CHILDREN'S PROCESSING OF ANAPHORA DURING READING COMPREHENSION

AN EYE TRACKING STUDY

Dissertation
zur Erlangung des akademischen Grades
Doktor der Philosophie (Dr. phil.)

Max-Planck-Institut für Bildungsforschung, Berlin

eingereicht bei der
Humanwissenschaftlichen Fakultät
der Universität Potsdam

von
Sarah Eilers, M.Ed.

Datum der Disputation:
1.11.2021

Die Arbeit wurde betreut von:

Prof. Dr. Sascha Schroeder, Universität Göttingen
(ehem. Max-Planck-Institut für Bildungsforschung, Berlin)

Prof. Dr. Reinhold Kliegl, Universität Potsdam

Erstgutachter: Prof. Dr. Sascha Schroeder

Zweitgutachterin: Prof. Dr. Holly Joseph, University of Reading (UK)

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-52714
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-527141

ZUSAMMENFASSUNG

Viele Kinder haben Schwierigkeiten, während des Lesens einen Textinhalt adäquat zu erfassen. Lesen ist eine komplexe kognitive Aufgabe, die verschiedene Unteraufgaben umfasst, darunter zum Beispiel das Dekodieren von Wörtern und das Verknüpfen mehrerer aufeinander folgender Sätze. Einen Teil dieser Verknüpfungen machen referenzielle Ausdrücke aus. Referenzen wie nominale Anaphern (Minky/*die Katze*) oder Pronomen (Minky/*sie*) signalisieren den Lesenden, wie die Protagonisten und Protagonistinnen in aufeinander folgenden Sätzen zusammenhängen. Die Lesenden verknüpfen diese Information in einem mentalen Modell des Textes, nachdem sie die Referenz aufgelöst haben. Besonders Personalpronomen (*er*/*sie*) können ohne einen solchen Auflösungsprozess nicht verstanden werden. Sie müssen mit einem passenden Antezedenten in Verbindung gebracht werden, oder das mentale Modell bleibt unvollständig. Gelungene Pronomenauflösung ist somit besonders bedeutsam für ein gutes Textverständnis. Die vorliegende Dissertation beschäftigt sich mit der Pronomenauflösung von Grundschulkindern im Alter von 8-9 Jahren und geht dabei der grundsätzlichen Frage nach, ob Kinder in diesem Alter Pronomen in natürlichen Lesesituationen spontan auflösen. Zudem wurde am Beispiel der Geschlechtsinformation des Pronomens untersucht, welchen Einfluss die Informationsdichte um die Pronomenregion auf die Blickbewegungen von Kindern hat. Dabei ging es auch um den Einfluss von Leseentwicklung und Lesefertigkeiten auf die Blickbewegungen beim Lesen, sowie auf das Verstehen eines Pronomens.

Die erste Studie untersuchte das Lesen kurzer Texte, die aus jeweils drei Sätzen bestanden. Der erste Satz führte einen Protagonisten mit Namen ein (*Mia*), auf den sich der zweite oder dritte Satz bezog, entweder mit einer Wiederholung des Namens (*Mia*) oder einem Pronomen (*sie*). Die Studie ging der Frage nach, ob Kinder bei solchen salienten Antezedenten ein Pronomen (*sie*) als referenziellen Ausdruck dem wiederholten Namen (*Mia*) vorziehen. In der Literatur zum Lesen Erwachsener ist dieser Befund als *repeated name penalty effect* bekannt: Der Lesefluss von geübten Lesenden wird durch die Wiederholung einer Nominalphrase deutlich beeinträchtigt. Für Kinder dagegen wurde die Hypothese aufgestellt, dass deren Lesefluss durch die Wiederholung verbessert werden könnte, und zwar aufgrund der sich überlappenden Wortform (*Mia – Mia*) die eine kognitiv aufwändige Auflösung des Pronomens (*Mia – sie*) überflüssig macht.

Die zweite Studie untersuchte die Verarbeitung von kongruenten und inkongruenten Geschlechtsinformation auf dem Pronomen. Die Kinder bekamen komplexe Sätze zu lesen, bei denen das Pronomen entweder passend zum Antezedenten gewählt war (*Mia – sie*) oder unpassend (*Mia – er*). Ergänzend wurden Leseverstehen und Leseflüssigkeit erhoben und mit der Fähigkeit der Kinder, spontan ein inkongruentes Pronomen während des Lesens zu erkennen, in Verbindung gebracht.

Die dritte Studie untersuchte die Blickbewegungen auf dem Pronomen in Abhängigkeit von variierender Geschlechtsinformationen genauer. Sie verglich den Lesefluss und das Leseverstehen von Kindern in Pronomenregionen, in denen das Pronomen anhand von der Geschlechtsinformation eindeutig einem Protagonisten zugeordnet werden kann (Peter und Paula…, *er…)* mit Lesesituationen, in denen der weitere Satzkontext zur Auflösung herangezogen werden muss (Peter und Paul, … *er*…). Dabei wurden die Blickbewegungen auf der Pronomenregion mit dem Leseverstehen, insbesondere dem Verstehen des Pronomens, in Verbindung gebracht. Dieses Experiment wurde im Sinne einer Longitudinalstudie in Klasse 3 und Klasse 4 mit der gleichen Gruppe von 70 Kindern durchgeführt.

Zusammengefasst belegen die Ergebnisse dieser Dissertation, dass Kinder im Alter zwischen 8 und 9 Jahren noch deutliche Schwierigkeiten mit dem Verstehen von Pronomen in Leseaufgaben haben. Die Antworten auf Verständnisfragen zum Pronomen zeigen insbesondere, dass Kinder die Kontextinformation in Sätzen nur unzureichend für die Pronomenauflösung nutzen, und dass ihr Verständnis eines Pronomens wesentlich davon abhängt, ob das Pronomen anhand der Geschlechtsinformation eindeutig einem Antezedenten zugewiesen werden kann. Dies zeigte sich bei Kindern im 3., aber auch noch im 4. Schuljahr.

Die Ergebnisse der Analyse von Blickbewegungsdaten, welche den wesentlichen Beitrag der vorliegenden Dissertation zum Forschungsfeld darstellen, zeigen zunächst, dass Kinder durchaus ein Pronomen erwarten, wenn der Antezent salient ist (Studie 1). Anders als angenommen gibt es keinen Beleg dafür, dass der kindliche Lesefluss von einer Wiederholung des Antezedenten profitiert. Der Befund eines *repeated name penalty effects* bei Kindern dieser Altersgruppe belegt im Gegenteil eine Sensitivität für die Diskursregeln, nach denen Pronomen auf saliente Antezedenten referieren. Allerdings kann daraus nicht abgeleitet werden, dass die Online-Pronomenauflösung von

Kindern mit denen von erwachsenen Lesenden vergleichbar ist. Die Ergebnisse der Analyse von Blickbewegungsdaten auf der Pronomenregion (Studien 2 und 3) belegen signifikante Unterschiede zwischen Kindern und Erwachsenen, sowie deutliche interindividuelle Unterschiede in Zusammenhang mit dem Leseverstehen und der Leseflüssigkeit der Kinder.

Die Ergebnisse der Studie 2 belegen einen Zusammenhang zwischen der Leseflüssigkeit eines Kindes und der Fähigkeit, eine Inkongruenz zwischen Pronomen und Antezedenten während des Lesens wahrzunehmen. Während alle Kinder längere *gaze durations* (erste Verweildauer) auf einem inkongruenten Pronomen hatten, zeigte sich nur bei Kindern mit hoher Leseflüssigkeit eine Tendenz zu regressiven Blickbewegungen aus der fraglichen Pronomenregion. Diese regressiven Blickbewegungen gelten als Signal für eine lokale Verarbeitungsschwierigkeit und werden als Versuch interpretiert, diese Schwierigkeit aufzulösen der zu „reparieren". Eine hohe Leseflüssigkeit war also korreliert mit dem Erkennen der Inkongruenz. Darüber hinaus war das Blickbewegungsmuster der Kinder, die die Inkongruenz erkannten, vergleichbarer mit dem der erwachsenen Kontrollgruppe. Die Befunde werden so interpretiert, dass Kinder mit einer höheren Leseflüssigkeit mehr kognitive Ressourcen für die Überwachung ihres eigenen Leseprozesses zur Verfügung stehen, und sie diese freien Ressourcen zur Pronomenauflösung auch in schwierigen Satzkontexten nutzen können.

Kinder unterscheiden sich stark in ihrem Leseverstehen, auch innerhalb einer Kohorte. Die Ergebnisse der vorliegenden Dissertation belegen, dass vorwiegend Kinder mit gutem Leseverstehen in der Lage sind, Pronomen während des Lesens anhand derer Geschlechtsinformation aufzulösen. Es lässt sich zeigen, dass sich gute Lesende nachweislich mehr Zeit in einer Pronomenregion nehmen, wenn das Pronomen anhand der Geschlechtsinformation direkt aufgelöst werden kann. Darin unterscheiden sie sich von schlechteren Lesenden, auch wenn diese insgesamt eine längere Lesedauer zeigen. Das Alter der Kinder war dabei weniger entscheidend als ihre individuelle Leistung im Leseverstehens- und Leseflüssigkeitstest. Zusammengefasst lässt sich sagen, dass gute Lesende unter den Kindern in der Lage sind, Pronomen während des Lesens spontan aufzulösen. Dabei ist das Leseverstehen ein entscheidender Faktor in beiden unter-

suchten Altersstufen. Für einen Einfluss des Alters der Kinder gab es dagegen kein Indiz.

Der Beitrag der vorliegenden Dissertation zum Forschungsfeld ist die Untersuchung und Darstellung der spezifischen Blickbewegungsmuster im Zusammenhang mit einer erfolgreichen Auflösung von Pronomen bei Kindern. Das Blickbewegungsverhalten in der Pronomenregion ist abhängig vom Leseverstehen und der Leseflüssigkeit der Kinder. Die vorliegenden Ergebnisse lassen vermuten, dass viele Kinder Pronomen während des Satzverstehens nicht spontan auflösen, was sich negativ auf ihr Leseverstehen auswirkt, und zwar vermutlich umso mehr in komplexeren Textzusammenhängen. Die vorliegende Arbeit verdeutlicht die kognitiven Anforderungen, die erfolgreiche Pronomenauflösung an Kinder stellt. Nicht zuletzt gibt sie wichtige Impulse für die Untersuchung von übergeordneten Leseprozessen in natürlichen Leseumgebungen mittels Eyetracking auch bei jüngeren Kindern.

ABSTRACT

Many children struggle with reading for comprehension. Reading is a complex cognitive task depending on various sub-tasks, such as word decoding and building connections across sentences. The task of connecting sentences is guided by referential expressions. References, such as anaphoric noun phrases (Minky/*the cat*) or pronouns (Minky/*she*), signal to the reader how the protagonists of adjacent sentences are connected. Readers construct a coherent mental model of the text by resolving these references. Personal pronouns (*he*/*she*) in particular need to be resolved towards an appropriate antecedent before they can be fully understood. Pronoun resolution therefore is vital for successful text comprehension. The present thesis investigated children's resolution of personal pronouns during natural reading as a possible source of reading comprehension difficulty. Three eye tracking studies investigated whether children aged 8-9 (Grade 3-4) resolve pronouns online during reading and how the varying information around the pronoun region influences children's eye movement behavior.

The first study investigated whether children prefer a pronoun over a noun phrase when the antecedent is highly accessible. Children read three-sentence stories that introduced a protagonist (*Mia*) in the first sentence and a reference to this protagonist in one of the following sentences using either a repeated name (*Mia*) or a pronoun (*she*). For proficient readers, it was repeatedly shown that there is a preference for a pronoun over the name in these contexts, i.e., when the antecedent is salient. The first study tested the repeated name penalty effect in children using eye tracking. It was hypothesized that in contrast to proficient readers, the fluency of children's reading processing profits from an overlapping word form (i.e., the repeated noun phrase) compared to a pronoun. This is because overlapping word forms allow for direct mapping, whereas pronouns have to be resolved towards their antecedent first.

The second study investigated children's online processing of pronominal gender in a mismatch paradigm. Children read sentences in which the pronoun either was a gender-match to the antecedent or a gender-mismatch. Reading skill and reading fluency were also tested and related to children's ability to detect a mismatching pronoun during reading.

The third study investigated the online processing of gender information on the pronoun and whether disambiguating gender information improves the accuracy of pronoun comprehension. Offline comprehension accuracy, that is the comprehension of the pronoun, was related to children's online eye movement behavior. This study was conducted in a semi-longitudinal paradigm: 70 children were tested in Grade 3 (age 8) and again in Grade 4 (age 9) to investigate effects of age and reading skill on pronoun processing and comprehension.

The results of this thesis clearly show that children aged 8-9, when they are in the second half of primary school, struggle with the comprehension of pronouns in reading tasks. The responses to pronoun comprehension questions revealed that children have difficulties with the comprehension of a pronoun in the absence of a disambiguating gender cue, that is when they have to apply context information. When there is a gender cue to disambiguate the pronoun, children's accuracy improves significantly. This is true for children in Grade 3, but also in Grade 4, albeit their overall resolution accuracy slightly improves with age.

The results from the analyses of eye movements suggest that the discourse accessibility of an antecedent does play a role in children's processing of pronouns and repeated names. The repetition of a name does not facilitate children's reading processing like it was anticipated. Similar to adults, children showed a penalty effect for the repeated name where a pronoun is expected. However, this does not mean that children's processing of pronouns is always adult-like. The results from eye movement analyses in the pronoun region during sentence reading revealed significant individual differences related to children's individual reading skill and reading fluency.

The results from the mismatch study revealed that reading fluency is associated with children's detection of incongruent pronouns. All children had longer gaze durations at mismatching than matching pronouns, but only fluent readers among the children followed this up with a regression out of the pronoun region. This was interpreted as an attempt to gain processing time and "repair" the inconsistency. Reading fluency was therefore associated with detection of the mismatch, while less fluent readers did not see any mismatch between pronoun and antecedent. The eye movement pattern of the "detectors" is more adult-like and was interpreted as reflecting successful monitoring and attempted pronoun resolution.

Children differ considerably in their reading comprehension skill. The results of this thesis show that only skilled readers among the children use gender information online for pronoun resolution. They took more time to read the pronoun when there was disambiguating gender information that was useful to resolve the pronoun, in contrast to the less skilled readers. Age was a less important factor in pronoun resolution processes and comprehension than were reading skill and reading fluency. Taken together, this suggests that the good readers direct cognitive resources towards pronoun resolution when the pronoun can be resolved, which is a successful comprehension strategy. Moreover, there was evidence that reading skill is a relevant factor in this task but not age.

The contribution of the present thesis is a depiction of the specific eye movement patterns that are related to successful and unsuccessful attempts at pronoun resolution in children. Eye movement behavior in the pronoun area is related to children's reading skill and fluency. The results of this thesis suggest that many children do not resolve pronouns spontaneously during sentence reading, which is likely detrimental to their reading comprehension in more complex reading materials. The present thesis informs our understanding of the challenge that pronoun resolution poses for beginning readers, and gives new impulses for the study of higher-order reading processes in children's natural reading.

TABLE OF CONTENTS

I. THEORETICAL FRAMEWORK

1 INTRODUCTION

From an early age, children are surrounded by written language. In contrast to spoken language, children do not simply acquire reading but need direct instruction to learn it. In Germany, reading instruction starts in primary school. After about one year of direct instruction, primary schoolers are expected to read short texts with words that are familiar to them. The complexity of their reading materials increases throughout the primary school years. While some children quickly learn how to read after only some instruction, others struggle immensely. According to the Progress in International Reading Literacy Study (IGLU/PIRLS 2016), as many as 18.6% of German children leave primary school after Grade 4 without being able to read fluently for comprehension (Hußmann et al., 2017).

As a society, we have a great interest to better understand children's reading development and its pitfalls because reading is a basic skill that massively affects educational outcomes. Perhaps unsurprisingly, a more recent large-scale study (Survey of Adult Skills; OECD, 2019) confirmed that many adults have under-developed reading skills. One interpretation of these results is that once children leave school with bad reading skills, they are unlikely to "catch up" later in life. Their limited career options aside, adults who cannot fully make sense of written news sources, contracts or instructions can be expected to have trouble navigating everyday life in our information-based society. It is imperative that the research community understands the basic problems regarding reading comprehension, in order to be able to develop intervention programs that foster reading across the board.

The relevance of reading as a basic skill is reflected in a long tradition of research into children's reading comprehension. We have come to understand that children differ immensely in their text comprehension and we can attribute these differences both to properties of the text and children's individual reading skill and developmental aspects. The reader may have to decode lengthy or infrequent words, bridge information gaps, apply sufficient world knowledge to the text, and integrate seemingly conflicting information – all tasks that are critical to reading that beginning readers may or may not master successfully (for an overview, see Nation, 2005). For developing readers, indi-

vidual differences in reading fluency and reading experience are also particularly relevant.

In much of the research on children's reading comprehension, the reading process has been treated as the metaphorical "black box". The majority of seminal studies on children's reading development concentrated on the outcome of reading: Can they answer comprehension questions accurately? Can they recall details from the text? Are they able to name the referent for a pronoun? The processes at work during reading have only recently come into focus, largely with the advent of techniques and methods that made reading process research less invasive and more feasible with children (see Blythe & Joseph, 2011; Schroeder, Hyönä, & Liversedge, 2015).

With the aim to add to our knowledge of children's higher-order reading processing and comprehension, the present thesis investigated children's processing of anaphoric personal pronouns in three experiments. The introduction will lay out how pronoun resolution is an exemplary case of a local inference process that is vital for successful text comprehension. First, a broad framework for text comprehension suitable for children's reading will be laid out. Then, the mechanisms that are assumed to underlie pronoun resolution in adults will be discussed, before closing the introduction by laying out what we know so far about children's pronoun resolution.

2 A FRAMEWORK FOR TEXT COMPREHENSION

It is generally agreed that during reading, a text is "translated" into a non-propositional, representational mental format. This format is called the *situation model* (van Dijk & Kintsch, 1989; Zwaan & Radvansky, 1998), or, in more general terms, the *mental model* (Johnson-Laird, 1983). Based on the text and additional information derived from general world knowledge, readers generate their individual mental representation of the written text. With this mental representation, readers track the order of events, connect causal links between events, relate protagonists to each other and integrate their world knowledge with text information. Understanding a text, it follows, means to successfully build a mental model, which incorporates all relevant textual information and incrementally adds new propositions in a way that is meaningful to the reader. In this way, proficient readers arrive at a coherent interpretation of a text (Sanford & Garrod, 1981).

Many propositions that have to be integrated into an existing mental model are not fully specified and therefore, readers have to draw inferences during reading. Inferences bridge under-specifications and information gaps in the text, ensuring the coherence of a mental model (van Dijk & Kintsch, 1983). Such gaps can be smaller and close to the text level, requiring local coherence inferences, or larger and more globally connected to the model, requiring global coherence inferences. In their seminal work, Graesser et al. (1994) define global coherence as the organization and interrelation of local chunks of information into higher order chunks. Global coherence inferences are not necessarily connected to the text level in a direct manner. While reading a novel, for example, readers may track a character's motivations based on past actions of that character going several chapters back or infer the setting of the novel by piecing together information given throughout the first chapters. For global coherence inferences, readers typically apply extraneous knowledge sources to the text such as knowledge about text genres or general world knowledge. Local coherence inferences, as the term suggests, are more directly connected to the local text level. Local coherence, in contrast, is defined in terms of processes that organize elements, constituents, and referents of adjacent or nearby clauses (Graesser et al., 1994). Local coherence inferences typically comprise the connection of two words across sentences through mapping, for example by exemplars of a category (*the poodle* and *the dog*), or the connection of entities in adjacent sentences through anaphora (*Phoebe, Prue, and Piper* and *the witches*).

One line of research has been investigating the amount and type of information being "bridged", or in other words, how far readers go in their effort to maintain coherence. Most research agrees that making the appropriate inferences during reading is key to text comprehension; however, there is no general consensus with respect to the automaticity and extent of inferences during reading (see Perfetti, 1994). Some research claims that inferences are drawn routinely and near-automatically wherever an information gap presents itself, while other research suggests that inferences are highly strategic, associated with a high cognitive cost and therefore only drawn when the discourse would be incoherent otherwise. Proponents of both the constructivist (or 'maximal') approach to inferences (e.g. Graesser, Singer, & Trabasso, 1994) and the minimalist approach (e.g. McKoon & Ratcliff, 1992) have generated experimental evidence for their claims. Further investigation into these mixed results lead some researchers to believe that readers draw inferences not as a general rule, but as a result

of situational processes that are connected to specific text-reader interactions (Perfetti, 1994; van den Broek, Lorch, Linderholm, & Gustafson, 2001; van den Broek, Risden, & Husebye-Hartmann, 1995). It has to be kept in mind that not only properties of the text, but also individual reader differences play a role in inferencing. This is particularly relevant for developing readers, who show large variances in individual variables that have been identified as determinants of reading comprehension, such as word reading fluency (e.g., Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003; Nation, 2005).

2.1 The role of pronouns in text comprehension

All three studies in this dissertation investigate local inference processes in developing readers by studying reading processing of the personal pronouns *he* and *she*. Pronouns are relevant from a discourse processing perspective because their interpretation depends on local inference processes that establish coherence. Pronouns are short and very frequent and consequently easy to process on the word level, however, they have to be enriched with meaning to be fully understood. Readers have to resolve pronouns in order to understand their meaning in the current sentence context. Often, this resolution process requires connections across adjacent sentences.

Pronouns and other types of anaphora are an integral part of the situation model approach to text comprehension. Zwaan & Radvansky (1998) dissociate five dimensions of the mental text representation regarding situation models: (1) space, (2) time, (3) causation, (4) intentionality, and (5) objects and protagonists. According to the model, readers keep track of the goals and plans of protagonists on the intention dimension and they track the protagonists' relation to the events denoted by each proposition on the object dimension. Therefore, protagonists are particularly important as anchors for the global coherence of a model and, subsequently, for text comprehension. The related *event-indexing model* (Zwaan, Langston, & Graesser, 1995) suggests that readers decompose every new proposition into 'indices' of the five dimensions to keep track of the series of events denoted by the proposition. The coherence of a text depends on the amount of changes on the dimensions: The more dimensions undergo changes from one proposition to the next, the more challenging it is to maintain coherence. Consider the set of sentences in (1) and (2).

(1) The man is reading a book. The girl is sitting on a bench. The dog is barking.

(2) The man is reading a book. His girl is sitting on a bench. Their dog is barking.

Arguably linguistic text definitions would agree that (1) does not qualify as a text as much as (2) because the sentences in (1) lack coherence (de Beaugrande & Dressler, 1981; Janich, 2008). Local coherence can be achieved on the text level in these sentences by exchanging the indefinite article *a* with the pronominal expressions *his* and *their*. The sentences are now coherence-marked on the local level by way of possessive pronouns. These pronouns clarify to the reader how the sentences are interconnected. The repeated reference to the same set of discourse entities through pronouns or other anaphora is an important cue for the construction of coherence in mental models and has been termed referential continuity, or topic continuity (Garnham, Oakhill, & Johnson-Laird, 1982; Givón, 1983). When a proposition continues to refer to the same referent or set of referents as the preceding one, this is explicitly marked by anaphoric expressions on the local text level. For example, while noun phrases with an indefinite article (e.g., *a mouse, a child*) typically introduce a new referent, definite noun phrases and pronouns (e.g., *the mouse*, *she*) pose a cue to retrieve a referent that has been mentioned before, i.e., a referent from memory (Kehler, 2002). This means that subtle differences on the text surface (*his* vs. *the*) can have important consequences for the mental model of the text. Proponents of the mental model approach assume that readers detect and incorporate these cues as they strive towards coherence of their model. In other words, anaphora signal to the reader how the sentence information is connected to the information that is already in their mental model.

In summary, anaphors are particularly relevant for text comprehension. Pronouns are a type of anaphor that play a special role in connecting sentences of a text by acting as coherence markers. The following paragraph discusses the conditions for successful pronoun resolution and its mechanisms in proficient readers, before moving on to what we know about pronoun resolution in beginning readers.

2.2 Pronoun resolution in proficient readers

Pronoun resolution has been studied in fields as diverse as psycholinguistics, developmental linguistics, computer linguistics, and language psychology. Theories of pronoun

resolution are therefore as numerous as they are diverse, depending on the research interests of the field they originated in and the aims and grain sizes of the associated models. The purpose of this section therefore is not to give a comprehensive overview of the available theories of pronoun resolution. Rather, it will give some background on the processing and comprehension of personal pronouns in adults from a psycholinguistic perspective as far as is necessary to understand where the experiments with children in this thesis are coming from. A note on terminology: The terms "pronoun resolution", "pronoun processing", and "pronoun comprehension" are not always used consistently in the literature. The terms are used here as follows: "Pronoun comprehension" means the way a reader reports their understanding of a pronoun after reading or listening, i.e. in response to a comprehension question. "Pronoun processing" refers to readers' online behavior in response to a pronoun, which can be depicted for example using eye movement measures. "Pronoun resolution" incorporates both processing and comprehension, as it refers to the mechanisms triggered by a pronoun and the comprehension outcome of these mechanisms. Lastly, "pronoun resolution skill" is used in reference to children's individual differences in the processing and comprehension of a pronoun, where "skilled resolution" equals adult-like processing and comprehension.

2.2.1 Cue-based pronoun resolution: gender and accessibility

The meaning of a pronoun has to be inferred from context. One obvious question is: if pronouns have to be inferred, why would we ever use them? Why take up the burden of inference when a speaker could just repeat the original referent? Cognitive theories of anaphor resolution give an answer. According to these theories, the type of anaphor that is chosen in a discourse is governed by general cognitive constraints on information retrieval (see Ariel, 2001, 2004; Almor, 1999; Lewis & Vasishth, 2005; Lewis, Vasishth, & van Dyke, 2006; Greene, McKoon, & Ratcliff, 1992). One such cognitive approach is the minimalist view on inference generation, or *resonance theory* (McKoon & Ratcliff, 2005). Resonance means the process by which information is retrieved from working memory. It is a mechanism in which cues in short-term memory activate information from long-term memory. In resonance theory, information processing is regarded as a parallel and automatic process, such that there is no single source of

information in long-term memory that takes precedence over others. The only determinant of the degree of activation is the strength of the association between the cue and the retrieved item. Note that this account is contrasted by staged theories of sentence processing where syntactic information is assumed to take precedence (see for example Friederici, 1995, 2002). While resonance theory has not been designed specifically for pronoun resolution or is in fact limited to language processing, it has been applied to the resolution of pronouns (Greene, McKoon, & Ratcliff, 1992; McKoon, Gerrig, & Greene, 1996; McKoon & Ratcliff, 2005). According to resonance theory, the form of an anaphor in a text, that is if it occurs as a full name or a pronoun, is not a random choice but determined by salience. Salience can be defined as the degree of accessibility of a discourse entity relative to others. The more cognitively salient a given entity, the higher its accessibility in memory. Salience is important for the activation of a specific discourse referent during pronoun resolution because pronouns usually refer to the most salient discourse entity with which they are compatible. This is in line with the *accessibility theory* (Ariel, 2001, 2004). It assumes that all referring expressions code a specific degree of salience, or "mental accessibility" of its referent. The accessibility marking scale defines a hierarchy of discourse referents that are common across languages:

Full name+modifier > **full name** > long definite description>
short definite description > last name > first name >
distal demonstrative + modifier > proximate demonstrative+ modifier >
distal demonstrative + NP > proximate demonstrative+ NP>
distal demonstrative (-NP) > proximate demonstrative (-NP) >
stressed pronoun + gesture > stressed pronoun > unstressed **pronoun**
> cliticized pronoun > verbal person inflections > zero

Figure 2.1. Accessibility marking scale from high accessibility to low accessibility (adapted from Ariel, 2001; highlights added)

Not all of the anaphoric expressions in Ariel's accessibility marking scale are available in German or relevant for reading processing (e.g., stressed/unstressed pronouns). Close to the top of the marking scale, we can find full names (e.g., *Mia*, *Max*) and definite

descriptions (e.g., *the children*, *my aunt*), marked as least accessible, whereas the pronoun is found at the bottom of the marking scale and is marked as highly accessible. Note that an item's position on the accessibility scale is also related to informative content: The pronoun contains very little information, whereas full names contain a lot of information. Full names can identify an entity unambiguously while pronouns can, in principle, refer to many different entities. Pronouns are therefore exclusively used in discourse contexts where their antecedent is salient.

The fact that readers expect a pronoun when the antecedent is salient, is demonstrated by the so-called *repeated name penalty effect* (Gordon, Grosz, & Gilliom, 1993). Numerous studies have shown that adults slow down during reading when the name of a salient discourse entity is repeated in the text, an effect that has been termed "repeated name penalty" (e.g., Fukumura & van Gompel, 2015; Kennison & Gordon, 1997; Shapiro & Milkes, 2004). The fact that an overlapping word form does not seem to pose the ideal retrieval cue for an accessible antecedent demonstrates the relevance of discourse expectations as is predicted by the accessibility marking scale. The repeated name penalty effect is also predicted by the *informational load hypothesis* (in the following ILH; Almor, 1999; Almor & Nair, 2007) which is closely related to accessibility. The ILH posits that anaphora differ in the weight of their semantic features, or "informational load". Proper names for example are informationally "heavy" because they carry all the information to denote exactly one entity in the model. Personal pronouns in contrast carry little informational load. Importantly, according to the ILH informational load is related to processing speed, where anaphora with high informational load (noun phrases, proper names) should take longer to process than anaphora with low informational load (reflexives, personal pronouns).

Salience is not the only cue in pronoun resolution. Among the linguistic constraints imposed upon pronoun resolution in German is their match in gender and number with respect to the antecedent (see Fagan, 2009). In this respect German is very similar to English. In German, singular male antecedents are referred to by the personal pronoun *er* (Eng.: *he*) and singular female antecedents by the pronoun *sie* (Eng.: *she*). Note that German *sie* is also the plural (Eng.: *they*), which has to be carefully considered in experimental stimuli. Grammatical number of course is also relevant in pronoun resolution, however, it has not been used in the experimental stimuli of the experiments in this

dissertation since *sie* (female singular) and *sie* (plural) are ambiguous in German.

It has been shown that adults rapidly and routinely use both gender information and accessibility to resolve pronouns during listening. Arnold, Eisenband, Brown-Schmidt, and Trueswell (2000) recorded adults' eye movements while they viewed a picture featuring two cartoon characters of the same or a different gender and listened to a text describing said picture. The text contained a pronoun which referred to either the first, more accessible, character or to the second. The authors found evidence for the use of both gender and accessibility about 200 ms after the offset of the pronoun. Their results further indicate no precedence of either information source, which is consistent with resonance and cognitive theories of pronoun resolution in general. Other evidence for the rapid and simultaneous use of gender cues comes from syntactic interference studies. Without going into too much detail here, interference experiments test the assumption that syntactically inaccessible referents interfere with the processing and comprehension of a (syntactically accessible) reflexive when both antecedent and non-antecedent match in gender, e.g., *The tough soldier* (antecedent) *that Fred* (non-antecedent) *treated in the hospital introduced himself* (reflexive) *to all the nurses*, versus *The tough soldier* (antecedent) *that Katie* (non-antecedent) *treated in the hospital introduced himself* (reflexive) *to all the nurses* (Cunnings & Felser, 2013; Patil, Vasishth, & Lewis, 2016). *Himself* refers to *the though soldier* in both sentences due to syntactic constraints, however the gender-matching *Fred* can result in online interference effects. Such interference effects have generally been interpreted in favor of simultaneous use of different information sources in sentence processing.

This section has so far discussed anaphor resolution as a cue-based process. Pronouns have to be resolved because they are semantically underspecified, however they do contain grammatical information and their interpretation is guided by morpho-syntactic cues. The accessibility marking scale and the informational load hypothesis, both rooted in cognitive theories of anaphor resolution, can explain where a pronoun is appropriate in a discourse as opposed to other types of anaphor, such as a (repeated) noun phrase. The concept that they rely upon is salience, or accessibility. The following section will move away from the discourse view on pronouns and towards the mechanisms that govern pronoun resolution.

2.2.2. TWO-STAGE MODELS OF PRONOUN RESOLUTION

Research on pronoun resolution with adults still leaves many open questions, one of which is the exact time course of pronoun resolution. Some authors have suggested that pronoun resolution is fast and nearly automatic (constructivist view; e.g., Graesser, Singer, & Trabasso, 1994), while others claim it is strategic. For example, Greene, McKoon and Ratcliff (1992) show in a series of experiments that readers do not always identify a unique referent for a pronoun during reading and reject the idea that pronouns are automatically resolved. This is especially the case in the presence of two relatively indistinguishable discourse entities, e.g., two male characters. They assume that in these contexts, pronouns are not resolved routinely but dependent on the task, among other factors. Without a respective task, pronouns may be processed in a "shallow" way, in other words, readers never fully integrate the pronoun and continue with an incomplete understanding of the sentence (see also Stewart, Holler, & Kidd, 2007).

Two-stage models of pronoun resolution can somewhat reunite these seemingly contrasting views by dissociating two stages of pronoun resolution. Common to these models is the view that the pronoun is first connected to the possible antecedents using retrieval cues, and then resolved towards the most appropriate antecedent. While the first stage may be automatic, the second stage can be partly or fully strategic.

In the bonding and resolution framework (Garrod & Terras, 2000; Sanford, Garrod, Lucas, & Henderson, 1983), two processes are dissociated: The point in time when a pronoun is connected to an appropriate antecedent, for example by way of gender-number-matching and syntactic constraints (bonding), and a slightly later point in time when that information is checked against the discourse context and general world knowledge (resolution). The dissociation of these two phases is consistent with both the assumption that anaphor resolution is cue-based, automatic and efficient, and the assumption that pronouns are resolved strategically rather than automatically. The cue is then relevant for the bonding stage, whereas resolution depends on additional information like textual context, reading experience, world knowledge or even demands of the experimental task.

In a similar two-stage account of pronoun resolution, Rigalleau, Caplan and Baudiffier (2004) attempt to reconcile the seemingly contradicting experimental findings on the

use of gender cues during pronoun resolution. They propose an automatic "coindexation process" in which the pronoun is coindexed with all its gender-matching, or cue-matching, antecedents. The second step is a "strategic disengagement process", in which the activation of the non-antecedent is suppressed. Importantly, this second step depends on the reader's motivation for resolution, in other words it is strategic.

To sum up, pronouns are used as cues, or informationally light "shortcuts" to refer to an accessible discourse entity. Pronoun resolution is a specific type of local inference that is not fully determined by text information, but also depends on the reader's individual effort in establishing a referent for the pronoun. In adults, pronoun resolution is a fast, cue-based process that is strategic in more ambiguous contexts. The pronoun is bonded with fitting discourse entities before it can be resolved towards a contextually plausible referent. The challenge with pronouns is not surface processing, but this resolution towards a discourse entity under the consideration of context information. Therefore, pronouns can serve as a proxy for the quality of children's information integration during reading.

3 THE DEVELOPMENT OF PRONOUN RESOLUTION

3.1 READING COMPREHENSION DEVELOPMENT

A model of reading comprehension that is frequently applied to children's reading is the reading systems framework (Perfetti & Stafura, 2014; Stafura & Perfetti, 2017). Put in broad terms, it assumes that reading comprehension relies on different knowledge sources: word knowledge, linguistic knowledge, and metacognitive knowledge. It further assumes that deficits in one of these knowledge sources lead to differences in reading comprehension. Linguistic knowledge for example would determine a child's effort towards pronoun resolution. Metacognitive knowledge determines whether a child monitors their understanding of the text during reading to detect inconsistencies between an antecedent and a pronoun. In short, readers can show weaknesses in different knowledge sources that are relevant for pronoun resolution, leading to different reading outcomes.

Children differ enormously in their ability to understand written text, however there is no definitive consensus about the determinants of these individual differences in reading comprehension. For beginning readers, differences in decoding ability, connected to word knowledge in the reading systems framework, have been shown to be of particular relevance (e.g., Best, Floyd, & McNamara, 2008; Cain, Oakhill, & Bryant, 2004; Perfetti & Hogaboam, 1975). Decoding ability is often measured in terms of reading fluency, for example, words read per minute (Moll & Landerl, 2010). Before beginning readers reach a level of automaticity that allows for wholistic word decoding, word reading is rather labored: As children begin to read, they decode a word in a letter-by-letter fashion, putting the letters together to form a word at a high cognitive cost. This sequential reading of words is exemplified by the word length effect that has repeatedly been found in young readers but not adults (e.g., Gagl, Hawelka, & Wimmer, 2015; Tiffin-Richards & Schroeder, 2015). Importantly, readers with low or under-developed decoding ability struggle at building a coherent mental model of what they have read. One line of explanation for this finding follows the *capacity theory* (Just & Carpenter, 1992), which assumes that there is a limited amount of information that can be actively retained in working memory). As the amount of information that beginning readers can retain during reading is limited, their working memory capacity may be taken up by the demands of word decoding, leaving little to no capacity for integration processes. As a consequence, children can be expected to show difficulties in inference generation tasks (Currie & Cain, 2015; Oakhill, 1994). It follows that pronouns pose specific challenges for beginning readers. As discussed above, a pronoun is easy to decode but needs to be resolved towards a discourse entity. As children begin to read, they may lack the necessary knowledge or processing capacity for dealing with pronoun resolution during reading online. This may leave them with an impoverished mental model of the text, which in turn may be responsible for poor performance in reading comprehension assessment and low comprehension skill.

Pronoun resolution in children has been mainly studied from a post-reading comprehension perspective. Comprehension studies typically use post-reading questions to assess children's pronoun resolution while processing studies use on-line measures such as reading time or eye tracking measures to investigate the time course of pronoun processing in children. It is a noticeable gap in the literature that these views have not yet been reconciled, for example by combining eye tracking with targeted post-reading

comprehension questions. Since the two lines of research have been developed largely separate from another, they will be reviewed in turn.

3.2 Children's comprehension of pronouns

One prominent line of research has used comprehension questions to investigate children's resolution accuracy after listening to a story or reading a text. Multiple studies have shown that children struggle to name the correct referent for pronouns during and after a reading task. Yuill and Oakhill (1988) investigated 7- to 8-year-old's comprehension of sentences containing a pronoun and one or two gender-matching antecedents. The authors found that children performed remarkably poor when naming the referent for a pronoun, even when there was only one referent present and the pronoun therefore entirely unambiguous. Depending on the difficulty of the pronominal inference, children had an error rate of up to 28%, which demonstrates that pronoun resolution during reading is a difficult task for them.

In a similar experiment, Oakhill and Yuill (1986) manipulated antecedent gender to investigate whether children resolve pronouns from context at all. They used sentences like *Liz lent ten pence to Tom because he was very poor,* or *Peter lent ten pence to Tom because he was very poor*. In the first sentence, the only possible match for the pronoun *he* is *Tom*. In the second sentence, although there are two grammatically possible antecedents, the correct antecedent *Tom* can be derived from context: People who need to borrow money are typically poor. As expected, the children performed significantly worse when there was no gender cue (16-27% error rate, depending on comprehension skill) compared to when there was a gender cue (2-14% error rate). This study shows, first, that while pronoun resolution during reading is challenging for many children, explicit cues such as the gender of the antecedents improve their resolution accuracy. Second, the study hints at the large interindividual differences in children's resolution accuracy.

Children may be unable or unwilling to draw local inferences, either because they lack the cognitive resources, or because they are unaware that they have to allocate attention to pronoun resolution during reading (poor linguistic knowledge), or because they

do not monitor their comprehension sufficiently (poor metacognitive knowledge; Stafura & Perfetti, 2014). In the literature, the processes associated with metacognitive knowledge have been termed *metacognitive monitoring* or *comprehension monitoring* (e.g., Ehrlich, 1996; Ehrlich, Remond, & Tardieu, 1999; Oakhill, Hartt, & Samols, 2005; Vorstius, Mayer, Radach, & Lonigan, 2013). When confronted with a difficult text, children may not monitor their comprehension sufficiently and instead resort to a good-enough reading strategy, which leaves them with an impoverished understanding of the text (Wonnacott, Joseph, Adelman, & Nation, 2016).

Importantly, studies targeting resolution accuracy do not depict how and when children resolve a pronoun. The distinction is important to determine whether children resolve pronouns during natural reading as well. Do they use a gender cue online or do they use it only after reading, when a comprehension question appears? It has been shown that comprehension questions themselves may act as a retrieval cue as they prompt the reader to retrieve the correct antecedent for a pronoun. Thereby they may induce a depth of resolution that is only applied off-line (Rupp, Ferne, & Choi, 2006). It is therefore conceivable that children's processing of pronouns remains shallow in more natural reading contexts. Processing studies are designed to investigate children's resolution of pronouns as they are reading a sentence or text.

3.3 Children's processing of pronouns and other anaphora

Studies using listening tasks have been investigating the time course of pronoun resolution development in young children. They have produced mixed results due to a variety of methods and languages studied (Hickmann, Schimke, & Colonna, 2015). Many studies show an early sensitivity towards establishing pronoun-antecedent-relationships. For example, Arnold, Brown-Schmidt, and Trueswell (2007) showed that children reliably use gender cues to resolve a pronoun during listening from five years of age. In an eye tracking study using the preferential-looking-paradigm, it has also been shown that even young children show a resolution bias towards the subject, which is similar to the bias found in adults (Song & Fisher, 2015). Therefore, we can assume that by the time children attend primary school, they are able to resolve pronouns while listening to a story (see, however, Francey & Cain, 2014). In a reading task, in contrast, children struggle with pronoun resolution well into the primary school years.

As a measure of processing time in children's reading, some researchers have employed self-paced reading tasks. In a self-paced reading task, sentences or texts are given word-by-word. The child presses a button each time they have read a word, and a new word (or several words, in different variations of the task) appears. This way the sentence is presented in a piecemeal manner. These experiments were vital in understanding children's processing of pronouns and other anaphora in different sentence contexts. A self-paced reading experiment with 10-year-olds (Ehrlich, Rémond, & Tardieu, 1999) showed that good readers, but not poor readers, take more time to read sentences containing a pronoun than a repeated name in anaphor position. Even though good readers read faster than the poor readers in general, the good readers seem to adjust their reading time to the demands of the text while poor readers do not.

A general problem of the self-paced-reading method for assessing reading processing in children shall be highlighted here: The time they spend on a word during first-pass reading may not be the best indicator for successful integration. Even while fast decoding is associated with good comprehension for the reasons outlined above, this may not always be the case when pronoun resolution is difficult. For example, if good young readers are allowed to reread specific parts of the text, they choose to reread more often than poor readers (Ehrlich, Rémond, & Tardieu, 1999). This has been interpreted as an indicator of good metacognitive knowledge: Only when children understand that a text passage is difficult for them, they can allocate more processing time to it. Further, the time course of processing may be relevant here, in other words, where and when during reading children allocate this extra processing time. For example, while rereading an ambiguous pronoun during first-pass reading of the sentence could be effective when resolution is difficult, children may also want to go back to the antecedent. This strategy would not be picked up with self-paced reading. One listening experiment with children exemplifies the need of using sensitive processing measures: Clackson, Felser and Clahsen (2011) showed that a specific effect in pronoun comprehension (the coreferential delay, see Reinhardt, 2011) does not necessarily surface in comprehension assessment but is visible in an on-line processing measure. In their experiment, children aged 6 to 9 answered the pronoun resolution questions in an adult-like manner, however, a processing delay was detectable during listening in the visual-world eye tracking paradigm. If it surfaces during reading as well, such a processing delay is potentially relevant for the comprehension of pronouns in written text.

More recently and with the advent of more accessible eye tracking technology, the method has been used to study the cognitive processes involved in reading as well as their connection to reading development (Blythe, 2014; Blythe & Joseph, 2011; Schroeder, Hyöna, & Liversedge, 2015). Research on reading processes using eye tracking operates under the general assumption that the differences in eye movements during reading are a reflection of the cognitive processes involved during the processing of written text (Just & Carpenter, 1980; Rayner, 1998). Compared to the vast amount of research on adult's eye movements during reading, the literature on children's reading using eye tracking is still young and rather limited. Further, most early studies focused on word reading. They consistently show that children make more fixations than adults, that their fixations are longer and that they engage in more unselective rereading (Blythe & Joseph, 2011; Schroeder, Hyönä, & Liversedge, 2015). Whether these differences in lower-level reading processing are connected to differences in processing at higher levels of reading, such as anaphoric processing or other integration efforts, has been of a more current interest.

In an early eye tracking study on children's referential processing, Murray and Kennedy (1988) showed that the eye movement behavior associated with pronoun resolution differed for good and poor readers. Good readers among the children made more selective, targeted regressions when reading sentences containing pronouns, while less-skilled comprehenders made shorter, less selective regressions. This unselective regressive reading was termed "backtracking" by the authors and is an eye movement feature of children's reading compared to that of adults also found in later studies. Moreover, their results using eye tracking confirm the observation that good readers among the children reread selected parts of the text (Ehrlich, Rémond, & Tardieu, 1999; see above) in a more natural reading situation.

In an eye tracking study on anaphoric noun phrases, Joseph, Bremner, Liversedge, & Nation (2015) had English-speaking children (10-11 years) read short paragraphs in which they manipulated the distance between antecedent and anaphor and the semantic typicality of the antecedent (typical: *a truck – the vehicle*, atypical: *a crane – the vehicle*). The authors were interested in the time course of anaphoric processing such as the immediacy of children's eye movements after having read the anaphor in the different conditions. They found an interaction of working memory skill and antecedent distance

such that children with high working memory skill showed longer first fixation times in the near, but not in the far anaphor condition which they interpreted as an inverse effect of typicality. There was a similar inverse effect of typicality in regressions out of the anaphor region, meaning that children made more regressions back to the antecedent when it was typical rather than atypical. Given these rather counter-intuitive findings, the authors argue that children showed more regressions out of the anaphor region when they resolved the anaphor. This interpretation is backed by the finding that it was the skilled working memory group who showed longer first fixations in the near, but not in the far condition, possibly indicating an immediate resolution of the anaphor in the easier condition. From these interpretations, it follows that many children – those with low working memory skills in the given study – did not resolve the anaphors at all. This concurs with the general observations from comprehension studies discussed above: Under certain circumstances, children may not engage in anaphoric resolution during reading at all and their situation model of the text remains shallow or underspecified. Whether this is the case, however, could be examined further by combining eye tracking and resolution accuracy measures such as post-reading questions.

In sum, there are several indications for the relevance of anaphoric resolution for children's reading comprehension in the literature. We know that children struggle with the comprehension of pronouns during and after reading and we have several indications for developmental and inter-individual differences in children's pronoun processing. However, little is known about children's processing of anaphoric pronouns in natural reading.

4 Open research questions

The mechanisms of children's pronoun processing during natural reading have not yet been studied in detail. There are less than a handful of eye tracking experiments on the subject (notably Murray & Kennedy, 1988) despite the fact that the method is established as state-of-the-art in reading processing research.

Some of the most relevant questions concerning children's online pronoun processing are whether beginning readers resolve pronouns online and how they do so, if there is

a developmental trajectory of pronoun resolution in beginning readers. Also, it is unclear how individual differences, such as reading ability or processing speed, influence pronoun processing. Further, any research into children's higher-order reading processing will aim for predicting children's pronoun comprehension in relation to their processing behavior, or, in other words, how eye tracking measures are related to pronoun comprehension in beginning readers. While these questions are highly inter-related, they will first be discussed in turn for clarity before laying out to what extent they have been addressed in the present dissertation.

4.1 Mechanisms of children's pronoun processing

Given what we know about children's reading on the one hand, and their resolution of pronouns on the other hand, it is conceivable that they process pronouns in a shallow way and that this is one of the reasons for poor reading comprehension in beginning readers. This may be particularly the case when the antecedent of a pronoun has to be inferred from context. For adults, eye tracking studies employing the preferential-looking-paradigm have shown that gender information is used rapidly online in pronoun processing during listening (Arnold, Brown-Schmidt, Eisenband, & Trueswell, 2000) and reading (Patil, Vasishth, & Lewis, 2016), and there is evidence that children use gender information online during listening as well (Arnold, Brown-Schmidt, & Trueswell, 2007). It has further been shown that children profit from gender cues for: Comprehension studies found that children are more likely to name the correct referent for a pronoun when it is gender marked and therefore unambiguous (Oakhill & Yuill, 1986; Yuill & Oakhill, 1988). However, whether and how children use the gender cue online during reading is entirely unclear. Eye tracking studies can address this question by recording moment-to-moment eye movements of children while they read sentences and texts with varying gender cues for pronominal anaphors.

The accessibility of a referent plays an important role in the choice of anaphora and their resolution. While a pronoun and a noun phrase may refer to the same antecedent, they have different discourse functions, driven by the accessibility of the referent. Given what we know about children's reading processing from the eye tracking literature, however, it is unclear whether and how the mechanisms found for adults transfer to children's pronoun processing. Children have been characterized as local readers who

do not make extensive connections across the text. As children have fewer processing resources they may profit from different retrieval cues than adults, both qualitatively and quantitatively.

Before the finding of the repeated name penalty effect, Gernsbacher (1989) suggested that repeated names actually foster anaphoric processing. In her structure building framework, accessibility is a direct function of the similarity between a retrieval cue and a "memory trace" for the antecedent. Therefore, she argues, the more explicit the anaphor, the more likely it should trigger suppression of non-antecedents and enhance the antecedent. She finds confirmation for her view in a series of experiments using probe verification tasks. These experiments essentially test the reaction times for the antecedent versus the non-antecedent (called "probe names") sometime after the anaphor was presented during a computerized sentence reading task. Although her view was later disputed, not least due to findings of the repeated name penalty effect, the reasoning behind it may still be relevant in children's reading. If children struggle with pronoun resolution, a noun phrase may be a more effective "retrieval cue" than a pronoun. As it contains more information, bonding of anaphor and antecedent is much more straightforward. Insights into these child-specific needs for online pronoun resolution are important in several ways. Not only can they shed light on children's local inference processes during reading. Equally important, any intervention efforts for children with reading difficulties require a solid understanding of the needs and requirements children have for a text. While it is plausible that the use of resolution cues and explicit repetitions in a text facilitate children's reading processing, there is currently no evidence for this claim.

4.2 Individual differences in children's pronoun processing

Children of the same age differ greatly in their ability to resolve pronouns. We also know of large interindividual differences in the component skills of reading, such as decoding fluency, working memory, listening comprehension skill, and metacognitive monitoring skill. An open question concerns the relationship between these component skills and pronoun processing. Are there specific component skills of reading which influence how children resolve pronouns online? Insight into the individual skill sets of

children who successfully process and resolve a pronoun may help in developing intervention programs for children who struggle with pronoun resolution, or, more generally, local inference generation. Arguably, training the specific skills that underlie pronoun resolution is more sustainable and effective than focusing on pronouns alone, as children may benefit from these improved skills also in other areas.

A closely related question is the locus of pronoun resolution difficulties in children. It was discussed above that a common line of argumentation for children's inability to resolve a pronoun are processing resource constraints in relation to slow word decoding. We cannot, however, simply assume that lower-level performance issues lead to higher-level comprehension issues, and that, once these lower-level issues (such as reading fluency, decoding ability, or word knowledge) are resolved, children automatically become better comprehenders. First, there is some evidence that while slow decoding is certainly a factor in children's comprehension difficulties, improved reading fluency does not guarantee improved reading comprehension. This manifests in the existence of so-called "poor comprehenders", who show poor comprehension skills, including pronoun resolution skills, despite adequate reading fluency (e.g., Cain & Oakhill, 1999; Cain, Oakhill, Barnes, & Bryant, 2001; Yuill & Oakhill, 1988). A single source of the reading impairments in poor comprehenders is unlikely (Cain & Oakhill, 2006). While this group is small, their existence alone is reason to believe that despite the general importance of fluent reading for reading comprehension, fluent reading does not automatically lead to better comprehension skill.

One of the more recent interests related to individual differences in children's reading comprehension has been the study of metacognitive monitoring skills and how these are related to text processing. As discussed above, the pronoun is easy to decode, but has to be properly connected to its antecedent. Therefore monitoring skills arguably influence a child's ability to resolve a pronoun, as they have to be both willing and able to go beyond decoding at the pronoun. If this process is less automatic, good metacognitive monitoring skills may be closely related to successful pronoun resolution. Children with good monitoring skills may be better able to (a) direct attention to morphosyntactic information on the pronoun, such as number and gender, and (b) use the pronoun as a cue for initiating processing strategies directed towards pronoun resolution, such as rereading or refixations of previous text materials.

4.3 Development of children's pronoun processing

Another open question concerns the development of pronoun resolution in children, and in particular whether there is a developmental trajectory discernible in children's pronoun processing. There is evidence from the literature on inference generation that young children are less likely to make inferences than older children (Cain & Oakhill, 1999; Casteel & Simpson, 1991; Omanson, Warren, & Trabasso, 1978)

It is likely that the development of the component skills of reading is related to the ability to process pronouns during reading. As we know that decoding speed and reading fluency develop with age, it may be hypothesized that children's ability to resolve pronouns online during reading develops alongside these component skills. However, this would again require a solid understanding of the functional relationship between the component skills of reading and pronoun resolution. To answer these questions, longitudinal studies using eye tracking are required which also test children's individual reading skill. As eye tracking investigations of children's word reading have generated new impulses in recent years (for reviews see Blythe, 2014; Blythe & Joseph, 2011; Schroeder, Hyönä, & Liversedge, 2015) the approach is certainly promising for studying higher-level reading processes as well.

It is yet unclear what we can expect of children at different ages. Consequently, although we are generally aware of the importance of inference generation for text comprehension, there is currently no reading curriculum in place that discusses children's ability to draw local inferences or uses specific materials to address the importance of inference generation with children in an age-appropriate way. The development of such materials for educational purposes requires more basic research into the development of children's processing and comprehension of pronouns and other local ambiguities in texts.

5 Aims of the experimental studies

The three studies in this thesis were conducted as part of the Developmental Eye Tracking Study, conducted at the Max Planck Institute for Human Development in Berlin. The study followed the reading development of around 80 primary school children from

Grade 2 to Grade 4 in a series of experiments targeting both lower-level and higher-level reading processes. The studies in this thesis all investigate the online processing of personal pronouns in children using eye tracking. All studies are essentially designed to investigate whether children resolve pronouns online during reading and how they do so, that is, how the varying information around the pronoun region influences children's eye movement behavior.

The potential influencing factors on children's processing of pronouns were addressed in the studies in varying degrees: (1) features of the text such as informativeness of the pronoun or anaphor, (2) developmental influences related to children's cognitive and reading skill, (3) individual differences in reading fluency or comprehension monitoring, and (4) the relationship of pronoun processing and comprehension. The mapping of the three studies to these research questions is described in the remainder of this chapter, where the general aims of the studies are discussed in turn. For a detailed description of the items and method, please be referred to the respective study in the studies section.

5.1 Study 1: The repeated name penalty effect in children's reading

The first study set out to investigate the possibility that discourse accessibility is fundamentally different for children compared to proficient readers. The study tested the repeated name penalty effect in children using eye tracking. Inspired by Gernsbacher (1989), it hypothesized that children's reading processing may profit from overlapping word forms (i.e. repeated noun phrases) in a text where pronominal anaphors would otherwise occur.

Three-sentence stories containing pronouns and anaphoric noun phrases in the form of repeated names were presented to children in year 3 (around 8 years old). We were interested in the processing differences associated with pronouns in contrast to repeated names. First, we expected these differences to be informative with respect to children's sensitivity to the accessibility hierarchy, meaning their sensitivity to the discourse function of pronouns during reading. The available evidence for the repeated name penalty effect in adults allowed us to investigate whether children have similar

discourse expectations as adults, or whether they have specific needs towards the informativeness of an anaphor, that is whether they profit from a more informative anaphor in the form of a repeated name.

The study investigated processing differences of pronouns and repeated names after the respective anaphors were presented. If children's reading processing is equally slowed down as adults' after the presentation of a repeated name, it can be concluded that children are sensitive to the accessibility hierarchy and subsequently to the discourse function of pronouns. Alternatively, children might not show a penalty effect but instead even speed up reading after a repeated name. This is inspired by Gernsbacher's (1989) original idea that repeated names should facilitate processing by way of direct mapping. Such a finding would not only suggest that children do not make full use of context information, but also that they have specific needs for anaphor resolution during reading that are different from adults'.

5.2 Study 2: Individual differences in children's pronoun processing

As was described in the previous section, the first study in this dissertation asked whether children expect a pronoun in the discourse similar to adult readers. However, it did not ask any questions about the mechanisms of pronoun processing. Strictly speaking, the first study cannot even determine beyond doubt that children resolved the anaphoric pronoun. Therefore, the second study was designed to investigate children's processing of the most prominent source of information on the pronoun itself, its grammatical gender, in a mismatch paradigm. If children realize that there is a gender mismatch between pronoun and antecedent, they have at least attempted to resolve the pronoun towards an antecedent. Their eye movement behavior then reveals whether they also "repair" the inconsistency and how they do so. One of the open research questions concerns individual differences in children's pronoun processing and whether these are connected to reading comprehension skill. Therefore, the replication of the initial study was conducted with a larger sample of children to test for effects of individual reading skill, reading fluency and auditory sentence comprehension on pronoun resolution processes.

In the second study, children in Grade 4 (around 9 years old) read complex sentences

like *Leon/Lisa shoed away the sparrow/the seagull and then he ate the tasty sandwich.* The sentences were constructed such that the male pronoun *he* was a match (*Leon*) or a mismatch (*Lisa*) to the antecedent in subject position. There was also a gender matching or mismatching entity in object position which was designed as a distractor. In a first step, children's eye movement behavior during the reading of matching and mismatching pronouns was monitored. Eye-movement measures were analyzed in the pronoun region and post-pronoun region to shed light on the immediacy of the mismatch effect in children. An immediate recognition of inconsistencies at the pronoun would point to an early use of gender information already at the bonding stage of pronoun resolution.

In the second part of the study, a direct replication with a larger sample of children, we assessed whether children detected a pronoun gender mismatch. This was used as a factor in the analysis of eye tracking measures. Besides aiming at replicating the effects found in the first study, the eye movement behavior of children who resolved the pronoun could be contrasted against the eye movement behavior of children who arguably did not resolve the pronoun. It was of particular interest how children react to a mismatching pronoun in the sentence context: If they attempt to resolve a pronoun online, the mismatching pronoun should initiate repair processes as were observed in other mismatch paradigms. Reading comprehension, reading fluency, and auditory sentence comprehension were also assessed to use as a factor in the analyses as possible determinants of successful pronoun resolution.

5.3 Study 3: Gender cue effects in children's pronoun processing

The third study takes a step towards bridging the gap between processing and comprehension of pronouns by not only focusing on the eye movement measures but also post-reading comprehension questions. The previous studies did not include comprehension questions targeted at the pronoun for methodological reasons: In study 1, comprehension questions were difficult to construct due to the nature of the experimental texts which included only one referent. In study 2, any comprehension questions would have been at odds with the integrity of the experiment because they would have given away the mismatch manipulation. Study 3 therefore used yet a different sentence paradigm to investigate whether a gender cue—meaning disambiguating gender information on

the pronoun, similar to study 2—improves the accuracy of pronoun comprehension offline. Moreover, the third study was conducted in a semi-longitudinal paradigm: We tested 70 children in Grade 3 and again in Grade 4 to investigate effects of age on pronoun processing and comprehension. We further investigated effects of reading skill on pronoun processing and resolution at both Grade levels.

The 70 children read sentences containing pronouns with or without a gender cue such as *Paul envied Tessa because she had a pool at home* (with gender cue) versus *Paul envied Theo because he had a pool at home* (without gender cue). Either the gender cue determines the antecedent or the antecedent has to be inferred from context: It is plausible that Paul envies Theo because *Theo* has a pool at home, rather than Paul himself. Processing measures at the pronoun and after the pronoun were recorded and pronoun comprehension accuracy was determined using comprehension questions such as *Who had a pool at home?* This is how study 3 targeted the relationship of children's pronoun resolution processes and comprehension.

II. EXPERIMENTAL STUDIES

6 THE REPEATED NAME PENALTY EFFECT IN CHILDREN'S NATURAL READING

ABSTRACT

We report data from an eye tracking experiment on the repeated name penalty effect in 9 year-old children and young adults. The repeated name penalty effect is informative for the study of children's reading because it allows conclusions about children's ability to direct attention to discourse-level processing cues during reading. We presented children and adults simple three-sentence stories with a single referent, who was referred to by an anaphor—either a pronoun or a repeated name—downstream in the text. The anaphor was either near or far from the antecedent. We found a repeated name penalty effect in early processing for children as well as adults, suggesting that beginning readers are already susceptible to discourse-level expectations of anaphora during reading. Furthermore, children's reading was more influenced by the distance of anaphor and antecedent than adults', which we attribute to differences in reading fluency and the resulting cognitive load during reading.

This section is identical to the accepted manuscript of the following published article:

Eilers, S. & Tiffin-Richards, S. & Schroeder, S. (2019a). The repeated name penalty effect in children's natural reading. Evidence from eye tracking. *Quarterly Journal of Experimental Psychology, 72*, doi: 10.1177/1747021818757712

Introduction

Reading comprehension processes have been characterized as an online effort to build a mental model of what a text is about. Abstracted from the text surface, readers integrate the content of each new proposition into the existing set of propositions in an incremental way (van Dijk & Kintsch, 1983; Zwaan, Langston, & Graesser, 1995; Zwaan & Radvansky, 1998). Readers therefore need to evaluate how every new proposition fits into the mental model that they have constructed thus far. Protagonists, in particular, serve as anchors for the global coherence of a text. The repeated reference to the same set of discourse entities has been termed "referential continuity" and constitutes a powerful coherence marker (Garnham, Oakhill, & Johnson-Laird, 1982; Givón, 1983). This is important because referential continuity is signalled explicitly on the text level by the use of anaphora; wherever there is a continuity of referents in the text, this is explicitly marked by anaphoric expressions. The fact that readers are sensitive to the fit of referring expressions in the discourse is demonstrated by the repeated name penalty (RNP) effect. In this study, we investigated children's sensitivity to the form of referential expressions in a discourse. The RNP is informative for the study of children's reading because it allows conclusions about children's ability to direct attention to discourse-level processing cues during reading. More precisely, we studied similarities and differences in children's and adults' processing of repeated names and pronouns when they are either near to their antecedent or far from their antecedent in a three-sentence story. Using an eye tracker, we recorded readers' eye movements to obtain a detailed picture of developmental differences in the time course of reading processes when resolving pronouns and repeated names.

The RNP effect

The RNP effect is a well-established finding in the literature on anaphoric processing. The term "repeated name penalty" was coined in the seminal paper by Gordon, Grosz, and Gilliom (1993). In their Experiment 1, they measured young adults' reading times for passages containing a proper name in the subject position of the first sentence, and either pronouns or repeated names in the subject position of three subsequent sentences. They observed decreased reading times for sentences containing pronouns compared with repeated names, while comprehension accuracy was comparable across

conditions. The authors interpret their findings within the Centering theory framework (Grosz, Joshi, & Weinstein, 1983; Grosz, Weinstein, & Joshi, 1995). Centering theory establishes a set of formal rules about the appropriate anaphor for an entity based on its relative prominence in the discourse. Following Centering theory, the most prominent entity in the discourse should be pronominalized. The experimental literature has shown repeatedly that adult readers prefer pronouns as a referent for prominent discourse entities (e.g., Fukumura & van Gompel, 2015).

An alternative explanation for the RNP effect is offered by the Informational Load Hypothesis (ILH; Almor, 1999). The ILH explains the effect in terms of memory interference between representation of the referent in the current situation model and the representation of the referential expression (Almor & Nair, 2007; Peters, Boiteau, & Almor, 2016). Repeated names are semantically rich, as they carry substantive information and associations. Pronouns, in contrast, are semantically uninformative and code only number and gender. Importantly, such an uninformative anaphor is expected when a referent is highly accessible in the discourse, for example, because the referent was mentioned recently or is the only available dis- course entity at present (Ariel, 2001; Kehler, 2002). The general idea behind the ILH was already formulated in Grice's cooperative principle of quantity, that is, to make a contribution as informative as required, but not more informative than required (Grice, 1975). Names are typically used to introduce new referents, which is inconsistent with their use as referring expressions for prominent discourse entities. In other words, repeated names in a phrase where a pronoun could be used clash with readers' discourse model and impede sentence processing because readers assume some added value of the repeated name and spend time trying to integrate superfluous information (Almor & Nair, 2007). In summary, the ILH stresses cognitive access to the referent, whereas Centering theory concentrates on the linguistic features of the discourse to determine the type of anaphor used. However, both make identical predictions regarding the RNP in adults; a repeated name slows discourse processing compared with a pronoun when it refers to a prominent, or accessible, discourse entity. Our aim in this study was to elicit the RNP effect in children using eye tracking measures and compare it with the effect in adults.

Self-paced reading time studies have repeatedly shown that for salient discourse entities, reading times increase when a repeated name is used instead of a pronoun (see

Lezama, 2015, for a review). Using eye tracking, it has further been shown that in contexts where a pronoun is expected, repeated names increase the likelihood of regressive eye movements (Kennison & Gordon, 1997). The RNP may even have effects on aspects of reading comprehension but only in highly skilled readers who make use of the pronoun as a local coherence marker (Shapiro & Milkes, 2004).

More generally, the implication of the RNP is that proficient readers are sensitive to the type of referring expression during reading processing because the type of referring expression chosen in the text is directly linked to discourse coherence. Proficient readers evaluate incoming information not only on the text surface level but also from a discourse representational point of view. The main aim of the current experiment was to investigate the RNP effect in beginning readers because it is currently unclear whether beginning readers use discourse context in a similar way to skilled adult readers.

Development of the RNP effect

For children, the RNP has been studied online in listening comprehension (Megherbi & Ehrlich, 2009). Engelen, Bouwmeester, de Bruin, and Zwaan (2014) studied 6- to 11-year-old children's eye movements in a visual world paradigm while they were listening to a complex story involving multiple characters. The probability that children fixated the target increased after the mention of a proper name, but not a pronoun. They also assessed com- prehension of these texts and found that good comprehenders were more likely to make anticipatory eye movements to the referent of a pronoun than poor comprehenders. The authors discuss the possibility that poor comprehenders lack the ability to make inferences during listening of complex stories. This has been suggested before in studies which assessed comprehension. When complex inferences are required to identify the referent of a pronoun, poor comprehenders among the children failed to name the correct referent in up to one-third of the presented items (Oakhill & Yuill, 1986). It can be concluded that referential processing could be facilitated for children when inferences are not required because of a text-level identity between antecedent and referential expression. This is particularly relevant for children's reading. Children read more slowly and spend more time on single words than adults (Blythe & Joseph, 2011). Their reading can be characterized as more effortful, associated with an

increase in cognitive load. The establishment of prominence in the discourse may be difficult for children, and they may not use discourse-level cues in the building of situation models as efficiently as adults. As the RNP effect has been explained in terms of the accessibility of referents in working memory, we predicted that children should not show the same RNP as adults. On the contrary, beginning readers' processing downstream from the repeated name might be facilitated because when a repeated name is used, antecedent and anaphor can be mapped directly at the word level (Gernsbacher, 1989). Children may profit from the repeated name because the referential expression and referent are identical on the text surface level, unlike the pronoun, which requires a local inference. Pronouns require inferences that span several words in the text, which arguably poses a challenge for beginning readers. We know of only one eye tracking study that focused specifically on children's comprehension of anaphora during reading. Joseph, Bremner, Liversedge, and Nation (2015) had children read short paragraphs in which (a) the distance between antecedent and anaphor and (b) the semantic typicality of the antecedent (typical: a truck—the vehicle, atypical: a crane—the vehicle) were manipulated. Although the authors did not find effects of distance in early online measures, children did make more regressions out of far than near anaphors in their study. We are further aware of two studies which directly tested the RNP effect in children's reading. Ehrlich, Rémond, and Tardieu (1999) conducted a self-paced reading study with 10-year-old children. The children read expository texts with a repeated noun phrase or a pronoun in subject position. The RNP effect was not a main focus in their study, but the authors report elevated reading times for sentences with a pronoun rather than a repeated name. Interestingly, this was particularly true for skilled readers. However, the texts in their study were rather complex and therefore, it may have been particularly difficult for the children to resolve the pronoun in these texts. In a study with 10-year-old German children, Schimke (2015) compared reading times of the verb downstream from the anaphor in sentences such as "John is sitting in the ground and John/he/Ø draws a picture," where the second noun phrase was a noun, a pronoun, or an ellipsis (i.e., it was omitted entirely, which is possible in German). She found a penalty for the repeated name compared with the elliptical subject, but not the pronoun. These results do not directly compare with the existing findings of the RNP in English reviewed above. It is not clear how the discourse integration of a repeated name or a pronoun relates to the discourse integration of an ellipsis. Taken together, previous

studies into children's reading of pronouns and repeated names suggest that children do show some sensitivity to Anaphor Type. However, these studies did not directly compare adults' and children's reading, so developmental differences have not yet been addressed.

The current study

In this article, we report a natural reading experiment with children using a repeated name manipulation of short, three-sentence stories. Children and adults read these para- graphs while their eye movements were recorded. We used a two-factorial design, contrasting pronouns and repeated names in three-sentence discourse contexts where the anaphor was either near or far from the antecedent. We included a distance manipulation based on findings of a prior eye tracking study on children's processing of anaphora during natural reading (Joseph et al., 2015). If it is the case that cognitive load is a relevant factor for children's processing of pronouns and repeated names, we argue that distance of anaphor and antecedent should play a role for the RNP in children. This is because a direct mapping of anaphor and antecedent may be particularly helpful for beginning readers when they are further apart in the story.

The RNP effect typically spills over to the region following the anaphor, which was identical in all conditions in our reading materials. We were therefore particularly interested in two regions: The anaphor itself (anaphor region) and the region directly following the anaphor (post-anaphor region). We analyzed first fixation times and gaze durations to tap into early processing effects in the post-anaphor region. Furthermore, we analyzed total reading times to pick up later effects of Anaphor Type in the post-anaphor region.

We expected to find differences in the way a repeated name affects adults' and children's reading processing which lead to two different sets of hypotheses for children and adults. For the adults, we expected to replicate the RNP effect with our materials using eye tracking. In line with the well-established RNP literature, we hypothesized a processing advantage for pronouns over repeated names. We expected longer first fixation times and gaze durations in the post-anaphor region following a repeated name than following a pronoun. Second, building on previous findings (Kennison & Gordon, 1997), we expected adults to make more regressions from repeated names than pro-

nouns in the anaphor region. Because the items were short and written for a primary school reading-level, we did not expect that distance to the antecedent would induce any difficulty for the adults.

For the children, in contrast to the adults, we hypothesized a processing advantage for repeated names over pronouns. We hypothesized shorter first fixation times and gaze durations after a repeated name than a pronoun in the post-anaphor region. This would suggest that children rely more on surface-level text information during reading (mapping of information) and do not use discourse-level cues for online situation model building as efficiently as adults do. Second, in contrast to the adults, we did not expect children to make more regressions out of repeated names than pronouns. If the pronoun is more difficult for the children to integrate than the repeated name, there should however be more regressions from the post-anaphor region following a pronoun than a repeated name. This hypothesis is the opposite of our expectations for the adults. In addition, for the children, we predicted longer gaze durations in the post-anaphor region after distant pronouns than near pronouns because of the added difficulty of connecting lexical information that spans longer distances of text (Joseph et al., 2015). Finally, if we found an interaction of Anaphor Type and Distance to the antecedent for reading time measures in children, we would expect the repeated name to ease processing of distant anaphors, indicated by shorter first fixation times and gaze durations. Such a finding would imply that distance to the antecedent affects the processing of pronouns and repeated names differentially in children, such that the pronoun is even more difficult to resolve when the ante- cedent is further away in the text. We did not expect such an interaction between Anaphor Type and Distance for the adults.

Method

Participants

We recruited 29 fourth graders from three Berlin schools who took part in two sessions. From these, 23 full datasets were obtained. Five children were excluded because of missing data due to technical issues, and the data from one child were excluded because he had learned German after the age of 6years. Of the remaining 23 children, 9 were

girls. In addition, 25 native German-speaking adults were recruited via university mailing lists.

The children were 9 to 10 years old (M = 9 years, standard deviation (SD) = 15 months). The adults were M = 25.2 years old, SD = 38.5 months, and 17 were women. All participants reported normal or corrected- to-normal vision. The participants completed a standardized reading fluency test (SLRT-II; Moll & Landerl, 2010). Children did not differ from the population mean in either word reading fluency, M = 53.0, SD = 24.4, $t(22) < 1$, p = .56, or non-word reading fluency, M = 55.5, SD = 22.6, $t(22)$ = 1.2, p = .25. Our adult sample did not differ from the population mean in word reading fluency, M = 51.1, SD = 29.9, $t(24) < 1$, p=.86, but was slightly above average in non-word reading fluency, M = 63.0, SD = 29.4, $t(24)$ = 2.13, p = .04. Children additionally completed a standardized reading comprehension test (ELFE; Lenhard & Schneider, 2006). Importantly, our sample did not differ significantly from the population mean on either the word comprehension subscale, M = −0.07, SD = 0.8, $t(22) < 1$, p = .68, or the text comprehension subscale, M = −0.15, SD = 1.0, $t(22) < 1$, p = .46.

Materials

Items

Materials consisted of 52 three-sentence stories. For each of the 52 items, four different stimulus versions were created in which the factors Anaphor Type (repeated name vs pronoun) and Distance (near vs far) were manipulated in a within-item design. The stories comprised 16 to 17 words (89-108 characters) and were structurally similar. The introductory sentence of each story contained a referent and an activity of the referent. The structure of the stories is demonstrated in Table 6.1.

The target sentence with our main regions of interest contained either a personal pronoun (pro) or a repeated name (rpn) to refer to the referent. All target sentences were of the form adverb—verb—subject (anaphor region)—direct object (post-anaphor region). The anaphor was in the middle of each target sentence, which corresponds to standard word order in German. The sentences were kept simple and contained age-appropriate topics for fourth graders. Word frequencies of the direct object in the post-anaphor region were derived from the German children's book corpus childLex

(Schroeder, Würzner, Heister, Geyken, & Kliegl, 2015). The mean normalized lemma frequency of the direct objects was high, *M* = 89.4, *SD* = 104.2.

Table 6.1. Structure of stimulus materials.

Condition	Story	Anaphor	Distance
Introduction (invariable)	Peter steigt aus dem Bett. Peter gets up from his bed.		
1	Sofort macht [**Peter**] [das Frühstück.] Right away, Peter prepares breakfast. Der Rest der Familie schläft noch friedlich. The rest of the family is still fast asleep.	rpn	near
2	Der Rest der Familie schläft noch friedlich. The rest of the family is still fast asleep. Sofort macht [**Peter**] [das Frühstück.] Right away, Peter prepares breakfast.	rpn	far
3	Sofort macht [**er**] [das Frühstück.] Right away, he prepares breakfast. Der Rest der Familie schläft noch friedlich. The rest of the family is still fast asleep.	pro	near
4	Der Rest der Familie schläft noch friedlich. The rest of the family is still fast asleep. Sofort macht [**er**] [das Frühstück.] Right away, he prepares breakfast.	pro	far

Note. The anaphor is written in bold face and square brackets indicate the regions of interest used in analyses; English translations (non-literal) are printed in grey; rpn = repeated name, pro = pronoun.

One extra sentence, designed to lengthen the distance between antecedent and anaphor, appeared either in the middle of the story or in final position. It never introduced a referent that could be confused with the target referent and was plausible within the story in both positions. Prior to the eye tracking experiment, we had 40 children of the same age group as our child sample (mean age *M* = 9.7 years, *SD* = .54, 20 of them girls) rate the items for plausibility and difficulty. The children took part in a 45-min paper-pencil group session, including breaks, at their school. Children were asked to read the stories silently and afterwards rate them on a 4-point scale, where 1 = *very implausible/very difficult to read* and 4 = *very plausible/very easy to read*. Our manipulations did not affect plausibility or difficulty: An analysis of variance (ANOVA) with the factors Anaphor and Distance yielded no significant effects for either comprehension or plausibility, all $F(1, 204) < 1$.

Apparatus

We used an EyeLink 1000 eye tracker (SR Research, Ontario, Canada) to record eye movements during reading at a rate of 1,000Hz. The stories were presented on an ASUS LCD monitor (21′′) with a refresh rate of 120Hz. The stories appeared in the middle of the screen in a 4:3 frame. The stories were presented using SR Research Experiment Builder (SR Research, 2009). All stories appeared continuously in two to three lines, in Courier New, font size 16, using black letters on a white background. Although line breaks occasionally occurred before or after a region of interest, this was then the case for all conditions of that item. Participants were seated at a monitor distance of 62cm in a head-and-chin rest. Recording of the eyes was monocular and only the left eye was tracked.

Procedure

Written informed consent was obtained from the children's parents ahead of the study, and oral consent was obtained from each child prior to testing. Adult participants signed an informed consent form. For each item, one of the story versions was assigned to one of four item lists according to a Latin square design. Participants were assigned to one of the lists based on their order of appearance. Children took part in two sessions at their school. The paper-pencil part of the test was administered in one group session. The individual sessions were conducted in a quiet room that was suitable for eye tracking provided by the school. Adults were tested in the facilities of the Max Planck Institute for Human Development in Berlin.

A 5-point calibration was conducted for each participant until calibration error reached a maximum of 0.5° of visual angle. After the first calibration, participants read three practice stories, each with a following comprehension question. They were instructed to read the stories silently, press a button on a gamepad after having finished reading and answer the comprehension question via button-press. Comprehension questions appeared randomly after 25% of trials. The questions never tapped comprehension of the pronoun but were designed to ensure attentive reading, for example, "Was the family wide awake?" (see Table 6.1).

Analysis

Data were inspected and y-axis drift corrections were applied as necessary using the DataViewer software (SR Research, 2011). Fixations were cleaned automatically using the DataViewer four-stage fixation cleaning: At Stage 1, fixations shorter than 80ms and within 0.5° from the neighboring fixation were merged with each other. At Stage 2, fixations shorter than 40ms and within 1.25° distance were merged with a neighboring fixation. At Stage 3, all interest areas were checked for at least three neighboring fixations of less than 140ms and if found, these were merged. At Stage 4, only fixations between 120 and 1,200ms (for children data) and between 80 and 1,000ms (for adult data) were kept. The cleaning removed about 13% of fixations of the children, and about 16% of fixations of the adults. Finally, before models for the dependent measures were calculated for each eye movement measure, all observations above 2.5 standard deviations from the person or item mean of each dependent measure were deleted from the fixation record (roughly 2% of observations).

We calculated four eye tracking measures for the anaphor and the post-anaphor region: *first fixation time* (duration of the first fixation that falls into the area of interest), *gaze duration* (summed duration of first-pass fixations), *total reading time* (summed fixations in a region), and *regression probability* (the likelihood of a leftward saccade out of a region).

Reading time data were analyzed with linear mixed-effects models and regression probability was analyzed with generalized linear mixed-effects models, using the lme4 package version 1.7 (Bates, Maechler, & Bolker, 2012) in R (R Development Core Team, 2016). We calculated individual models for each region of interest and each dependent variable with Anaphor (repeated name vs pronoun), Distance (near vs far), and Age (child vs adult) as fixed effects, and participants and items as crossed random intercepts.

All reading time measures were log-transformed to achieve a more normal distribution. To ease interpretation, the back-transformed results are reported in milliseconds. The significance of the fixed effects was determined using effects coding and type-II model comparisons in the ANOVA function in the car package (Fox, Friendly, & Weisberg, 2013). Post hoc comparisons were estimated using cell-means coding and single-de-

gree-of-freedom contrasts as implemented in the glht function in the multcomp package (Hothorn et al., 2015).

Results

Global measures

Mean comprehension accuracy for the adults was high, *M* = 97%, *SD* = 18, and slightly lower for children, *M* = 92%, *SD* = 27, but consistently above chance level. Adults and children differed in mean text reading time, which amounted to an averaged *M* = 8.7s, *SD* = 4.2, for children, whereas adults took *M* = 3.9s, *SD* = 1.6, to read the stories. Consequently, we found a large effect of Age group for all our dependent reading time measures (see Table 6.2). As children's reading is characterized by more and longer fixations compared with adults' (Blythe & Joseph, 2011), this was to be expected and we will concentrate on interactions of Age with Anaphor and Distance in the remainder of this article.

Table 6.2. Results of mixed-effects models.

	First fixation time		Gaze duration		Total reading time		Regression probability	
	Anaphor	Anaphor +1	Anaphor	Anaphor +1	Anaphor	Anaphor +1	Anaphor	Anaphor +1
Distance	0.10	12.45***	9.80**	15.00***	0.05	17.31***	9.64**	325.75***
Anaphor	1.62	0.42	121.02***	13.87***	165.10***	1.16	12.14***	13.24***
Age	38.66***	12.80***	71.32***	73.57***	48.75***	71.13***	3.49	3.88*
Distance × Anaphor	3.83	0.07	2.21	1.12	0.02	0.07	0.51	1.88
Distance × Age	3.38	3.94*	9.46**	0.01	3.13	0.53	6.66**	89.49***
Anaphor × Age	0.01	0.53	74.08***	0.86	59.19***	0.29	1.16	0.62
Distance × Anaphor × Age	0.51	1.89	1.35	2.57	0.20	1.33	2.31	1.90

Note. ANOVA: Analysis of variance. F values for first fixation time, gaze duration and total reading time. χ^2 values for regression probability. *p < .05; **p < .01; and ***p < .001.

Regions of interest

We will report our results by region, starting with the post- anaphor region. Note that

we will not report effects of Anaphor for reading time measures in the anaphor region itself because these cannot be separated from word length and frequency of pronouns (short, frequent) and names (longer, less frequent). The model means for all dependent measures can be found in Table 6.3.

Post-anaphor region

There was no effect of Anaphor for first fixation time. For gaze duration, however, there was a main effect of Anaphor such that regions following repeated names took longer to read, M = 527ms, SE = 25, than regions following pronouns, M = 484ms, SE = 23. We found no interaction of Anaphor and Age in gaze duration, indicating an RNP effect for both adults and children. For regression probability, we found a main effect of Anaphor such that both groups were more likely to make regressions out of the post-anaphor region following pronouns, M = 0.37, SE = 0.03, than repeated names, M = 0.29, SE = 0.03.

Table 6.3. Model means for dependent measures.

	First fixation time		Gaze duration		Total reading time		Regression probability	
	Anaphor	Post-anaphor	Anaphor	Post-anaphor	Anaphor	Post-anaphor	Anaphor	Post-anaphor
Adults								
pro								
far	185(5)	195(5)	194(9)	363(23)	213(12)	447(31)	.09(.02)	.77(.04)
near	183(5)	187(5)	193(9)	336(21)	204(12)	427(29)	.04(.01)	.12(.02)
rpn								
far	192(5)	188(5)	198(9)	390(25)	229(13)	466(32)	.12(.02)	.71(.04)
near	183(5)	188(5)	198(9)	356(23)	225(12)	433(30)	.06(.01)	.09(.02)
Children								
pro								
far	217(5)	214(6)	253(11)	725(48)	290(17)	1028(73)	.07(.01)	.47(.05)
near	229(6)	203(5)	267(12)	623(41)	302(17)	922(66)	.09(.02)	.26(.04)
rpn								
far	228(6)	218(6)	336(15)	757(50)	424(24)	1011(72)	.16(.02)	.42(.05)
near	225(6)	200(5)	389(17)	736(49)	437(25)	953(68)	.12(.02)	.14(.03)

Note. Back-transformed, rounded model means for the dependent measures in the anaphor and post-anaphor regions of interest. Standard errors are given in parentheses.

There was further a main effect of Distance in the post-anaphor region, such that gaze durations were longer after anaphora that were far from the antecedent, M = 528ms, SE = 25ms, than those that were near their antecedent, M = 484ms, SE = 23ms. There was

no interaction of Distance and Anaphor in gaze duration. Furthermore, there were early main effects of Distance and an interaction of Distance and Age in first fixation times. Planned contrasts revealed that children showed longer first fixation times in the post-anaphor region after far anaphors, M = 216ms, SE = 5ms, than after near anaphors, M = 201ms, SE = 5ms, t = −3.89, p < .001. In contrast, the effect of Distance for first fixation time in adults was not significant, t = −1.14, p = .25.

Furthermore, there was no effect of Anaphor but a significant main effect of Distance for total reading time: The post-anaphor region following far anaphors was read for longer, M = 682ms, SE = 35ms, than following near anaphors, M = 635ms, SE = 32ms. We further found a large main effect of Distance and an interaction of Distance and Age for regression probability. Although the Distance effect was significant for both adults, t = −19.5, p < .001, and children, t = −8.9, p < .001, post hoc contrasts showed that it was significantly larger for the adults compared with the children, t = 9.5, p < .001.

Two additional analyses were conducted to investigate the unexpected effect of Distance. First, we analyzed regression probability for the region in paragraph-final position, which was the post-anaphor region in the far condition and the last region of the extra sentence in the near condition, respectively. We included Distance and Age as fixed effects, and subjects and items as random effects in a generalized linear mixed-effects model with regression probability as the dependent variable. There was a main effect of Age, t = 22.8, p < .001, but no effect of Distance, t < 1, and importantly no interaction of Distance and Age, t = 1.8, p = .175. In the final region of the paragraph, regression probability was higher for adults, M = 80%, SE = 4%, than children, M = 47%, SE = 6%.

Second, we conducted a similar analysis including first- pass skipping probability on all words preceding the post- anaphor region as dependent variable. The main effects of Distance and Age were not significant, both t < 1, but there was an interaction of Distance and Age, t = 15.4, p < .001, such that adults' first-pass skipping probability was higher in the far condition when there was more text between anaphor and referent, M = 0.08, SE = 0.01, than in the near condition, M = 0.07, SE = 0.01, t = −2.7, p < .01, whereas for children, it was higher in the near condition, M = 0.08, SE = 0.01, than in the far condition, M = 0.06, SE = 0.01, t = 2.9, p < .01.

These two complementary analyses suggest two differences in the overall patterns of

eye movements between children and adults which were unrelated to the anaphor manipulation. In contrast to children, adults showed a greater increase in regression probability in the final region of the paragraph regardless of where the anaphor was positioned, as well as slightly increased skipping probability when more text was between referent and anaphor.

Anaphor region

There was a main effect of Anaphor for regression probability in the anaphor region, such that regressions were more likely from repeated names, M = 11%, SE = 1%, than pronouns, M = 7%, SE = 0.1%. There was no interaction of Anaphor and Age for regression probability, which suggests that the effect of Anaphor was the same for adults and children.

We further found effects of Distance in the anaphor region. For gaze durations, planned contrasts revealed that Distance had an effect on children's gaze durations, $t = 4.21$, $p < .0001$, but not adults', $t < 1$, $p = .90$. Children read near anaphors, M=322ms, SE=13ms, slower than far anaphors, M = 291 ms, SE = 12 ms, independent from Anaphor Type. There were no effects of Distance, nor interactions of Distance and Age, in first fixation time or total reading time.

For regression probability, we found a main effect of Distance and an interaction of Distance with Age. Post hoc contrasts revealed that the Distance effect in regression probability was driven solely by the adults. There was a simple main effect of Distance in adults, $t = -4.0$, $p < .001$, but not in children, $t < 1$, $p = .77$, such that adults made more regressions when the anaphor was far from its antecedent than when it was close to the antecedent.

Finally, higher skipping rates for pronouns than repeated names were expected in the anaphor region, as pronouns are often skipped during reading (e.g., Drieghe, Desmet, & Brysbaert, 2007). We found that pronouns were skipped more often than repeated names by adults and children. Adults had a skipping rate of 40% (SD = 49%) for pronouns and 21% (SD = 41%) for repeated names. Children skipped considerably less, with a skipping rate of 16% (SD = 37%) for pronouns and 4% (SD = 19%) for repeated names. A generalized linear mixed-effects model over skipping rate in the anaphor region showed significant main effects of Anaphor, $t = 51.0$, $p < .001$, and Age, $t = 18.0$, $p < .001$.

Discussion

We conducted an eye tracking experiment to compare the RNP effect in children's and adults' natural reading. Both groups read short, three-sentence stories with a single, salient discourse entity introduced in the first sentence. We used a two-factorial design varying Anaphor Type (repeated name vs pronoun) and Distance to the antecedent (near vs far). The aim was to replicate the RNP for adults using eye tracking and contrast adults' and children's sentence processing. Our hypotheses for the adults based on the existing literature were that they show longer reading times after a repeated name than a pronoun, and more regressions from repeated names than pronouns. For the children, in contrast, we expected to see longer reading times for regions following pronouns than repeated names because pronouns add the necessity to connect information across several words in the paragraph for local inference. We expected children to benefit from a repetition of text surface information during reading processing. Surprisingly, we saw more similarities than differences in children's and adults' processing of the paragraphs, and conclude that 9-year-old children already show sensitivity to discourse-level information during text reading.

The RNP effect manifested in children and adults as longer gaze durations in the post-anaphor region after the repeated name compared with after the pronoun, indicating increased integration difficulty when the name is repeated (Lezama, 2015). We observed more regressions directly out of repeated names than pronouns for both age groups. In line with effects reported by Kennison and Gordon (1997), the repeated name induced more regressions as soon as it was encountered. This suggests that adults and children initiate immediate repair strategies when faced with unexpected information at the discourse-level. Importantly, our results suggest that children, despite the fact that their reading is much slower and more effortful than adults', already anticipate the appropriate form of discourse referent during reading. Both age groups seem to expect a pronoun when the referent is salient in the text, and their processing is disrupted when a repeated name is used, that is, when textual information clashes with discourse expectation. Note that since we are comparing a pronoun with a repeated name, the regression behavior in the anaphor region may to some extent be driven by lexical characteristics of the anaphor. In research designs comparing pronouns and names, it is not possible to control for lexical features. Length and frequency are not the

only lexical differences between pronouns and names. As we have discussed in the introduction, the ILH essentially assumes that the RNP is a result of the semantic richness of content words (repeated name) compared with function words (pronoun). Put differently, the ILH predicts that the processing difference for pronouns and repeated names follows from lexical characteristics, which are a defining feature of the two types of anaphor.

Although distance to the antecedent clearly had a detrimental effect on anaphor processing, distance did not modulate the RNP. With only one referent in the discourse, the extra sentence between antecedent and anaphor may not have introduced sufficient intervening linguistic material to license the use of a repeated name. Presumably, readers still anticipated a recurrent entity and the appropriate pronominal reference. Further studies may want to investigate whether children and adults process a repeated name similarly in contexts where there is more than one referent present.

We expected an interaction of Anaphor Type and Distance to the antecedent for the children and no effect of distance for the adults. Instead, we found that Distance to the antecedent influenced processing independently from Anaphor Type for adults and children. First, we found an early effect of distance to the anaphor for children's first fixation time in the post-anaphor region. Given that the lexical content of the post-anaphor region does not differ between the conditions, and given that the effect is too early to reflect integration effort in the post-anaphor region, we interpret it as a spillover effect from the anaphor region. Longer first fixation durations in children may reflect an increased processing load for referring expressions when their antecedent is further away. Converging effects emerged for both age groups in gaze duration and total reading time, which may indicate that adults and children need more effort to integrate anaphors that are far from their antecedent. The fact that we did not find an interaction of Distance to the antecedent and Anaphor Type in the two early measures suggests that it is not pronoun resolution or the repeated name in particular, but more generally the integration of a distant referring expression which leads to delayed processing in readers of both age groups. In line with this interpretation, total reading times were longer in the post-anaphor region when these were far from their antecedents. This suggests that children and adults take more time to integrate anaphors when these are further away from their antecedents. This finding is generally consistent with prior

work on the processing of near and distant typical and atypical anaphors in children (Joseph et al., 2015).

We found that both age groups made more regressions from the post-anaphor region when it is further away from its antecedent. This distance effect was stronger for the adults than the children, which is unexpected and contradicts our initial hypothesis. Because the Distance effect did not interact with Anaphor Type, we assume that it is independent from Anaphor Type and therefore not attributable to the RNP. The fact that the effect was stronger for adults than children and surfaced in late processing measures puts its connection to processing difficulty into question. Two options will be considered here which may explain the unexpected effect of the distance between referent and anaphor in regression probability. It has been shown that in skilled readers, regressions are more likely from sentence-final regions than mid-sentence regions (Rayner et al., 2000). Note that in the far condition, the whole paragraph has been presented when readers reach the post-anaphor region and therefore the likelihood to regress to earlier sections may be greater. The results from the exploratory analysis of regression probability from final regions suggest that the position of the region in the paragraph is a critical determining factor for regression probability in adults. Moreover, in the far condition, adults were more likely to skip a word on first-pass and revisit it in a second pass. The high skipping rate of the anaphor region for adults in particular may be explained by the high predictability of the referring expression in the given discourse (e.g., Drieghe, Brysbaert, Desmet, & De Baecke, 2004). Finally, there was a main effect of Anaphor Type on regression probability in the post-anaphor region which went in the opposite direction from the effect in the anaphor region. This finding directly contradicts our hypotheses, as we have predicted more regressions after the repeated name than the pronoun. Note, however, that we found more regressions immediately in the anaphor region, which we have interpreted in terms of an RNP. As pronouns require local inferences, readers may make regressions to allow for additional processing time following a pronoun. It is a partial limitation of this study that we cannot fully disentangle wrap-up effects from effects that are purely related to anaphor processing. Future studies in this direction may want to use sentences with additional linguistic material between the pronoun and the sentence-final region to enable a better distinction between anaphor processing and end-of-sentence effects.

Taken together, we can conclude from the results of our study that adults' and children's processing of pronouns and repeated names is more similar than expected. We replicated the RNP using eye tracking with adults and saw generally similar anaphor type effects in 9-year- old children. The finding of an RNP in children suggests sensitivity to discourse-level information during online reading processing in beginning readers.

7 INDIVIDUAL DIFFERENCES IN CHILDREN'S PRONOUN PROCESSING DURING READING

Abstract

In two eye tracking experiments, we tested fourth graders' and adults' sensitivity to gender feature mismatches during reading of pronouns, and their susceptibility to interference of feature-matching entities in the sentence. In Experiment 1, we showed children and adults two-phrase sentences such as "Leon{m}/Lisa{f} shooed away the sparrow{m}/the seagull{f} and then he{m} ate that tasty sandwich". Eye tracking measures showed no qualitative differences between children's and adults' processing of the pronouns. Both age groups showed longer gaze durations on subject mismatching than matching pronouns and there was no evidence of interference of a gender-matching object. Strikingly, in contrast to the adults, not all fourth graders reported detection of the subject gender mismatch. In Experiment 2, we replicated earlier results with a larger sample of children (*N* = 75), and found that only half of the fourth graders detected the gender mismatch during reading. The detectors' reading pattern at the pronoun differed from that of the non-detectors: Children who reported detection of the mismatch showed a reading pattern more similar to the adults. Children who did not report detection of the mismatch had comparably slower gaze durations, and were less likely to make regressions directly at the pronoun. We conclude that children who read more fluently use their available processing resources to immediately repair grammatical inconsistencies encountered in a text.

This section is identical to the accepted manuscript of the following published article:

Eilers, S. & Tiffin-Richards, S. & Schroeder, S. (2018). Individual differences in children's pronoun processing during reading: Detection of incongruence is associated with higher reading fluency and more regressions. *Journal of Experimental Child Psychology, 173*, doi: 10.1016/j.jecp.2018.04.005

Introduction

Reading is a complex task which involves not only word decoding, but also linking pieces of information across longer text passages. Proficient readers use coherence markers to integrate new information into the current situation model (Zwaan & Radvansky, 1998). Pronouns and other anaphora are important markers for coherence because they link entities across sentences and, thus, serve as a cue for the way phrases are interconnected (Ariel, 2004). Online processing and integration of pronouns, therefore, is important for situation model building and ultimately for text comprehension (Garnham, Oakhill, & Johnson-Laird, 1982). However, pronouns are semantically underspecified because they carry only number and gender information. Therefore, they often have to be inferred based on the sentence context (Kehler, 2002; Kehler, Kertz, Rohde, & Elman, 2008).

In the current study, we investigate whether the online processing of pronominal gender information is a possible source of reading difficulty for children. In Experiment 1, we tested adults' and children's sensitivity to gender feature mismatches on the pronoun and their susceptibility to interference effects when a gender-matching object is present. In Experiment 2, we compared the online pronoun processing of children who reported detection of the gender mismatch to those who did not report detection. We were interested in the eye movement patterns associated with the report of mismatch detection, and inter-individual differences that may contribute to successful mismatch detection in children.

Online pronoun resolution in proficient readers

Proficient readers infer the antecedent of a pronoun online by combining lexical information (e.g., gender of the pronoun) and contextual information (e.g., verb meaning, disambiguating sentence information). In a self-paced reading experiment, Garnham and Oakhill (1985) showed that adults need more time for the integration of a subclause when there is no gender cue on the pronoun and the antecedent needs to be inferred entirely from context. This shows that readers use the gender cue on the pronoun for resolution online during reading. Gender mismatches of pronoun and antecedent,

therefore, should disrupt the reading process. In a self-paced reading study with proficient readers, Carreiras, Garnham, Oakhill, and Cain (1996) found longer reading times for the last sentence of a story when it contained a mismatching pronoun for the stereotypical gender of a referent (i.e., female for nurse, male for doctor). They concluded that adults use gender information as soon as it becomes available, and their results show that adult readers form expectations for the gender of a pronoun, such that gender mismatches result in longer processing times. In a study with a stronger manipulation, Rigalleau, Caplan, and Baudiffier (2004) presented adults with sentences such as "Wendy complimented Nancy because she/he made an effort" and found that reading times on the subordinate clause were significantly longer when the pronoun did not match the two antecedents. Their results also show that the gender feature is a strong determinant for the identification of an antecedent. When there is no available gender-matching antecedent, processing of the pronoun is made difficult to a point where a proficient reader does not engage in resolution at all even if enough context information is available to infer the correct antecedent. In the experiment by Rigalleau et al., however, response accuracy on the comprehension questions of the gender-mismatch sentences was equal to that of the gender-match sentences. However, response latencies were significantly longer for the mismatch sentences. This suggests that although readers understood the sentences despite the mismatching pronoun, a gender-mismatching pronoun has both immediate and substantial effects on the efficiency of sentence comprehension.

Children's comprehension and processing of anaphora

Pronoun processing and comprehension have been studied extensively in very young children using listening tasks. Although this line of research has produced mixed results due to a variety of methods, materials, age groups, and languages studied (Hickmann, Schimke, & Colonna, 2015), many studies provide evidence that children can use gender cues effectively to establish pronoun–antecedent relationships during listening from 3 years of age (e.g., Arnold, Brown-Schmidt, & Trueswell, 2007; see also Sekerina, 2015, for a review). Therefore, we have reason to assume that by the time children reach the end of primary school, they are able to identify the correct referent for a pronoun online in spoken language. However, this is not the case for reading. It has repeatedly been

shown that children struggle to name the correct referent for pronouns when reading a text. Yuill and Oakhill (1988) tested 7- and 8-year-olds' comprehension of sentences containing a pronoun and one or two gender-matching antecedents. Depending on the difficulty of the pronominal inference, children had an error rate of up to 28%. In a similar experiment, Oakhill and Yuill (1986) showed that 7- and 8-year-olds have difficulties in naming the correct referent for the personal pronoun after having read sentences such as "Peter lent ten pence to Tom [Liz] because he [she] was very poor." The children performed significantly worse when there was no gender cue (16–27% error rate) compared with when there was a gender cue (2–14% error rate). These studies show that pronoun resolution is difficult for children and that they do rely on gender cues for comprehension. However, studies targeting comprehension cannot clarify whether children use the cue spontaneously online or whether they use it for offline comprehension only when prompted by a question.

It is conceivable that children's online processing of pronouns remains ''shallow" because reading is more effortful for children in general, given that their word decoding is less automatized than that of proficient readers (e.g., Gagl, Hawelka, & Wimmer, 2015). Contrasting the reading process in adults and children using eye tracking, it has been shown that children make more and longer fixations than adults and engage in more unselective rereading (for reviews, see Blythe & Joseph, 2011; Schroeder, Hyönä, & Liversedge, 2015). In studies of inter-individual differences in children's reading, slow decoding has been associated with poor comprehension (Nation, 2005; Nation & Snowling, 1998). Children may struggle in particular when they need to integrate information that spans longer distances of text online. In a seminal eye tracking experiment with 10-year-olds, Joseph, Bremner, Liversedge, and Nation (2015) compared the processing of anaphora with typical antecedents (a truck–the vehicle) and atypical antecedents (a crane–the vehicle) in short texts where the anaphor (the vehicle) was either near or far from its antecedent. The authors showed that when resolution is most difficult (i.e., atypical anaphors that are far from the antecedent), children might not resolve the anaphor during reading at all. In other words, in the difficult condition, children might not understand that "the crane" and "the vehicle" refer to the same entity in their situation model, resulting in impoverished comprehension. Studies on the resolution of pronouns in children are still rare. Recent results from a study contrasting pronouns with repeated names indicate that 8-year-olds already show a repeated name penalty

effect during online reading. This suggests that beginning readers show the same discourse-level expectations for pronouns as adults (Eilers, Tiffin-Richards, & Schroeder, 2018). However, it is still unclear how children resolve the pronoun, specifically which type of information they take into account and whether resolution happens online in children's reading.

Because there are large inter-individual differences in children's reading comprehension and related component skills (e.g., Cain, Oakhill, & Bryant, 2004; Oakhill, 1982, 1984; see also Nation, 2005, for a review), children may differ in their ability to resolve pronouns online. A recent reading comprehension study with children indeed demonstrated that the ability to resolve pronouns correctly may account for individual variance in reading comprehension (Elbro, Oakhill, Megherbi, & Seigneuric, 2017).

As one of few studies targeting online processing of pronouns in children, Ehrlich, Rémond, and Tardieu (1999) investigated 10-year-olds' resolution of pronouns in expository text passages. They conducted a self-paced reading experiment in which the children could decide to reread previous parts of the text via button press. The authors demonstrated that less skilled comprehenders—that is, children who read as fluently as their peers but perform poorly in a standard reading comprehension test—struggled when encountering personal pronouns as opposed to repeated names in expository texts. Reading times were longer for less skilled comprehenders than for skilled comprehenders over- all; however, reading times on sentences containing a pronoun were longer for skilled comprehenders but not for less skilled comprehenders. The skilled comprehenders further chose to reread earlier parts of the text more often than the less skilled comprehenders. Ehrlich et al.'s study demonstrates that skilled readers among the children may display reading patterns that are qualitatively different from those of less skilled readers. However, because the study did not address the time course of the resolution process, it is unclear where the specific problems of less skilled comprehenders arise during pronoun processing.

In an early eye tracking study with 8-year-old children, Murray and Kennedy (1988) further showed that the eye movement behavior associated with pronoun resolution in skilled comprehenders differs from that in less skilled comprehenders. Skilled comprehenders made regressions more selectively when reading sentences containing pronouns, whereas less skilled comprehenders made shorter, less selective regressions

(termed “backtracking” by the authors). The study suggests that selective regressions in children may be associated with better comprehension. In addition, if children use gender information for the association of pronoun and antecedent online, they may be distracted by an interfering gender-matching antecedent. In German, all nouns carry linguistic gender (for an overview, see Fagan, 2009). Linguistic gender is indicated by its preceding article, der for male nouns (e.g., *der Spatz* [the sparrow], *der Brief* [the letter]) and die for female nouns (e.g., *die Möwe* [the seagull], *die Karte* [the card]). Reference to a mouse in German requires the personal pronoun *sie*, whereas reference to a hamster requires the pronoun *er*. The gender of a noun is internalized early during language acquisition in German. A number of studies have shown that pronoun processing may be inhibited by interfering gender–feature matching discourse entities in proficient readers even when these interfering entities are excluded as the antecedent of the pronoun for structural reasons (see Jäger, Engelmann, & Vasishth, 2017, for a review). This shows that the gender feature is a strong determinant for the association of pronoun and antecedent in proficient readers. Children have been shown to resort to shallow processing when confronted with reading material that is difficult for them (e.g., Joseph et al., 2015). For beginning readers, sentences with a mismatching pronoun may be particularly misleading when there is a gender-matching distractor present even if this distractor does not match the pronoun based on sentence context. In other words, children’s reading may be sufficiently shallow to allow processing of a mismatching pronoun when the distractor is a gender match for the pronoun.

In summary, our interests in the current study were (a) whether children process mismatching pro- nouns similarly to adults and the efficiency of their processing, (b) whether children are susceptible to interference effects by a gender-matching distractor, and (c) differences in pronoun processing in children who reported detection of the mismatch compared with children who did not report detection of the mismatch.

Experiment 1

Rationale

We studied children’s and adults’ processing of pronouns in a mismatch paradigm with

gender-mismatching and gender-matching pronouns and an intervening discourse entity of the same or different gender. Our main interests were effects in the pronoun region and the sentence-final region, that is, the region directly following the pronoun. We used gaze duration in the pronoun region as an early indicator of processing difficulty, and regression probability and regression path duration as late indicators of processing difficulty, associated with "repair" processes. We also report total reading time, a measure that incorporates gaze duration and rereading time. In the final region, we were interested in the integration processes typically found at the end of a sentence. Effects in this region may be related to pronoun resolution if resolution processes affect regions downstream from the anaphor. More important, however, readers are expected to engage in sentence-final meaning integration. This process is informative in mismatch paradigms because it allows conclusions about whether readers repair local inconsistencies online or whether such inconsistencies disrupt their ability to integrate sentence meaning.

We presented two-phrase sentences, such as "Leon/Lisa shooed away the sparrow/the seagull and then he ate the tasty sandwich," to children and adults. In all stimuli, the subject of the first clause (Leon) was the contextually plausible antecedent for the pronominal subject of the second clause (he). In the example above, he must be co-referential with Leon. The gender–feature mismatch prohibits bonding of the pronoun with Lisa (subject gender mismatch). In the pronoun region, we expected longer gaze durations for gender-mismatching pronouns than for gender-matching pronouns for the adults. We further hypothesized that adults would initiate repair processes at the mismatching pronoun, such that regression probability increases with mismatching pronouns and regression path duration is prolonged. For the children, we also hypothesized longer gaze durations for mismatching pronouns than for matching pronouns, indicating mismatch detection. However, for children the effect may be spatially delayed and occur only in the region following the pronoun when meaning integration is expected. Such a delayed effect would be consistent with earlier findings for children's processing of implausible thematic relations (Joseph et al., 2008). Moreover, we hypothesized that children differ from adults in later processing stages. Integration of a mismatching pro- noun requires repair of the inconsistent information, and beginning readers might not be able to engage in online repair due to cognitive resource constraints.

In this study, we were also interested in whether children's processing of a pronoun may be influenced by a second gender-matching discourse entity. We expected that the gender-matching object is a potential source of confusion for pronoun processing during reading. We hypothesized that in children the effect of a mismatching subject may be modulated by an object in the sentence that is a gen- der match to the pronoun (the sparrow{m} vs. the seagull{f}). Such confusion of the appropriate antecedents during processing would indicate that children rely heavily on gender cues in pronoun integration, arguably because gender information is available directly at the word level, which makes it easier to integrate than context information.

Method

Participants

In Experiment 1, we recruited 29 children from three schools in Berlin. Of these 29 children, 5 were excluded because of missing data due to technical issues, and data from 1 child were excluded because the child was a late immigrant (i.e., arrived in Germany after 5 years of age). Of the remaining 23 children, 2 were early immigrants (arrived in Germany before age 5). All participating children received reading instruction in German only. Of the participating children, 17 were girls. The children were 9 years old (*SD* = 15 months) on average, and all attended fourth grade. All children had normal or corrected-to-normal vision. We further recruited 25 young adults (M_{age} = 25.24 years, *SD* = 3.2) from universities in Berlin via mailing lists. Of these adults, 17 were women. All adults were native speakers of German and reported normal or corrected-to-normal vision.

To verify that the samples did not differ from typically developed readers in their age group, we administered a reading fluency test and a reading comprehension test. All participants completed the SLRT-II standardized test of reading fluency (Moll & Landerl, 2010). Children did not differ from the population mean standard score of 50 in word reading fluency (M = 53.0, SD = 24.4), $t(22) < 1$, p = .56; neither did the adults (M = 51.1, SD = 29.9), $t(24) < 1$, p = .86. Children further completed the ELFE 1–6 standardized reading comprehension test (Lenhard & Schneider, 2006). Because of technical difficulties during data collection, the sentence comprehension subtest scores and

summed z scores could not be calculated. The obtained z scores, however, indicate that the children did not differ significantly from the population mean in the word comprehension subtest (M = 0.07, SD = 0.80), $t(22) < 1$, p = .68, or text comprehension subtest (M = 0.15, SD = 1.0), $t(22) < 1$, p = .46.

Table 7.1. Two examples of stimulus materials in the four conditions.

					Region of interest		
					Pronoun	Final	Condition
Leon{m}	verjagte	den Spatz{m}	und dann	aß	[er das]	[leckere Brötchen.]	
Leon	*shooed away*	*the sparrow*	*and then*	*ate*	*he the*	*tasty sandwich.*	match -
Max{m}	schieb	den Brief{m}	und dann	lief	[er zur]	[nächsten Post.]	match
Max	*wrote*	*the letter*	*and then*	*ran*	*he to-the*	*the next post office.*	
Lisa{f}	verjagte	den Spatz {m}	und dann	aß	[er das]	[leckere Brötchen.]	
Lisa	shooed away	the sparrow	and then	ate	he the	tasty sandwich.	mismatch
Mia{f}	schieb	den Brief{m}	und dann	lief	[er zur]	[nächsten Post.]	- match
Mia	*wrote*	*the letter*	*and then*	*ran*	*he to-the*	*the next post office.*	
Leon{m}	verjagte	die Möwe{f}	und dann	aß	[er das]	[leckere Brötchen.]	
Leon	*shooed away*	*the seagull*	*and then*	*ate*	*he the*	*tasty sandwich.*	match –
Max{m}	schieb	die Karte{f}	und dann	lief	[er zur]	[nächsten Post.]	mismatch
Max	*wrote*	*the card*	*and then*	*ran*	*he to-the*	*the next post office.*	
Lisa{f}	verjagte	die Möwe{f}	und dann	aß	[er das]	[leckere Brötchen.]	mismatch
Lisa	*shooed away*	*the seagull*	*and then*	*ate*	*he the*	*tasty sandwich.*	- mis-
Mia{f}	schieb	die Karte{f}	und dann	lief	[er zur]	[nächsten Post.]	match
Mia	*wrote*	*the card*	*and then*	*ran*	*he to-the*	*the next post office.*	

Note. Literal English translations are given to ease interpretation. Brackets indicate regions of interest for analysis. {m}, male gender; {f}, female gender.

Materials

Materials consisted of 48 sentences as depicted in Table 7.1. The sentences were written in 4 experimental versions, resulting in 192 stimuli. All items were written specifically for primary school children and contained concepts to which they could be expected to relate. All items consisted of a complex clause composed of a main clause and a subordinate clause. The main clause always introduced a boy or girl by name in subject position and either an animal or an artifact in object position. The protagonists in subject position could be of either male or female gender (e.g., *Leon* vs *Lisa*, *Max* vs *Mia*; see below). The animal or artifact in object position could be of either male or female linguistic gender (e.g., *der Spatz* vs *die Möwe*, *der Brief* vs *die Karte*).

The sentences consisted of 11 or 12 words, with 2 words in each region of interest (see Table 7.1). The pronoun region contained a pronoun followed by a preposition or an

article to increase the likelihood of a fixation in this region. Note that the subordinate clause was identical across conditions. Importantly, it always used the male singular pronoun (German *er*; English *he*) to refer to the subject of the main clause. This resulted in two subject gender-mismatch conditions and two subject gender- match conditions. We avoided the female singular pronoun (German *sie*; English *she*) because it is identical to the plural pronoun in German (German *sie*; English *they*), which would have resulted in resolution ambiguity. The materials further contained an interference manipulation. Although it is possible to bind the pronoun to the object in the mismatch-match condition grammatically, it cannot be resolved in this manner in the given context. Note that in all items, the subject of the main clause was the only contextually plausible referent for the pronoun.

To ensure that all children knew the names in subject position, we drew these from a ranking of children's names for 2006, the birth year of the children in our sample (Gesellschaft für Deutsche Sprache, 2006). We selected the 48 most frequent names for girls and 48 most frequent names for boys, excluding unisex names. The boys' names in the subject match conditions and the girls' names in the subject mismatch conditions were matched for length. Each group had a mean length of 5 letters (SD = 1.7). To ensure that all children were familiar with the direct object of the main clause, objects were chosen from the childLex corpus (Schroeder, Würzner, Heister, Geyken, & Kliegl, 2015). All objects had a normalized lemma frequency larger than 5 occurrences per million. An omnibus analysis of variance (ANOVA) showed that there was no significant difference, $F(1,91) = 0.29$, $p = .594$, between the frequency of the male objects ($M = 43$, $SD = 54$) and that of the female objects ($M = 35$, $SD = 39$). Furthermore, the length of male objects was identical to the length of female objects, with both having a mean length of 6 letters ($SD = 2$).

Procedure

We used an EyeLink 1000 eye tracker (SR Research, Ottawa, Ontario, Canada) to record eye movements at a sampling rate of 1000 Hz. The stimulus sentences were presented on an ASUS 21-inch LCD monitor with a refresh rate of 120Hz. Participants were seated at a monitor distance of 62cm in a head-and-chin rest to minimize head movements. Sentences were presented using the SR Research Experiment Builder (SR Research,

2009). All sentences appeared in one or two lines in Courier New font size 16. Four item lists were created to ensure that every participant read each item in only one of the four conditions. For each of the 48 items, a different sentence version was assigned to each of the four lists, and participants were assigned one of the four experimental lists in the order of their attendance.

For the children, testing took place during school hours. The paper-and-pencil tests were administered in one group session in a quiet room. The individual sessions were conducted in a separate quiet room provided by the school. Written informed consent was obtained from the children's parents, and oral consent was obtained from each child prior to testing. For the adults, testing took place in the lab facilities of the Max Planck Institute for Human Development. Adults were tested in a single session of about 60min, with written informed consent. The study was approved by the ethics committee of the Max Planck Institute for Human Development.

A five-point calibration procedure (a moving black dot on a white backdrop) was conducted and validated until calibration accuracy reached at least 0.5° of visual angle. The calibration routine was repeated every 25 trials or after head movements. Tracking was monocular. The right eye was tracked unless tracking of the left eye considerably improved calibration. After the first calibration, the participants were presented with two practice items, each followed by a comprehension question. Participants were instructed to read the sentences silently before pushing a button on a gamepad. Upon button press, a comprehension question appeared to ensure attentive reading. The questions did not direct participants' attention to the pronoun and were designed to be answerable in all experimental conditions (e.g., ''Did they run to the bank?", ''Was it a tasty sandwich?"; cf. Table 7.1). Comprehension questions appeared randomly after 25% of trials. Along with the 48 target items, we presented 52 unrelated, structurally dissimilar fillers from a different experiment. The fillers did not contain any mismatches. After the experiment, the participants were debriefed. Those participants who did not spontaneously report detection of the mismatch during the experiment were informed about the mismatch and asked whether they had noticed it during reading. All adults had noticed the mismatch during reading, whereas roughly half of the children had noticed.

Analysis

Data were inspected and cleaned for children and adults separately using SR Research Data Viewer (SR Research, 2011). Fixations were cleaned automatically using the Data Viewer's implemented four-stage fixation cleaning. In a first step, short fixations with a maximum distance of 0.5° from a neighboring fixation were merged. In a second step, remaining fixations were merged with a neighboring fixation if they were shorter than 40ms and within 1.25° distance. In a third step, all regions were checked for at least three neighboring fixations of less than 140ms. If regions that matched this condition were found, the respective fixations were merged. In the fourth step, fixations outside a 120- to 1200-ms threshold for children, or an 80- to 1000-ms threshold for adults, were deleted from the fixation record. This step removed 7.5% of fixations of the children and 6.4% of fixations of the adults. Lastly, observations above 2.5 standard deviations from the person and item means of each dependent reading time measure were deleted for adults and children separately. For each measure, this affected roughly 2.5% of the data.

Reading time data were analyzed with linear mixed-effects models, and regression probability was analyzed with generalized linear mixed-effects models, using the lme4 package (Version 1.7) (Bates, Maechler, & Bolker, 2012) in R (R Development Core Team, 2016). Separate models were calculated for two regions: the pronoun region (always *er* and a following function word) and the sentence-final region (always a modifier and a noun). Each model included Subject Gender (Subject Gender Match vs Subject Gender Mismatch), Object Gender (Object Gender Match vs Object Gender Mismatch), and Age (Child vs Adult) as fixed effects and included participants and items as crossed random intercepts. For these regions, we calculated four reading measures: gaze duration (summed duration of first-pass fixations), total reading time (summed fixations in a region), regression path duration (fixations in a region, including regressions, until the region is left to the right for the first time, also called gopast time), and regression probability (percentage of saccades leaving the region to the left). All reading time measures were log-transformed to achieve a near-normal distribution. Note that the back- transformed model results are reported in milliseconds. Effect coding and Type II model comparisons were used to determine the significance of the fixed effects using

the Anova function of the car package (Fox, Friendly, & Weisberg, 2013). Post hoc comparisons were estimated using cell means coding and single degree of freedom contrasts as implemented in the glht function from the multcomp package (Hothorn, Bretz, & Westfall, 2008).

Results and Discussion

Global results

A 2 × 2 ANOVA with Subject Gender and Age as crossed factors showed that comprehension accuracy was high in both adults (M = 94%, SD = 23) and children (M = 88%, SD = 33), whereas adults were more accurate on average, $F(1, 92) = 8.0$, $p < .05$. The difference in comprehension accuracy for subject gender match and subject gender mismatch was not significant, $F(1, 92) = 3.8$, $p = .054$, nor was the Subject Gender × Age interaction, $F(1, 92) = 0.7$, $p = .416$.

Analyses of eye tracking measures on the trial level showed that adults made fewer fixations (M = 19, SD = 9) than children (M = 33, SD = 19) and, therefore, had shorter mean trial reading times (M = 2941ms, SD = 1261) than children (M = 5782ms, SD = 2698). As a consequence, we found large main effects of age for all dependent reading time measures and so concentrate on interactions of age with subject gender and object gender, respectively. In the following paragraphs, we report results from the analyses in the two regions of interest. The observed means for the regions are reported in Table 7.2, and results from the linear mixed-effects models are summarized in Table 7.3.

Table 7.2. Experiment 1: Observed means of eye tracking measures.

	Pronoun region				Final region			
	Subject male		Subject female		Subject male		Subject female	
Age group	Obj female	Obj male	Obj female	Obj male	Obj female	Obj male	Obj female	Obj male
				Gaze duration				
Children	612 (408)	587 (401)	647 (432)	625 (439)	918 (495)	888 (497)	879 (517)	884 (462)
Adults	290 (150)	293 (146)	309 (156)	322 (169)	427 (179)	413 (194)	432 (169)	422 (185)
				Total reading time				
Children	809 (535)	831 (570)	883 (568)	881 (553)	1100 (483)	1071 (540)	1115 (540)	1137 (527)
Adults	384 (226)	413 (233)	441 (262)	461 (274)	587 (279)	608 (271)	596 (244)	592 (254)
				Regression path duration				
Children	826 (622)	872 (685)	981 (432)	898 (672)	1268 (565)	1264 (653)	1367 (517)	1352 (610)
Adults	419 (347)	444 (364)	458 (348)	462 (359)	851 (370)	841 (353)	933 (347)	941 (400)
				Regression probability				
Children	.13 (.34)	.17 (.38)	.22 (.41)	.20 (.40)	.25 (.43)	.22 (.42)	.29 (.45)	.24 (.43)
Adults	.22 (.42)	.24 (.43)	.31 (.46)	.33 (.47)	.55 (.50)	.53 (.50)	.65 (.48)	.62 (.49)

Note. Subject male, pronoun match; Subject female, pronoun mismatch; Obj female, object female; Obj male, object male (distractor). Standard deviations are given in parentheses.

Table 7.3. Experiment 1: ANOVA *F* values for reading time measures. χ^2 values for regression probability.

	Gaze duration		Total reading time		Regression path duration		Regression probability	
Region	Pronoun	Final	Pronoun	Final	Pronoun	Final	Pronoun	Final
Subject	7.49 **	0.30	20.26 ***	0.62	13.96 ***	15.60 ***	17.66 ***	14.99 ***
Object	0.23	2.55	2.91	0.01	0.00	2.27	0.64	3.09
Age	81.27 ***	110.27 ***	48.95 ***	74.43 ***	45.44 ***	21.02 ***	6.37 *	33.09 ***
Subject × Object	0.42	3.00	0.07	0.01	2.42	0.02	0.93	0.24
Subject × Age	0.28	2.62	0.38	0.07	0.03	1.19	0.17	2.30
Object × Age	2.59	0.07	0.91	0.40	1.17	0.02	0.03	0.05
Subj × Obj × Age	0.05	0.30	0.33	3.49	0.24	0.18	0.98	0.01

Note. * = $p < .05$, ** = $p < .01$, *** = $p < .001$.

Regions of interest

Pronoun region

In the pronoun region, we found an early main effect of subject gender in gaze durations for both adults and children but found no effects of object gender. Interactions with age were not significant. Pronoun regions in subject gender-mismatch sentences elicited significantly longer gaze durations (M = 376ms, SE = 15) compared to pronoun regions in subject gender-match sentences (M = 353ms, SE = 15). In total reading time, there was a converging effect of subject gender in the pronoun region; mismatching pronouns took longer to read (M = 530ms, SE = 26) than matching pronouns (M = 480ms, SE = 24).

In regression probability, there were again main effects of subject gender but no effects of object gender and no interactions with age. In subject gender-mismatch sentences regressions were initiated in the pronoun region in 25% (SE = 2) of observations, whereas in subject gender-match sentences regressions were initiated in only 18% (SE = 2) of the observations. For regression path duration, we found a main effect of subject gender, where pronouns in subject gender-mismatch sentences elicited longer regression path durations (M = 518ms, SE = 29) than pronouns in subject gender-match sentences (M = 467ms, SE = 26).

In summary, the results in the pronoun region suggest that the mismatch between pronoun and subject gender disrupts bonding of pronoun to antecedent in adults and children alike. Both age groups showed longer gaze durations in the pronoun region of subject mismatch sentences than in that of subject match sentences. Furthermore, both groups showed immediate regressions out of the pronoun region when the pronoun was a mismatch to the subject antecedent. The effect of subject gender in total reading time in the pronoun region is likely a consequence of readers' regression behavior.

Final region

In the final region, there were main effects of subject gender in regressions out of the

pronoun region and regression path duration, but there were no effects of object gender. Interactions with age were not significant in any of the measures. We did not find any effects of subject gender or object gender in gaze duration, suggesting no early effects in the final region. With respect to the late measures, regression path durations in the final region were significantly longer in the subject gender- mismatch sentences (*M* = 1066 ms, *SE* = 50) than in the subject gender-match sentences (*M* = 978ms, *SE* = 46). This finding converges with the effects found in the pronoun region. The probability of regressions out of the final region was significantly higher in subject gender-mismatch sentences (*M* = 44%, *SE* = 4) than in subject gender-match sentences (*M* = 35%, *SE* = 4). We did not find any effects of subject gender or object gender in total reading times. Furthermore, there were no interactions of subject gender with object gender in any of the reported measures, nor did we find any interactions of subject gender and age. Taken together, we saw converging effects of the gender mismatch between subject and pronoun for regression probability and regression path duration, which mirrored the effects from the pronoun region. With respect to the object interference, our data do not confirm our hypotheses for the children because there were no discernible effects of object gender interference.

Experiment 2

Rationale

The results of Experiment 1 suggest that children and adults process pronouns in much the same way. This similarity was surprising because only roughly half of the children reported that they saw a pronoun mismatch in the sentences during testing when asked. We did not formally record reports of mismatch detection in Experiment 1 because the sample size would not have allowed a comparison between children who reported detecting a mismatch and those who did not. Therefore, we replicated the experiment to investigate inter-individual processing differences among the children depending on their awareness of the gender mismatch in the stimulus sentences. We used the same set of items as in Experiment 1. We were interested in two groups of children who emerged in Experiment 1: children who report mismatch detection (de-

tectors) and those who do not report mismatch detection (non-detectors). We used reporting of the mismatch detection as an indicator of children's successful reading comprehension monitoring. We hypothesized that mismatch detection is associated with reading processing and investigated which specific reading processes are associated with successful and unsuccessful mismatch detection. We hypothesized that the detectors process the mismatching pronoun region comparably to the adults, but we expected to see differences in non-detectors. We further hypothesized that differences in the report of mismatch detection between the detectors and non-detectors are associated with individual differences in the component skills of reading. Therefore, we compared differences in reading comprehension, reading fluency, and efficiency of auditory sentence comprehension between the two groups.

Method

Participants

The sample in Experiment 2 was a subset of the Berlin Developmental Eye Tracking Study (*DevTrack*), for which a total of 92 children from two primary schools in Berlin were tested. Of these, 75 successfully completed the experimental sessions for the current study. Of the participating children, 2 were early immigrants who arrived in Germany before 5 years of age. All children received their reading instruction in German only. Of the participating children, 41 were girls. The study was con- ducted in the winter term of the children's fourth school year, when they were 9 years old (M = 119 months, SD = 6). All children reported normal or corrected-to-normal vision.

Materials

The sentence materials for the eye tracking session were identical to those in Experiment 1 (see Table 6.2.1). We again assessed reading comprehension with the ELFE 1–6 (Lenhard & Schneider, 2006) and assessed reading fluency with a subtest of the SLRT-II (Moll & Landerl, 2010). Furthermore, we tested children's auditory syntactic integration skill using the computerized ProDi-L (Richter, Isberner, Naumann, & Kutzner, 2012).

Reading comprehension

The German reading comprehension test ELFE 1–6 contains three subtests that target word comprehension, sentence comprehension, and text comprehension. In the word com- prehension subtest, children receive a list of pictures and need to mark the word for the depicted item from a list of five given words. In the sentence comprehension subtest, children are asked to insert the appropriate word in a sentence context from a choice of five words. The ELFE 1–6 text comprehension subtest comprises a list of short texts, each of which has a set of questions tapping various levels of comprehension.

Reading fluency

The SLRT-II reading fluency subtest contains a list of words that need to be read out loud. Every child is given 1min to read as many words as possible until the test terminates. The fluency score is calculated by the number of words read, corrected for misses and omissions.

Syntactic integration

An auditory version of the syntactic integration subtest of the German ProDi-L was used to assess the efficiency of syntactic integration. Children listened to a list of 40 sentences via headphones and were asked to press a green button when the sentence was correct and a red button when the sentence was incorrect. Half of the sentences contained morpho-syntactic errors such as a wrong word order and faulty case marking. Children's response accuracy was analyzed along with reaction time, the latter as a marker for the efficiency of morpho-syntactic processing.

Procedure

The participating children completed the ELFE 1–6 reading comprehension test as part of a group session in their classroom, and they completed the SLRT-II and ProDi-L computerized tests in separate individual sessions. The eye tracking setup was the same as in Experiment 1 except that for technical reasons the presentation software was exchanged for the University of Massachusetts' EyeTrack (Stracuzzi & Kinsey, 2006). The sentences appeared in one or two lines in the middle of the screen in a monospaced font

(Courier New size 14). Upon arrival, the children were assigned to one of the four stimulus lists by order of appearance and were asked for oral consent prior to testing. The procedure was the same as in Experiment 1.

In Experiment 2, we also assessed mismatch detection during testing by asking children a series of questions after the first block (approximately 20 trials). If a child had not reported the mismatch spontaneously by the first pause, the experimenter would first ask whether the child had noticed some- thing weird in the sentences. If the response was negative, the experimenter continued to ask, "You know, sometimes one does not understand a word during reading or a word seems wrong in the sentence. Did that happen in what you have just read?" If the response remained negative, the experimenter would just make an encouraging remark ("You're doing a really good job. Ready to move on?") and continue to the next block. If a child reported the mismatch spontaneously or after inquiry, the experimenter would say, "You spotted a mistake, good job! That may happen again, but you can just continue reading quietly." The children were not prompted again during the experiment, but if they reported the gender mismatch anywhere during the remainder of the session, this was rated as positive report of mismatch detection.

Analysis

Data were cleaned as follows. First, fixations of less than 80ms were combined with a neighboring fixation if it was within one character. Short fixations of 40ms or less were deleted if they occurred within three characters of a neighboring fixation. Second, only fixations within a threshold of 120–1200ms were kept for analysis. This cleaning procedure removed less than 2% of fixations of each measure. Before models were calculated, observations above 2.5 standard deviations from the person or item mean of each dependent measure were removed. This removed less than 3% of the data in each measure.

The cleaned data were transformed and analyzed using the same methods and dependent variables as in Experiment 1 except that in Experiment 2 the models included incongruence detection (detector vs non-detector) as a fixed effect. We further used Welch's two-sample t-test to compare reading comprehension, reading fluency, and auditory sentence comprehension of the detector and non-detector groups.

Table 7.4. Experiment 2: Observed means of eye tracking measures.

	Pronoun region				Final region			
	Subject male		Subject female		Subject male		Subject female	
Group	Obj female	Obj male	Obj female	Obj male	Obj female	Obj male	Obj female	Obj male
				Gaze duration				
Non-detectors	438 (241)	407 (233)	474 (287)	487 (312)	724 (385)	754 (401)	699 (395)	720 (399)
Detectors	381 (219)	391 (213)	431 (268)	421 (251)	736 (410)	731 (407)	651 (379)	677 (399)
				Total reading time				
Non-detectors	603 (358)	561 (358)	672 (422)	667 (418)	976 (501)	961 (482)	967 (401)	1017 (515)
Detectors	631 (420)	632 (369)	790 (477)	784 (499)	1061 (534)	1051 (505)	1031 (510)	1049 (534)
				Regression path duration				
Non-detectors	538 (413)	526 (401)	601 (462)	582 (444)	1204 (718)	1201 (710)	1288 (738)	1330 (763)
Detectors	488 (404)	493 (415)	582 (494)	554 (442)	1501 (760)	1549 (770)	1646 (812)	1616 (796)
				Regression probability				
Non-detectors	.13 (.34)	.15 (.36)	.17 (.38)	.15 (.38)	.37 (.48)	.37 (.48)	.42 (.49)	.45 (.50)
Detectors	.23 (.42)	.21 (.41)	.30 (.46)	.34 (.47)	.65 (.48)	.65 (.48)	.73 (.45)	.73 (.44)

Note. Subject male, pronoun match; Subject female, pronoun mismatch; Obj female, object female; Obj male, object male (distractor). Standard deviations are given in parentheses.

Table 7.5. Experiment 2: ANOVA *F* values for reading time measures and χ^2 values for regression probability.

	Gaze duration		Total reading time		Regression path dur.		Regression probability	
	Pronoun	Final	Pronoun	Final	Pronoun	Final	Pronoun	Final
Subject	29.25 ***	19.40 ***	103.71 ***	0.04	32.57 ***	22.93 ***	23.93 ***	27.08 ***
Object	0.94	1.57	1.25	0.30	1.19	0.22	0.10	0.34
Detection	1.54	0.94	2.83	1.01	0.79	16.93 ***	15.71 ***	31.64 ***
Subject × Object	0.00	0.73	0.07	2.73	0.41	0.00	0.58	0.03
Subject × Detection	1.13	1.58	6.84 **	1.81	0.79	0.26	5.00 *	0.64
Object × Detection	1.32	0.04	1.34	0.10	0.41	0.09	0.02	0.12
Subj × Obj × Detection	3.39	0.00	1.66	0.59	0.41	1.18	3.99 *	0.07

Note. * = $p < .05$, ** = $p < .01$, *** = $p < .001$

Results and Discussion

In Experiment 2, we replicated our main findings for the children from Experiment 1. The observed means are summarized in Table 7.4, and the effects from the linear mixed-effects models are summarized in Table 7.5.

Our inquiry during testing resulted in 43 children (57%) who reported mismatch detection and 32 children who never reported mismatch detection. Children who did not report the mismatch, however, may still have been aware of it on some level without explicitly verbalizing it. For simplicity, we refer to the children who reported mismatch detection as detectors and the children who did not report mismatch detection as non-detectors, but it is important to keep in mind that not reporting a mismatch does not necessarily presuppose absence of awareness.

Table 7.6. Analyses of group differences.

	Detectors		Non-detectors	
Measure	*M*	*SD*	*M*	*SD*
Reading comprehension	90.3	12.8	84.5	15.0
Reading fluency	83.8	17.6	73.2	14.4
Sentence comprehension accuracy	.88	.07	.88	.07
Sentence comprehension reading time	3011	323	2896	232

Note. Raw scores for the reading comprehension test, raw scores for the reading fluency test, and accuracy and reaction time in the sentence comprehension test.

Global results

Mean accuracy in the comprehension questions was high (M = 93%, SD = 26). An ANOVA over response accuracy with subject gender and detection as crossed factors showed that accuracy was not affected by subject gender match, $F(1, 146) < 1$, $p = .588$. There was a main effect of detection, such that accuracy was slightly higher in the detectors (M = 94%, SD = 10) than in the non-detectors (M = 90%, SD = 13), $F(1, 146) = 4.23$, $p < .05$. There was no interaction of subject gender and detection, $F(1, 146) < 1$, $p = .674$. The results of the analyses of individual differences for the two detector groups are summarized in Table 7.6.

Detectors and non-detectors did not differ in reading comprehension skill, $t(62) = 1.8$, $p = .078$. Our analyses of accuracy in the sentence processing task showed that both

groups correctly rejected sentences with morpho-syntactic errors, such that there was no difference between the groups for accuracy, $t < 1$, $p = .968$. Furthermore, we did not find a significant difference for the efficiency of auditory processing between the two groups, $t = 1.8$, $p = .075$. However, the two groups differed in reading fluency, with the detectors being the more fluent readers, $t(74) = 2.9$, $p = .005$. In the reading fluency task, the detectors read significantly faster than the non-detectors (about 10 words more per minute). The results of the reading fluency test concur with our observations in the eye tracking measures. Analyses of eye tracking measures on the trial level showed that children who reported detection of the mismatch spent less time on first-pass reading than children who did not report the mismatch. The observed mean gaze duration per region for the detectors (M = 480ms, SD = 121) was significantly shorter than that for the non-detectors (M = 552ms, SD = 129), $t(65) = 2.5$, $p = .016$. However, total reading times did not differ between the detectors and non-detectors. The detectors spent as much time reading each region (M = 786ms, SD = 224) as the non-detectors (M = 766 ms, SD = 216), $t(69) < 1$, p=.708.

In summary, eye movement measures on the trial level revealed that the detectors read faster during the first pass than the non-detectors. However, there was no difference between the groups in total reading time, from which we may conclude that the detectors spent more time rereading the sentences. The non-detectors made regressions less frequently, and their gaze durations were correlated more closely with their total reading times.

Regions of interest

Pronoun region

In the pronoun region, there was a main effect of subject gender in gaze durations, such that subject gender-mismatch sentences elicited significantly longer gaze durations (M = 379ms, SE = 14) than subject gender-match sentences (M = 347ms, SE = 14). For total reading time, we also found a significant effect of subject gender in the pronoun region. The region was read longer in subject gender-mismatch sentences (M = 594ms, SE = 25) than in subject gender-match sentences (M = 497ms, SE = 21). Detectors showed a significant 131-ms effect of subject gender, which was significantly larger than the 67-

ms effect of subject gender in non-detectors, t = 2.61, p = .009. A converging effect emerged in regression path duration. In the subject gender-mismatch sentences, children took significantly longer to pass the pronoun region (M = 452ms, SE = 19) than to pass the same region in the subject gender-match sentences (M = 406ms, SE = 17). For regression probability, we again found a main effect of subject gender, such that children were more likely to initiate regressions from the pronoun region in the subject gender-mismatch sentences (M = 21%, SE = 2) than from the pronoun region in the subject gender-match sentences (M = 15%, SE = 2). Furthermore, there was a main effect of detection, such that, in general, the detectors were more likely to make regressions out of the pronoun region (M = 24%, SE = 2) than the non-detectors (M = 13%, SE = 2). Post hoc contrasts revealed that in fact, the subject gender effect in the pronoun region was entirely driven by the detectors. There was a significant simple main effect of subject gender in the detector group, t = 5.4, p < .001, but not in the non-detector group, t = 1.1, p = .268. For the detectors, the likelihood to make a regression out of the pronoun region in subject gender-mismatch sentences was higher (M = 30%, SE = 3) compared with subject gender-match sentences (M = 19%, SE = 2), whereas the non-detectors made equally few regressions out of the pronoun region in subject gender-match sentences (M = 14%, SE = 2) and subject gender-match sentences (M = 12%, SE = 2).

Lastly, we found a three-way interaction of incongruence detection, subject gender, and object gender in the pronoun region. The detectors made numerically more regressions out of the pronoun region of subject gender-mismatch sentences in the object gender-match condition than in the object gender-mismatch condition. The difference, however, was nonsignificant, t = 1.83, p = .067. The non-detectors, in contrast, made numerically more regressions in the subject gender mismatch sentences when the object was a mismatch as well, although this difference was also nonsignificant, t = 1.18. p = .237.

In summary, the results in the pronoun region suggest that children use pronoun gender for rapid pronoun–antecedent bonding during online reading given that the reading processing of both groups of children was disrupted immediately in the subject gender-mismatch conditions. Our analyses of the eye movement behavior of detectors and non-detectors revealed important differences in late measures, such that the detectors make more regressions. These resulted in longer total reading times in the pronoun region

for the detectors compared with the non-detectors.

Final region

In the final region, there were main effects of subject gender and main effects of detection but no interactions. There were no effects of object gender. Gaze durations in the final region showed an early effect of subject gender in the opposite direction from the pronoun region. Final regions of sentences with a matching subject elicited significantly longer gaze durations (*M* = 628ms, *SE* = 24) than final regions of sentences with a mismatching subject (*M* = 578ms, *SE* = 22). Note that in Experiment 1 we did not find any effect of subject for gaze duration in the final region. There were no effects for total reading time in the final region, consistent with our findings in Experiment 1.

The delayed effects of subject gender went in the same direction as in the pronoun region. There were prolonged regression path durations for the final region of subject mismatch sentences (*M* = 1286ms, *SE* = 54) compared with subject match sentences (*M* = 1183ms, *SE* = 49). Moreover, the detector group showed significantly longer regression path durations in the final region (*M* = 1428 ms, *SE* = 72) compared with the non-detector group (*M* = 1066ms, *SE* = 60). This result converges with the longer trial reading times of the detectors and their regression behavior; for regression probability, we found that subject mismatch sentences elicited more regressions in general (*M* = 61%, *SE* = 3) than subject match sentences (*M* = 51%, *SE* = 3). We also found a large main effect of incongruence detection in the final region, such that the detectors made more regressions out of this region (*M* = 73%, *SE* = 4) than the non-detectors (*M* = 38%, *SE* = 4). Notably, there was no interaction of subject and incongruence detection in the final region, suggesting that both groups show sentence-final mismatch effects.

In summary, Experiment 2 revealed significant differences in the processing of children who report mismatch detection and children who do not report it. The fact that the detectors made more regressions overall explains the disparity of gaze duration and total reading time between the two groups in the regions of interest. With respect to the object gender manipulation, we replicated our null effect from Experiment 1. Lastly, we saw a novel effect in the final region that we did not obtain in Experiment 1. In the final region, there were prolonged gaze durations for subject gender-match sentences compared with subject gender-mismatch sentences. We

discuss this finding in more detail in the General Discussion.

General discussion

The main focus of the current study was on children's sensitivity to gender feature mismatches during the reading of pronouns and their susceptibility to interference of feature-matching entities during reading. Our materials consisted of two-clause sentences with a male pronoun in the second clause such as ''Leon/Lisa shooed away the sparrow/the seagull and then he ate the tasty sandwich." The pronoun always referred to the subject of the main clause, which was either a gender-matching or gender-mismatching name (Leon{m}/Lisa{f}). Furthermore, the sentences contained an interfering direct object that was either a gender match or mismatch to the pronoun (the sparrow{m}/the seagull{f}). We recorded children's and adults' eye movements while they read the sentences, focusing our analyses on the pronoun and sentence-final regions, the latter of which directly followed the pronoun region. In Experiment 1, results suggested no qualitative differences between children's and adults' processing of the pronouns. Both age groups showed immediate sensitivity to the subject gender mismatch and no effects of interference from the object. However, in contrast to the adults, not all children seemed to detect the subject gender mismatch. In Experiment 2, we replicated our finding from Experiment 1 with a larger sample of children and found that 43% of the children reported that they were unable to detect the mismatch during the experiment. Although those children who detected the gender mismatch processed the mismatching pronoun comparably to the adults, there were important differences in pronoun processing between the ''detectors" (i.e., children who did report the mismatch when prompted) and the ''non-detectors" (i.e., children who did not report the mismatch).

Processing of the pronoun region in children and adults

In our experiments, we aimed to tap early and late reading processes using different eye tracking measures: gaze duration for early stages of processing and total reading time, regression behavior, and regression path duration for later stages of processing. We hypothesized that children's processing efficiency may differ from adults' and that

the eye movement measures should be affected by these processing differences. In summary, our results suggest that pronoun–antecedent association based on gender cue is indeed as automatic in children as in adults (Rigalleau et al., 2004). This was reflected in the longer gaze durations on the pronoun for mismatching subjects than for matching subjects. In Experiment 1, we did not find any differences between children and adults with respect to gaze durations. Likewise, in Experiment 2, both children who reported mismatch detection and children who did not report it showed longer gaze durations for gender-mismatching pronouns than for gender-matching pronouns. Whether children resolve a pronoun successfully, however, seems to be subject to individual processing differences.

Both groups of children slowed down during the first pass of the region when a mismatching pronoun was presented. This suggests that gender information was automatically registered as incongruent even by those children who were unable to report a mismatch after reading. We assume that children who reported the mismatch also understood the correct resolution of the pronoun. It is important to note that, in contrast, absence of reporting does not necessarily imply absence of comprehension by the non-detectors. Therefore, the group of non-detectors is defined less clearly. Our results from gaze durations indeed show that they too are sensitive to the gender mismatch. We did not ask comprehension questions tapping pronoun resolution in this study because such questions would have prompted the participants to explicitly pay attention to pronoun inconsistencies. This would likely have interfered with natural reading. Successful monitoring and reading comprehension, however, have been shown to be closely related in children (van der Schoot, Reijntjes, & van Lieshout, 2012).

Although both children who reported the mismatch and children who did not report the mismatch slowed down during gaze duration at a mismatching pronoun compared with a matching pronoun, the detectors were more likely to reread the mismatching area. Successful integration, as evidenced by the report of mismatch detection, was associated with higher regression probability for mismatching pronouns than for matching pronouns in the pronoun region. This is what led to longer total reading times of the mismatching pronoun. We interpret this finding in terms of processing depth of the pronoun. As originally observed by Rayner (1998), readers may make very short regressions (up to one word to the left) when the currently fixated word disrupts fluent

text processing. It has been suggested that in this way readers may delay new input from upcoming words in order to allow for more processing time in a conflictive sentence region. The eye movement pattern we found in the adults and detectors may constitute a ‘‘coping mechanism” in the face of local processing difficulty, leading to an increase in processing depth. Thus, more processing time in the critical region—that is, directly at the pronoun—is associated with mismatch detection by way of deeper processing of the pronoun.

The subject gender mismatch was also evident in regression probability from the sentence-final region, where we saw a main effect of subject gender and of mismatch detection. The detectors made more regressions overall at the end of a sentence. Note that this matches the adult reading behavior. The result is also in line with earlier findings connecting rereading probability and pronoun comprehension (Ehrlich et al., 1999). Our results show that whereas immediate regressions are associated with the reporting of mismatch detection in children, the detectors also make more regressions overall. Arguably, monitoring of comprehension is closely connected to rereading behavior in children.

Individual differences in children’s reading processing

The eye movement behavior associated with successful comprehension monitoring in the detectors can be described as faster first-pass reading, evident as shorter gaze durations, combined with more regressive saccades in the mismatching pronoun region compared with the non-detectors. In addition to the processing data, we tested children’s reading comprehension, reading fluency, and syntactic integration skill. Contrary to our expectations, we did not find any differences in the non-detectors’ and detectors’ reading comprehension skill or in their accuracy or efficiency of syntactic integration. Although the striking similarities between the detectors’ and adults’ reading processing may suggest that the non-detectors are less developed readers, we would expect significant differences in the reading comprehension test if this were the case. Although reading comprehension and syntactic integration were comparable between the groups, the non-detectors were significantly slower in the reading fluency test. This result concurs with our processing data, which showed that the detectors had shorter gaze durations than the non-detectors at the trial level. We may conclude that

processing speed is a main determinant of reading comprehension and monitoring in children. The association of slow decoding speed and comprehension monitoring failure in our study is in line with the lexical quality hypothesis (Perfetti, 1994; Perfetti & Hart, 2001). The lexical quality hypothesis assumes that slow word reading is a sign of inefficient and effortful extraction of word information such as ortho- graphic representation and semantic meaning. If this process is effortful, children may lack the cognitive resources to engage in monitoring of comprehension. Therefore, we conclude that children who read more fluently can make the necessary resources available for comprehension monitoring and, as a result, can report the mismatch when prompted.

The pattern of our findings is in line with a prior study on children's detection of semantic anomalies. Connor et al. (2014) studied fifth graders' eye movements in two-sentence stories that contained local semantic anomalies, for example, ''truck" when ''plane" would be appropriate: ''Last week Kyle flew to visit his family in another city. The large plane/truck was spacious and quickly transported him." The authors reported that children with stronger literacy skill read more fluently and made more regressions at the semantic anomaly than children with weaker literacy skill, although both groups' first-pass reading was slowed down by the anomaly. The authors concluded that slowing down during first-pass reading at a semantic incongruence in the text is largely automatic and that successful comprehension is determined by the extent of a child's engagement with the text after a semantic incongruence was detected. Our results show that this pattern transfers to pronoun processing. The mismatching pronoun induced a local processing disruption given that children who reported the mismatch made regressions immediately. In addition, there was a reverse effect of subject gender mismatch for gaze durations in the final region. Contrary to our hypotheses, we found prolonged gaze durations for matching pronouns rather than mismatching pronouns. We interpret this finding in terms of a wrap-up for congruent sentences but not for incongruent sentences. It is conceivable that children do not attempt meaning integration at the sentence level for incongruent sentences at all. This interpretation of our results suggests that even surface-level incongruences can lead to unresolved sentence processing in children. Because we did not see this reverse effect of subject gender mismatch in the adults, we would hypothesize that the processing disruption induced by a mismatching pronoun has a more lasting effect on less proficient readers than on proficient readers. It has previously been shown that incongruent pronouns affect reaction

time to comprehension questions, but not comprehension accuracy, in adults (Rigalleau et al., 2004). This effect may be individually different in developing readers, who might not recover from conflicting surface information in text as quickly as proficient readers.

Null effect of object gender

Lastly, we turn to the null effect of the object gender manipulation. We did not obtain any main effects of object gender match for pronoun processing in our experiment. In contrast to our hypotheses, the intervening object did not influence processing of the matching pronoun regions or the mismatching pronoun regions. Findings for the online reading pattern associated with feature- based interference are mixed in the literature (Jäger et al., 2017), and our null results do not exclude the existence of cue-based interference effects during pronoun resolution in children. We suggest that future research pursuing such interference effects use paradigms more closely related to those used with adults (e.g., Badecker & Straub, 2002; Cunnings & Felser, 2013; Patil, Vasishth, & Lewis, 2016).

Conclusion

Our experiments show that examining local processing strategies at key points in a sentence may inform our understanding of children's reading comprehension and potential sources of difficulties. The results from Experiment 1 suggest no developmental differences between children's and adults' pronoun processing. This indicates that there are no overall qualitative differences in pronoun processing between beginning and proficient readers. Experiment 2, however, showed that there are inter-individual differences in children's processing of mismatching pronouns. These processing differences were associated with children's reporting of the mismatch when prompted. Importantly, total reading time was not itself related to mismatch detection; rather, those children who reported detection of the mismatch allocated additional time to rereading only. Therefore, we may also conclude that total time on task during reading is not a sufficient indicator of reading performance. This is important because educators tend to assess children's reading performance in classroom settings based on how much time the children need to finish reading a passage. To further understand how pronoun resolution influences children's text comprehension at large, we suggest that whole text

pas- sages be taken into account. Because pronouns serve as an anchor for textual coherence, their resolution is important for text comprehension. Our findings using tightly experimentally controlled sentences, therefore, are likely to transfer to less constrained natural text reading.

8 GENDER CUE EFFECTS IN CHILDREN'S PRONOUN PROCESSING

Abstract

Children struggle with the resolution of pronouns during reading, but little is known about the sources of their difficulties. We conducted a longitudinal eye tracking experiment with 70 children in the final years of primary school. The children read sentences with a contextual resolution preference in which gender was either an informative resolution cue for the pronoun or not. We were interested in children's processing of the pronoun and their resolution preferences, as well as the effects of individual differences of Grade level and reading skill. Children's resolution ability improved with age, and good readers were more accurate than poor readers. In the eye tracking measures, we found strong individual differences related to reading skill: Children with good reading skill took more time to read the pronoun region when pronoun gender was informative, suggesting that good readers make better use of the available information at the pronoun than poor readers.

This section was published open access:

Eilers, S. & Tiffin-Richards, S. & Schroeder, S. (2019b). Gender cue effects in children's pronoun processing: A longitudinal eye tracking study. *Scientific Studies of Reading, 23*, doi: 10.1080/10888438.2019.1617293

Introduction

Many beginning readers struggle with text comprehension even after having mastered fluent word reading. This suggests that word reading is necessary but not sufficient for text comprehension. Proficient readers make inferences during reading, which is one determinant of successful text comprehension (e.g., Oakhill, Berenhaus, & Cain, 2015). One example of a local inference process is pronoun resolution: Pronouns are ubiquitous in texts and easy to process by themselves as they are short and carry very little semantic meaning. On the level of word reading, pronouns are therefore not particularly challenging for beginning readers. In order to be fully understood, however, the pronoun has to be bound to an appropriate antecedent. Proficient readers routinely infer the correct antecedent using morpho-syntactic information such as gender markers (Patil, Vasishth, & Lewis, 2016). This requires the integration of information from memory across several words in a sentence or text. It has been suggested that one source of children's reading comprehension difficulties is the failure to make such inferences (Megherbi & Ehrlich, 2005; Wykes, 1981; for reviews see Nation, 2005; Oakhill, Berenhaus, & Cain, 2015; Perfetti, Landi, & Oakhill, 2005). More recently, it has been shown that children's ability to specify referents in texts accounts for unique variance in reading comprehension skill (Elbro, Oakhill, Megherbi, & Seigneuric, 2017). The ability to resolve referential relations is one of the key steps to sentence and text comprehension.

In the present longitudinal study, we examine how children at different ages and varying reading skill take different types of information into account when processing and resolving pronouns. Specifically, we investigated German children's processing and comprehension of pronouns in sentences where adults show a clear contextual resolution preference. We manipulated whether pronominal gender was an informative resolution cue or not by introducing two antecedents of either the same or a different gender.

Children's pronoun comprehension

Previous studies of children's comprehension of pronouns have produced mixed re-

sults with respect to the developmental trajectory of pronoun resolution accuracy, presumably due to differences in methods, age groups, materials and languages studied (for a review see Hickmann, Schimke, & Colonna, 2015). One study showed that children use gender information to guide online pronoun resolution during listening from 3 years of age (Arnold, Brown-Schmidt, & Trueswell, 2007), and we can assume that most children resolve pronouns correctly during listening by the time they attend primary school. However, comprehension skill moderates pronoun resolution in primary school students: In a cross-modal naming task with French 7- and 8-year-olds, Megherbi and Ehrlich (2005) demonstrated that poor comprehenders do not resolve pronouns systematically using gender information. Instead, they may resort to a default strategy where recency "overrides" other available cues.

Studies of children's reading have also shown that pronoun resolution is a source of comprehension errors. Yuill & Oakhill (1986) tested 7- to 8-year-old's comprehension of sentences such as "On Saturday morning, Bill was going on a fishing trip with his Uncle. [...] As he carried his rod to the bus stop [...]." Children were then asked, "Who carried his rod to the bus stop?" Good comprehenders answered these questions with an error rate of 10% while poor comprehenders had an error rate of 28%. Further, Oakhill & Yuill, 1988) showed that 7- and 8-year-old children have difficulties finding the correct referent for the personal pronoun in sentences such as "Peter lent ten pence to Tom [Liz] because he [she] was very poor". The children performed worse in the condition without an informative gender cue (16–27% error rate) compared to the condition with an informative cue (2–14% error rate). Thus, while good comprehenders performed better than poor comprehenders, both groups of children benefited from disambiguating gender information when answering the resolution question. These studies also show clearly that children struggle with the comprehension of pronouns, but they do not inform about the reading processes that are associated with resolution difficulty.

Children processing of pronouns and referential relations

Children's reading is slower and more effortful than that of adults (e.g., Gagl, Hawelka, & Wimmer, 2015). They invest extensive cognitive resources in word identification, because the translation of orthographic information into semantic representations is

slower than in proficient readers. As lower-level reading requires their attention, children can invest fewer resources in higher-level processing, such are inference making and comprehension. Pronouns are very easy to process on the word level because they are both short and frequent, but they also require a higher-level integration effort, involving the retrieval of antecedent features from memory. Cue-based approaches to memory retrieval suggest that morphosyntactic cues (e.g., gender, number, grammatical case) are routinely used for resolution of pronouns (Lewis, Vasishth, & Van Dyke, 2006). Studying the use of such cues during pronoun processing can inform our understanding of how the processing demands of higher-order reading affect children of different ages and reading skill.

Pronouns have indeed been shown to be a source of difficulty in children's sentence processing, and reading ability determines pronoun processing. A self-paced reading experiment with 10-year-olds (Ehrlich, Rémond, & Tardieu, 1999) demonstrated that good comprehenders had longer reading times in clauses with a personal pronoun compared to clauses with a repeated name. In addition to reading the pronoun for a longer period of time, good comprehenders chose to press a button to display previous text more often, indicating that they adjust their rereading behavior to pronoun resolution demands. This shows that during processing, pronouns pose a specific challenge for children, arguably because they have to be resolved towards an antecedent.

Recently, eye tracking has been established as a method of choice in studying children's reading processes (for reviews see Blythe & Joseph, 2011; Schroeder, Hyönä, & Liversedge, 2015). It is favored over self-paced reading or priming methods because it allows the uninterrupted recording of multiple measures at specific points in a text.

In a pioneering eye-tracking study with 8-year-old children, Murray and Kennedy (1988) showed that good readers make more regressions in sentences that contain pronouns. Selective regressions were associated with a better comprehension of sentences with pronouns. While poor readers make more regressions in general during reading, good readers make more regressions at the pronoun than elsewhere. In a more recent eye-tracking study, Joseph, Bremner, Liversedge, and Nation (2015) examined 10-year-old children's processing of nominal anaphors. The authors compared the processing of nominal anaphors (the vehicle) with typical antecedents (a truck) and atypical antecedents (a crane) in stories where the anaphor was either near or distant. The authors

observed more regressions when the antecedent was typical compared to when it was atypical. This finding suggests that children invest resolution effort when they are establishing a connection between anaphor and antecedent. In line with this interpretation, the authors argue that children may not resolve nominal anaphors in the distant/atypical condition at all, i.e., when resolution is most difficult. Since the study did not examine children's anaphor comprehension, however, it is still largely unclear how differences in anaphor processing are related to comprehension.

The current study

We investigated pronoun processing and comprehension in 70 German primary-school children of different reading skill in a longitudinal study. We presented sentences of the following form (see Table 8.1): "Paul beneidete Tessa, weil sie zu Hause einen Pool hatte" (Engl.: Paul envied Tessa because she had a pool at home) vs. "Paul beneidete Theo, weil er zu Hause einen Pool hatte" (Engl.: Paul envied Theo because he had a pool at home). We manipulated the gender of the subject and object in the main clause, resulting in sentences where pronominal gender was informative for pronoun resolution or not. In the first sentence, the gender of the pronoun is an informative resolution cue because she can only refer to Tessa, not to Paul. In the second sentence, gender is not informative for resolution because he could refer to both Paul and Theo. In the given example, however, it is plausible that Paul envied Theo because Theo had a pool at home. While the reading that Paul envied Theo because Paul had a pool at home is not strictly ruled out, it is rather laborious and less plausible. Therefore, there is a resolution preference towards Paul even in the absence of a gender cue. Similar rationales have been used in experiments with English-speaking adults (e.g., McDonald & MacWhinney, 1995; Vonk, 1984). Note that while gender-marking in German differs from English in several ways (Fagan, 2009), singular pronouns (he/she) are marked for natural gender as in English. The syntactic particularities of German (see the example in Table 8.1) do not interfere with our manipulation. In the following, we will therefore refer to our materials using English translations.

We asked a forced-choice pronoun resolution question (e.g., "Who had a pool at home?") after every sentence to obtain resolution preference and response time (offline measures). We also recorded children's eye movements during reading (online

measures). The children further completed a standardized reading comprehension test. The main research question of this study was how children of different ages and reading skill use gender and context information during pronoun processing and towards pronoun resolution.

Comprehension of the pronoun (offline measures)

We predicted that children would answer the resolution questions more accurately after sentences that contain pronouns with an informative gender cue than no informative cue (e.g., Yuill & Oakhill, 1986). We further predicted that as children gain more reading experience with age, they should depend less on explicit gender information for resolution and instead show a more adult-like resolution preference based on the integration of sentence context. Similarly, reading skill was expected to influence resolution preferences such that better readers among the children answer the resolution questions faster and more accurately. Lastly, an interesting question concerns the relationship between reading development and individual reading skill: As children become more experienced readers, individual differences in reading skill may become less important for pronoun resolution. Such a trend would suggest that in the final years of primary school a threshold is reached such that children resolve pronouns more automatically.

Processing of the pronoun (online measures)

We analyzed reading time measures on the pronoun itself and the subsequent region. The sub- sequent region was taken into account to pick up effects from the pronoun that occur after it has been read. Since it is very short, effects from the pronoun may spill over onto the following word. Such a "delay" of effects has been observed in children's syntactic processing before (Wonnacott, Joseph, Adelman, & Nation, 2016) and was shown to be developmentally relevant, as the delay reduces with age (Joseph & Liversedge, 2013). We expected to find more regressions from the pronoun region in the informative gender cue condition, i.e., when the pronoun can be resolved (Joseph et al., 2015). This would indicate that the children use the disambiguating gender information immediately for pronoun resolution. We do not expect the children to engage in

resolution effort in the non-informative condition, where the pronoun can only be resolved at the end of the sentence. Besides gaze duration, we analyzed total reading time and gopast time to obtain a detailed picture of children's rereading of the pronoun. While gaze duration is indicative of processing ease and reading fluency, total reading time and gopast time incorporate rereading following a regression. Rereading in the informative cue condition would indicate additional processing effort when disambiguating gender information is available. Longer gopast times would further indicate that children do not only regress but engage in more extensive rereading of earlier regions.

Our second research question concerned individual differences of the resolution processes in children. Children in the same Grade level differ dramatically in their individual reading ability. It is plausible that reading skill determines if and how beginning readers use gender information as a processing cue. Assuming that reading behavior at the pronoun and reading comprehension are related, we expected to see longer processing times and more regressions in the pronoun region in good readers than in poor readers. We further investigated the possibility that delayed effects occur in poor readers and therefore appear in the post-pronoun region.

Table 8.1. Structure of stimulus materials.

Gender	Item	Resolution question
Contrast	Paul beneidete ***Tessa***, [weil ***sie***] [zu] Hause einen Pool hatte. *Paul envied Tessa because she had a pool at home.*	Wer hatte einen Pool? *Who had a pool?*
Identity	Paul beneidete ***Theo***, [weil ***er***] [zu] Hause einen Pool hatte. *Paul envied Theo because he had a pool at home.*	Wer hatte einen Pool? *Who had a pool?*

Note. Bold text indicates the referent of the pronoun (the gender of the pronoun was counterbalanced). Square brackets indicate regions of interest for analysis. English translations are given in italics.

Method

Participants

The children who participated in the current experiment attended two primary schools in Berlin. From the 92 original participants, we included all children who participated in both Grade 3 and Grade 4. One child was excluded because their response accuracy to comprehension questions in Grade 3 was below the chance level. The remaining 70

children completed the experiment in Grade 3, at age 8.3 years (*SD* = 0.5), and again 1 year later in Grade 4, aged 9.4 years (*SD* = 0.5). Of these 70 children, 42 were girls. All children had normal or corrected-to-normal vision.

Materials

Materials for this study comprised 24 items like the one depicted in Table 8.1. The study was conducted in German, but for simplicity we will illustrate the materials using English translations that leave the integrity of our stimuli intact. The sentences contained 9–12 words. Each sentence appeared in one of the two conditions (Informative Gender Cue vs. Non-Informative Gender Cue). The condition was altered by changing the names in the sentences and by adapting the pronoun accordingly. The gender of the pronoun was counterbalanced to prevent habituation effects. We took care to construct sentences with topics familiar to primary school children. For every sentence, a forced-choice resolution question was constructed from the subclause.

To support the resolution preference for the pronoun in the condition without an informative gender cue, we used implicit causality verbs that bias the resolution of the pronoun (e.g., Koornneef & Van Berkum, 2006). Only implicit causality verbs that occur in the childLex corpus, a corpus of German children's books (Schroeder, Würzner et al., 2015), were used in this experiment to ensure that the children know them. As the occurrence of these verbs in the childLex corpus is limited, the resolution preference for subject and object was counterbalanced, as were male and female pro- nouns. All sentences continued bias-congruent, in other words, the subordinate clause supported the preferred reading induced by the verb and there were no sentences with a conflict between verb bias and gender information. Consider the example Clara admired Anne because she could draw so nicely, where Anne is likely admired because she can draw nicely, or Felix bored Pete because he always told the same stories, where Felix is likely boring because he tells the same stories repeatedly.

To check this resolution preference, the sentences were presented to a sample of 25 adults who were recruited from local universities via mailing lists. The results from the comprehension task showed that the adults conformed to the intended resolution preference in 97% of questions, and an ANOVA with the dependent variable accuracy and the two-level factor Gender (Informative vs. Non-Informative Gender Cue) resulted in

no significant effect, $F(1,48) = 2.16$, $p = .147$.

Children's reading skill was tested with the standardized German reading comprehension test ELFE 1–6 (Lenhard & Schneider, 2006). This test comprises three subtests targeting word, sentence and text comprehension. The raw scores for each subtest are first transformed to standardized scores and then summed up to serve as an overall indicator of children's reading skill.

Procedure

Written informed consent was collected from the children's parents ahead of the study, and oral consent was obtained from each child prior to testing. The study was approved by the ethics committee of the Max Planck Institute for Human Development, Berlin, and conforms with the Declaration of Helsinki.

Children were tested individually in a quiet room at their school during school hours. In addition, the children participated in a group session in their classroom, during which the reading comprehension test was administered. Children were tested under the same conditions in Grade 3 and Grade 4. In each session, they were assigned to one of two item lists to ensure that they read every item in only one of the cue conditions.

We used an EyeLink 1000 table-mounted eye tracker (SR Research) to record eye movements at 1000Hz. The eye tracker was positioned under an ASUS LCD monitor (21'', 120Hz) at a 65cm viewing distance to the child. The sentences appeared in random order, on a single line at the center of the monitor. They were presented in 14pt Courier New using the UMass EyeTrack software (Stracuzzi & Kinsey, 2006b). The right eye was tracked unless tracking of the left eye considerably improved calibration. The eye tracker was calibrated using a 5-point calibration routine until calibration error reached a maximum 0.5° of visual angle. Calibration was repeated after breaks or when calibration drifts were detected. After the first calibration, all children completed three practice trials. They were instructed to read the sentences at their own pace and indicate via button press when they have finished reading. Upon pressing the button, the forced-choice pronoun resolution question appeared. To avoid confusion, the assignment of buttons to names consistently followed their position in the sentence (subject left, object right). Forty filler sentences from an unrelated experiment, including simple yes/no-comprehension questions after 25% percent of filler trials, were interspersed

randomly (for details see Tiffin-Richards & Schroeder, 2015). The children answered a total of 34 comprehension questions in this experiment.

Analysis

The eye movement data were cleaned step-wise: First, each trial was inspected visually using the software EyeDoctor (Stracuzzi & Kinsey, 2006a), and y-axis drift corrections were applied to groups of fixations as necessary. Next, we applied an automatic fixation cleaning procedure as implemented in EyeDoctor. Fixations of less than 80ms were combined with a neighboring fixation if it was within 1 character. Fixations of 40ms or less were deleted if within 3 characters of the nearest fixation. Finally, trials with less than 5 fixations were removed (2 trials in Grade 3, 4 trials in Grade 4) and fixations under 60ms or above 1200ms were discarded (1.1% in Grade 3, 1.0% in Grade 4).

Four eye tracking measures were calculated for each region: gaze duration (sum of all fixations on a region before leaving it), total reading time (sum of all fixations on a region), gopast time (sum of all fixations from the first visit of a target region until it is left to the right), and the probability of regression out (likelihood that the region is exited to the left). For each measure, data points deviating more than 2.5 standard deviations from the word and subject mean were deleted (less than 2.0% of data in each group). A Pearson product-moment correlation coefficient was computed to assess the relationship of reading measures in Grade 3 and Grade 4.

We used generalized linear mixed-effects models for binomially distributed data as implemented the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) in R (R Core Team, 2016) to analyze response accuracy, and linear mixed-effects models to analyze response time and eye movement measures. Gender (Informative Gender Cue vs Non-Informative Gender Cue) and Grade (Grade 3 vs Grade 4) were included as effect-coded fixed effect. Reading Skill was included as a centered continuous variable. Participants and items were entered as crossed random effects in the models to allow for random intercepts for participants and items. Duration measures were log-transformed to make the distribution more normal. To ease interpretation, back-transformed model means are reported in milliseconds and probabilities, respectively. The significance of the fixed effects was determined using type-II model comparisons as implemented in

the Anova function in the package car (Fox, Friendly, & Weisberg, 2013). Planned comparisons were estimated using cell-means coding and single-degree-of-freedom contrasts as implemented in the glht function in the package multcomp (Hothorn, Bretz, & Westfall, 2008).

Results

Offline measures

Resolution accuracy was positively correlated between Grade 3 and Grade 4, r = .52, $t(68)$ = 4.97, p < .001, and response time was highly correlated, r = .77, $t(68)$ = 10.10, p < 001. The correlation of reading skill in Grade 3 and reading skill in Grade 4 was also high, r = .75, $t(68)$ = 9.46, p < .001. The model results for response accuracy and response time are summarized in Table 8.3, and the distributions are depicted in Figure 8.1.

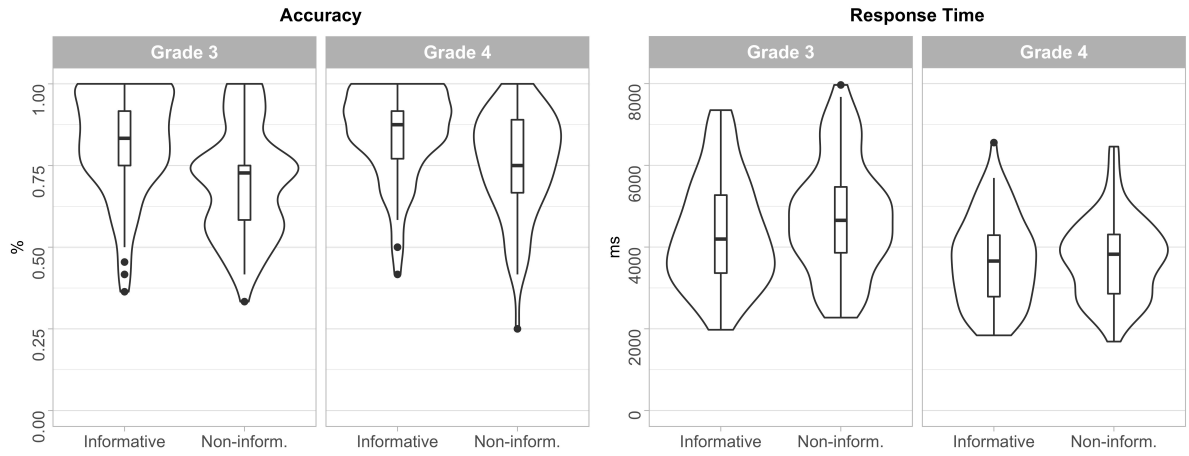

Figure 8.1. Distribution of resolution accuracy in percentage (panel 1) and response time in milliseconds (panel 2) for children in Grade 3 (left) and in Grade 4 (right), separately for the gender conditions. Error bars represent 2 standard errors.

In resolution preference, there was a main effect of Gender: As we had predicted, children were more successful in identifying the plausible antecedent in the Informative Gender condition, M = 87%, SE = .02, than in the Non-Informative Gender condition, M = 76%, SE = .03. Further, there was a main effect of Grade: Children were better at identifying the antecedent on average in Grade 4, M = 84%, SE = .02, than they were in Grade

3, M = 80%, SE = .03. There was also a main effect of Reading Skill on response accuracy: Good readers (1 SD above the mean), M = 92%, SE = .03, were better on average than poor readers (1 SD below the mean), M = 63%, SE = .09.

In response time, we found a main effect of Gender: Responses were given faster in the Informative Gender condition, M = 3615ms, SE = 128, than in the Non-Informative Gender condition, M = 3819ms, SE = 135. In addition, there was a main effect of Grade: Children were faster to respond on average in Grade 4, M = 3275ms, SE = 117, than they were in Grade 3, M = 4216ms, SE = 150. Finally, there was a main effect of Reading Skill: Good readers answered significantly faster, M = 2780ms, SE = 20, than poor readers, M = 4966ms, SE = 355. In addition, the Grade × Gender interaction was significant in response time: Post-hoc analyses showed that the simple main effect of Gender was significant only in Grade 3, $t = 4.85$, $p < .001$, but not in Grade 4, $|t| < 2$, $p = .146$. The Grade × Reading Skill interaction was also significant: Post-hoc comparisons showed that the simple main effect of Grade was smaller in good readers, Δ = 574ms, $t = 4.97$, $p < .001$, than in poor readers, Δ = 1412ms, $t = 10.44$, $p < .001$, with a significant difference effect, $t = -4.42$, $p < .001$. In summary, the effect of Gender on response accuracy remained stable with age. An unexpected effect of Gender emerged in response time, such that the informative gender cue had a facilitative effect on children's response times in Grade 3, but not in Grade 4. This may be explained by a ceiling effect such that the gender cue manipulation did not affect response times in the same way as in Grade 3.

Online measures

Gaze duration in the pronoun region in Grade 3 was positively correlated with gaze duration in Grade 4, $r = .78$, $t(68) = 10.41$, $p < .001$, as was total reading time, $r = .79$, $t(68) = 10.62$, $p < .001$ and gopast time, $r = .84$, $t(68) = 12.90$, $p < .001$. Regression probability was moderately correlated in Grade 3 and Grade 4, $r = .45$, $t(68) = 4.20$, $p < .001$.

Descriptive statistics for eye tracking measures in the pronoun and post-pronoun region are given in Table 8.2, and the results from the mixed-effect models are given in Table 8.4. To describe the effect of individual differences in reading skill on the eye movement measures, we quantified the effect of reading skill at 1 *SD* above and 1 *SD* below the mean reading score using contrasts.

Table 8.2. Overview of observed means.

	Pronoun region		Post-pronoun region	
	Inf. gender	Non-inf. gender	Inf. gender	Non-inf. gender
		Gaze duration		
Grade 3	569 (376)	557 (375)	364 (245)	364 (223)
Grade 4	455 (291)	453 (279)	305 (191)	299 (158)
		Total reading time		
Grade 3	819 (552)	795 (567)	516 (378)	516 (356)
Grade 4	662 (450)	614 (427)	428 (315)	401 (273)
		Gopast time		
Grade 3	761 (609)	761 (713)	495 (435)	482 (402)
Grade 4	643 (630)	595 (523)	413 (365)	377 (303)
		Regression probability		
Grade 3	.20 (.02)	.20 (.02)	.18 (.02)	.17 (.02)
Grade 4	.20 (.02)	.15 (.02)	.20 (.02)	.17 (.02)

Note. Standard deviations are given in parentheses. Inf. gender = Informative gender, Non-inf. Gender = non-informative gender.

Table 8.3. Results from the linear mixed-effects models over offline measures. χ^2 values for response accuracy. ANOVA *F* values for response time.

	Offline measures	
	Accuracy	Time
Gender	66.76 ***	19.11 ***
Grade	20.37 ***	243.39 ***
Reading	8.92 **	26.78 ***
Gender × Grade	0.33	4.05 *
Gender × Reading	0.59	0.90
Grade × Reading	0.07	6.52 *
Gen. × Grade × Reading	2.47	0.01

Note. * = $p < .05$, ** = $p < .01$, *** = $p < .001$.

Table 8.4. Results from the linear mixed-effects models over online measures in the pronoun region. ANOVA *F* values for reading time measures. χ^2 values for probability of regression out.

	Gaze	Total	Gopast	Regr. prob.
		Pronoun region		
Gender	0.08	9.08 **	2.35	4.31 *
Grade	139.76 ***	190.77 ***	123.76 ***	3.25
Reading	23.24 ***	40.03 ***	21.66 ***	2.36
Gender × Grade	0.01	1.35	1.24	4.70 *
Gender × Reading	1.58	3.85 *	5.95 *	6.27 *
Grade × Reading	22.54 ***	3.05	6.90 **	0.15
Gen. × Grade × Reading	3.60	0.00	0.02	0.26
		Post-pronoun region		
Gender	0.04	0.76	2.69	3.63
Grade	120.89 ***	147.00 ***	101.28 ***	0.52
Reading	36.47 ***	46.76 ***	23.51 ***	0.08
Gender × Grade	0.03	0.26	0.00	0.67
Gender × Reading	1.17	1.32	2.01	0.02
Grade × Reading	33.78 ***	13.10 ***	6.89 **	1.56
Gen. × Grade × Reading	0.12	0.03	1.67	0.50

Note. * = $p < .05$, ** = $p < .01$, *** = $p < .001$. Regr. prob. = Regression probability.

Pronoun region

In gaze duration, there was a main effect of Grade: Children in Grade 4 showed shorter gaze durations than in Grade 3 (Δ = 58ms). There was also a main effect of Reading Skill: Good readers had significantly shorter gaze durations in the pronoun region than poor readers (Δ = 178ms). In addition, the Grade × Reading Skill interaction was significant. The simple main effect of Grade was not significant in the good readers, $|t| < 1$, $p = .476$, but was significant in the poor readers, Δ= 146 ms, $t = 10.83$, $p < .001$. Neither the main effects of Gender nor any interaction involving Gender were significant.

In total reading time, there were main effects of Gender and Grade: Total reading time was higher in the Informative condition, M = 608ms, SE = 24, than in the Non-Informative condition, M = 577ms, SE = 23. The main effect of Grade showed that children became faster readers in Grade 4 (Δ = 107ms). There was also a main effect of Reading Skill: Good readers spend less time in the pronoun region than poor readers (Δ = 675ms). In addition, the Gender × Reading Skill interaction was significant: The simple main effect of Gender was significant in good readers, Δ= 59ms, $t = 2.89$, $p < .01$, but not poor readers, $|t| < 1$, $p = .562$.

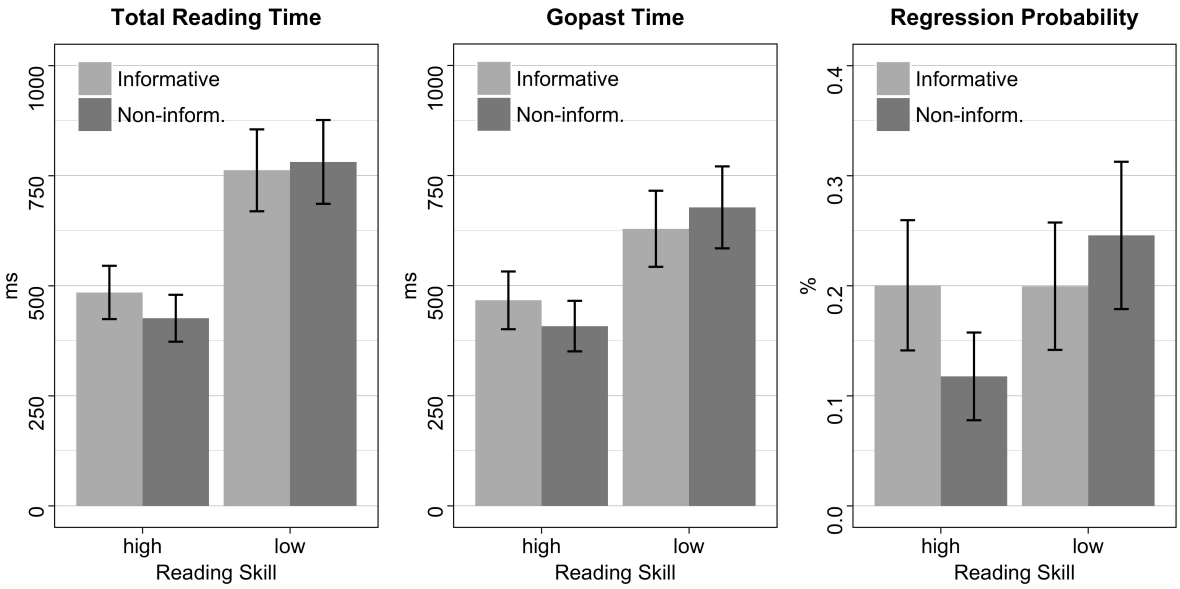

Figure 8.2. Means for total reading time (left panel), gopast time (mid panel), and regression probability (right panel), back-transformed to milliseconds and probability, respectively, at 1 *SD* above the mean (good reading skill) and 1 *SD* below the mean (poor reading skill), in the two Gender conditions. Error bars represent 2 standard errors.

The means of good (+1 *SD*) and poor (–1 *SD*) readers in the two gender cue conditions are depicted in Figure 8.2 (left panel). In addition, the full distribution of total reading times in the two gender cue conditions as a function of reading skill is provided in Figure 8.3 (top left panel).

In gopast time, we found no main effect of Gender but a main effect of Grade: Children had shorter gopast times in Grade 4, M = 503ms, SE = 20, than in Grade 3, M = 697ms, SE = 28. In addition, there was a main effect of Reading Skill and interactions of Gender × Reading Skill (see Figure 8.2, mid panel and Figure 8.3, top right panel). The simple main effect of Gender was significant in good readers, Δ = 59ms, t = 2.70, p < .01, but not poor readers, $|t|$ < 2, p = .118. There was also an interaction of Grade × Reading Skill. From Grade 3 to Grade 4, good readers significantly reduced their gopast times, Δ = 70ms, t = 2.79, p < .01. In poor readers, this reduction was significantly larger, Δ = 269ms, t = 8.06, p < .001. In summary, our findings suggest that children with better reading skill spend more processing time in the pronoun region during the second pass when it contains useful information for resolution. Against our expectations, these effects do not change from Grade 3 to Grade 4.

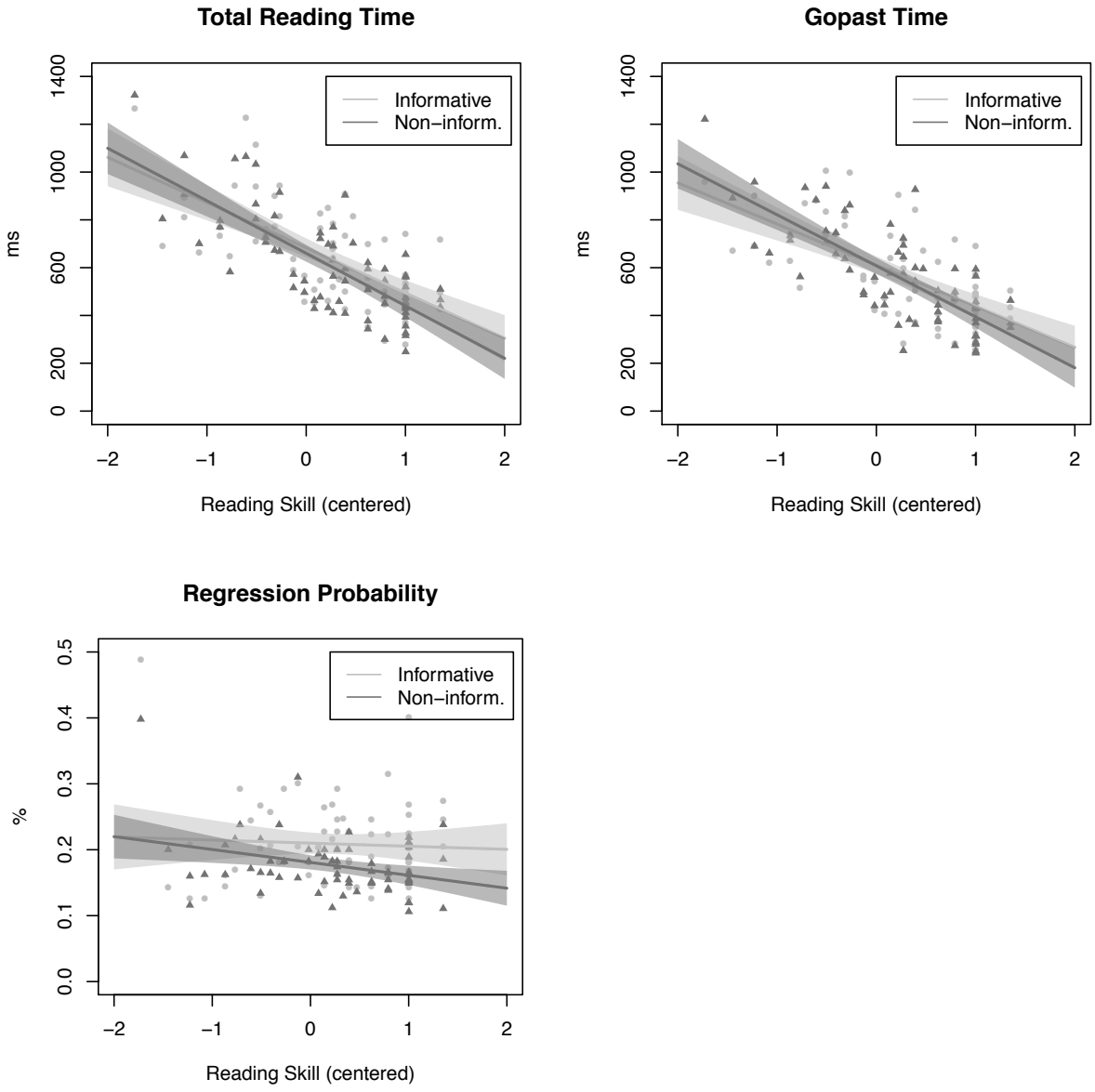

Figure 8.3. Distribution of total reading time (top left panel), gopast time (top right panel), and regression probability (bottom left panel) in the two Gender conditions as a function of reading skill (centered), back-transformed to milliseconds and probability, respectively. Confidence intervals represent 2 standard errors.

In regression probability, there was a main effect of Gender, as well as a significant Gender × Reading Skill interaction. In addition, the interaction of Gender × Reading Skill was significant (Figure 8.2, right panel and Figure 8.3, bottom left panel). The simple main effect of Gender was only significant in good readers, Δ =.08, t = 2.87, p < .01, but not in poor readers, $|t| < 2$, p = .187. Good readers made more regressions when Gender was informative, M = 20%, SE = .03, than when it was not informative, M = 12%, SE = .02. In addition, the Gender × Grade interaction was also significant: The simple main

effect of Gender was significant in Grade 4, $t = 2.74$, $p < .01$, but not in Grade 3, $|t| < 1$, $p = .884$.

Post-pronoun region

In the post-pronoun region, we found no significant effects of Gender or interactions with Gender in any of the reported measures (see Table 8.4). There were however main effects of Grade and Reading Skill, as well as interactions of Grade × Reading Skill in gaze duration, total reading time and gopast time: In gaze duration the simple main effect of Grade was not significant for good readers, $|t| < 1$, $p = .476$, but was significant for poor readers, $\Delta = 164$ms, $t = 10.83$, $p < .001$. Similarly, in total reading time, the simple main effect of Grade was not significant for good readers, $|t| < 2$, $p = .072$, but was significant for the poor readers, $\Delta = 229$ms, $t = 9.78$, $p < .001$. In gopast time, the simple main effect of Grade was significant for good readers, $\Delta = 34$ms, $t = 2.07$, $p < .05$, as well as the poor readers, $\Delta = 157$ms, $t = 7.32$, $p < .001$. The difference of the effect of Grade for good and poor readers was also significant, $t = 2.63$, $p < .01$. In regression probability, we found no effects at all in the post-pronoun region.

In summary, as there were no effects of Gender, or interactions of Gender × Grade or Gender × Reading Skill in the post-pronoun region, we may conclude that there were no spill-over effects of Gender information from the pronoun region. The effects of Grade are similar to those found in the pronoun region, indicating that the children become faster, more fluent readers with age.

Exploratory analyses of antecedent position

To further explore how children resolved the pronoun in our experiment, we conducted an additional analysis of the effects of antecedent position. Recall that the resolution preference was counterbalanced in the sentences. In each Gender condition therefore half of the antecedents were in subject position and therefore mentioned first, while the other half were in object position and mentioned second. We calculated a set of additional models in which we added the factor Antecedent Position (Mentioned First vs. Mentioned Second), everything else being equal.

The effects of Gender, Age and Reading Skill reported above remained significant in the offline measures after the addition of Antecedent Position into the model. There was a

main effect of Position on resolution ability: Children were better at selecting the plausible antecedent in the Mentioned Second condition (object antecedent), M = 75%, SE = .02, than in the Mentioned First condition (subject antecedent), M = 67%, SE = .02, t = 31.45, p < .001. In response time, there was also a main effect of Antecedent Position, such that questions in the Mentioned Second condition were answered significantly faster, M = 3495ms, SE = 132, than in the Mentioned First condition, M = 3950ms, SE = 150, t = 9.48, p < .01. In both measures, there were no interactions with other factors, all t < 2.2. In the online measures, i.e., gaze duration, total reading time, gopast time and regression probability, there were no effects of Antecedent Position, all t < 2. To summarize these results: Children are faster and more accurate, i.e., conform to the plausible context more often, when the pronoun refers to the second-mentioned, or last-mentioned antecedent. However, effects of Gender, Age and Reading Skill remain robust after accounting for Antecedent Position. We saw no indication that Position influences online reading behavior at the pronoun.

Discussion

The present study investigated how children use gender and context information when resolving pronouns. We presented sentences containing pronouns with informative and non-informative gender cues to children in Grade 3 and again in Grade 4. We found that disambiguating gender information had a positive effect on children's ability to determine the correct referent after reading. While children's general resolution ability improved from Grade 3 to Grade 4, the effect of gender information on resolution remained stable. We further showed that disambiguating gender information on the pronoun affected late processing measures in children, but this effect was moderated by reading skill: Only children with high reading skill used the gender information immediately during reading, such that they invest more processing time when an informative gender cue can be used to resolve the pronoun on the spot. We conclude that children with higher reading skill invest available processing resources towards local inference generation. We discuss the findings from the offline and online measures separately in the remainder of the Discussion.

Comprehension of the pronoun

Results from our offline measures showed that children clearly struggled with the assignment of an antecedent for the pronoun in our study, particularly in the absence of gender information as a resolution cue. When the pronoun could only be resolved on account of context information, given by the main verb and the subclause, children's accuracy dropped significantly. This is in line with earlier observations (Yuill & Oakhill, 1986). We interpret these findings to suggest that German children in Grade 3 and Grade 4 typically need explicit resolution cues for the resolution of pronouns when reading. Although they allowed themselves more response time when there was no informative gender cue, many children seem unable to find a plausible antecedent for the pronoun using context information alone. Further, while children's overall resolution ability improved with age, the effect of the gender information cue on resolution remained stable. This indicates that children in the final phase of primary school may not have developed the necessary inference skills to resolve pronouns in the absence of explicit cues, regardless of Grade level or reading skill.

But how do children decide on an antecedent? The results from our additional analysis of antecedent position showed that children often chose the last-mentioned person as the antecedent for the pronoun, even when this interpretation is not supported by the sentence context. This might indicate that children resort to default resolution strategies, as has previously been shown in listening studies (e.g., Megherbi & Ehrlich, 2005). Further, the study of children's comprehension of relative clauses has shown that children predominantly interpret object relative clauses as subject relative clauses (e.g., Adani, Van der Lely, Forgiarini, & Guasti, 2010). The authors suggest that the children fail to interpret the syntactic dependencies. This supports our interpretation that children do not sufficiently take the sentence context into account when resolving the pronoun. Since this additional analysis is based on exploratory results, it should be treated with some caution. Our results, however, certainly warrant further investigation into children's strategies of pronoun resolution during reading, and their effects on sentence and text comprehension.

Online processing of the pronoun region

In addition to the offline comprehension measures, we recorded children's eye movements in the pronoun area to obtain a detailed picture of the incremental reading processes at the pronoun. While the offline measures reflect children's response behavior after having read the whole sentence, the online measures provide information on the moment-to-moment processing of the pronoun when it is encountered. We were interested in the processing of the pronoun region because it indicates whether the children use information from the gender cue immediately during reading. The results are clear-cut. First, we found that when children initiate regressions and rereading, they did so directly from the pronoun region and not the post-pronoun region. This is true even for the poor readers, who did not show the delayed effects of processing we had hypothesized. Our results are consistent with Joseph et al. (2015), who found evidence of anaphoric processing in children beginning directly on the anaphor itself. Second, effects of gender information only occurred in the good readers, and only in late processing measures, specifically regression probability and total reading time. Because there were no effects in gaze duration, the effects in total reading time are entirely attributable to rereading. Based on the results for gopast times, we can say that the informative pronoun does not induce extensive rereading of the previous sentence regions. This indicates that good readers, but not poor readers, adjust rereading time of the pronoun region. The individual differences in online processing were substantial: Only children with good reading skill had longer total reading times in the pronoun region and made more regressions from the pronoun in the informative cue condition, when the antecedent was unambiguous. Our results are compatible with cue-based approaches to memory retrieval in sentence processing (e.g., Lewis, Vasishth, & Van Dyke, 2006; Patil et al., 2016), which assume that proficient readers use different types of information, including morpho-syntactic gender information, towards pronoun resolution immediately when it becomes available. It appears that when reading processing is effortful for children, they may not allocate attention to these retrieval cues. Another explanation is that children lack the necessary reading experience to identify morpho-syntactic information, such as pronoun gender, as a relevant cue during online reading. In both scenarios, beginning readers may then resort to default strategies for pronoun resolution.

Good readers among the children use gender information immediately when it is informative to resolve the pronoun, hence their longer processing times. This suggests that children with good reading comprehension skill process key areas in a sentence differently from children with poor reading comprehension skill: Children with good comprehension skill reread selectively and adjust their processing time to the informative content of the pronoun. This is in line with earlier findings for individual differences in children's regression behavior (Murray & Kennedy, 1988).

Studies of children's reading development have repeatedly found that faster word decoding does not necessarily lead to successful comprehension (for a review see Nation, 2005). It is noteworthy that although children's reading fluency improved considerably from Grade 3 to Grade 4, the effect of reading skill on the use of the disambiguating gender information remained stable in our study. Thus, despite faster word reading, children with poor reading skill did not automatically "catch up" in their pronoun comprehension or the way in which they process the pronoun region during reading.

Considering our offline and online results together, we conclude that many children are unable to resolve a pronoun during sentence reading when they cannot do so immediately. In other words, when the children cannot resolve the pronoun on the spot based on an explicit, informative gender cue, they are unlikely to do so later in the sentence or after reading. It seems that when resolution is difficult because it requires integration of context information, many children do not invest the necessary effort to construct a coherent representation of what they have read.

In sum, the results of our study show that German children at the end of primary school still struggle with the resolution of pronouns during reading, particularly when they need to take the sentence context into account to identify a plausible antecedent. While the accuracy of pronoun resolution generally improved from Grade 3 to Grade 4, children in Grade 4 still benefit from an explicit, informative gender cue and have not yet reached adult resolution efficiency.

III. General Discussion

9 SUMMARY OF RESULTS

The present thesis investigated pronoun resolution in 8- to 9-year-old children by recording their eye movements while reading sentences and short texts. In three studies, children's online pronoun resolution and post-reading pronoun comprehension were studied by using different, selected textual paradigms. Importantly, the studies took into consideration children's reading skills, their reading fluency, and their reading development with age. The results of the studies were interpreted within the bonding-and-resolution framework, and they show how incomplete pronoun resolution may result in incoherent mental models in children's reading. In the following, the main findings of each study will be summarized.

Study 1 investigated whether beginning readers benefit when the pronoun is replaced by a repeated name. The hypothesis was that a repeated name would preempt the need for a local inference at the anaphor and therefore improve reading fluency. This is in contrast to the finding of the so-called repeated name penalty effect in adults. The results from this first study suggest that, contrary to our hypothesis, children also show a repeated name penalty effect. This means that children, like adults, are sensitive to the salience, or accessibility, of an anaphor and expect that accessible discourse entities are referenced by a pronoun. In the texts with only one highly accessible antecedent, both adults and children made more regressions from a repeated name than a pronoun. Unrelated to the pronoun, adults and children also made more regressions when the anaphor had a greater distance to the antecedent. This reading behavior occurred regardless of whether the anaphor was a repeated name or a pronoun. In turn, this suggests that the distance to the antecedent influences anaphor resolution regardless of anaphor type.

Study 2 used a mismatch paradigm to investigate the type of information that children rely on during online pronoun resolution. It investigated beginning readers' sensitivity to gender mismatches of pronoun and antecedent. In the first experiment with fourth graders, there were no qualitative differences between children's and adults' processing of the pronouns. Both children and adults had longer gaze durations on the pronoun in the mismatch condition. In the first experiment, surprisingly few children signaled their detection of an error in the sentences. Consequently, the experiment was

replicated with a larger group of children in order to further investigate differences between these "detectors" and the children who did not recognize an error, termed "non-detectors".

In that second experiment of the study, the main results were replicated: The children had longer gaze durations in the pronoun area when the pronoun was a mismatch to the antecedent. This is despite the fact that only 57% of the fourth graders detected the gender mismatch during the course of the experiment. The eye movements of these "detectors" differed from that of the non-detectors and were more similar to the adults' eye movements in experiment 1. The non-detectors had comparably slower gaze durations overall and were less likely to make a regression from the pronoun region. This suggests that more fluent readers among the children, as signaled by their shorter overall gaze durations, have more processing resources available for pronoun resolution. They apparently can direct resources to repair grammatical inconsistencies immediately. Further, the results of experiment 2 fit with the bonding and resolution framework, which assumes that a first step in pronoun resolution is the (automatic) bonding of pronoun and antecedent via grammatical features, followed by the (strategic) resolution of said pronoun by taking into account additional information from inside and outside the text. The results from the analysis of gaze durations on the pronoun, a rather early measure, show that bonding is disrupted for all readers – even for those who do not report a mismatch after reading. The report of a mismatch correlates with regressions, a late eye movement measure, indicating a more strategic process.

Study 3 complements the results of the earlier studies by taking into account further the interindividual differences between the children of the same age groups that determine successful pronoun resolution. In this study, 70 children read sentences containing a pronoun either with or without a gender cue for resolution. The experiment was conducted in a semi-longitudinal paradigm with children in Grade 3 (8 years of age) and the same children in Grade 4 (9 years of age). The focus of this study was on the differences in pronoun processing and comprehension between the two Grade levels, and the effects of reading skill. The results of the comprehension testing showed that children's pronoun resolution ability improved with age, but good readers were more accurate than poor readers across Grade levels. The eye tracking measures showed strong individual differences related to reading skill. Children with good reading skill

had longer total reading times in the pronoun region when there was a gender cue. This means they took more time to read the pronoun region when the gender information on the pronoun was useful for resolution and the pronoun was easier to resolve. Good readers, it seems, make better use of the available gender information on the pronoun than poor readers. This is true for children in Grade 3 and also Grade 4.

The aim of the remainder of this section is a discussion of the results in reference to the research questions established in section 4. While the studies section includes an in-depth discussion of the respective findings of each experiment, the aim here is to reconcile these results and give a more comprehensive picture of the findings of the present thesis.

10 Pronoun resolution processes in children

The main question of the present thesis concerned children's online pronoun processing and resolution. Given that pronoun resolution is a complex cognitive process involving inference generation, it was of interest whether children resolve pronouns online during reading and how they do so.

The results of this thesis clearly show that children aged 8-9, when they are in the second half of primary school, struggle with the comprehension of pronouns in reading tasks. This result is right in line with the literature on children's comprehension of pronouns using post-reading questions and other reading tasks (Ehrlich & Rémond, 1997; Ehrlich, Rémond, & Tardieu, 1999; Oakhill & Yuill, 1986; Yuill & Oakhill, 1988; Megherbi & Ehrlich, 2005). Nearly half of the children in study 2 were unimpressed with a pronoun without a fitting antecedent in the sentence, which suggests that they did not attempt to resolve the pronoun towards an antecedent in the first place. Given that the sentences in study 2 were relatively short, with two possible and only one plausible antecedent, it is conceivable that these children would equally struggle with the resolution of pronouns in longer texts. Children's responses to pronoun comprehension questions of the form *Who did [action in the sentence]?* in study 3 revealed that even fourth Graders struggle significantly when they have to infer a pronoun from context and without a resolution cue that is locally available at the pronoun. When there is such

a cue in the form of gender information that disambiguates the antecedent for a pronoun, their performance improves significantly. One might therefore assume that children do not attempt to resolve the pronoun in more difficult contexts, when there is no direct resolution cue. The results also show that children profit from such cues and are able to use them, at least after reading and when prompted by a question. This is what we learn from the response data that were recorded as part of the experiments in the present thesis. The specific contribution of this thesis to the field, however, is the analysis of children's eye movements and the insights gained into moment-to-moment reading processes. This is where we turn from children's pronoun comprehension to children's processing of pronouns.

Children are sensitive to pronouns as discourse-level cues in the text. They seem to understand that a pronoun is appropriate when an antecedent is highly accessible, and that a repeated name in its place is rather anomalous. This is based on the results found in study 1, where children's eye movements at the repeated name were very similar to those of the adults. Note, however, that in this first experiment, children were not tested on their reading skill or reading fluency. These individual skills proved to influence pronoun processing in study 2 and study 3. While study 1 makes an important contribution to our understanding of children's processing of repeated names, it is limited with respect to how the component skills of reading may influence this process. Given what we know from the consecutive studies in this thesis, the results from study 1 may apply only to children with better reading skill and/or reading fluency, whereas children with poorer reading skill may show a different eye movement pattern. Further, the question remains how the processing of pronouns and repeated names is related to pronoun comprehension, and text comprehension at large. To understand how children's text comprehension is affected by a repeated name, future studies should consider using longer texts with more instances of pronouns and repeated names. This would open up the possibility of following up with comprehension questions that are directly related to the text at hand, which in turn might make the relationship between eye movement measures and reading outcomes more transparent.

One of the main research questions of the present thesis regarded the time course of pronoun processing in children's reading. The temporal pattern of eye movements vis-

ible in study 2, that is, longer gaze durations at the mismatching pronoun for both detectors and non-detectors, but longer reading times only for the detectors, can also be interpreted in terms of the bonding and resolution framework (see section 2.2.2). The famework describes the time course of pronoun resolution as a two-step process: first the association of pronoun and antecedent by way of feature match (bonding) and subsequently the alignment of this information with the given discourse context (resolution). Children had longer gaze durations at mismatching than matching pronouns, conceivably showing a bonding failure at the mismatching pronoun. Some children then make an immediate regression, probably in order to reread the offending region and resolve the pronoun. This temporal pattern of eye movements may be taken to show automatic bonding (longer gaze durations for detectors and non-detectors in study 2) and resolution that is decidedly strategic (extended total reading times only in children who report an incongruence). Our results therefore corroborate the assumption that bonding is automatic, but resolution requires strategy, as the eye movements of all children, even those who did not realize that there was a mismatch, slowed down on the mismatching pronouns (longer gaze durations). One might argue that the results from mismatch paradigms, as was used in study 2, do not perfectly align with pronoun resolution, but may also reflect more general processes of "repair" in the face of inconsistent text information. Therefore, study 3 investigated further the time course of pronoun resolution in grammatical sentences, varying only the informativeness of the pronoun by using gender information as a resolution cue. The results of study 3 support the view that regressions at the pronoun indicate resolution, as regressions were significantly more frequent in the informative condition (see also Joseph, Bremner, Liversedge, & Nation, 2015). Note however that there were significant interindividual differences with respect to the reading behavior at the pronoun, which are discussed in a separate section below.

The informativeness of the pronoun, that is, whether there is disambiguating gender information, is very relevant for children's processing behavior. When the pronoun can be resolved on the spot, children take more time to read it, presumably in an effort to integrate the referential information into their mental model. As discussed above, children's response accuracy to the pronoun comprehension questions improves significantly when there is a gender cue, also indicating that regressions at the pronoun are indeed a resolution strategy. As in study 3, the results obtained from the analysis of

gopast times indicate that children reread only the pronoun region, and do not extensively revisit previous text regions. While it is conceivable that this is partly due to the brevity of the experimental sentences (a decision that was taken to keep the amount of reading material manageable for children) it can also be interpreted in terms of the cognitive processes related to pronoun resolution.

In sum, it seems that during natural reading, children either resolve a pronoun immediately or not at all. If they cannot make the resources for resolution available on the spot, it is unlikely that they resolve the pronoun later during reading, or invest additional reading time in the resolution process. Other than adults, children do not use the discourse context (i.e., semantic plausibility) for online resolution, but resort to the structural default of taking the last-mentioned antecedent as referent, even if such an interpretation is implausible. It is important to note that there are important interindividual differences in the processing and comprehension of pronouns. The remainder of the discussion addresses these differences.

10.1 Development of pronoun resolution

The question how children use the available gender information at the pronoun during natural reading was addressed in a semi-longitudinal experiment: Children took part in study 3 in Grade 3 (approx. age 8) and one year later in Grade 4 (age 9). In this study, comprehension questions were used to assess children's pronoun resolution accuracy after having read sentences that contained pronouns with and without a gender cue. The results from this study indicate that pronoun comprehension improves in this time span, as reflected by the resolution questions. This is expected because as children get faster readers, they are assumed to improve their integration ability as they have more cognitive resources available for comprehension monitoring, and consequently make more inferences as they read. Following the situation model approach, this leads to better reading comprehension.

While the children's resolution accuracy improved in Grade 4, the disambiguating gender information influenced pronoun resolution accuracy in children of all ages, suggesting that primary school children generally benefit from resolution cues. The children in

Grade 4 still perform significantly worse than the adult control group of university students. It can be concluded that the adults integrate additional text information to resolve the pronoun even when there is no direct cue. The adults name the more plausible referent for the pronoun based on the sentence context. For example, in *Paul envied Theo because he had a pool at home*, the adults name *Theo*, the object of the main clause, as the referent for the pronoun because Theo having a pool is likely the source of Paul's envy. By using context information, in *Felix bored Florian because he always told the same stories,* adults conversely name *Felix*, the subject of the main clause, as the referent because people who repeat their stories over likely bore others. Note that the pronoun can be resolved in these examples only after the whole sentence has been read. This is in contrast to the same items in the gender cue condition, where the pronoun can be resolved much earlier: Consider *Felix bored Fatima because he…*, where the referent is already disambiguated upon encountering *he*. Successful integration of the pronoun in the condition without the gender cue requires a delay of the resolution process downstream from the pronoun and therefore retaining the sentence information in memory until the pronoun can be resolved. One explanation for our results therefore may be that children in Grade 4 still have problems retaining enough sentence information for this task, such that they resort to the last-mentioned entity when asked to resolve the pronoun. It is clear that further investigation into the development of children's pronoun resolution is warranted. The results from the present thesis suggest it would be worthwhile to extend this study with children in Grade 5 and 6 to follow up on their pronoun comprehension development and find out when exactly they reach a competence level that is similar to that of the adults. The children that were studied for this thesis, somewhat unexpectedly, do not seem to be at the key age for pronoun resolution development. The question remains when children make the transition from requiring local resolution cues to incorporating context information for the resolution of pronouns.

The major interest of the present thesis was in children's online reading behavior at the pronoun, however, there were no substantial differences from Grade 3 to Grade 4 in the eye movement measures. Again, this might indicate that the relevant developmental steps occur only later, and future studies might want to investigate pronoun and inferential processing with children in higher Grade levels. However, it might also be the case that age group simply is not the relevant factor when it comes to eye movements

in reading comprehension tasks. The interindividual differences in children's reading skill are very large and may, in practice, span up to three Grade levels. It is such a well-known fact that in most German schools, reading instructors and German teachers follow a practice called "internal differentiation" (e.g., von Brand & Brandl, 2017) within the cohort, meaning they use teaching materials tailored to the needs of up to four different reading levels for children *at the same Grade level*. Indeed, the results of the present thesis suggest that there are interindividual differences with respect to reading behavior, however, these are not as much connected to age than reading skill and reading fluency. These results are discussed in section 7.4 below.

To sum up, while the children improve their resolution accuracy from Grade 3 to Grade 4, we can assume that they reach adult competence only later. It should be noted that these results are important for educational practice. First, direct reading instruction is not a part of the curriculum after Grade 4 in many schools, even though it is clear that many children still struggle with basic inference generation at that point. Second, in many German schools, children in Grade 4 are expected to make the transition from *learning to read* to *reading to learn*. Third, the results discussed above suggest that many children are not yet ready for this transition, and that they might profit from more direct reading practice and specialized instruction even in Grade 4 (see Francey & Cain, 2014). The children may be able to fluently read words by then, but they do not always build a coherent situation model. If educators ignore these findings, children may be at risk of falling behind in other subject areas, simply because of their inability to read coherently.

10.2 Individual differences in pronoun resolution

The results from the present dissertation strongly suggest that pronoun resolution is not uniform in children. Particularly the strategic processes in resolving a pronoun are subject to individual differences. In experiment 2, all children showed an early disruption of the reading flow, evidenced by gaze durations, at a mismatching pronoun. However, only some children follow this up with a regressive eye movement, resulting in longer total reading times. Regressive eye movements at key areas of the text have previously been interpreted in terms of comprehension monitoring of good readers (Connor et al., 2014; Ehrlich, Rémond, & Tardieu, 1999; Murray & Kennedy, 1988). The

difference in late reading time measures between the detectors of the mismatch and the non-detectors in the present study was interpreted such that the detectors may delay the incoming of new information for a moment at the mismatching pronoun in order to make sense of it first. In the present thesis, reading skill and reading fluency were taken into account to help explain how the detectors differ from the non-detectors.

As part of the DevTrack study, reading skill and reading fluency were measured as follows: Reading skill was measured using ELFE 1-6, a standardized German reading comprehension test (Lenhard & Schneider, 2006). This test comprises three subtests targeting word comprehension, sentence comprehension, and text comprehension. Reading fluency was measured using the SLRT-II (Moll & Landerl, 2010). In this test, children read a list of words and then a list of non-words as fast and accurately as possible. A score is calculated from the number of words per minute that they read accurately. Primary-school children generally show large variances in both reading skill and reading fluency, and in the present thesis these were correlated with children's online reading behavior, specifically in late online measures, which capture the more strategic processes at hand.

Reading fluency was correlated with detection of incongruent pronouns. In study 2, it was found that children who read more fluently were more likely to report a mismatching pronoun. These results can be interpreted in terms of better comprehension monitoring: Children who monitor their reading comprehension should detect an incongruent pronoun because these children would want to link the pronoun to an antecedent. Studies on comprehension monitoring often use mismatch paradigms, like it was done in study 2. Comprehension monitoring failure is then exemplified by an inability to detect anomalies in the text. Presumably, this leads to an impoverished mental model of the text. The pattern of results in study 2 was very similar to that in a study on comprehension monitoring by Connor et al. (2014). They show that only children with good academic language skills (i.e., vocabulary skill, background knowledge and understanding of text structure) attempt at repairing an inconsistency online by spending more time rereading. However, all children, including those with weak academic language skills who did not detect the inconsistency, had longer gaze durations on anomalous words. The authors conclude that a disruption of the reading flow (longer gaze durations) is necessary, but not sufficient for incongruence detection: In addition, there

needs to be a deliberate, strategic reanalysis. This is the same pattern that we saw in study 2 regarding pronoun resolution. There was, however, no correlation between detection of the incongruence and reading skill in study 2, rather, the discriminating factor here was reading fluency. It is not entirely clear where the difference originates from, but note that the academic language skills tests in the study by Connor et al. (2014) and the ELFE 1-6 in our study measure rather different component skills of reading comprehension such that they are not directly comparable.

Study 3 of this thesis extended the study of individual differences in pronoun resolution and asked which children make use of a resolution cue online. It was discussed above that children can use the gender cue online for pronoun resolution when it was informative. However, the effect of gender information was only found in the good readers (ELFE score +2 *SD* from the mean; see study 3 for details). This means that the good readers made more regressions from the pronoun region when it was informative for resolution. Note that the good readers generally read more fluently and make less regressions in the course of a sentence, but do make more regressions presumably when higher-order reading processes require them. In their pioneering eye tracking study with children, Murray and Kennedy (1988) observed these "selective reinspections" in good readers and distinguished them from what they called the "backtracking" of poor readers, which were more erratic regressive eye movements across all regions of the text. In the present thesis, good readers had longer total reading times on the pronoun; however, not in the more difficult condition, but in the easier condition in which the pronoun could be resolved on the spot. Presumably, this reflects resolution in good, but not poor readers. Similar temporal patterns have been observed in other studies on local anaphor resolution and have been interpreted in the same direction (Joseph, Bremner, Liversedge, & Nation, 2015; Joseph, Wonnacott, & Nation, 2021). Specifically, in a recent study, Joseph et al. (2021) directly compared the processing of inconsistent words and local anaphors, and found that passage reading times were not increased when encountering an inconsistent word over a local inference. Passage reading times were even longer in the inference condition compared to the inconsistent condition when a comprehension question, encouraging inferencing during reading, was presented before the reading task. In light of these results, it is likely that in study 2, longer total reading times on the pronoun region indeed reflect resolution processes, also a type of local inference process, only by the good readers, while the poor readers do not

seem to engage in resolution processes.

Taken together, the findings of the present thesis show that children's reading skill and fluency impact pronoun resolution processes online, and that these processes can be observed using eye movement measures. The discussion above shows that additional studies are needed to investigate how individual differences are related to pronoun resolution, for example: What is the mechanism linking fluency to successful use of resolution cues? It was suggested that the link might be better memory capacity by fluent readers, however, this should be investigated further. A better understanding of the relationship between reading skill, component skills of reading, and pronoun resolution would also help in identifying those children in need of specialized reading instruction to foster coherent reading.

11 FINAL CONCLUSIONS

Children in the second half of primary school still struggle with the resolution of pronouns and have not yet reached adult competence. This can be seen both in offline comprehension measures and also, which was the main endeavor of this thesis, in online resolution processes. Children do not resolve pronouns spontaneously online. They do expect pronouns in a discourse to refer to accessible antecedents; however, in texts with more than one antecedent, only the good readers among the children make use of a gender cue online to resolve a pronoun. Where there is no local, disambiguating resolution cue, children may resort to rough heuristics, like recency, and as a result fail to resolve the pronoun correctly. They do not routinely take into account more global text information, such as plausibility or verb meaning, although this slightly improved from Grade 3 to Grade 4. The use of a gender cue towards pronoun resolution could be observed in the good readers among the children, linking reading skill to online pronoun resolution. The fluent readers among the children further engaged in a targeted reanalysis of mismatching pronouns, which was linked to the detection of the mismatch. Taken together, the studies in this thesis show that gaze duration can be related to bonding of pronoun and antecedent, while regressive eye movements and total reading time are linked to resolution. The results from this thesis suggest that only good readers among the children resolve pronouns online in Grades 3 and 4 (aged 8-9), leaving many

children with an incoherent representation of the text due to incomplete pronoun resolution.

IV. References

Adani, F., Van der Lely, H. K., Forgiarini, M., & Guasti, M. T. (2010). Grammatical feature dissimilarities make relative clauses easier: A comprehension study with Italian children. *Lingua*, *120*(9), 2148–2166.

Almor, A. (1999). Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review*, 106(4), 748–765

Almor, A., & Nair, V. A. (2007). The form of referential expressions in discourse. *Language and Linguistics Compass,* 1(1-2), 84–99.

Ariel, M. (2001). Accessibility Theory: An overview. *Text representation: Linguistic and psycholinguistic aspects*, 8, 29–87.

Ariel, M. (2004). Accessibility marking: Discourse functions, discourse profiles, and processing cues. *Discourse Processes*, *37*(2), 91–116.

Arnold, J. E., Brown-Schmidt, S., & Trueswell, J. (2007). Children's use of gender and order-of-mention during pronoun comprehension. *Language and Cognitive Processes*, *22*(4), 527–565.

Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., Trueswell, J. C. (2000). The rapid use of gender information: evidence of the time course of pronoun resolution from eye tracking. *Cognition*, 7*6*(1), 13–26.

Badecker, W., & Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *28*, 748–769.

Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes (2011). R package version 0.999375-42.

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

Best, R. M., Floyd, R. G., & McNamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology*, *29*(2), 137–164.

Blythe, H. I. (2014). Developmental changes in eye movements and visual information encoding associated with learning to read. *Current Directions in Psychological Science, 23*(3), p. 201–207.

Blythe, H. I. and Joseph, H. (2011). Children's eye movements during reading. In S. Liversedge, I. Gilchrist, & S. Everling (Eds.): *Oxford Handbook on Eye Movements* (pp. 643–662). Oxford: Oxford University Press.

Cain, K., & Oakhill, J. (2006). Profiles of children with specific reading comprehension difficulties. *British Journal of Educational Psychology*, *76*(4), 683–696.

Cain, K., & Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing*, *11*(5), 489–503.

Cain, K., Oakhill, J. V., Barnes, M. A., & Bryant, P. E. (2001). Comprehension skill, inference-making ability, and their relation to knowledge. *Memory & Cognition*, *29*(6), 850–859.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology*, *96*(1), 31–42.

Casteel, M. A., & Simpson, G. B. (1991). Textual coherence and the development of inferential generation skills. *Journal of Research in Reading*, *14*(2), 116–129.

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

Connor, C. M. D., Radach, R., Vorstius, C., Day, S. L., McLean L., & Morrison, F. J. (2014). Individual differences in fifth graders' reading and language predict their comprehension monitoring development: An eye-movement study. *Scientific Studies of Reading 19*(2), 114–134.

Cunnings, I., & Felser, C. (2013). The role of working memory in the processing of reflexives. *Language and Cognitive Processes*, *28*(1/2), 188–219.

Currie, N. K., & Cain, K. (2015). Children's inference generation: The role of vocabulary and working memory. *Journal of Experimental Child Psychology*, *137*, 57–75.

de Beaugrande, R., & Dressler, W. U. (1981). *Einführung in die Textlinguistik* [*Introduction to text linguistics*]. Tübingen: Niemeyer Verlag.

Drieghe, D., Brysbaert, M., Desmet, T., De Baecke, C. (2004). Word skipping in reading: On the interplay of linguistic and visual factors*. European Journal of Cognitive Psychology 16* 79–103*.*

Drieghe. D., Brysbaert, M., Desmet, T. (2007). How important are linguistic factors in word skipping during reading? *British Journal of Psychology 98*, 157–171.

Ehrlich, M. F., Remond, M., & Tardieu, H. (1999). Processing of anaphoric devices in young skilled and less skilled comprehenders: Differences in metacognitive monitoring. *Reading and Writing*, 11(1), 29–63.

Eilers, S., Tiffin-Richards, S. P., & Schroeder, S. (2018). Individual differences in children's pronoun processing during reading: Detection of incongruence is associated with higher reading fluency and more regressions. *Journal of Experimental Child Psychology*, *173*, 250–267.

Eilers, S., Tiffin-Richards, S. P., & Schroeder, S. (2019a). The repeated name penalty effect in children's natural reading: Evidence from eye tracking. *Quarterly Journal of Experimental Psychology*, *72*(3), 403–412.

Eilers, S., Tiffin-Richards, S. P., & Schroeder, S. (2019b). Gender cue effects in children's pronoun processing: A longitudinal eye tracking study. *Scientific Studies of Reading*, *23*(6), 509–522.

Elbro, C., Oakhill, J., Megherbi, H., & Seigneuric, A. (2017). Aspects of pronominal resolution as markers of reading comprehension: The role of antecedent variability. *Reading and Writing*, *30*(4), 813–827.

Engelen, J. A., Bouwmeester, S., de Bruin, A. B., & Zwaan, R. A. (2014). Eye movements reveal differences in children's referential processing during narrative comprehension. *Journal of Experimental Child Psychology*, 118, 57–77.

Fagan, S. M. (2009). *German: A Linguistic Introduction*. Cambridge: Cambridge University Press.

Fox, J., Friendly, M., & Weisberg, S. (2013). Hypothesis tests for multivariate linear models using the car package. *The R Journal*, *5*(1), 39–52.

Francey, G., & Cain, K. (2015). Effect of imagery training on children's comprehension of pronouns. *The Journal of Educational Research*, *108*(1), 1–9.

Friederici, A. D. (1995). The time course of syntactic activation during language processing: A model based on neuropsychological and neurophysiological data. *Brain and Language*, *50*, 259–281.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. Trends in Cognitive Science, *6*(2), 78–84.

Fukumura, K., & van Gompel, R. P. (2015). Effects of order of mention and grammatical role on anaphor resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(2), 501-525.

Gagl, B., Hawelka, S., & Wimmer, H. (2015). On sources of the word length effect in young readers. *Scientific Studies of Reading*, *19*(4), 289–306.

Garnham, A., & Oakhill, J. (1985). On-line resolution of anaphoric pronouns: Effects of inference making and verb semantics. *British Journal of Psychology*, *76*(3), 385–393.

Garnham, A., Oakhill, J., & Johnson-Laird, P. N. (1982). Referential continuity and the coherence of discourse. *Cognition*, 11(1), 29–46.

Garrod, S., & Terras, M. (2000). The contribution of lexical and situational knowledge to resolving discourse roles: Bonding and resolution. *Journal of Memory and Language*, 42(4), 526–544.

Gernsbacher, M. A. (1989). Mechanisms that improve referential access. *Cognition* 32, 99-156.

Gesellschaft für Deutsche Sprache (Ed.) (2006). *Liste der beliebtesten Vornamen*. [List of most popular given names.] Wiesbaden: Gesellschaft für Deutsche Sprache.

Givón, T. (1983). *Topic continuity in discourse*. Amsterdam: John Benjamins.

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

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*(3), 371–395.

Greene, S. B., McKoon, G., & Ratcliff, R. (1992). Pronoun resolution and discourse models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(2), 266–283.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.): *Syntax and Semantics III: Speech Acts*, (p. 41–58). New York: Academic Press.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. *Proceedings of the 21st annual meeting on Association for Computational Linguistics*, 44–50.

Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, *21*(2), 203–225.

Hickmann, M., Schimke, S., & Colonna, S. (2015). From early to late mastery of reference. L. Serratrice & S. E. M. Allen (Eds.), *The Acquisition of Reference* (pp. 181–211). Amsterdam: John Benjamins.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*(3), 346–363.

Hothorn, T., Bretz, F., Westfall, P., Heiberger, R. M., Schuetzenmeister, A., Scheibe, S., & Hothorn, M. T. (2015). Package 'multcomp'. Version 1.4-0.

Hußmann, A., Wendt, H., Bos, W., Bremerich-Vos, A., Kasper, D., Lankes, E.-M., McElvany, N., Stubbe, T. C., Valtin, R. (Eds.) (2017). *IGLU 2016. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich.* [*IGLU/PIRLS 2016. Reading skills of German primary school children compared to international standards*.] Münster, New York: Waxmann.

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, 94, 316–339.

Janich, N. (Ed.) (2008). *Textlinguistik: 15 Einführungen* [*Fifteen introductions to text linguistics*]. Tübingen: Gunter Narr Verlag.

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology, 95*(4), 719–729.

Joseph, H. S., & Liversedge, S. P. (2013). Children's and adults' on-line processing of syntactically ambiguous sentences during reading. *PloS one*, *8*(1), e54141.

Joseph, H. S., Bremner, G., Liversedge, S. P., & Nation, K. (2015). Working memory, reading ability and the effects of distance and typicality on anaphor resolution in children. *Journal of Cognitive Psychology*, *27*(5), 622–639.

Joseph, H. S., Liversedge, S. P., Blythe, H. I., White, S. J., Gathercole, S. E., & Rayner, K. (2008). Children's and adults' processing of anomaly and implausibility during reading: Evidence from eye movements. *The Quarterly Journal of Experimental Psychology*, *61*(5), 708–723.

Joseph, H. S., Wonnacott, E., & Nation, K. (2021). Online inference making and comprehension monitoring in children during reading: Evidence from eye movements. *Quarterly Journal of Experimental Psychology*, *Advance online publication*. doi: 10.1177/1747021821999007.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological Review*, *99*(1), 122–149.

Kehler, A. (2002). *Coherence, reference, and the theory of grammar*. Stanford: CSLI Publications.

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

Kennison, S. M., & Gordon, P. C. (1997). Comprehending referential expressions during reading: Evidence from eye tracking. *Discourse Processes*, *24*(2-3), 229–252.

Koornneef, A. W., & Van Berkum, J. J. (2006). On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye tracking. *Journal of Memory and Language*, *54*(4), 445–465.

Lenhard, W., & Schneider, W. (2006). *ELFE 1-6: Ein Leseverständnistest für Erst-bis Sechstklässler* [ELFE 1-6: A reading comprehension test for first through sixth grade]. Göttingen, Germany: Hogrefe.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29,* 375– 419.

Lewis, R. L., Vasishth, S, & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences, 10*, 447–454.

Lezama, C. G. (2015). A review of the repeated name penalty: Implications for null subject languages. *Revista LinguiStica*, *8*(2), 22–34.

McDonald, J. L., & MacWhinney, B. (1995). The time course of anaphor resolution: Effects of implicit verb causality and gender. *Journal of Memory and Language*, *34*, 543–543.

McKoon, G. & Ratcliff, G. (1992). Inference during reading. *Psychological Review, 99*(3), 440–466.

McKoon, G., & Ratcliff, R. (2005). "Meaning Through Syntax" in sentence production and comprehension: Reply to McRae et al. (2005). *Psychological Review, 112*(4), 1032–1039.

McKoon, G., Gerrig, R. J., & Greene, S. B. (1996). Pronoun resolution without pronouns: Some consequences of memory-based text processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(4), 919–932.

Megherbi, H., & Ehrlich, M. F. (2005). Language impairment in less skilled comprehenders: The on-line processing of anaphoric pronouns in a listening situation. *Reading and Writing*, *18*(7), 715–753.

Megherbi, H., & Ehrlich, M. F. (2009). The on-line interpretation of pronouns and repeated names in seven-year-old children. *Current Psychology Letters. Behaviour, Brain & Cognition*, *25*(2), doi: cpl.revues.org/4895.

Moll, K. & Landerl, K. (2010). *SLRT-II: Lese- und Rechtschreibtest. Weiterentwicklung des Salzburger Lese- und Rechtschreibtests (SLRT).* Göttingen: Hogrefe.

Murray, W. S., & Kennedy, A. (1988). Spatial Coding in the Processing of Anaphor by Good and Poor Readers: Evidence from Eye Movement Analyses. *The Quarterly Journal of Experimental Psychology, 40*(4), 693–718.

Nation, K. (2005). Children's reading comprehension difficulties. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 248–265). Oxford, UK: Blackwell.

Oakhill, J. (1994). *Individual differences in children's text comprehension.* In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (p. 821–848). New York: Academic Press.

Oakhill, J. V. (1982). Constructive processes in skilled and less-skilled comprehenders. *British Journal of Psychology*, *73*, 13–20.

Oakhill, J. V. (1984). Inferential and memory skills in children's comprehension of stories. *British Journal of Educational Psychology*, *54*, 31–39.

Oakhill, J. V., Berenhaus, M. S., & Cain, K. (2015). Children's reading comprehension and comprehension difficulties. In A. Pollatsek & R. Treiman (Eds.), *The Oxford Handbook of Reading* (pp. 344–359). Oxford: Oxford University Press.

Oakhill, J., & Yuill, N. (1986). Pronoun resolution in skilled and less-skilled comprehenders: Effects of memory load and inferential complexity*. Language and Speech*, *29*(1), 25–37.

Oakhill, J., Hartt, J., & Samols, D. (2005). Levels of comprehension monitoring and working memory in good and poor comprehenders. *Reading and Writing*, *18*(7), 657–686.

Omanson, R. C., Warren, W. H., & Trabasso, T. (1978). Goals, inferential comprehension, and recall of stories by children. *Discourse Processes*, *1*(4), 337–354.

Patil, U., Vasishth, S., & Lewis, R. L. (2016). Retrieval interference in syntactic processing: The case of reflexive binding in English. *Frontiers in Psychology*, *7:329*, doi: 10.3389/fpsyg.2016.00329.

Perfetti, C. A., & Hart, L. (2001). The lexical basis of comprehension skill. In D. S. Gorfein (Ed.): *On the Consequences of Meaning Selection: Perspectives on Resolving Lexical Ambiguity*, (pp. 67–86). Washington, DC: American Psychological Association.

Perfetti, C. A., & Hogaboam, T. (1975). Relationship between single word decoding and reading comprehension skill. *Journal of Educational Psychology, 67*(4), 461–469.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*(1), 22–37.

Perfetti, C., Landi N, & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The Science of Reading: A Handbook* (pp. 227–247). Oxford, UK: Blackwell.

Perfetti, C. A. (1994). Psycholinguistics and reading ability. In M. A. Gernsbacher (Ed.), *Handbook of Psycholinguistics* (pp. 849–894). Amsterdam: Elsevier.

Peters, S. A., Boiteau, T. W., & Almor, A. (2016). Semantic relations cause interference in spoken language comprehension when using repeated definite references, not pronouns. *Frontiers in psychology, 7*, 214.

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.

Rayner, K., Kambe, G., & Duffy, S. A. (2000). The effect of clause wrap-up on eye movements during reading. *Quarterly Journal of Experiment Psychology, 53*(4), 1061–1080.

Reinhart, T. (2011). Processing or pragmatics? Explaining the coreference delay. In E. Gibson & N. J. Pearlmutter (Eds.), *The Processing and Acquisition of Reference* (pp. 157–194), Cambridge: MIT Press.

Richter, T., Isberner, M.-B., Naumann, J., & Kutzner, Y. (2012). Prozessbezogene Diagnostik von Lesefähigkeiten bei Grundschulkindern [Process-based measurement of reading skills in primary school children]. *Zeitschrift für Pädagogische Psychologie*, *26*, 313–331.

Rigalleau, F., Caplan, D., & Baudiffier, V. (2004). New arguments in favour of an automatic gender pronominal process. *Quarterly Journal of Experimental Psychology*, Section A, *57*(5), 893–933.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, *23*(4), 441–474.

Sanford, A. J., Garrod, S., Lucas, A., & Henderson, R. (1983). Pronouns without explicit antecedents? *Journal of Semantics*, *2*(3-4), 303–318.

Schimke, S. (2015). Die rezeptive Verarbeitung von Markierungen der Diskurskohärenz bei Grundschulkindern mit Deutsch als Erst- oder Zweitsprache [The processing of markers of discourse coherence of primary school children with German as their first or second language]. In G. Pagonis & H. Klages (Eds.), *Linguistisch fundierte Sprachförderung und Sprachdidaktik: Grundlagen, Konzepte, Desiderate* (pp. 303–324). Berlin; Boston: De Gruyter.

Schroeder, S., Hyönä, J., & Liversedge, S. P. (2015). Developmental eye-tracking research in reading: Introduction to the special issue. *Journal of Cognitive Psychology*, *27*, 500-510.

Schroeder, S., Würzner, K. M., Heister, J., Geyken, A., & Kliegl, R. (2014). childLex: A lexical database of German read by children. *Behavior Research Methods*, *47*(4), 1085–1094.

Sekerina, I. A. (2015). Online evidence for children's interpretation of personal pronouns. In L. Serratrice & S. E. M. Allen (Eds.): The acquisition of reference, (p. 213–240). Amsterdam: John Benjamins.

Shapiro, A., & Milkes, A. (2004). Skilled readers make better use of anaphora: A study of the repeated-name penalty on text comprehension. *Electronic Journal of Research in Educational Psychology*, 2(2), 161–180.

Song, H. J., & Fisher, C. (2005). Who's "she"? Discourse prominence influences preschoolers' comprehension of pronouns. *Journal of Memory and Language*, *52*(1), 29–57.

SR Research (2009). *Experiment Builder* (Version 1.10) [Computer software]. Mississauga, Ontario, Canada: SR Research Ltd.

SR Research (2011). *Data Viewer* (Version 1.11.9) [Computer software]. Mississauga, Ontario, Canada: SR Research Ltd.

Stafura, J., & Perfetti, C. (2017). Integrating word processing with text comprehension. Theoretical frameworks and empirical examples. In K. Cain, D. L. Compton, & R. K Parilla (Eds.), *Theories of Reading Development* (pp. 9–31). Amsterdam/New York: John Benjamins.

Stewart, A. J., Holler, J., Kidd, E. (2007). Shallow processing of ambiguous pronouns: Evidence for delay. *Quarterly Journal of Experimental Psychology*, *60*(12), 1–18.

Stracuzzi, D., & Kinsey, J. (2006a). *EyeDoctor* (Version 0.6.5) [Computer software]. Retrieved from http://www.psych.umass.edu/eyelab/software/

Stracuzzi, D., & Kinsey, J. (2006b). *EyeTrack* (Version 7.10m) [Computer software]. Retrieved from http://www.psych.umass.edu/eyelab/software/

Tiffin-Richards, S. P., & Schroeder, S. (2015). Children's and adults' parafoveal processes in German: Phonological and orthographic effects. *Journal of Cognitive Psychology, 27*(5), 531–548.

van den Broek, P., Lorch, R. F., Linderholm, T. & Gustafson, M. (2001). The effects of readers‘ goals on inference generation and memory for texts. *Memory & Cognition, 29*, 1081–1087

van den Broek, P., Risden, K., & Husebye-Hartman, E. (1995). The role of readers’ standards for coherence in the generation of inferences during reading. In R. F. Lorch & E. J. O’Brien (Eds.), *Sources of Coherence in Reading* (p. 353–373). Amsterdam: Lawrence Erlbaum Associates.

van der Schoot, M., Reijntjes, A., & van Lieshout, E. C. D. M. (2012). How do children deal with inconsistencies in text? An eye fixation and self-paced reading study in good and poor comprehenders. *Reading & Writing*, *25*, 1665–1690.

Van Dijk, T. A., Kintsch, W., & Van Dijk, T. A. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Von Brand, T., & Brandl, F. (2017). *Deutschunterricht in heterogenen Lerngruppen*. [*Teaching German in heterogenous classrooms*.] Hannover: Klett Kallmeyer.

Vonk. W. (1984). Eye movements during the comprehension of pronouns. In A. G. Gale & F. Johnson (Eds.), *Theoretical and Applied Aspects of Eye Movement Research* (pp. 203–212). Amsterdam: North-Holland.

Vorstius, C., Radach, R., Mayer, M. B., & Lonigan, C. J. (2013). Monitoring local comprehension monitoring in sentence reading. *School Psychology Review*, *42*(2), 191–206.

Wonnacott, E., Joseph, H. S., Adelman, J. S., & Nation, K. (2016). Is children's reading "good enough"? Links between online processing and comprehension as children read syntactically ambiguous sentences. *The Quarterly Journal of Experimental Psychology*, *69*(5), 855–879.

Wykes, T. (1981). Inference and children's comprehension of pronouns. *Journal of Experimental Child Psychology*, *32*(2), 264–278.

Yuill, N., & Oakhill, J. (1988). Understanding of anaphoric relations in skilled and less skilled comprehenders. *British Journal of Psychology*, *79*(2), 173–186.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185.

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological science*, *6*(5), 292–297.

V. APPENDIX

Allgemeine Erklärungen zur Dissertationsschrift

1. Ich erkläre, an keiner anderen Hochschule ein Promotionsverfahren eröffnet zu haben.

2. Ich erkläre, dass die Dissertation in der gegenwärtigen Fassung keiner anderen Hochschule zur Begutachtung vorgelegen hat oder vorliegt.

3. Ich erkläre, dass die Arbeit selbstständig und ohne Hilfe Dritter verfasst wurde und bei der Abfassung alle Regelungen guter wissenschaftlicher Standards eingehalten wurden.

Datum Unterschrift