ingrid met badminton.
37. Billy werd aangevallen door Ted en Sonya vloog hem aan in de avond.

Critical Items Experiment 2

1. Oscar sloeg Susan, en Kim schopte hem erg hard tegen de schenen.
2. Julia besloop Kevin, en Floor bespioneerde hem stiekem vanaf de derde verdieping.
3. Pieter volgde Femke, en Beau liep hem achterna op weg naar school.
4. Iris versloeg Aaron, en Gwen verpletterde hem in de volgende ronde volkomen.
5. Martin bedreigde Karin, en Britt mepte hem met de krant van gisteren.
6. Daphne hielp Edwin, en Roos ondersteunde hem tijdens de volgende twee lessen.
7. Patrick bespotte Manon, en Maud ontmoedigde hem met gemene en persoonlijke opmerkingen.
8. Emma overtuigde Albert, en Claire praatte hem daarna om in de trein.
9. Berend e-mailde Lisa, en Noor belde hem tot diep in de nacht.
10. Maaike mistte Johan, en Joyce troostte hem met een lekker kopje koffie.
11. Stefan verveelde Petra, en Els negeerde hem de rest van de dag.
12. Hannah verstopte Richard, en Paul zocht hem tevergeefs door het hele gebouw.
13. Otto trakteerde Chantal, en John bedankte hem de volgende dag erg hartelijk.
14. Wendy ondervroeg Ruben, en Mike interviewde hem buiten onder de hoge boom.
15. Thomas verzorgde Agnes, en Bram bezocht hem later die week twee keer.
16. Ilse verwelkomde Robert, en Mark kuste hem als begroeting op de wang.
17. Casper alarmeerde Dorien, en Frank waarschuwde hem voordat het te laat was.

18. Marleen corrigeerde Jeroen, en Lars vermaakte hem om het gezellig te houden.

19. Michiel begeleidde Eva, en Sven bracht hem de volgende dag naar huis.

20. Nienke droeg Hendrik, en Nick kietelde hem om een beetje te plagen.

21. Willem verwarde Judith, en Niels misleidde hem zodat het allemaal mis ging.

22. Sandra groette Jacob, en Rolf zwaaide hem toe vanuit het open raam.

23. Maarten schilderde Esther, en Tom fotografeerde hem met een oude zwart-wit camera.

24. Lotte overhoorde Hugo, en Daan testte hem met een paar simpele vragen.

25. Lucas arresteerde Tessa, en Kees meldde hem bij de politie wegens diefstal.

26. Rosa geloofde Boris, en Henk vertrouwde hem genoeg om mee te komen.

27. Erik registreerde Anouk, en Frits schreef hem in op de nieuwe website.

28. Carmen duwde Wouter, en Roy ramde hem met zijn nieuwe mountain bike.

29. Sander telde Anna, en Thijs rekende hem mee in het totale aantal.

30. Linda stoorde Dennis, en Max onderbrak hem met een compleet ongerelateerde vraag.

31. Jeffrey bevrijdde Laura, en Bas trok hem weg van de verongelukte auto.

32. Sophie verwees Vincent, en Dirk stuurde hem door naar het andere gebouw.