Towards Commonsense Reasoning in Automated Anaphora Resolution: Investigating Reference Resolution Acquisition in Children

An Honors Thesis for the Department of Computer Science

Jasmine Falk

Tufts University, 2020

Abstract

The ability to resolve anaphoric expressions is a crucial task in natural language processing (NLP). The accuracy of anaphora resolution models can have a major impact of the performance of other NLP tasks, such as sentiment analysis, summarization, and machine translation. The most advanced anaphora resolution algorithms use deep learning techniques, relying on learning statistical patterns from vast amounts of data to make referent decisions. Yet, even today's state-of-the-art algorithms perform poorly on anaphoric instances that require commonsense reasoning or real-world knowledge to resolve.

Anaphora resolution algorithms must therefore incorporate commonsense reasoning to improve accuracy. However, existing research has not yet defined the minimal reasoning and inference abilities a model would need to incorporate, making improvements in this area difficult. In this thesis, we propose that reasoning and inference mechanisms children acquire to resolve anaphora could be applied to existing models of automated anaphora resolution to address this challenge.

Through (1) surveys of current literature on both state-of-the-art anaphora resolution algorithms and anaphora resolution acquisition in children, and (2) a preliminary analysis of a corpus of infant eye-tracking data, this thesis investigates the applicability of a psycholinguistic framework for making future improvements to anaphora resolution algorithms for resolving inference-based anaphoric instances.

To Iverson, who believed in me before I believed in myself,

and my parents, who never stop celerbating me.

Acknowledgements

I have been fortunate to have had the opportunity to work at the Human-Robot Interaction Lab at Tufts for several semesters. The work that I have been able to be a part of and the students and mentors at the HRI Lab have been incredibly influential to my interests in cognitive science and computer science. This thesis bridges these disciplines and would not have been possible without the interdisciplinary mentorship I have received and projects that have inspired me.

I would like to thank my thesis and major advisor, Dr. Matthias Scheutz, for his support in my research endeavors and interests throughout my time at the HRI Lab and during this thesis. His guidance in the process of determining a research direction made this thesis possible.

I am also indebted to Yayun Zhang and Dr. Chen Yu at the Computational Cognition and Learning Lab at Indiana University. Yayun guided me through the research and analysis process, and gave me much of her time to explain the data and turn my coding into results. The data used in this thesis was all collected and processed by Dr. Yu's lab, and I am incredibly grateful to them for sharing it.

Contents

List of Figures

Chapter 1

Introduction

As a ubiquitous phenomenon in natural language, anaphora has attracted the attention of researchers in the fields of natural language processing (NLP) and psycholinguistics for decades. The term *anaphora*, which comes from the Greek *ana*, meaning “back", and *pheri*, meaning “to carry", refers to the use of a word or phrase to reference an entity mentioned earlier in a discourse [Lust, 1986, Sukthanker et al., 2020]. The expression that refers, or “points back", to a preceding entity is called the *anaphor* [Sidner, 1979]. The entity being referred to by the anaphor is known as the *antecedent* [Yang et al., 2008].

Understanding anaphora therefore requires some computation; that is, an anaphor cannot be understood without first determining its antecedent [Lust, 1986, Palmović et al., 2018]. This process of identifying the entity that an anaphoric expression refers to, known as *anaphora resolution* [Sukthanker et al., 2020], is a uniquely challenging and crucial task in natural language processing. To understand the challenges of anaphora resolution, we open with an overview of the linguistic fundamentals of anaphora and anaphora resolution and a survey of the current state-of-the-art anaphora resolution algorithms.

1.1 Linguistic Background

A core component of natural language is the act of referring [Arnold, 2010]. Without the ability to refer to things, we would often find discourse redundant. Consider the following sentences from [White et al., 2017]:

> **Example 1.1:** The bee landed on the flower because it wanted pollen.
>
> **Example 1.2:** The bee landed on the flower because it had pollen.

The anaphora in **Example 1.1** and **Example 1.2** have been underlined. If each anaphor was replaced with its antecedent noun phrase (NP), the sentences would sound unnatural and repetitive. In the course of a conversation, interlocutors constantly use anaphora, some research suggesting perhaps 50 times per minute [Arnold, 2010]. Consequently, successful language comprehension hinges on the ability to correctly determine an anaphor's antecedent [Järvikivi et al., 2014]. For this reason, anaphora resolution is a critical task in NLP, with the potential to greatly increase accuracy in several other NLP tasks when it is used as a component in the pipeline of processing discourse [Sukthanker et al., 2020].

1.1.1 Types of Anaphora

Anaphora resolution is a challenging task in part due to the number of different forms a reference may take [Sukthanker et al., 2020]. This particular challenge is sometimes termed the "coverage issue", meaning that most reference resolution algorithm are designed to focus only on certain types of references [Sukthanker et al., 2020]. We will briefly discuss what are thought to be some of the most common forms of anaphora in natural language [Mitkov, 2007].

1.1.1.1 Pronominal Anaphora

The most prevalent type of anaphora by far, occurring often in everyday speech and text, is the *pronominal anaphora* [Sukthanker et al., 2020]. With pronominal anaphora, the antecedent is referred to by a pronoun, such as in the following sentence:

> **Example 1.3:** I do not like green eggs and ham. I do not like them, Sam-I-am (Dr. Seuss, *Green Eggs and Ham*).

Not all pronouns, however, are anaphoric. For instance, the pronoun "it" may also be non-anaphoric when it does not refer to a specific entity. Such pronouns are termed *pleonastic* [Mitkov, 2007].

> **Example 1.4:** The sun did not shine. It was too wet to play (Dr. Seuss, *The Cat in the Hat*).

Between a quarter and half of “it” instances are estimated to be pleonastic [Bergsma et al., 2008]. Because of the frequency of non-anaphoric pronouns, the NLP task of finding spans of text that make up possible antecedents, *mention detection*, is still an important open area of investigation [Jurafsky and Martin, 2000].

1.1.1.2 One Anaphora

For *one anaphora*, as the name suggests, the antecedent is referred to by a “one" NP, as in the following sentences:

> **Example 1.5:** This one has a little star. This one has a little car (Dr. Seuss, *One Fish Two Fish Red Fish Blue Fish*).

1.1.1.3 Split Anaphora

With *split anaphora*, the anaphor refers to more than one antecedent [Sukthanker et al., 2020]. For example, in the sentence below, the pronoun (3) refers to the union of antecedent (1) and antecedent (2):

> **Example 1.6:** Thing Two(1) and Thing One(2)! They(3) ran up! They(3) ran down! (Dr. Seuss, *The Cat in the Hat*).

1.1.2 Interpreting Anaphoric Expressions

The process of anaphora resolution typically involves identifying all NPs preceding the anaphor as potential candidates for the antecedent, and then applying linguistic constraints to filter out incorrect or unlikely references. To narrow down the search space and choose the correct antecedent, anaphora resolution relies on several *constraint* and *preference* rules [Mitkov, 2007]. Constraints are eliminating factors, ruling out certain candidate antecedents, while preferences rank candidate antecedents [Poesio et al., 2016]. The delineation between these two mechanisms is an important factor in many computational models of anaphora resolution [Poesio et al., 2016].

1.1.2.1 Constraints

Constraints are morphological and syntactic rules that eliminate certain NPs from the set of possible antecedent candidates [Poesio et al., 2016]. In this section, we provide a brief overview of some of the most commonly used constraints.

Gender and Number Agreement

Anaphora and their corresponding antecedents must agree in both number and grammatical gender [Jurafsky and Martin, 2000]. For instance, an antecedent representing multiple entities must be followed by a plural anaphor.

Person Agreement

In English, which delineates between first (e.g. “I” and “me”), second (e.g. “you”), and third person (e.g. “she” and “them”), pronouns must agree with their antecedent in person [Jurafsky and Martin, 2000, Sukthanker et al., 2020]. In the sentences below, the antecedent (1) agrees in person with the anaphor (2). If (2) were replaced with “they”, (1) would be eliminated as a possible antecedent.

> **Example 1.7:** And Sally and I(1) did not know what to say. Should we(2) tell her the things that went on there that day? (Dr. Seuss, *The Cat in the Hat*).

However, there are exceptions, such as in the following example where entities (1), (2), and (3) are all coreferent, but (2) and (3) do not agree in person:

> **Example 1.8:** Then our mother(1) came in and she(2) said to us two, “Did you have any fun? Tell me(3). What did you do?” (Dr. Seuss, *The Cat in the Hat*).

Semantic Consistency

An anaphor and its antecedent must be semantically sensible within the context of the sentence [Mitkov, 2007]. **Example 1.1** and **Example 1.2** demonstrate this type of constraint.

1.1.2.2 Preferences

In contrast with constraints, preferences do not always hold, and therefore do not eliminate antecedent candidates. Rather, they give greater salience to certain candidates [Mitkov, 2007].

Recency

Entities introduced most recently in a discourse are typically more salient (i.e. are more likely to be the correct antecedent) than entities that occurred earlier in the discourse [Jurafsky and Martin, 2000, Sukthanker et al., 2020].

Grammatical Role

Entities mentioned in the subject role of a sentence are more more salient than entities mentioned in the object position [Jurafsky and Martin, 2000]. **Example 1.1** follows the grammatical role preference. "The bee" as the subject, is the antecedent of the anaphor "it".

Repetition

Entities that are mentioned repeatedly are given greater salience than other candidates [Sukthanker et al., 2020].

1.2 Anaphora vs. Coreference Resolution

In the reference resolution literature, there is a great deal of overlap and confusion between the terms *anaphora* and *coreference* [Poesio et al., 2016], so it is necessary to discuss their usages. Anaphoric expressions, as noted above, are words or phrases that refer back to entities previously mentioned in discourse. Coreferent expressions are words or phrases that refer to the same real-world entity [Poesio et al., 2016, Stylianou and Vlahavas, 2019]. From these definitions, it is clear that anaphora and coreference resolution tasks are closely related [Poesio et al., 2016], but their relation is debated within the literature. Some research argues that anaphora resolution can be thought of as a subtask of coreference resolution [Stylianou and Vlahavas, 2019], and vice versa [Sukthanker et al., 2020]. However, there are cases where an anaphoric referent is not coreferent [Stylianou and Vlahavas, 2019, Sukthanker et al., 2020]. [Sukthanker et al., 2020] proposes a Venn diagram model of these two terms: in this model, 'anaphora resolution' and 'coreference resolution' overlap in cases of definite pronominal anaphora and identity references. Despite the debate, 'coreference' has become virtually synonymous with 'anaphora' [Poesio et al., 2016].

Because of the ambiguity surrounding these terms and this work's focus on the common challenge anaphora resolution and coreference resolution algorithms face–namely an inability to resolve references requiring commonsense reasoning–we will use the terms 'anaphora' and 'coreference' interchangeably within this thesis. For the interested reader, [Poesio et al., 2016], [Stylianou and Vlahavas, 2019], and [Sukthanker et al., 2020] provide detailed discussions of this terminological inconsistency.

1.3 Deep Learning Reference Resolution Algorithms

The field of anaphora resolution in NLP research has a rich history dating back to the 1960s, with a multitude of algorithms for resolving references being proposed over the decades. Existing algorithms can be broadly grouped into three categories: rule-based algorithms, statistical and machine learning algorithms, and deep learning algorithms. Many of the earliest resolution algorithms were rule-based, relying on a set of hand-crafted rules and heuristics drawn from linguistic knowledge [Sukthanker et al., 2020]. These algorithms were knowledge-intensive and tended to be inflexible [Poesio et al., 2016, Sukthanker et al., 2020]. Statistical and machine learning-based algorithms began emerging in the late '90s, offering more robust, knowledge-poor methods, a trend driven by greater availability of corpora and other NLP resources [Mitkov et al., 2001].

More recently, with advancements in hardware that can run complex neural models and the advent of word embeddings, reference resolution algorithms have demonstrated performance strides by employing deep neural networks [Stylianou and Vlahavas, 2019, Sukthanker et al., 2020]. Consequently, the focus in anaphora resolution research has shifted to developing and advancing deep learning algorithms in the last several years. Investigating psycholinguistic mechanisms of anaphora acquisition that can be applied to computational anaphora resolution algorithms requires a foundational understanding of the current state-of-the-art anaphora resolution models. To that end, we provide a survey of the existing seminal work on deep learning algorithms for reference resolution, as well as the limitations of these state-of-the-art algorithms.

The neural network-based models are typically classified into several categories based on differing approaches: (1) mention-pair classifiers, (2) entity-based models, (3) latent-tree or latent-structure models, (4) mention-ranking models, and (5) span-ranking models [Gu et al., 2018, Stylianou and Vlahavas, 2019]. Two of the most influential models, [Wiseman et al., 2015] and [Lee et al., 2017], have been mention-ranking and span-ranking algorithms, respectively. These models outperformed all existing approaches to reference resolution and inspired several improved algorithms that used them as baseline models. [Wiseman et al., 2015], [Lee et al., 2017], and the models they inspired have proven to be the current state-of-the-art algorithms in reference resolution. As such, we survey the literature on mention-ranking and span-ranking models, the core of the current state of research in deep learning reference resolution.

1.3.1 Mention-Ranking Models

The central approach of mention-ranking models is to score pairs of mentions for their likelihood of being coreferent [Clark and Manning, 2016a]. In other words, the goal is to identify how likely a mention refers to the same entity as an antecedent mention [Wiseman et al., 2015]. *Mentions* are words or phrases that either refer to another word/phrase or are referred to [Wiseman et al., 2015]. Mention-ranking models have a number of advantages in that they are fast, scalable, and relatively straightforward to train [Clark and Manning, 2016a].

[Wiseman et al., 2015] introduced a non-linear mention-ranking model for coreference resolution, meaning that they used a non-linear scoring function to rank mention pairs. During training, their model was able to learn feature representations that are useful for anaphoricity detection (whether or not a mention is anaphoric) and antecedent ranking (predicting the correct antecedents for mentions known to be anaphoric). Their model outperformed all other models at the time*.

Using [Wiseman et al., 2015] as a starting point, [Clark and Manning, 2016a] built a mention-ranking model, initially described in [Clark and Manning, 2016b]. [Clark and Manning, 2016a] demonstrated that reinforcement learning could be applied to a neural mention-ranking model to successfully optimize for evaluation metrics. When tested on the standard CoNLL 2012 English corpus, their model outperformed [Wiseman et al., 2015], making it one of the current-state-of-the-art approaches.

1.3.2 Span-Ranking Models

In contrast to mention-ranking models, span-ranking models do not rely on syntactic parsers and hand-crafted features to detect mentions and propose them as possible antecedents [Lee et al., 2017,Zhang et al., 2018]. Rather, they consider all *text spans*, a series of consecutive words [Subramanian and Roth, 2019], as possible mentions and learn distributions over possible antecedents, removing the need for hand-engineered mention detection systems [Lee et al., 2017, Zhang et al., 2018]. [Lee et al., 2017] proposed the first span-ranking, end-to-end neural reference resolution model. An *end-to-end system* must be able to detect mentions, detect anaphoricity, and link or cluster references together [Zhang et al., 2018]. The model introduced in [Lee et al., 2017], the first to perform all of these tasks required in end-to-end reference resolution, demonstrated that the use of syntactic parsers for mention detection actually hinders performance and that using vector embeddings to represent spans

*There are a number of standard metrics that are used to evaluate reference resolution models. Most commonly seen are B^3, MUC, and CEAF. Several coreference resolution corpora, such as CoNLL 2012 and MUC-6/MUC-7, are used to provide a standardized test set on which to evaluate different models. Although this survey will not explore the specifics of evaluation metrics and datasets, and instead will discuss algorithmic performance at a higher level, a detailed review can be found in [Poesio et al., 2016] and [Sukthanker et al., 2020].

of text in a document allows a model to jointly learn which spans are mentions and how to best cluster them. [Lee et al., 2017] demonstrated that their model outperformed all existing models in all standard evaluation metrics, including that of state-of-the-art mention-ranking models [Wiseman et al., 2015] and [Clark and Manning, 2016a].

[Zhang et al., 2018] then proposed an improvement to the model in [Lee et al., 2017] by using a different scoring algorithm for ranking possible antecedents. Rather than using the feed forward neural network used by [Lee et al., 2017], they proposed a biaffine attention model[†]. The change in how antecedent scores were computed optimized mention detection and mention clustering, improving on the performance of [Lee et al., 2017], making [Zhang et al., 2018] a state-of-the-art end-to-end resolution model.

[Gu et al., 2018] also build on the work of [Lee et al., 2017], introducing a coreference clustering modification algorithm to modify span clusters built by the model in [Lee et al., 2017]. The algorithm helps to rule out mentions that are not coreferential within clusters to increase the liklihood that spans in a cluster refer to the same entity [Gu et al., 2018]. While their model does not outperform the most advanced state-of-the-art systems in some metrics, [Gu et al., 2018] notes that their model's scores are still competitive, and their approach can be seen as a way of post-processing the output given by [Lee et al., 2017].

1.4 Limitations of Current Algorithms

Despite major advancements using deep learning, reference resolution models continue to struggle with the uphill battle of successfully resolving anaphoric expressions that require commonsense reasoning or real-world knowledge [Emami et al., 2018,Lee et al., 2017, Sukthanker et al., 2020]. *Commonsense reasoning* can be broadly defined as physical and behavioral knowledge about the world, and *world knowledge* can be thought of as connecting words and phrases to their semantic meaning in the

[†] The reader can refer to [Dozat and Manning, 2016] for a discussion on deep biaffine attention models.

real world [Sukthanker et al., 2020]. For instance, consider again the anaphora in **Example 1.1** and **Example 1.2**:

> **Example 1.1:** The bee(1) landed on the flower(2) because it wanted pollen.
>
> **Example 1.2:** The bee(1) landed on the flower(2) because it had pollen.

Although seemingly simple, sentences like these are difficult anaphoric cases for automated systems–whether rule-based, statistical, or neural–to resolve since they require reasoning about real-world entities [Emami et al., 2018]. Even state-of-the-art deep learning algorithms, such as [Lee et al., 2017], are unable to differentiate that the anaphor "it" in **Example 1.1** refers to antecedent (1) ("the bee"), while the same anaphor in **Example 1.2** refers to antecendent (2) ("the flower")[‡]. Sentences such as **Example 1.1** and **Example 1.2** are considered difficult cases because resolving their anaphora requires complex inference; syntactic and morphological cues alone do not indicate the correct antecedent here.

Because many of the standard corpora used to train and evaluate reference resolution algorithms contain relatively few complex instances that require inference or reasoning, automated systems can perform well on these datasets by simply exploiting statistical patterns and surface cues in the data [Emami et al., 2018]. When tested on corpora specifically constructed to target commonsense reasoning and world knowledge, such as the KnowRef corpus introduced in [Emami et al., 2018], state-of-the-art-models do not perform as well [Emami et al., 2018]. This leads to the question of what exactly these deep learning models are learning from training data: syntax patterns or semantic compatibility of referents based on world knowledge and context [Emami et al., 2018].

This performance gap between automated techniques and humans in the task of resolving ambiguous or difficult anaphoric instances can be explained by the failure of even state-of-the-art models to capture context, instead relying on morphosyntactical patterns and statistical knowledge of candidate antecedents, such as gender and

[‡] This was tested on AllenNLP's Coreference Resolution Demo, which uses an implementation of [Lee et al., 2017].

number constraints, to make reference decisions [Emami et al., 2018]. For instance, [Hobbs, 1979] notes that ambiguous pronouns in naturally-occurring text can often be resolved correctly using a preference based on grammatical role of the antecedents. Of course, this does not always work, and a system exploiting statistical patterns like this will consistently make incorrect decisions on realistic examples of anaphora [Levesque et al., 2011]. The reliance of reference resolution systems on statistical patterns in the training data to resolve anaphora can also be seen in the algorithmic gender biases of many models[§]. Natural language data inevitably reflects social stereotypes, and reference resolution systems that learn purely by finding patterns in syntax and morphology cues will make biased decisions based on biased data [Emami et al., 2018, Zhao et al., 2018].

The challenge of successfully resolving complex anaphora cases with commonsense reasoning and world knowledge has been addressed throughout the history of anaphora resolution research. Even some of the earliest rule-based algorithms attempted to account for the necessity of computational inference in reference resolution [Charniak, 1972]. More recently, this issue has been addressed through probabilistic reasoning [Liu et al., 2016], extraction of world knowledge from large-scale databases [Plu et al., 2018], knowledge attention mechanisms [Zhang et al., 2019], and more challenging benchmark corpora and tasks [Emami et al., 2018, Levesque et al., 2011]. Here we survey several seminal works that attempt to employ reasoning and real-world knowledge for anaphora resolution, as well as their limitations.

1.4.1 Charniak's Deep Semantic Processing

One of the first systematic attempts to account for inference in anaphora resolution was a knowledge-intensive system called called Deep Semantic Processing (DSP) introduced in [Charniak, 1972]. The work was motivated by the broader goal of creating a model that could use commonsense knowledge to comprehend

[§] Reference resolution algorithms are considered gender biased if the system links pronouns to occupations dominated by the gender of the pronoun more accurately than it links occupations not dominated by the gender of the pronoun [Zhao et al., 2018]. For instance, a biased model may be more likely to resolve the pronoun "he" to the NP "the doctor" and "she" to "the secretary".

children's stories, but a core component the model would need for comprehension would be an anaphoric resolver that could make decisions about complicated reference cases. [Charniak, 1972] proposed that DSP would take as input hand-coded assertions and statements as factual knowledge, and would then carry out deductive inferences using this real world knowledge [Poesio et al., 2016]. As a byproduct, anaphoric references would be resolved [Poesio et al., 2016]. However, DSP had a number of known issues: it was only partially implemented, never systematically evaluated, relied heavily on hand-crafted rules, and the inference mechanisms suggested were not particularly convincing [Poesio et al., 2016].

1.4.2 Benchmark Tasks

Corpora and tasks that accurately measure a system's ability to resolve difficult anaphoric cases is crucial to the advancement of research in this area. The standard corpora for evaluating the accuracy of a reference resolution algorithm have few cases of anaphora requiring inference [Emami et al., 2018]. When most cases in a reference resolution task can be resolved based solely on statistical knowledge, the task is not an accurate benchmark of reasoning and knowledge abilities. As alternatives tasks, the Winograd Schema Challenge and KnowRef corpora have been proposed.

1.4.2.1 The Winograd Schema Challenge

Perhaps the most influential work in encouraging research into commonsense reasoning and world knowledge for reference resolution has been the proposal of the the Winograd Schema Challenge (WSC), first presented in [Levesque et al., 2011] as an alternative to the Turing Test. [Levesque et al., 2011] defines a *Winograd schema* to be a pair of sentences that differ in at most two words and contain an ambiguous reference that is resolved in different ways in the two sentences. A Winograd schema should be designed such that a human reader would find the resolution of the references obvious, but an automated system could not easily resolve the reference through constraints and statistical patterns learned from text corpora. Instead, world knowledge and reasoning are required. For instance, **Example 1.1** and **Example 1.2** would be considered a Winograd schema pair. Only a system achieving human-level accuracy, presumably near 100%, would pass the WSC [Levesque et al., 2011].

1.4.2.2 KnowRef

[Emami et al., 2018] presents KnowRef, a corpus of more than 8,700 annotated text passages with ambiguous pronominal anaphora cases that require significant commonsense reasoning and knowledge. The corpus is meant to address the size limitation of the Winograd Schema Challenge corpus by [Levesque et al., 2011], which has only 273 instances, and offers a new benchmark task that targets a system's ability to reason in context.

1.4.3 Probabilistic Reasoning

[Liu et al., 2016] proposes a deep learning approach for probabilistic reasoning in artificial intelligence, called the neural association model (NAM). *Probabilistic reasoning*, a method of handling knowledge uncertainty in reasoning based on probability theory, can be used to predict conditional probability of one event given another. NAM uses non-linear activations in deep neural nets to model the conditional probabilities between any two possible events in a domain. [Liu et al., 2016] suggests that the commonsense knowledge required to resolve Winograd schema problems could be framed as an association relationship between separate events which NAM could predict. The work reported about 60% accuracy on Winograd schema problems, demonstrating that probabilistic reasoning using deep learning does solve some commonsense reasoning problems successfully, but still has a long way to go to achieve automatic reasoning.

1.4.4 World Knowledge Sources

One approach to incorporating real-world knowledge in reference resolution algorithms is to leverage semantic knowledge from external data sources, typically large-scale knowledge bases, such as Wikipedia, YAGO, WordNet, and FrameNet, and combine this semantic information with statistical or learning methods of reference resolution [Bryl et al., 2010, Ponzetto and Strube, 2006, Rahman and Ng, 2011,Uryupina et al., 2011]. The most recent attempt to incorporate semantic knowledge is the Sanaphor++ model proposed by [Plu et al., 2018]. Sanaphor++ improved upon the performance of the state-of-the-art deep neural network model from [Clark and Manning, 2016b] by adding semantic features extracted from the knowledge base DBpedia. Although [Plu et al., 2018] successfully demonstrated that using semantic knowledge extracted from knowledge bases offers improvements in performance even to state-of-the-art reference resolution models, it is important to note that the Sanaphor++ model, like all other advanced reference resolution systems, was evaluated on the standard CoNLL-2012 corpus, which has infrequent cases of complex or ambiguous anaphora and may not provide an accurate benchmark in measuring a system's knowledge-use and reasoning abilities [Emami et al., 2018].

1.4.5 Knowledge Attention Mechanisms

Most recently, [Zhang et al., 2019] presented an end-to-end state-of-the-art neural model that learns to resolve pronouns with general knowledge graphs (KGs). Moreover, since not all knowledge is always helpful, they introduced a knowledge attention module, which learns to select the most relevant and informative knowledge for pronoun resolution based on context. The proposed model that uses KGs and a knowledge attentional module proved to outperform the most recent state-of-the-art models by a large margin.

1.5 Applications of Anaphora Resolution in NLP

Within NLP research, anaphora resolution algorithms have received a great deal of focus. Part of this interest is due to anaphora resolution's *extrinsic evaluation*, or the impact an anaphora resolution module has when applied to a larger NLP system [Mitkov et al., 2007]. NLP tasks such as text summarization, opinion extraction, and text classification have shown improvement in accuracy with the incorporation of an anaphora resolution system. The experiments described in this section demonstrate that anaphora resolution is a key component of the processing pipeline of text that is used for a variety of NLP tasks, and that improvements to anaphora resolution algorithms will have wide-ranging impacts in NLP research.

1.5.1 Text Summarization

The NLP task of *summarization* involves extracting simplified discourse models from text [Steinberger et al., 2007]. One technique for summarization is Latent Semantic Analysis (LSA), which extracts 'hidden' dimensions of the semantic representation of terms and sentences based on their use [Steinberger et al., 2007]. [Steinberger et al., 2007], using an improved version of the anaphoric resolver GUITAR developed by [Poesio and Kabadjov, 2004], demonstrated that an LSA-based summarizer using anaphoric information was able to achieve significantly better performance than a summarizer that did not incorporate an anaphora resolution system. By checking for anaphora that might have been interpreted incorrectly in the summary and replacing them, the addition of an anaphoric resolver made the text produced by the summarizer more coherent [Steinberger et al., 2007].

1.5.2 Opinion Extraction

Opinion extraction, sometimes referred to as *opinion mining*, is a branch of NLP that aims to extract user perceptions from text to create a summary of the writer's opinions [Kobayashi et al., 2005, Saqia et al., 2018]. [Saqia et al., 2018] found that the application of anaphora resolution to an opinion extraction algorithm improved results in accuracy measurements of precision, recall, and f-score.

1.5.3 Text Classification

A wide range of supervised machine learning algorithms have been applied to *text classification*, also known as *text categorization*, the task of classifying documents into a set of defined categories [Yeh and Chen, 2003]. [Mitkov et al., 2007] experimented with four of these text classification methods (k nearest neighbors (kNN), naïve Bayes, maximum entropy, and support vector machines) to investigate the impact the application of an anaphoric resolution system would have on their performance. They found that the incorporation of MARS, the anaphora resolution algorithm introduced in [Mitkov et al., 2002], into the kNN method yielded a statistically significant improvement [Mitkov et al., 2007].

Although the other classification methods that were tested did not show a significant improvement with the application of MARS, it is worth noting that the success rate of MARS has been reported to range from 45% to 65%, depending on the evaluation data [Mitkov et al., 2007]. These accuracy rates have been exceeded in recent years by current state-of-the-art deep learning models, such as [Lee et al., 2017]. It is possible that these text classification algorithms could improve more significantly with the application of more accurate anaphora resolution systems.

[Yeh and Chen, 2003] also applied an anaphora resolution algorithm to text that was then fed into a text classification system. The results of the experiment demonstrated that incorporating the anaphora resolution algorithm improved the accuracy of the text classification system from 79% to 85%. On the whole, both [Yeh and Chen, 2003] and [Mitkov et al., 2007] found statistically significant improvements to their existing text classification methods with the addition of an anaphoric resolver.

1.6 Focus of the Study

Despite impressive performance improvements achieved by applying deep learning techniques to reference resolution algorithms, even today's state-of-the-art algorithms, namely those proposed by [Clark and Manning, 2016a], [Clark and Manning, 2016b], [Lee et al., 2018], and [Zhang et al., 2018], struggle to accurately resolve anaphoric instances that require reasoning or external world knowledge beyond morphosyntactic cues and statistical knowledge that can be learned from patterns in training data. Given proven applications of anaphora resolution in other NLP tasks, such as text summarization, opinion extraction, and text classification, advancements in the accuracy of reference resolution algorithms could yield downstream performance improvements across NLP tasks.

Attempts at addressing the need to incorporate commonsense reasoning and real-world knowledge using deep learning and external knowledge bases have made progress in recent years. However, these current approaches have limitations and have not been evaluated on tasks that accurately assess inference abilities. Consequently, the most effective method for improving performance on complex anaphora cases remains an open question.

The goal of this thesis is to explore a novel approach to improving current state-of-the-art anaphora resolution algorithms. We propose that inference mechanisms that children acquire to resolve anaphora may guide future improvements to automated reference resolution. Children, as developing systems, may provide a framework for how an automated system may acquire reasoning abilities for anaphora resolution.

In Chapter Two, we discuss motivations for investigating a developmental linguistics approach to automated anaphora resolution and provide a literature review on psycholinguistic methods and studies that have been used to understand anaphora resolution acquisition in young children. Chapter Three describes the use and analysis of infant eye-tracking data for investigating how children acquire anaphora resolution mechanisms, and offers a discussion of how this work and future analyses on this corpus can provide evidence for the applicability of developmental linguistics to making future improvements to anaphora resolution systems. Chapter Four highlights the contributions of this thesis within the field of NLP and examines directions for future work in using psycholinguistic methods for improving performance of anaphora resolution algorithms.

Chapter 2

Psycholinguistic Methods of Studying Anaphora Resolution Acquisition

"Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's?"

- Alan Turing, *Computing Machinery and Intelligence*

2.1 A Developmental Linguistics Approach to NLP

Artificial intelligence (AI) has long been inspired by architectures and cognitive mechanisms of the human brain. The very basis of the event that established AI as a field, the 1956 Dartmouth Summer Research Project on Artificial Intelligence, was the idea that all aspects of human intelligence could be described precisely enough for a machine to be able to simulate it [McCarthy et al., 1955]. Human intelligence is characterized by an ability to learn, and to learn quickly from relatively few "training" examples. Alan Turning noted that truly imitating an adult mind would involve simulating education and experiences applied to the initial state of the mind at birth, similar to a child's cognitive development [Turing, 1950].

This process of cognitive development can offer a potential framework for guiding research and algorithms that attempt to imitate human intelligence in some regard. Research in developmental psychology has shown that even preschoolers are capable of accomplishing certain learning tasks the most advanced machine learning algorithms cannot [Smith and Slone, 2017]. Notably, infants acquire extensive knowledge through self-directed action with minimal supervision, a remarkable fact when viewed in contrast to the quantity of labeled data that current machine learning methods require to accomplish highly specific tasks [Stojanov et al., 2019,Turek, 2018].

Consequently, there has been a recent push in AI research toward taking inspiration from and incorporating research from developmental psychology into designing automated systems. In 2018, the Defense Advanced Research Projects Agency (DARPA) announced the Machine Common Sense (MCS) program. The stated goal of MCS is to address the challenge of commonsense reasoning in AI through creating "a service that learns from experience, like a child, to construct computational models that mimic the core domains of child cognition" [Turek, 2018]. [Smith and Slone, 2017] and [Smith et al., 2018] explore the question of how to incorporate developmental insights into machine learning. They propose that the visual experiences of infants and toddlers create a developmentally ordered curriculum, essentially optimized training data, that allows human learners to gain object recognition abilities. [Yi et al., 2020] introduces a CLEVRER, a video reasoning benchmark for systematic evaluation of computational models that draws inspirations from developmental psychology. [Haber et al., 2018], using a neural network that implements curiosity-driven intrinsic motivation, mathematically formalizes the ability of infants to generate new behaviors in an unstructured environment.

Our work follows in this line of AI research that draws inspiration from developmental psychology. This thesis aims to demonstrate the relevance of developmental linguistics in moving forward research in the field of anaphora resolution, and more broadly, NLP, by exploring commonsense reasoning and inference mechanisms children use to resolve anaphora.

One challenge to improving performance of anaphora resolution algorithms using world knowledge and commonsense reasoning is that the scope of these domains is extremely broad [Sukthanker et al., 2020]. In the discussion of their state-of-the-art neural model, [Lee et al., 2017] note that their "model does little in the uphill battle of making coreference decisions requiring world knowledge," pointing out that a vastly larger corpus of data or external sources of knowledge would be required to integrate commonsense reasoning into their model. Incorporating vast amounts of data or external knowledge is inefficient. Data containing enough training examples of complex anaphora cases to overcome the sparsity of these patterns may be difficult to acquire, and large amounts of external knowledge may not always be helpful in many cases. [Zhang et al., 2019], discussed in Chapter 1, attempt to tackle this issue by incorporating a knowledge attention module into a deep learning anaphora resolver that learns to select the most relevant information based on context.

In the current work, we propose that the scope of knowledge and reasoning capabilities required for anaphora resolution can be narrowed by investigating mechanisms that children use to resolve complex cases of anaphora use. Studying children as developing systems offers insight into the general path of how reference resolution is acquired, as well as what abilities a fully functional reference resolution system requires [Järvikivi et al., 2014]. For instance, if the use of certain cues to resolve anaphora develops late and cannot be observed in young children, this could be an indication that these cues are less important for a mature system as well [Järvikivi et al., 2014], making it less likely that they would be useful for an automated resolution system.

In the next section, we survey the existing studies of anaphora resolution acquisition in children to understand the current psycholinguistic theories of how children develop reference resolution abilities, focusing on studies employing eye-tracking techniques due to their acknowledged value in studying developmental linguistics [Huettig et al., 2011, Sekerina, 2014].

2.2 A Survey of Anaphora Resolution Acquisition Studies

Some of the earliest research on children's developing use and understanding of anaphora began in the 1970s and '80s, with most of the studies focusing on the anaphoric function of pronouns [Charney, 1980, Chipman and de Dardel, 1974, Fine, 1978, Loveland, 1984, Maratsos, 1973, Umstead and Leonard, 1983]. For instance, [Umstead and Leonard, 1983] explored factors that influence children's understanding of pronouns, specifically pronoun and referent distance, pronoun location relative to its referent, and repetition of the pronoun. [Charney, 1980] and [Loveland, 1984], focusing on speech addressed to children, found that children first comprehend second person pronouns, then first person, and finally third person. [Lust, 1986] provides a comprehensive survey of the literature on these early experimental studies of anaphora resolution acquisition.

More recently, eye-tracking, the continuous monitoring of eye movements, has increasingly become an important tool for experimental psycholinguistic studies. Head-mounted cameras and eye-trackers worn by infants and toddlers can capture everyday visual environments from the perspective of the child learner [Smith et al., 2018], providing an opportunity to study language acquisition in real time [Sedivy, 2010]. Several studies have employed eye-tracking techniques to observe children's use and comprehension of anaphora. In the following section, we delve into the relationship of eye-tracking methodologies and studying language development, and the applications of these techniques in existing anaphora resolution acquisition studies.

2.2.1 Application of Eye-Tracking Methods

The use of eye-tracking in spoken language studies on children provides the opportunity to examine language processing in real time. These eye-tracking techniques involve monitoring eye movements such that a fine-grained temporal analysis of the eye gaze position with respect to critical words or sounds in the speech stream can be conducted [Sedivy, 2010].

2.2.1.1 Assumptions Behind Eye-Tracking Methods

Several basic assumptions underlying the use of eye-tracking demonstrate the validity and relevance of eye-tracking techniques in psycholinguistic study. An excellent discussion of the literature behind these assumptions can be found in [Sedivy, 2010], so we will only discuss them in brief here.

Attention-Bound Gaze

People tend to direct their eye gaze to the objects or people they are attending to in their visual environment. As we view a scene, we make a number of discrete eye movements, called *saccades*, averaging about five saccades per second [Sedivy, 2010]. Saccades are known to occur due to both external factors (e.g. movement or salience) and internal, cognitive factors that inform where we attend to. Because eye movement is attention-bound, we can use eye gaze to infer where a subject's attention is directed.

Referentially-Driven Movement

Eye movements are known to be mediated by referential links between linguistic expressions and their real-world referents. Interpreting language, with the goal of establishing reference in particular, is thought to be one of the primary internal cognitive factors that drive saccadic behavior.

Incremental Comprehension

Eye movements reflect incremental interpretation of spoken language as they appear to be easily triggered by only partial commitments to interpretation of linguistic input. Consequently, tracking eye gaze provides real time, moment-by-moment information on the processes behind language comprehension.

Reflective of Linguistic Computation

Eye movements reflect the output of complex linguistic interpretation, a crucial assumption to inferring psycholinguistic information from saccades.

Developmental Continuity

For all the above assumptions, there is reasonable developmental continuity. To use eye-tracking to study developmental trajectories in language processes, the above assumptions should apply at all points in a child's development.

2.2.1.2 Relevance to Anaphora Resolution Studies

[Sedivy, 2010] notes that because eye movements reflect a referentially-driven interpretation of language, they can be used to gain insight into the process of reference resolution. The first eye-tracking study of anaphora interpretation, [Arnold et al., 2000] monitored eye movements and found that the gaze of listeners reliably shifted to likely referents of a spoken pronoun. In child development research, several studies along these lines have been published in the last two decades.

[Sekerina et al., 2004], one of the first of these studies, investigated online processing of referentially ambiguous pronouns in children age 4-7 years old, finding that their eye movements revealed implicit awareness of referential ambiguity despite lack of explicit knowledge. Similarly, [Järvikivi et al., 2014] found in a study on 4-year-old German-speaking children that children are sensitive to some structural cues in reference resolution, but may not fully be able to apply them as constraints to help them interpret anaphora. These studies appear to suggest that awareness of relevant cues for anaphora resolution at the perceptual level, measured through eye movements, precedes explicit knowledge of these cues in children as young as 4 years old.

Additionally, several studies have focused on investigating the question of whether children's comprehension of pronouns is affected by the grammatical role preference (discussed in Chapter 1), the observation that people tend to resolve an ambiguous pronoun to the subject of a sentence or the first-mentioned subject. It is known that adults exhibit this preference in pronominal resolution, but there have been conflicting studies on whether children do as well.

[Song and Fisher, 2005,Song and Fisher, 2007] show that children as young as 2.5 years old interpret pronouns using some of the same factors that adults use by ranking candidate referents based on subject and first-mention bias. Both [Pyykkönen et al., 2010] and [Hartshorne et al., 2011] report results in line with this finding. [Pyykkönen et al., 2010], investigating verb transitivity's effect on referent preference, found that 3-year-old children demonstrated a bias in resolving anaphora to the sentence's subject, although this preference was not as strong in sentences with high verb transitivity. [Hartshorne et al., 2011] reported that English-speaking 5-year-old exhibited a preference to resolving pronouns to first-mention or subject character. In contrast to this work, [Arnold et al., 2007] found that in children between the ages of 3 and 5 years old employed gender cues, but not first-mention bias to resolve anaphora.

A study done on older English-speaking children, ages 6-9 years old, by [Clackson et al., 2011] also investigated the subject bias. They found that children's final interpretation of anaphora showed similar use of constraints as adults, but eye movement data revealed that during processing, children were temporarily more distracted than adults when confronted with multiple cues, perhaps accounting for some of the discrepancy between the cited studies on subject bias.

Most recently, [Palmović et al., 2018] conducted a study on Croatian-speaking first graders (mean age was 7 years old), examining the relative role of information structure and visual context in anaphora resolution. Their results suggested that children rely more on visual cues, but when these cues are not available, children seem to perform at chance.

Although eye-tracking studies on the development of anaphora resolution in children are still at a relatively preliminary stage, in all cases, their findings suggest that children consistently make attempts to resolve references as anaphoric expressions arise in speech [Sedivy, 2010], and that children use preferences and constraints similar to those observed in adults [Järvikivi et al., 2014].

2.3 Limitations of Existing Work

One common aspect of the current eye-tracking studies investigating infant and toddler anaphora resolution is that eye movements were primarily measured using static eye-tracking tasks that were typically screen-based. In all the studies discussed in the previous section, children listened to audio, such as a short stories or disconnected sentences, while viewing pictorially or physically represented antecedents. During these tasks, children sat in chairs or a parent's lap in front of a screen or an experimental set-up in a laboratory setting. These studies fail to capture the dynamic visual interactions children have with their environment [Slone et al., 2018], nor do they accurately reflect a child's typical speech input, namely interactive dialogue between them and a parent or other adult. Consequently, the results from task-specific, laboratory studies should be investigated in comparison to studies conducted in more naturalistic environments with stimuli children may find in everyday experiences in order to validate their findings.

In Chapter 3, we discuss the current study, an analysis of a corpus of infant eye-tracking data collected in a naturalistic environment by [Schroer et al., 2019]. The study addresses the limitations noted here, as children are free to move around in a carpeted area with an assortment of toys, and parents are instructed to play with and speak to their child as they might at home. Not only does this capture the dynamic visual perspective of the child, the complexity of speech generated by the parent in dialogue with their child is likely to be reflective of and tailored to the language abilities of the child. As a result, the language input to each child in the study can be assumed to match their comprehension abilities, which may be less consistent in studies that use prerecorded or scripted speech, as language abilities frequently do not develop at the same rate in very young children. Thus the speech and eye movement data from this experiment is collected using a naturalistic, unscripted task with spontaneous generation of anaphoric instances in natural parent-child dialogue. Additionally, none of the existing studies has analyzed eye-tracking data within the context of NLP and automated anaphora resolution systems. The current study seeks to evaluate child eye-tracking data through the lens of NLP, a novel application of developmental eye-tracking studies.

Chapter 3

Child Eye-Tracking Data Analysis

3.1 Present Research

To investigate cognitive mechanisms of inference that children use to resolve ambiguous anaphoric instances in naturalistic speech, we coded and analyzed a corpus of speech transcripts and eye-tracking data collected by [Schroer et al., 2019] at Indiana University's Computational Cognition and Learning Laboratory. The goal of our analysis was to attempt to identify patterns in child anaphora resolution mechanisms that were expected based on the literature reviewed in Chapter 2.

Mechanisms are consistent across subjects. Because the literature demonstrates that children and toddlers are aware of both anaphora use and relevant cues in speech [Järvikivi et al., 2014, Sekerina et al., 2004], we expected to be able to identify consistent patterns in the mechanisms children use for anaphora resolution in the current study. Eye-tracking studies demonstrate that children consistently use heuristics such as the subject and first-mention biases [Hartshorne et al., 2011, Pyykkönen et al., 2010, Song and Fisher, 2005, Song and Fisher, 2007], gender cues [Arnold et al., 2007], and visual cues [Palmović et al., 2018]. Collectively, the literature indicates that young children employ identifiable mechanisms and consistently attend to specific cues, leading us to expect to find patterns in anaphora resolution strategies of the child subjects in the current study as well.

Resolution accuracy improves with age. We expected to find a positive correlation between the age of the child participants and their anaphora resolution accuracy scores. Because children can be studied as developing systems of cognition [Järvikivi et al., 2014], and adults are able to resolve instances of anaphora in speech more consistently and more accurately than children [Palmović et al., 2018], we hypothesized that older children would resolve anaphora more accurately than younger children.

Parent speech reflects child language abilities. We expected to find that the complexity of speech generated by a parent in dialogue with their child would reflect the language abilities of the child. Because parents spend a great deal of time everyday interacting with and conversing with their children, they are attuned to the language comprehension capabilities of the child. For instance, [Hetzroni and Ohn, 2012] found that parents of children with autism spectrum disorder perceived their children as less competent communicators, and consequently tailored their speech by using more directives. Thus, we hypothesized that the language directed toward the children in the study would be more complex for older children as they age and become more competent interlocutors.

3.2 Corpus

Although originally collected to examine the effects of parent speech on infant attention by [Schroer et al., 2019], the data naturally lends itself to further studies on language acquisition. The corpus' naturalistic environment, unscripted and spontaneous discourse, and eye-tracking information make it a valuable source of data for developmental linguistics analyses.

3.2.1 Methods

[Schroer et al., 2019] collected data from thirty-five toddlers (mean age = 18.75mos [range: 12.3-25.3]) and their parents who participated in a study on naturalistic parent-child interactions during free play while wearing head-mounted eye trackers. During the play session, parent speech and parent and child eye gazes were recorded.

3.2.1.1 Data Collection

In [Schroer et al., 2019], parents and infants played with 24 toys on a carpeted floor for an average of 8.46 minutes (range: 3.10-15.54). Parents and children could sit in any orientation on the carpet, but children were seated on the floor (as opposed to a parent's lap) because of the eye tracker's cable.

The experiment was run by two researchers, who placed and adjusted the eye trackers on both the parent and infant. After both parent and child were were fitted with eye trackers, the researchers ran a calibration procedure. The researchers then left the room and monitored the experiment from an adjacent room, reentering only to adjust and re-calibrate the infant's eye tracker if it was bumped or moved during the play session.

3.2.1.2 Materials

Eye-Tracking Equipment

Prior to the play session, both parent and child with fitted were head-mounted eye trackers from Positive Science LLC. The eye-tracking system consisted of a scene camera on the wearer's forehead that recorded images from the wearer's perspective with a visual field of 108°, as well as an infrared camera pointed at the participant's right eye movements and fixations. For stability, the child's eye tracker was attached to a hat that the child wore. The parent wore the eye tracker like a pair of glasses. To capture third-person views of the participants, additional cameras were placed throughout the room (Figure 3.1).

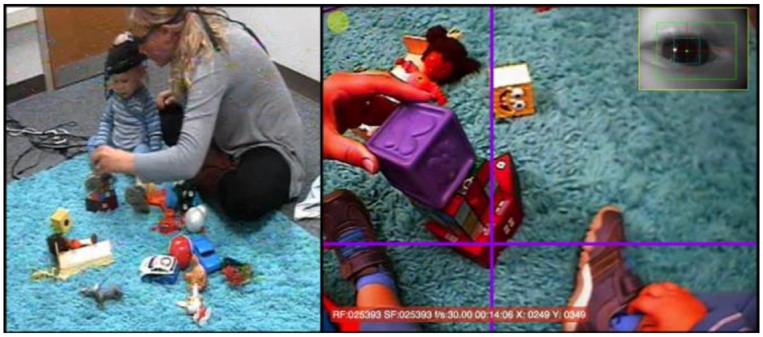


Figure 3.1: Third-person camera view (left), child's first-person view with cross-hairs indicating child gaze (right). Reprinted from [Schroer et al., 2019].

Toys

At the start of the play session, the 24 toys were spread out across the carpet at random. The toys consisted of blocks, animal and insect figurines, dolls, toy cars, and a variety of other play objects (Figure 3.2). Parents were instructed to play with their child using these toys as they might at home.

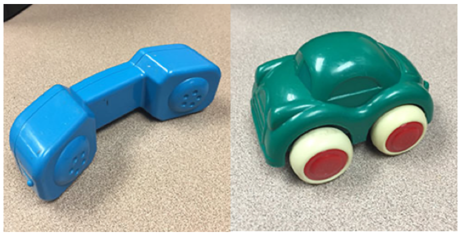


Figure 3.2: Examples of toys

3.2.2 Data Processing and Coding

The corpus we used in the current study consisted of the audio and visual data processed by [Schroer et al., 2019]. After the experiment was completed, the eye-tracking videos from the scene camera and eye camera were synchronized using a software program. The program also generated cross-hairs that indicated where the participant was looking during each video frame (Figure 3.1). Using the first-person scene camera with cross-hairs overlaid, the visual gaze of the parent and child were each manually coded by annotating which *region of interest* (ROI) the cross-hairs overlapped with during a gaze fixation. Each of the 24 toys and the social partner's face were coded as ROIs (25 ROIs in total).

The recordings of the parents' speech was transcribed using Audacity by [Schroer et al., 2019]. Utterances were considered separate if they were at least 400ms apart. Because of the separation criteria, speech that would be considered sentences could be split apart in the transcriptions, and separate sentences might be counted as a single utterance.

3.2.2.1 Coding Protocols

For the current study, we were interested in anaphoric instances spoken by the parent to their child. We manually coded each parent speech transcription for anaphoric instances and their correct NP antecedent(s), as well as information about anaphora used. In total, we coded for four variables: **anaphoric expression**, **referent ID**, **anaphora type**, and **cue variable**.

Anaphoric Expression

For each utterance that contained at least one anaphoric instance referencing an object or person in the room, we annotated the transcript to note the anaphoric word or words used by the parent (e.g. "it", "one", etc.). We did not code for pleonastic expressions or for anaphoric expressions that referenced entities that were not physically present in the experiment room.

Referent ID

Each of the 25 ROIs was given a **referent ID**, a number between 1 and 25. The referent IDs were used to identify each of the ROIs (toys and social partner's face) that were referenced by a parent using an anaphoric expression. Using the speech transcriptions and the video data collected from the recording cameras, the coder manually determined the antecedent of each anaphor and recorded the antecedent's ID based on the referent ID key shown in Figure 3.3. For instance, if a parent referred to the green car in Figure 3.2 with the anaphor "it", the utterance would be coded with the referent ID 13. A referent ID of 26 indicated that the parent used an anaphoric expression to refer to an object or entity in the room that was not one of the ROIs.

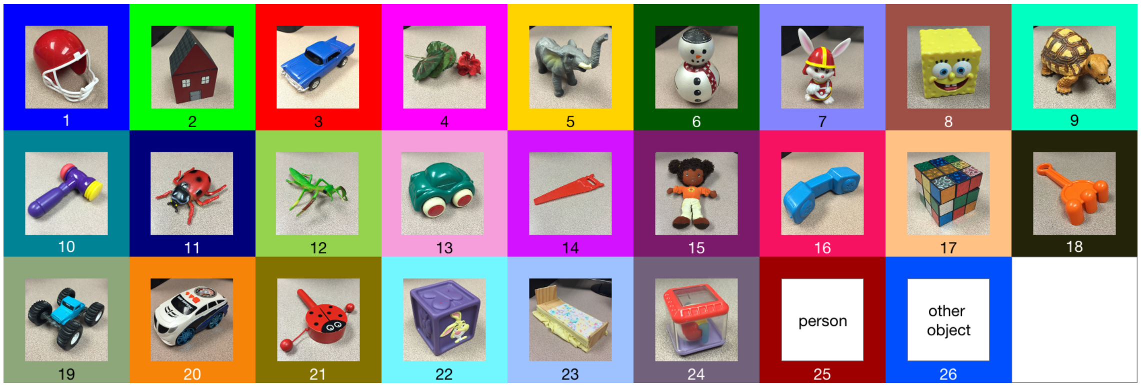


Figure 3.3: Referent ID key

Anaphora Type

As mentioned in Chapter 1, some of the most common types of anaphora used in natural discourse are the *pronominal anaphora*, *one anaphora*, and *split anaphora*. We coded each anaphoric expression in the transcriptions as one of these types. If an anaphoric expression was coded as a split anaphor, we recorded all referent IDs the anaphor referenced. For instance, the anaphoric expression "they" referencing the toy phone and the green car in Figure 3.2 would be coded as a split anaphor referring to referents `13` and `16` (see ID key in Figure 3.3).

Cue Variable

We also recorded the type of cue (speech and/or visual) that was required to resolve an anaphoric instance. This was coded as a binary variable: any given anaphor either required only speech cues or both speech and visual cues to resolve.

3.2.2.2 Visualizing with Data Streams

After the coding of the parent speech transcripts was completed, data streams of infant visual attention, parent visual attention, and anaphoric references were created to visualize the coded speech information.

Understanding the Data Streams

Our visualizations have five data streams: infant gaze (`ceye`), parent gaze (`peye`), parent speech utterances (`utt`), naming instances (`naming`), and anaphoric instances (`a-ref`, `a-cue`, `a-type`). To view an example, see Figure 3.4. Each block in a stream represents the occurrence and duration of a speech or visual event (e.g. a new utterance, switching gaze to a different object). Each color represents a different object, with the colors corresponding to the referent ID key. The length of the color blocks indicate duration of an event.

`ceye` and `peye` streams represent what object the child and parent are each looking at, respectively. The `utt` stream represents parent speech utterances, with the color indicating the object being named (dark red indicating no naming in that utterance). `naming` indicates *naming instances*, utterances that include explicit mention of an object's name (e.g. "Look at the green car!"). The streams `a-ref`, `a-cue`, and `a-type` give information about anaphoric instances. When a parent's utterance uses an anaphoric expression, `a-ref` shows a color block representing the object being referenced. `a-cue` shows that anaphor's cue variable (blue represents speech only cues, green represents speech and visual cues). `a-type` represents the anaphor's type (blue for one anaphora, green for split anaphora, and red for pronominal anaphora). The data streams are aligned along the x-axis by time, measured in seconds.

3.3 Results and Discussion

3.3.1 Analysis of Parent Anaphora Usage

As discussed in Chapter 2, spontaneous parent speech is assumed to reflect a parent's perception of their child's language capabilities. To an extent, we can therefore infer a child's actual language competence, and more specifically, their ability to resolve anaphora, from the complexity of the speech a parent directs toward their child. We expected to see that parents of older children would tend to use anaphoric expressions more frequently, and use anaphoric expressions of greater complexity (e.g. expressions with competing potential antecedents).

Anaphora frequency. As expected, parents tended to speak more utterances containing anaphoric expressions relative to total utterances the older their child was (Figure 3.5). The data supports our hypothesis. Parents of older children tended to speak utterances containing anaphora with greater frequency than parents of younger children.

Anaphora complexity. The data shows an unexpected trend in parents' use of complex anaphora in relation to their child's age. We operationalized anaphora complexity by measuring the percentage of utterances containing multiple referents per total utterances that contained anaphora (Figure 3.6a), as well as the percentage of utterances containing split anaphora per total utterances that contained anaphora (Figure 3.6b). Parents of children age 17-22 months old used complex anaphoric expressions more often: these parents most frequently (relative to their total utterances containing anaphora) referred to multiple referents in the same utterance, and most frequently used split anaphora. Interestingly, this did not correspond linearly with age as expected. In fact, parents of children nearest the median age (18.2 months old) used complex anaphora the most frequently, while parents of both the oldest and youngest children in the study barely used complex anaphoric expressions at all. The trends of complex anaphora usage seen in Figures 3.6a and 3.6b do not match with the positive correlation of anaphora frequency and age seen in Figure 3.5. We hypothesized that not only would parents use more anaphoric expressions as children got older, they would also use more complex cases of anaphora. The data did not support this hypothesis.

However, Figures 3.6a and 3.6b show a correlation in parents' use of multiple referents and use of split anaphora: parents who more frequently used multiple referents in the same utterance also tended to use more split anaphora, and similarly, parents who infrequently spoke utterances with multiple referents also infrequently used split anaphora. In this regard, the data suggests that the use of multiple referents and the use of split anaphora are tied to similar assumptions a parent makes about their child's anaphora resolution capabilities. In other words, parents who perceived their child as able to resolve utterances with competing referents also assumed ability to resolve split anaphora, indicating the validity of operationalizing anaphora complexity as the combination of these two measures.

3.3.2 Analysis of Child Reference Resolution Accuracy

We hypothesized that older children participating in the study would more accurately resolve anaphoric expressions than younger children. We measured each child's reference resolution accuracy by determining how how accurate the child's gaze fixations were based on the ROI being referenced by the parent. If a child's eye movements indicated a fixation on the correct ROI at any point during their parent's utterance containing the anaphoric expression, the utterance was marked as correct. By calculating the number of utterances the child resolved correctly in comparison to the total number of utterances containing an anaphoric expression, we obtained each child's resolution accuracy score. When plotted against age, no correlation was observed between a child's anaphora resolution accuracy score and their age (Figure 3.7).

One explanation for the lack of correlation is that the children participants of the study were possibly too young to observe reliable differences in their ability to resolve anaphora. Previous developmental eye-tracking studies on anaphora resolution have only examined children as young 2.5 years (30 months) old [Song and Fisher, 2007], with most studies focusing on children 4 years and older. The mean age of child participants in our study was 18.75 months, the oldest participant being 25.3 months old, still 5 months younger than participants in [Song and Fisher, 2007]. It is possible that children at this stage of development have not yet developed reliable mechanisms for anaphora resolution, and consequently perform at chance.

Another possible explanation is that the children's age range in our study was too narrow to observe differences in anaphora resolution based on age, resulting in the lack of correlation seen in Figure 3.7. Existing eye-tracking studies of anaphora resolution have not compared anaphora resolution accuracy between children who differ by only 1 year in age as our work does; indeed, comparison studies were typically done between children of a particular age group and adults. Future studies could examine a wider age range of participants to investigate reliable differences in anaphora resolution abilities.

Interestingly enough, our data shows that parents often did not look at the objects they were referring to when using anaphoric expressions. Across all subjects, parents only fixated on the same object they referred to in an utterance with anaphora on average 60% of the time, possibly because parents often attended to other ROIs in the environment even while speaking to select new toys to present to their child. Again, this could offer insight into the lack of correlation in children's age with their anaphora resolution accuracy. Infants and young children often use visual and social cues to choose where to attend to in their visual environment, and consequently may be sensitive to a parent's gaze as a cue guiding where to look. If parents are not looking at the correct referent, children might not as well.

3.3.3 Study Limitations and Suggestions for Future Analysis

3.3.3.1 Participant Age

As mentioned, one possible challenge in determining reliable patterns in the eye-tracking data was the age of the child participants in the study. The size of the age range (13.0 months) and the ages studied (12.3-25.3 months) pose potential limitations in determining significant language ability differences between the participants. It was unclear from the distribution of children's anaphora resolution accuracy scores (Figure 3.8) whether children performed at chance when resolving anaphora due to how young they were, or if they applied consistent mechanisms to make guesses better than just at chance. For more conclusive results, this distribution would need to be compared to a null model of eye-tracking data in the current paradigm. It could be assumed that a child merely resolving anaphora at chance in this study would identify the correct object in 50% of instances. However, this may be a flawed assumption as the child was given the option of looking at 24 different objects, so perhaps a child looking at the correct referent 50% of the time is actually much better than chance.

3.3.3.2 Language Comprehension Assessment

Another limitation to the study was that we could not measure the general language capabilities of each child, which might be a better indicator than age for anaphora resolution ability. Rather than comparing anaphora resolution accuracy across age, as we did in Figure 3.5, perhaps a more telling measure might be to conduct a developmental assessment of general language comprehension on the participants, and analyze anaphora resolution skills based on language capability test scores. A positive correlation of spoken language comprehension scores and anaphora resolution scores would be expected. Our analysis was limited in that it is unclear how much the children in the study truly comprehended their parent's speech. Without an accurate assessment of their overall language abilities to ensure comprehension, it remains difficult to study the responses of toddlers and infants to a specific language phenomenon like anaphora resolution.

Another possible solution to ensuring that data collected this way indeed shows children consciously comprehending and resolving anaphora would be to perform an assessment on anaphora resolution prior to collecting the data using experimental paradigms in studies discussed in Chapter 2, such as [Song and Fisher, 2005] or [Arnold et al., 2007].

3.3.3.3 Experimental Paradigm

An important question that arises from the experimental paradigm of the study is whether or not the children participating looked at the ROIs that they did as a result of resolving an anaphoric expression and attending to the correct object (or object that they thought was correct). In many instances, parents could be observed both visually and physically presenting objects to their children, and then using an anaphoric expression to comment on the object. For instance, one parent's transcript reads (each line indicates a separate utterance):

```
what is that
ladybug
yes
if you shake it, it makes noise, there you go
```

When paired with the video data, the parent can be observed handing the child a toy. Similarly, children were observed picking up and looking at objects, with parents then using anaphora to refer to the object the child was already looking at. In the same parent's transcript, the parent says:

what did you find

rake

car

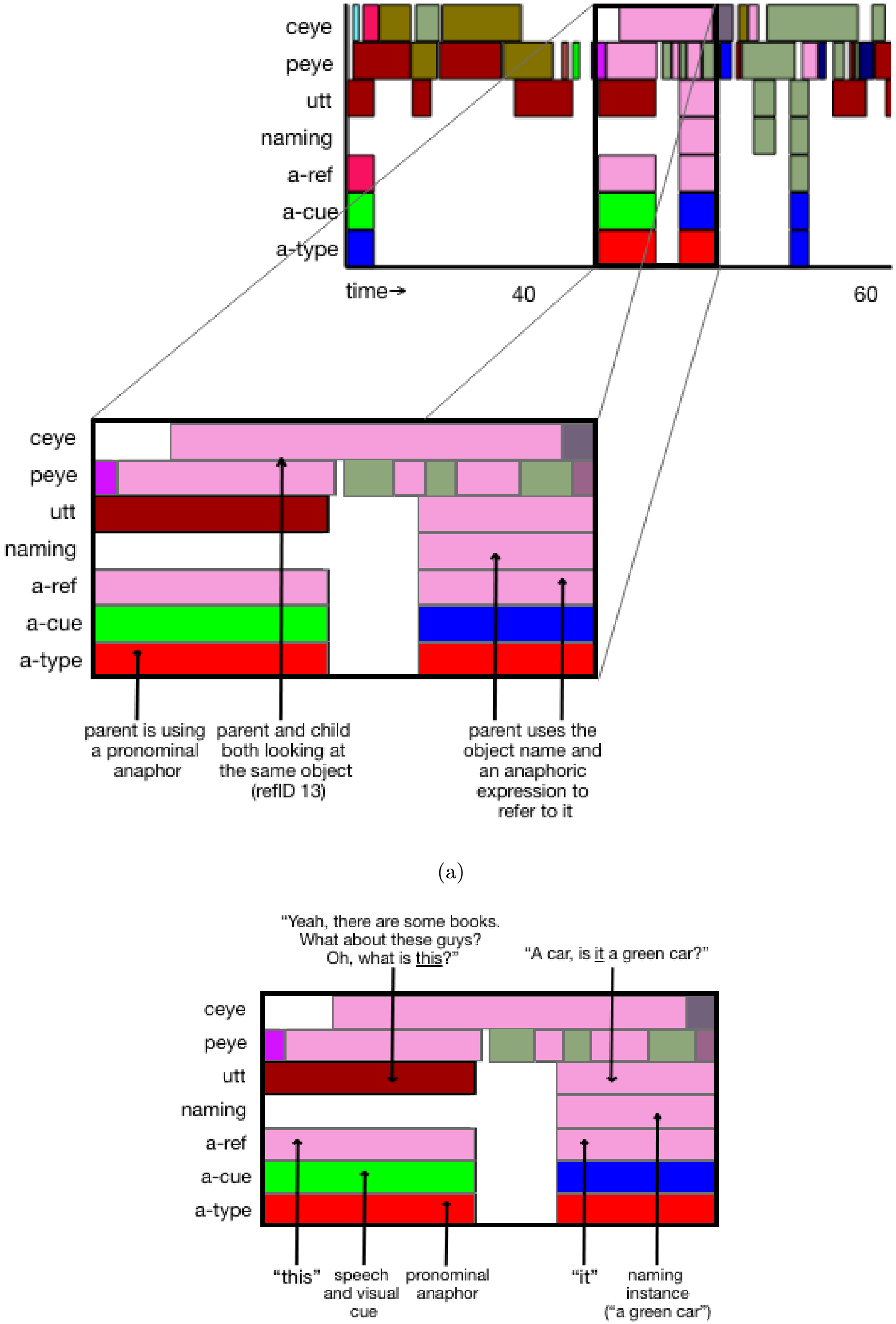
oh it has got buttons

Because of the intentionally unstructured nature of the experimental paradigm, it is difficult to determine causality for eye movements in the data. Perhaps children sometimes did not look at objects parents referred to with anaphora as a result of an anaphora resolution process, and instead looked at the object because a parent showed them. Or perhaps, as the previous example suggests, parents used anaphora to refer to objects that were already in a child's visual focus.

In order to draw meaningful conclusions from the eye-tracking data within this paradigm, it would be necessary to determine causation of each anaphoric instance. Cases where parents acknowledged toys children were already looking at, or where parents spoke about objects that children then attended to due to other visual and social cues (physically being handed an object, a parent pointing, etc.), would have to be eliminated from data analyzed for anaphora resolution accuracy and processes. This would be time-consuming to do, however, because this elimination process would likely involve manually coding the data for indicators of these instances.

Additionally, we had hoped to perform fine-grained analysis on individual cases of relatively complex or ambiguous anaphoric expressions to assess mechanisms of reasoning in anaphora resolution, but because of the uncertainties regarding whether children actively attempted to resolve anaphora, analyzing individual instances was inconclusive, such as the excerpt of one of the data streams in Figure 3.9.

Although the corpus we used had an important advantage in that the data was collected from children and parents participating in a naturalistic unscripted task with spontaneous generation of anaphoric instances in natural parent-child dialogue, the data was difficult to analyze to find meaningful insight into anaphora resolution processes because it was not collected for this purpose. The study's paradigm may not have been suited for our analyses since the analysis in [Schroer et al., 2019] was not on a particular speech phenomenon, but rather on the children's general attentional patterns as their parents spoke. We may be able to gain more conclusive results from a study specifically designed for this research.



(a)

(b)

Figure 3.4: An example of our data streams. a) Magnified view of subsection of the data streams with annotations on how to interpret them. Figure adapted from [Schroer et al., 2019]. b) This section of the data streams corresponds to two utterances: 1) "Yeah, there are some books. What about these guys? Oh, what is <u>this</u>?", 2) "A car, is <u>it</u> a green car?".

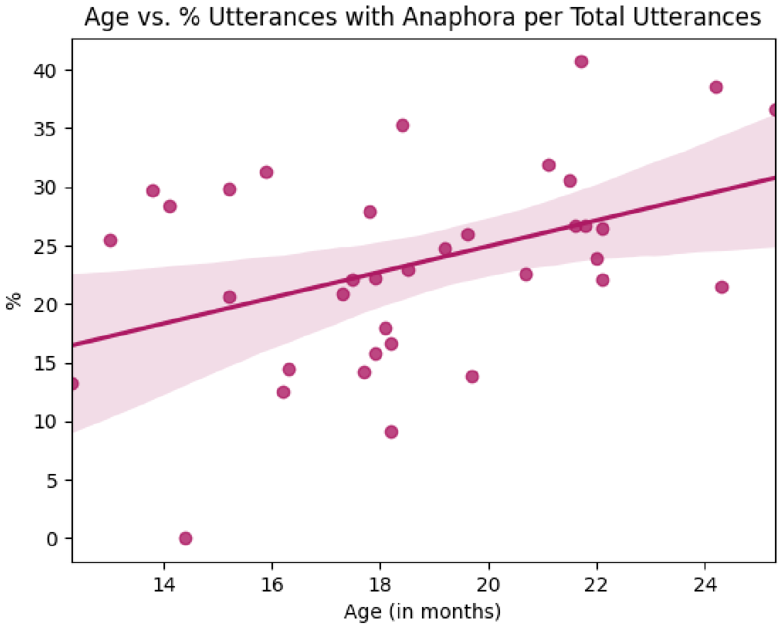
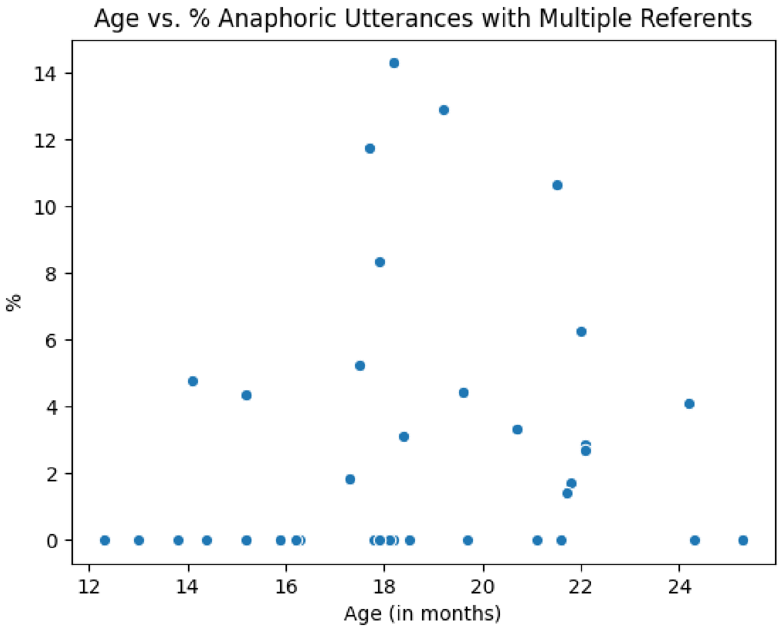
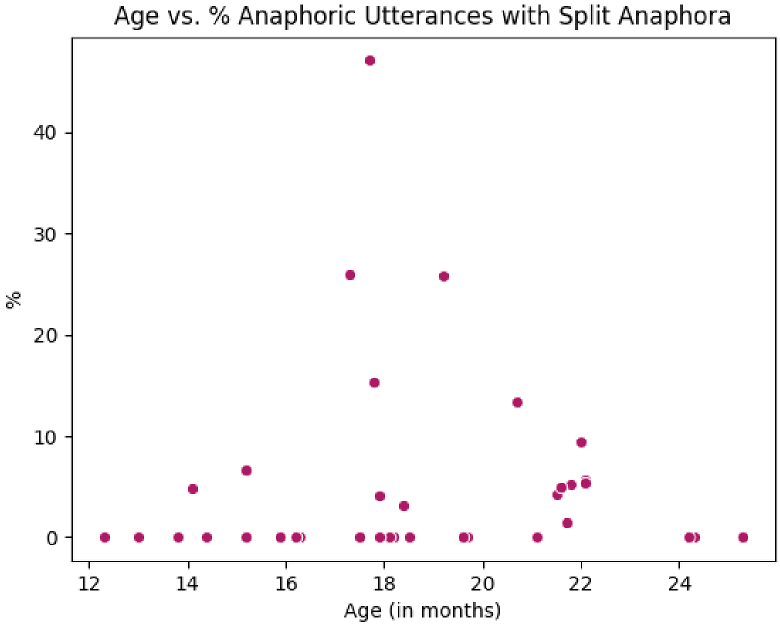


Figure 3.5: Percentage of utterances with anaphora per total utterances as a function of age. Parents tended to speak more anaphoric utterances relative to the total number of utterances spoken to children who were older. This was expected since parents' speech is assumed to reflect their perceptions of their child's language capabilities. A child of 14 months may exhibit inability to interpret anaphora, while a child of 24 months may demonstrate a better grasp.



(a)



(b)

Figure 3.6: Anaphora complexity with respect to age. a) Percentage of total utterances with anaphora that contained multiple referents. b) Percentage of total utterances with anaphora that contained split anaphora.

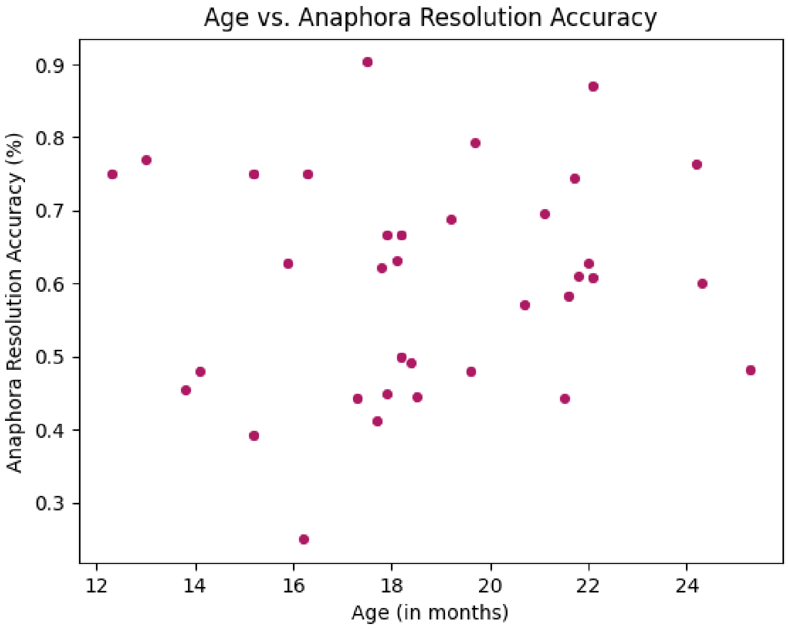


Figure 3.7: Anaphora resolution accuracy with respect to age. Positive correlation was expected, no was correlation found.

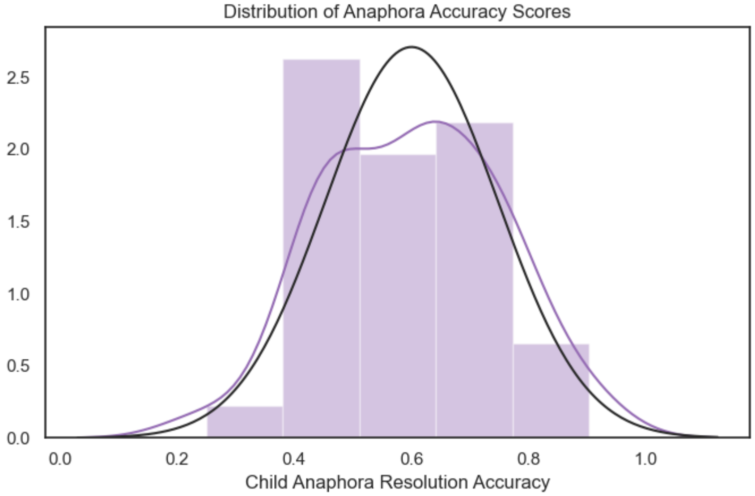


Figure 3.8: Distribution of accuracy scores of child anaphora resolution. The maximum possible score was 1.0 (100%). Mean=0.600, standard deviation=0.149, range=[0.250,0.903].

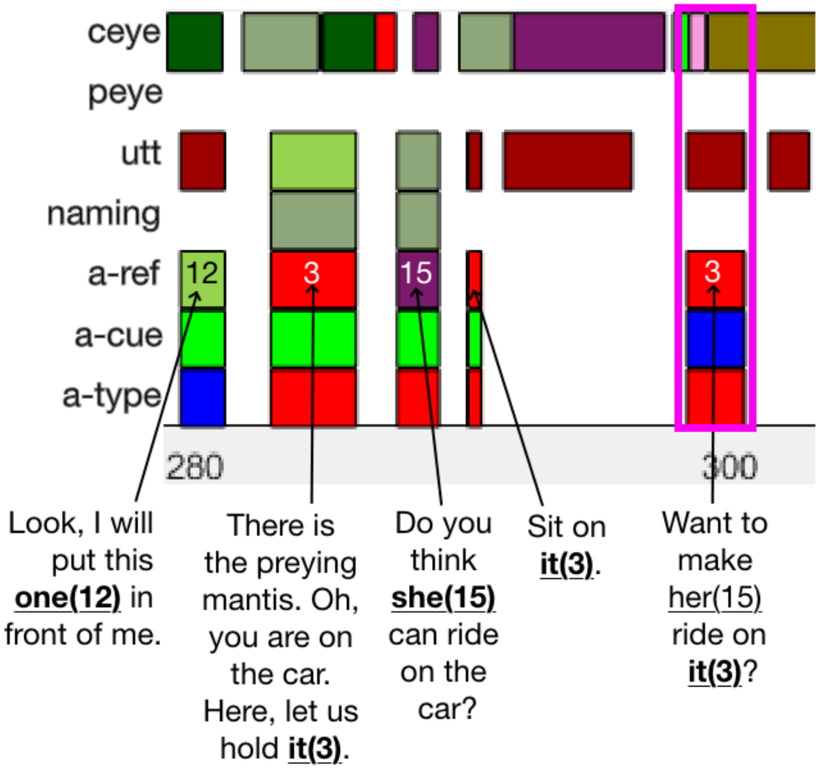


Figure 3.9: Fine-grained analysis of one anaphora instance ("Want to make her ride on it?"). The child fixated on several objects during the utterance, mainly on Toy 21, which was not referred to in any of the prior utterances or looked at by the parent or child, so it is unlikely the child confused the parent's anaphoric reference for Toy 21.

Chapter 4

Contributions and Conclusion

This thesis explores a novel approach to improving anaphora resolution which has promise for advancing several other areas of natural language processing as well. We proposed that inference mechanisms that children acquire to resolve anaphora can guide advances in incorporating commonsense reasoning to improve accuracy in automated anaphora resolution. Children, as developing systems, may provide a framework for how an automated system may acquire reasoning abilities for anaphora resolution.

To that end, our contributions include surveys of the literature in Chapters 1 and 2 that provide evidence and rationale for taking a developmental linguistics approach to reasoning in anaphora resolution systems. We also performed a preliminary analysis of a corpus of infant eye-tracking data collected by [Schroer et al., 2019]. Through the analysis, we hoped to investigate the applicability of developmental linguistic research to NLP. Unfortunately, our results were inconclusive due to several limitations of the corpus for our analyses, namely the experimental paradigm's lack of control for factors that could influence anaphora resolution accuracy. Consequently, we were unable to identify consistent or discrete mechanisms that children used to resolve anaphora. Thus this thesis does not offer conclusive evidence to the question of whether a developmental linguistics approach could be taken to improve anaphora resolution algorithms.

Future work will focus on addressing the limitations of the corpus used in the study discussed in Chapter 3, such as conducting a naturalistic study involving language comprehension assessments, an older age group, and a more structured experimental paradigm tailored toward anaphora resolution to control for other factors and cues that could hinder conclusive analysis on the eye-tracking data. Additionally, a further literature review on earlier studies of child anaphora resolution acquisition that do not involve eye-tracking paradigms should be conducted, moving towards a more cohesive theory of development of cognitive anaphora resolution that could be applied to natural language processing algorithms. Lastly, in this study, our notion of commonsense reasoning mechanisms was not precisely defined. It would be crucial to define these mechanisms to search for in future corpus analyses.

Our contributions are primarily theoretical. Our hope is that our suggestion for a new approach to understanding reasoning for ambiguous and complex anaphoric cases will initiate further research in this area.

Bibliography

[Arnold, 2010] Arnold, J. E. (2010). How speakers refer: The role of accessibility. *Language and Linguistics Compass*, 4:187–203.

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

[Arnold et al., 2000] Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., and Trueswell, J. C. (2000). The rapid use of gender information: Evidence of the time course for pronoun resolution from eyetracking. *Cognition*, 76(1):B13–B16.

[Bergsma et al., 2008] Bergsma, S., Lin, D., and Goebel, R. (2008). Distributional identification of non-referential pronouns. In *Proceedings of ACL-08: HLT*, pages 10–18, Columbus, Ohio. Association for Computational Linguistics.

[Bryl et al., 2010] Bryl, V., Giuliano, C., Serafini, L., and Tymoshenko, K. (2010). Supporting natural language processing with background knowledge: Coreference resolution case. In *The Semantic Web–ISWC 2010*, pages 80–95, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Charney, 1980] Charney, R. (1980). Speech roles and the development of personal pronouns. *Journal of Child Language*, 7(3):509–528.

[Charniak, 1972] Charniak, E. (1972). *Toward a Model of Children's Story Comprehension.* PhD thesis, Massachusetts Institute of Technology.

[Chipman and de Dardel, 1974] Chipman, H. H. and de Dardel, C. (1974). Developmental study of the comprehension and production of the pronoun "it". *Journal of Psycholinguistic Research*, 3:91–99.

[Clackson et al., 2011] Clackson, K., Felser, C., and 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.

[Clark and Manning, 2016a] Clark, K. and Manning, C. D. (2016a). Deep reinforcement learning for mention-ranking coreference models. *arXiv e-prints.*

[Clark and Manning, 2016b] Clark, K. and Manning, C. D. (2016b). Improving coreference resolution by learning entity-level distributed representations. *arXiv e-prints.*

[Dozat and Manning, 2016] Dozat, T. and Manning, C. D. (2016). Deep biaffine attention for neural dependency parsing. *arXiv e-prints.*

[Emami et al., 2018] Emami, A., Trichelair, P., Trischler, A., Suleman, K., Schulz, H., and Kit Cheung, J. C. (2018). The knowref coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. *arXiv e-prints.*

[Fine, 1978] Fine, J. (1978). Conversation, cohesive and thematic patterning in children's dialogues. *Discourse Processes*, 1(3):247–266.

[Gu et al., 2018] Gu, J.-C., Ling, Z.-H., and Indurkhya, N. (2018). A study on improving end-to-end neural coreference resolution. *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 159–169.

[Haber et al., 2018] Haber, N., Mrowca, D., Wang, S., Fei-Fei, L., and Yamins, D. L. (2018). Learning to play with intrinsically-motivated, self-aware agents. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8398–8409. Curran Associates Inc.

[Hartshorne et al., 2011] Hartshorne, J. K., Nappa, R., and Snedeker, J. (2011). Ambiguous pronoun processing development: probably not u-shaped. In *Proceedings of the 35th Annual Boston University Conference on Language Development*, volume 1, pages 272–282.

[Hetzroni and Ohn, 2012] Hetzroni, O. and Ohn, R. (2012). Children with ASD and parents: Parent perception of child's communication skills influencing nature of interaction. *Journal of Intellectual Disability Research*, 56(7-8):739–739.

[Hobbs, 1979] Hobbs, J. R. (1979). Coherence and coreference*. *Cognitive Science*, 3(1):67–90.

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

[Jurafsky and Martin, 2000] Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction for Natural Language Processing, Computational Linguistics, and Speech Recognition.* Prentice Hall.

[Järvikivi et al., 2014] Järvikivi, J., Pyykkönen-Klauck, P., Schimke, S., Colonna, S., and Hemforth, B. (2014). Information structure cues for 4-year-olds and adults: tracking eye movements to visually presented anaphoric referents. *Language, Cognition and Neuroscience*, 29(7):877–892.

[Kobayashi et al., 2005] Kobayashi, N., Iida, R., Inui, K., and Matsumoto, Y. (2005). Opinion extraction using a learning-based anaphora resolution technique. In *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts.*

[Lee et al., 2017] Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

[Lee et al., 2018] Lee, K., He, L., and Zettlemoyer, L. (2018). Higher-order coreference resolution with coarse-to-fine inference. *arXiv e-prints.*

[Levesque et al., 2011] Levesque, H., Davis, E., and Morgenstern, L. (2011). The Winograd Schema Challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pages 552–561.

[Liu et al., 2016] Liu, Q., Jiang, H., Evdokimov, A., Ling, Z.-H., Zhu, X., Wei, S., and Hu, Y. (2016). Probabilistic reasoning via deep learning: Neural association models. *arXiv e-prints*.

[Loveland, 1984] Loveland, K. A. (1984). Learning about points of view: spatial perspective and the acquisition of 'i/you'. *Journal of Child Language*, 11(3):535–556.

[Lust, 1986] Lust, B. (1986). *Studies in the Acquisition of Anaphora.* D. Reidel Publishing Company.

[Maratsos, 1973] Maratsos, M. (1973). The effects of stress on the understanding of pronominal co-reference in children. *Journal of Psycholinguistic Research*, 2:1–8.

[McCarthy et al., 1955] McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. (1955). A proposal for the dartmouth summer research project on artificial intelligence.

[Mitkov, 2007] Mitkov, R. (2007). Anaphora resolution: The state of the art. In *ACL 2007*.

[Mitkov et al., 2001] Mitkov, R., Boguraev, B., and Lappin, S. (2001). Introduction to the special issue on computational anaphora resolution. *Computational Linguistics*, 27(4):473–477.

[Mitkov et al., 2002] Mitkov, R., Evans, R., and Orasan, C. (2002). A new, fully automatic version of mitkov's knowledge-poor pronoun resolution method. In *Proceedings of CICLing-2002*, pages 168–186, Mexico City, Mexico.

[Mitkov et al., 2007] Mitkov, R., Evans, R., Orasan, C., Ha, L. A., and Pekar, V. (2007). Anaphora resolution: To what extent does it help NLP applications? In Branco, A., editor, *Anaphora: Analysis, Algorithms and Applications, 6th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2007*, volume 4410, pages 179–190. Springer-Verlag Berlin Heidelberg.

[Palmović et al., 2018] Palmović, M., Matić, A., and Kovačević, M. (2018). Resolution of anaphoric expressions in children and adults: Evidence from eye movements. *Suvremena lingvistika*, 44:139–154.

[Plu et al., 2018] Plu, J., Prokofyev, R., Tonon, A., Cudré-Mauroux, P., Difallah, D. E., Troncy, R., and Rizzo, G. (2018). Sanaphor++: Combining deep neural networks with semantics for coreference resolution. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

[Poesio and Kabadjov, 2004] Poesio, M. and Kabadjov, M. A. (2004). A general-purpose, off-the-shelf anaphora resolution module: Implementation and preliminary evaluation. In *Proceedings of LREC, Lisbon, Portugal*, Lisbon, Portugal.

[Poesio et al., 2016] Poesio, M., Stuckardt, R., and Versley, Y., editors (2016). *Anaphora Resolution: Algorithms, Resources, and Applications.* Theory and Applications of Natural Language Processing. Springer-Verlag Berlin Heidelberg.

[Ponzetto and Strube, 2006] Ponzetto, S. P. and Strube, M. (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 192–199, New York City. Association for Computational Linguistics.

[Pyykkönen et al., 2010] Pyykkönen, P., Matthews, D., and Järvikivi, J. (2010). Three-year-olds are sensitive to semantic prominence during online language comprehension: A visual world study of pronoun resolution. *Language and Cognitive Processes*, 25(1):115–129.

[Rahman and Ng, 2011] Rahman, A. and Ng, V. (2011). Coreference resolution with world knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 814–824, Portland, Oregon. Association for Computational Linguistics.

[Saqia et al., 2018] Saqia, B., Khan, K., Khan, A., Khan, W., Subhan, F., and Abid, M. (2018). Impact of anaphora resolution on opinion target identification. *International Journal of Advanced Computer Science and Applications*, 9:230–236.

[Schroer et al., 2019] Schroer, S. E., Smith, L. B., and Yu, C. (2019). Examining the multimodal effects of parent speech in parent-infant interactions. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, pages 1015–1021.

[Sedivy, 2010] Sedivy, J. C. (2010). *Experimental Methods in Language Acquisition Research*, chapter Using eyetracking in language acquisition research, pages 115–138. John Benjamins Publishing Company.

[Sekerina, 2014] Sekerina, I. (2014). Visual world eye-tracking paradigm. In Brooks, P. J. and Kempe, V., editors, *Encyclopedia of Language Development*, pages 657–658. Sage Publications, Inc.

[Sekerina et al., 2004] Sekerina, I. A., Stromswold, K., and Hestvik, A. (2004). How do adults and children process referentially ambiguous pronouns? *Journal of Child Language*, 31(1):123–152.

[Sidner, 1979] Sidner, C. L. (1979). Towards a computational theory of definite anaphora comprehension in english discourse. Technical report, Massachusetts Institute of Technology.

[Slone et al., 2018] Slone, L. K., Abney, D. H., Borjon, J. I., Chen, C.-h., Franchak, J. M., Pearcy, D., Suarez-Rivera, C., Xu, T. L., Zhang, Y., Smith, L. B., and Yu, C. (2018). Gaze in action: Head-mounted eye tracking of children's dynamic visual attention during naturalistic behavior. *Journal of Visualized Experiments*, 141.

[Smith et al., 2018] Smith, L. B., Jayaraman, S., Clerkin, E., and Yu, C. (2018). The developing infant creates a curriculum for statistical learning. *Trends in Cognitive Sciences*, 22(4).

[Smith and Slone, 2017] Smith, L. B. and Slone, L. K. (2017). A developmental approach to machine learning? *Frontiers in Psychology*, 8.

[Song and Fisher, 2005] Song, H.-j. and Fisher, C. (2005). Who's "she"? discourse prominence influences preschoolers' comprehension of pronouns. *Journal of Memory and Language*, 52(1):29–57.

[Song and Fisher, 2007] Song, H.-j. and Fisher, C. (2007). Discourse prominence effects on 2.5-year-old children's interpretation of pronouns. *Lingua*, 117(11):1959–1987.

[Steinberger et al., 2007] Steinberger, J., Poesio, M., Kabadjov, M. A., and Ježek, K. (2007). Two uses of anaphora resolution in summarization. *Information Processing and Management*, 43:1663–1680.

[Stojanov et al., 2019] Stojanov, S., Mishra, S., Thai, N. A., Dhanda, N., Humayun, A., Yu, C., Smith, L. B., and Rehg, J. M. (2019). Incremental object learning from contiguous views. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Stylianou and Vlahavas, 2019] Stylianou, N. and Vlahavas, I. (2019). A neural entity coreference resolution review. *arXiv e-prints*.

[Subramanian and Roth, 2019] Subramanian, S. and Roth, D. (2019). Improving generalization in coreference resolution via adversarial training. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 192–197, Minneapolis, Minnesota. Association for Computational Linguistics.

[Sukthanker et al., 2020] Sukthanker, R., Poria, S., Cambria, E., and Thirunavukarasu, R. (2020). Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162.

[Turek, 2018] Turek, M. (2018). Machine common sense (MCS).

[Turing, 1950] Turing, A. M. (1950). I.—Computing Machinery and Intelligence. *Mind*, LIX(236):433–460.

[Umstead and Leonard, 1983] Umstead, R. S. and Leonard, L. B. (1983). Children's resolution of pronominal reference in text. *First Language*, 4(11):73–84.

[Uryupina et al., 2011] Uryupina, O., Poesio, M., Giuliano, C., and Tymoshenko, K. (2011). Disambiguation and filtering methods in using web knowledge for coreference resolution. In *Florida Artificial Intelligence Research Society Conference*, pages 317–322.

[White et al., 2017] White, A. S., Rastogi, P., Duh, K., and Van Durme, B. (2017). Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005. Asian Federation of Natural Language Processing.

[Wiseman et al., 2015] Wiseman, S., Rush, A. M., Shieber, S., and Weston, J. (2015). Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for*

Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1416–1426, Beijing, China. Association for Computational Linguistics.

[Yang et al., 2008] Yang, X., Su, J., and Tan, C. L. (2008). A twin-candidate model for learning-based anaphora resolution. *Computational Linguistics*, 34(3):327–356.

[Yeh and Chen, 2003] Yeh, C.-L. and Chen, Y.-C. (2003). Using zero anaphora resolution to improve text categorization. In *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation*, pages 423–430, Sentosa, Singapore. COLIPS Publications.

[Yi et al., 2020] Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., and Tenenbaum, J. B. (2020). CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations*.

[Zhang et al., 2019] Zhang, H., Song, Y., Song, Y., and Yu, D. (2019). Knowledge-aware pronoun coreference resolution. *arXiv e-prints*.

[Zhang et al., 2018] Zhang, R., Santos, C. N. d., Yasunaga, M., Xiang, B., and Radev, D. (2018). Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 102–107. Association for Computational Linguistics.

[Zhao et al., 2018] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv e-prints*.