Theory of mind in language, minds, and machines: a multidisciplinary approach

Dijk, B.M.A. van

Theory of Mind in Language, Minds, and Machines

A Multidisciplinary Approach

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op vrijdag 17 januari 2025
klokke 13.00 uur
door
Bram Manuel Alejandro van Dijk
geboren te Bogotá, Colombia
in 1991

Promotor:
Prof.dr. M.R. Spruit

Co-promotor:
Dr. M.J. van Duijn

Promotiecommissie:
Prof.dr. M.M. Bonsangue
Prof.dr. S. Verberne
Prof.dr. A.P.J. van den Bosch (Utrecht University)
Prof.dr. C.P.A. Tiberius
Dr. B.F. Beekhuizen (University of Toronto)

The research in this dissertation was partly made possible by collaboration with Max van Duijn's NWO-funded research project 'A Telling Story', project number VI.Veni.191C.051.

Cover art by Pieter Verduijn.

Contents

Chapter 1

Introduction

This dissertation aims to deepen our understanding of the relation between Theory of Mind and language. It combines computational, qualitative, and experimental approaches and proposes to study Theory of Mind and language through a new language resource consisting of narratives. The empirical studies comprising this dissertation also focus on Theory of Mind and language in novel artificially intelligent models of language and cognition, and are complemented by broader reflection on how we can better understand Theory of Mind and language in the context of such models.

1.1 Background

Minds everywhere

> *"Lovers in the two-dimensional world, no doubt; little triangle number-two and sweet circle. Triangle-one (hereafter known as the villain) spies the young love."*[1]

Theory of Mind (ToM) is commonly understood as the ability to reason about one's own and others' *mental states* such as beliefs, desires and intentions (Apperly et al., 2009). ToM is a remarkable ability: typically developed humans exercise it time and again in navigating social interactions to get things done in their everyday lives, and because ToM is so seamlessly embedded in our actions, we often take it for granted. ToM as a concept was first coined in ethology regarding primates' anticipation of caretaker behaviour (Premack and Woodruff, 1978), but is probably best known for

[1]Quote from a participant describing the experimental scene (illustrated in Figure 1.1) as employed by Heider and Simmel (1944).

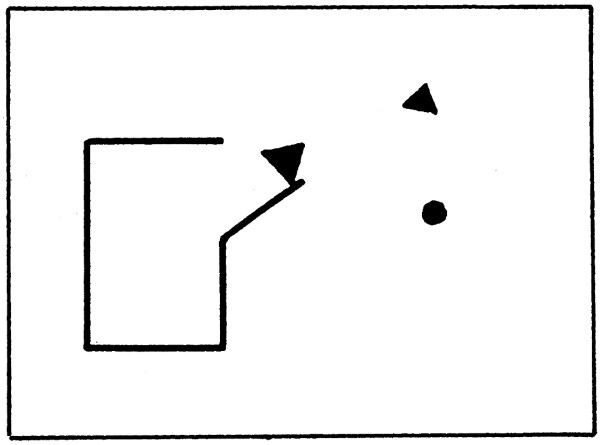


FIG. 1. EXPOSURE-OBJECTS DISPLAYED IN VARIOUS POSITIONS AND CONFIGURATIONS FROM THE MOVING FILM.
Large triangle, small triangle, disc and house.

Figure 1.1: Frame from the short movie used in Heider and Simmel (1944).

motivating a host of experiments in early child development that tests whether children understand that others may have beliefs that differ from their own (e.g. Baron-Cohen et al., 1985; Perner and Wimmer, 1985).

Already in the 1940s, a famous experiment by Fritz Heider and Marianne Simmel showed that when presented a short movie of plain geometrical shapes interacting (see Figure 1.1), participants readily attributed complex mental states to these shapes (Heider and Simmel, 1944). An illustration is given in a participant's description of the movie in the quote that opens this chapter: the use of *spying* implies coordination of mental states where 'triangle-one' (large triangle) makes some observation and *desires* that this observation remains *unknown* to 'little triangle number-two' and 'sweet circle' (a lexical phenomenon known in narratology as a *viewpoint package*, see Van Duijn and Verhagen, 2018).

Among other things, Heider and Simmel's experiment illustrated that when humans are triggered to tell a story, they will describe actions and events in terms of the intentional states of story characters (even when these are non-human). Thus, their experiment highlights the natural connection between ToM and narrative language, the focus of this dissertation. In addition, their findings precipitated the view developed by Sellars (1956) that humans in their early development continuously refine ideas about the minds of others in interaction with experiences in the world, hence, really develop a *Theory* of Mind that is constantly tested. A corollary idea was that

the maturation of ToM in children could be tested accordingly in experimental setups (Baron-Cohen et al., 1985; Gurney et al., 2021). Such setups presented for example a social scenario that tested children's ability to reason about the false beliefs of a story character (see for a recent example Figure 1.2), and found strong overlap between emerging ToM competence on the one hand, and language competence on the other (for reviews see Milligan et al., 2007; Wellman, 2018), leading some scholars to conclude that language is not only a key representational tool for ToM, but also provides the scaffold for this capacity to emerge (for a review see De Villiers and De Villiers, 2014).

Although progress has been made over the past decades, ToM is still a topic of ongoing debate in the scientific community. For instance, in developmental psychology the validity of experimental ToM test setups is at issue, as some scholars argue that ToM is essentially a *social* ability that should also be studied in social settings beyond the lab (Beauchamp, 2017; Beaudoin et al., 2020; Quesque and Rossetti, 2020), which would warrant utilising different types of data in unravelling ToM. If ToM is a social ability to coordinate other humans' behaviour, attention, and grasp their communicative intents, as Tomasello (2003) proposes, it may look different in lab settings with standardized ToM tests that have no obvious social relevance. Related to this problem is the finding that ToM competence depends on task-specific features such as memory or decision making, and varies in both children and adults (Barone et al., 2019; Flobbe et al., 2008; Van Duijn, 2016), implying that further work on ToM should include a broader range of tasks and sample populations than those typically included. And in Artificial Intelligence (AI), researchers have turned to experimental tests that have been used for decades in developmental psychology to evaluate to what extent contemporary Large Language Models (LLMs) as AI-models of language and cognition have ToM-like ability (for an overview see Ma et al., 2023c). though it remains unclear what LLM performance on these tests entails about ToM in both humans and machines (Trott et al., 2023; Ullman, 2023).

These unresolved issues provide fertile grounds for further analysis of the relation between ToM and language, that given the multidisciplinary nature of the discussions, should employ a corresponding multidisciplinary approach. This dissertation develops new perspectives along the following axes:

- **Methods –** This dissertation employs computational, qualitative, and experimental methods. We show that it is the combination of these methods that reveals patterns in unravelling the relation between children's ToM development and language use.

- **Resources –** The relation between ToM and language has long been investigated in children (for reviews see Milligan et al., 2007; Wellman, 2018), with traditional approaches often relying on ToM tasks in experimental settings, while other scholars have pioneered and called for more research on ToM in social contexts (e.g. Nicolopoulou and Ünlütabak, 2017; Quesque and Rossetti, 2020). We compiled ChiSCor, a new corpus of Dutch children's narratives, freely told in classroom settings, that enables analysis of children's ToM and cognition through their language use.

- **Empirical work –** We empirically examine the language children use when they render character minds in narratives, as proxy for their ToM competence, but also how LLMs can be meaningfully used in the context of language development, and how LLMs as AI-models deal with the character minds found in experimental ToM tests. In doing so we take a constructive position in the debate on (socio-)cognitive abilities of LLMs (Bubeck et al., 2023; Contreras Kallens et al., 2023; Kosinski, 2024; Ullman, 2023).

- **Reflection –** LLM performance on various (socio-)cognitive tests including ToM tests has sparked debate about LLMs' implications for human cognition, language development, and language understanding (e.g. Mahowald et al., 2024; Ullman, 2023; Warstadt and Bowman, 2022; Wilcox et al., 2022). We critically analyse recurring claims in this debate and develop a pragmatic perspective on (socio-)cognitive abilities of LLMs.

In the remainder of this introduction, given the prominence of narrative in the resources, methods and empirical work in the studies comprising this dissertation, we provide in Section 1.1.1 more background on the relation between ToM, narrative and child development. Thereafter, in Section 1.1.2, given our reliance on computational tools and models in analysing ToM and language use, we provide more background on the relation between ToM and narratives on the one hand and LLMs on the other. We then discuss our research questions in Section 1.2.1, methods in Section 1.2.2, datasets in Section 1.2.3, dissertation outline in Section 1.2.4, and lastly the contributions of this dissertation in Section 1.3.

1.1.1 Theory of Mind, Narrative, and Development

Narratives are part of our everyday lives, in many different forms (e.g. novels, oral stories, songs, advertisements), and with many different functions (e.g. educating,

motivating, entertaining, and persuading us). Hence, a single definition that covers all narrative manifestations is unattainable, as any narrative object depends on the perspective from which it is seen and used (Yamshchikov and Tikhonov, 2023; Zeman, 2016). In this dissertation, we adopt a liberal view on narratives and consider transcendence a key feature, which means a departure from the immediate here-and-now of the narrator (Zeman, 2016). Other definitions in studies may stress other aspects dependent on the goal of the investigation, for instance a narrative as a sequence of events, revolving around a particular protagonist, plot or issue (Botvin and Sutton-Smith, 1977; Ganti et al., 2022). Propp (1968) famously decomposed fairy tales into general sets of acts that have some bearing on the course of action, and a sequence of such acts constitutes a plot. Examples are 'Villainy', where a villain harms another character, 'Mediation' where the villainy becomes known to the hero of the story, and 'Beginning counteraction', where a hero thinks of possible solutions that will shape its future actions.

From a general perspective, narratives can be seen as culture-specific carriers of shared beliefs, values, norms and practices, that people continuously use to make sense of themselves and the world (Bruner, 1990). Narratives are present in virtually every society and constitute one of the oldest means for sharing human experience (Heath, 1986); some scholars even argue that every culture at its core is buttressed by narrative (Niles, 1999). Moving to the viewpoint of an individual 'consuming' fictional narratives, many (but not all) narratives constitute what Zunshine (2006) calls a cognitive experiment: with little linguistic cues we can typically 'try on' a variety of mental states like beliefs, desires and intentions of characters that are different from us. In doing so we learn to see the world through different characters' eyes, hence, activate and rely on our ToM competence. As a growing body of research demonstrates, exposure to fictional narratives tends to boost ToM of individuals by various measures (see e.g. Eekhof, 2024; Kidd et al., 2016). While such effects have not been studied for creating narratives, there is evidence that ToM and storytelling competence have interlinked developmental trajectories (Nicolopoulou and Ünlütabak, 2017), work that this dissertation builds upon and aims to further.

Besides this role of ToM in narrative, narrators can employ linguistic devices at the morphological, lexical, and syntactic levels to represent ToM and other discursive functions (Fernández, 2013). In narratology the linguistic representation of characters' perceptual, cognitive, and emotional states with respect to a particular situation is known as viewpoint (Eekhof et al., 2020), and viewpoint can be seen as the literary counterpart of ToM. We provide three illustrations in Table 1.1, that show that

Line		Function
1.	*..a girl **went** to the zoo and she **saw** a lot of tigers and other animals..*	– In line 1, the indefinite article 'a' in *a girl* signals the introduction of the girl as a new referent. Hereafter the girl (sister) is in the common ground between both narrator and audience, as indicated by the use of definite article 'the' in *the sister* in line 7. This pragmatic use of articles arguably relies on (a precursor to) ToM (Rubio-Fernández, 2021).
5&6.	*..and she **went** home all alone. But her brother **was** left behind he **was** sitting on a monkey..*	– The use of the **past tense** in lines 5&6 (but also in lines 1 and 7) suggests departure from the immediate here-and-now of the narrator (Clement, 1991; Zeman, 2016).
7.	*..then **said** the sister of the little boy where is my little brother now..*	– The use of the present tense and a first person pronoun in line 7 indicates Direct Speech and shifts attention to the girl's perspective (Leech and Short, 2007).

Table 1.1: Example functions of specific linguistic forms in a narrative context. The lines correspond to excerpts of a story with ID 072201 from the ChiSCor corpus (see Chapter 3).

narratives are 'sandboxes' for children for 1) learning how common ground can be indicated; 2) demarcating irrealis from realis; and 3) managing access to a (fictional) character's perspective. This all happens via using language in specific ways.

Thus, storytelling may challenge children to employ their ToM competence in creating and managing character minds, and their linguistic competence in rendering characters' perspectives in various ways (Frizelle et al., 2018; Nicolopoulou et al., 2015; Nicolopoulou, 1993; Southwood and Russell, 2004). Hence, researchers have turned to narratives produced by children to study their (socio-)cognitive and linguistic competences. For example, the mental complexity of the character minds children create, was used as proxy for their ToM ability (Nicolopoulou and Richner, 2007; Nicolopoulou, 2016); the lexical and syntactic properties of narratives as proxy for their linguistic competence (Miller, 1991; Nicolopoulou et al., 2022; Southwood and Russell, 2004); and narrative plot structure as proxy for their cognitive development (Botvin and Sutton-Smith, 1977; Shapiro and Hudson, 1991; Wardetzky, 1990). Such work employs narratives as windows on children's linguistic and (socio-)cognitive development, as an ecologically valid and complementary way to study children's development, since narratives lie at the root of many speech acts in childhood (Botting, 2002). Narratives have as further advantage that they provide a platform for contextualising actions and thoughts of (fictional) characters, i.e. why

they do or think certain things, context that in experimental ToM setups may be lacking, ambiguous, or irrelevant (Bloom and German, 2000).

This dissertation has as *methodological* contribution the demonstration that computational tools are valuable in complementing existing work on ToM and language. Earlier studies relied on manually labelling linguistic items, for example children's use of words referring to cognitive and emotional states (Fernández, 2013), and children's use of evaluative language (Nicolopoulou et al., 2022), which are both closely related to ToM.

Text classification algorithms as common in Natural Language Processing (NLP) can be helpful here. With a text labelling that indicates the different levels of ToM manifest in a story, such algorithms can retrieve specific words and other lexical and syntactic properties that are associated with each label. Such algorithms can also partly automate the extraction of viewpoint phenomena in narrative, i.e. the representation of characters' perspectives. As an example, in letting a text classifier distinguish Wikipedia text (arguably more viewpoint-neutral) from novels (where viewpoint is arguably more prevalent), the classifier will likely retrieve the lexical indicators of viewpoint in novels, and use them to distinguish novels from Wikipedia text. This ties in with existing manual approaches for identifying lexical items that indicate viewpoint as developed by Eekhof et al. (2020). In this dissertation, we aim to demonstrate how computational and qualitative methods can further reinforce each other in analysing the intersection of language and ToM.

Besides this methodological contribution in studying the relation between ToM and language, this dissertation contributes a new *resource*. The narrative datasets collected in earlier work were typically smaller in scale regarding the number of unique children and age range included (e.g. Fernández, 2013; Nicolopoulou and Richner, 2007), or were elicited not in interactive, social contexts (Tellings et al., 2018a). This is why this dissertation introduces ChiSCor (Children's Story Corpus). ChiSCor is a new corpus of 619 fantasy narratives freely told by 442 Dutch children aged 4-12 years in their natural classroom, day care, and community centre environments. ChiSCor constitutes the main resource for the various studies in this dissertation that employ (mixes of) computational, qualitative and experimental methods.

1.1.2 Theory of Mind, Narrative and Large Language Models

Although not immediately obvious, narratives are highly relevant for Large Language Models (LLMs) as novel type of AI-models of human language use. LLMs

are deep neural networks with a Transformer architecture (Vaswani et al., 2017), pre-trained with cloze objectives on large text corpora. BERT as developed by Devlin et al. (2019) is one of the first well-known language models and since BERT, ever larger descendants in terms of number of parameters and training data size have been developed, including e.g. BLOOM (Scao et al., 2022), LLaMA (Touvron et al., 2023) and GPT-4 (Achiam et al., 2024).[2] All these later LLMs have the capacity to produce fluent language, mostly acquired in the pre-training phase (Lin et al., 2024; Zhou et al., 2023a).

If human interaction with LLMs (and AI systems more generally) is to proceed smoothly and successfully, it is critical that these systems can deal with the mental states of the user, for example what the user knows about the world (e.g. background beliefs) and wants (e.g. desires) (Andreas, 2022; Cuzzolin et al., 2020; Kouwenhoven et al., 2022; Rabinowitz et al., 2018a). Note that this is not to say that this information must always be explicit in LLMs: just as in our own interactions with other humans, information about others' mental states typically remains implicit until requests for explication are made, for example in case of a misunderstanding.

Hence, it is intuitive to train LLMs at least partly on narrative datasets (Eldan and Li, 2023), as narratives contain information regarding ToM and its linguistic representation as explained above. Indeed, the datasets used for training LLMs often include narratives, although their inclusion is often not explicitly motivated. For example, the training dataset of the vanilla GPT-3 model (Brown et al., 2020) includes the BookCorpus, that contains almost 1 billion tokens scraped from self-published books on the web (Zhu et al., 2015). In addition, data from the large web crawl Common Crawl[3], a frequent component of LLM train data, includes many more (parts of) narratives in various forms from web fora and other sites where people share experiences, entertain each other, and so on, often by drawing on narrative.

Still, Sap et al. (2022) doubt whether the text in books, newspapers, Wikipedia and so on provides enough information for LLMs to learn to model mental states, as this

[2]Here we note that by current standards, BERT is typically not considered a *large* language model any more. Besides scale differences (BERT-large with 340M parameters (Devlin et al., 2019) is more than 500 times smaller than GPT-3 with 175B parameters (Brown et al., 2020)), also differences regarding architecture and capacities play a role. More recent LLMs like GPT-3 are typically unidirectional models with a decoder-only architecture, compared to BERT that is bidirectional and has a separate encoder. Also, more recent LLMs have been shown capable of doing downstream tasks like classification without further training, which BERT-models cannot (Brown et al., 2020). We discuss BERT-like language models in (Chapter 6 and Chapter 8) and will refer to them as Language Models to acknowledge this difference with LLMs. Still, both BERT-like language models and LLMs are Transformer networks that can be employed as powerful distributional learners to model cognitive and linguistic phenomena through language exposure.

[3]https://commoncrawl.org/the-data.

information could more often than not be implicit in such texts. Further, Van Eecke et al. (2023) argue that LLMs lack the human experience and world knowledge necessary to properly decode mental content in narratives. Yet, others show that LLMs are at least to some extent able to represent user intent (Andreas, 2022), and argue that LLMs encode a lot of world knowledge in their internal vector representations of input text, particularly in the relations between them (Piantadosi and Hill, 2022). Also, evidence is emerging that smaller LLMs trained with smaller sets of narrative data retain at least some of the fluency and reasoning capabilities of their larger counterparts (Eldan and Li, 2023), and that narrative formats provide a useful structure for LLMs to retrieve common sense knowledge (Bian et al., 2024). All this work draws on the general idea that narratives underlie how we store and transmit knowledge (Schank, 1995).

Apart from the value of training LLMs on narratives, narratives also provide useful opportunities for evaluating LLMs. Alabdulkarim et al. (2021) and Yamshchikov and Tikhonov (2023) argue that generating narratives constitutes a challenging task for such systems, as they may struggle to generate longer stories that are compelling in human eyes. Related work by Zhao et al. (2023) trains smaller LLMs with limited amounts of data and uses a storytelling task to assess model fluency, coherence and creativity. In addition, Stammbach et al. (2022) use narrative texts and prompt a LLM with reading comprehension questions to see if the model can correctly identify key Proppian roles such as Villain, Victim and Hero (Propp, 1968). In the medical domain, patient narratives regarding experiences with particular diseases were used to test LLMs' capacity to distinguish narrative text from other text types in social media posts (Ganti et al., 2022). In addition, fine-tuned BERT-like language models were used to extract patient coping strategies for dealing with adversarial drug effects through processing online forum posts (Dirkson et al., 2023).

In all the work mentioned above, ToM-related content plays a key role: the beliefs, desires and intentions of protagonists, characters, and patients that make a story compelling, creative and coherent, that render them typical Proppian characters, or that constitute the feelings and thoughts of what it is to deal with a particular disease or drug. That is, as Brahman et al. (2021) recognise, in both literary scholarship and computational approaches to understanding narratives, understanding characters and their mental states is vital, and the latter depends on properly modelling ToM in narratives.

Although LLMs can be successfully leveraged on a host of downstream text-based tasks like translation and question-answering (e.g. Brown et al., 2020), it is still de-

bated how valuable they are in other contexts such as human language acquisition (Warstadt and Bowman, 2022; Wilcox et al., 2023) and cognition (Browning and LeCun, 2022; Frank, 2023; Hu and Frank, 2024; Mitchell and Krakauer, 2023). This dissertation adopts a constructive position in the debate on the role of LLMs in studying human development and cognition. Since LLMs constitute, due to their unprecedented fluency in outputting language, arguably our current best models of language understanding (Sahlgren and Carlsson, 2021), and perhaps also further cognitive ability (Binz and Schulz, 2024), it is worthwhile to examine how well these models deal with (socio-)cognitive content present in narratives. In addition, LLMs provide possibly valuable 'benchmark' representations of mature language use, against which we can compare development in children's natural language samples. This dissertation provides further *empirical work* to explore these topics and contribute to the debate.

Besides the empirical work, this dissertation also provides broader theoretical *reflection* on pressing issues surrounding LLMs: do these models understand language like humans do, do they have acquired (socio-)cognitive abilities as a byproduct of their training objective (Bisk et al., 2020; Kosinski, 2024; Mitchell and Krakauer, 2023)? These questions have been addressed with empirical work, but theoretical reflection is lagging behind. This is why we critically analyse arguments regarding the (lack of) language understanding and other cognitive abilities in LLMs, and offer a different perspective on these issues. A related debate revolves around the claim that we as humans tend to fall in the trap of anthropomorphising LLMs (e.g. Bender and Koller, 2020; Floridi, 2023). In our reflection, we propose a philosophical pragmatic position that argues that simple anthropomorphisation does not adequately describe or explain how we as humans deal with unobservable entities such as mental states, that we infer from observable behaviour for pragmatic reasons, regardless of whether this behaviour is displayed by LLMs or humans.

1.2 Dissertation Design

1.2.1 Research Questions

With the background on ToM, narrative, development, and LLMs in place we can now formulate the following main research question (MRQ):

MRQ – How can we unravel the relation between Theory of Mind and language using computational methods and narratives?

This dissertation develops two complementary perspectives that are united under the MRQ, but differ in how they address it. The first perspective (employed in Chapters 2 through 5) focuses on unravelling ToM in children through narratives. The computational tools employed are feature engineering and classification, which are well-established in Computational Linguistics and NLP. The manual annotation of language data is also central to these chapters. The second perspective (employed in Chapters 6 through 8) also focuses on ToM and narrative, but in the context of modern AI. It employs a LLM as representation of mature language use to benchmark children's language use, employs LLMs as subjects in ToM tests themselves, and reflects on similar developments in current research. In sum, the computational aspect manifests in various ways: in text classification and feature engineering, but also in employing LLMs as novel computational models of language and cognition. Its multidisciplinary twist is that it is complemented by manual annotation, experimental data, psychological and narratological theory, and so on.

We break down the MRQ in seven research questions that correspond to seven self-contained chapters, that were originally published as research papers at various international, peer-reviewed conferences and workshops. These papers were included mostly as-is in this dissertation, apart from minor edits for consistent use of terminology, formatting, additional clarifications and information, etc. An advantage of this format is that each chapter can be read and understood on its own. A drawback is that there is some redundancy in the introductory sections of some chapters, for which we ask the reader's lenience. In the remainder of this section we introduce and motivate each research question.

RQ1 – How can we predict the mental complexity of story characters with computational tools?

Storytelling challenges children to employ their ToM in creating and managing character minds. Beyond that, it challenges children's linguistic competence regarding lexicon and syntax, and cognitive competences such as memory and planning to deliver a narrative that is interesting to an audience (Ebert and Scott, 2014; Frizelle et al., 2018; McKeough and Genereux, 2003; Nicolopoulou et al., 2015; Nicolopoulou, 1993;

Southwood and Russell, 2004). In this dissertation we show that these competences can be analysed through the language children use in narratives they tell freely. Linguistic analysis, however, is still often done manually in natural language samples of children (e.g. in Karlsen et al., 2021; Nicolopoulou et al., 2022; Nicolopoulou, 2016; Southwood and Russell, 2004), whereas recent developments in computational linguistics hold promise for the automatic analysis of children's language use (Harmsen et al., 2021; Hoeksema et al., 2022). Hence, the first exploration of the MRQ is obtaining evidence that computational tools disclose linguistic properties of children's narratives that can be linked to their ToM competence. As a proxy for ToM competence we annotate for each story the mentally most developed story character created by a child, a labelling originating from Nicolopoulou and Richner (2007) that we call Character Depth (CD). Regarding linguistic properties we focus on the lexical and syntactic complexity of the language used in the stories.

> **RQ2** – What is the contribution of narrative language data to research in (social) cognition and (computational) linguistics?

Research on the relation between ToM and language competence in children is typically done in controlled settings (for overviews see Milligan et al., 2007; Wellman, 2018). Still, scholars call for complementary, ecologically more valid ways to study language and ToM (Beauchamp, 2017; Beaudoin et al., 2020; Rubio-Fernández, 2021), as language and ToM as social competences can be thought of as two sides of the same coin following Tomasello (2003), and hence should also be studied in social contexts. Inspired by the work of Nicolopoulou (1993, 2007, 2016), Nicolopoulou and Richner (2007), and Nicolopoulou et al. (2022) on collecting children's narratives in social contexts to study ToM and language we introduce ChiSCor. ChiSCor (Children's Story Corpus) is a new Dutch resource of 619 narratives freely told by 442 children in social settings for research in (social) cognition and (computational) linguistics. ChiSCor drives much of the empirical work in this dissertation, and here we present three case studies that illustrate and underscore ChiSCor's broader potential for research on language, cognition, and in NLP.

> **RQ3** – How can a text classification task complement existing experimental work on the relation between Theory of Mind and language in children?

Text classification algorithms can assign labels to texts in explainable ways. We extract linguistic features from a large set of ChiSCor's narratives and train a classifier that assigns ToM labels to stories, based on the mental depth of their story characters (Character Depth as mentioned under RQ1). These features encode linguistic competences known to predict ToM in experimental settings, hence, allow us to see whether we can extend insights from experimental settings to more social settings by drawing on a data-driven approach. Although related work on classifying children's natural language responses on experimental ToM tests exists (Devine et al., 2023; Kovatchev et al., 2020), this work does not unravel the language children use when dealing with (character) minds, hence is less informative about their development.

> **RQ4** – What different types of Character Perspective Representation occur in ChiSCor's narratives and what is their relation to children's age and language use?

Character Perspective Representation (CPR) concerns the representation of what characters think, perceive, and say, that is, their perspective. Thus, CPR is closely related to ToM, but as concept more commonly found in narratology and stylistics. Here we focus on all possible instances of CPR in a story following a CPR framework from stylistics (Leech and Short, 2007), instead of on Character Depth as in previous RQs. Although the acquisition of Direct and Indirect Speech as specific CPR types has been studied in children (see e.g. Köder, 2016), the full range of CPR types children employ in storytelling has not been explored. Also, little is known about the linguistic contexts of different CPR types, which we analyse in this chapter with computational tools.

> **RQ5** – In what way can we meaningfully employ Language Models in studying children's language development?

There is discussion about whether LLMs are relevant in the context of children's language development, for example given LLMs' different learning mechanisms and advantages regarding language exposure (e.g. Bisk et al., 2020; Warstadt and Bowman, 2022). To illustrate how LLMs can be meaningfully employed in language development, we employ a Dutch language model as a representation of mature language use and analyse discursive meanings in children's use of Dutch perception verb *zien* ('to see') in ChiSCor. *See* can have a straightforward, denotational meaning that indi-

cates that 'entity X visually perceives object or event Y' as in *he saw the red car*. Beyond that, *see* can also have complex meanings involving further attentive aspects as in *he saw/evaluated the movie and did not like it*, and cognitive aspects as in *she saw/understood what he was up to* (San Roque et al., 2018; San Roque and Schieffelin, 2019). We predict masked occurrences of *see* in children's language use with a language model to quantify the distance to mature use and to explore the occurrence of complex meanings.

> **RQ6** – To what extent do Large Language Models show behaviour that is consistent with having Theory of Mind-like competence?

Scholars increasingly look to LLMs as subjects with cognitive abilities instead as mere 'autocompleters' of given inputs, that may have been learned as byproducts of training (Blank, 2023; Hagendorff, 2023). Especially testing ToM-like ability in LLMs has spurred discussion, for example about LLMs' generalisation capacity and what good or bad performance on ToM tests entails (e.g. Kosinski, 2024; Ullman, 2023; Trott et al., 2023). To address these issues, we set up a large-scale evaluation of ToM with various tests from developmental psychology that are presented to various recent LLMs. Different from similar work on this topic (e.g. Sap et al., 2022; Shapira et al., 2024), we evaluate LLM responses on open questions, include child performance on the same tests as a benchmark, and explain our findings with reference to human language evolution and development.

> **RQ7** – What are the implications of Large Language Models' complex behaviour for studying human language understanding and cognition?

LLMs have sparked debate on how they model human language understanding (Bender and Koller, 2020; Piantadosi and Hill, 2022; Wilcox et al., 2023) and cognition (Binz and Schulz, 2023; Blank, 2023; Frank, 2023). As Bowman (2022) has argued, strong claims about the limitations of NLP systems can be dangerous in that they are quickly picked up on by the research community but often lead to erroneous inferences. This in turn hampers properly understanding LLMs in broader academia, but also in the public sphere. We survey the debate and extract and critically analyse three recurring arguments regarding LLMs as models that i) are simple autocompleters; ii) cannot model the function of language; and iii) are irrelevant in the context of human language acquisition. In addition, we develop a pragmatic philosophical

framework to rethink what 'real' language understanding and intentionality mean in the context of LLMs.

1.2.2 Methods

ToM is studied differently in different fields and various practices inspire our combinations of computational, qualitative and experimental methods, on which we elaborate below.

In developmental psychology, *manual annotation* of the mental complexity of story characters (Character Depth, Section 1.2.1, RQ1) that children create provides insight in their capacity to create mental agents as proxy for ToM (Nicolopoulou and Richner, 2007). In narratology, the analysis of characters' perspectives (Character Perspective Representation, Section 1.2.1, RQ4) is concerned with what characters think, perceive and say, which is similar to ToM. The linguistic representations that realise such perspectives (Leech and Short, 2007; Van Duijn et al., 2015) may invite *manual analysis*: for example of the use of deictic terms and first- vs. third-person pronouns in story retellings of neurodivergent populations to analyse their perspective management (Van Schuppen et al., 2020). But analysing perspective can also involve *computational modelling*, for example character extraction from linguistic features that capture speech representation (Karsdorp et al., 2012), or *classification* of types of perspective representation from linguistic features (Brunner, 2013). Lastly, in AI there may be more reliance on *benchmarking* as a method for comparing systems' cognitive or ToM-like ability against some set standard. In the case of ToM, this amounts to assessing a system's behaviour in response to prompts containing ToM-related information (as in Sap et al., 2022).

In this dissertation we adopt combinations of the computational and qualitative methods mentioned above. For example, manual annotation may provide gold labels for text classification, and manual analysis of natural language may inform the features we want to extract from a text. An overview of the methods employed per chapter is given in Table 1.2. To guide the reader, we elaborate on these methods in the remainder of this section.

- **Manual annotation –** Refers to the theory-informed annotation of (parts of) ChiSCor's narratives. Examples are Character Depth (CD) annotation, where human annotators label the mental complexity of the story characters children create, and Character Perspective Representation (CPR) annotation, where human annotators label the ways children represent characters' perspectives. CD and CPR are used

Method	RQ1	RQ2	RQ3	RQ4	RQ5	RQ6	RQ7
Manual annotation	•		•	•			
Statistical modelling	•	•		•	•		
Feature engineering	•	•	•	•	•		
NLP		•	•			•	
Child experiments						•	
Theoretical analysis							•
Chapter	2	3	4	5	6	7	8

Table 1.2: Overview of the main methods employed for each RQ/chapter.

as dependent variables in statistical models (RQ1, RQ4) and CD constitutes the gold labels in a story classification setup (RQ3).

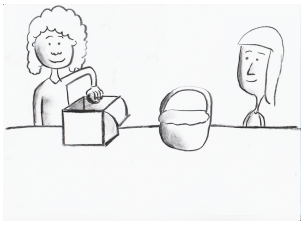
- **Statistical modelling –** Refers to statistical models such as linear models used for testing hypotheses regarding ToM (RQ1), CPR (RQ4), and language development (RQ2, RQ5).

- **Feature engineering –** Refers to the extraction of information from text using computational tools, for further use in statistical models or text classification setups. We extract linguistic features such as syntactic complexity (RQ1 through RQ4), but also more abstract features such as surprisal from a LLM (RQ5).

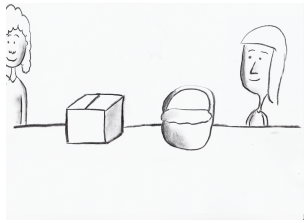
- **NLP –** Refers to experimental setups using vector models for inducing word semantics from text (RQ2), text classification based on linguistic features (RQ3), and a baseline of human ToM test performance against which LLMs are benchmarked (RQ6).

- **Child experiments** – Refers to ToM experiments carried out with children alongside ChiSCor's compilation, with the goal of creating a benchmark for comparing LLM performance on the same experiments (RQ6). These ToM tests presented children with a social scenario in text and (audio)visual format on a screen (see Figure 1.2), and asked comprehension questions about what characters (falsely) believe, want, intend, etc.

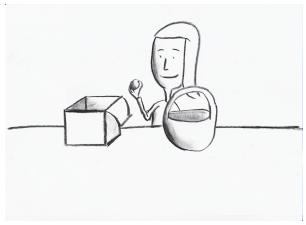
- **Theoretical analysis** – Refers to critical analysis of arguments in current debates on LLMs and their implications for studying human cognition and language understanding (RQ7). This also refers to the development of a pragmatic philosophical perspective on these debates.



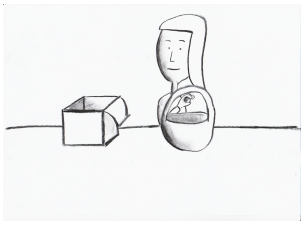
(a) This is Sally **(left)** and Anne **(right)**. They are playing. Sally has a box and Anne has a basket, and there is a ball. Sally puts the ball in her box...



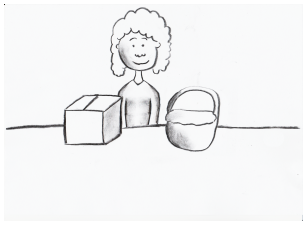
(b) Then Sally goes to play somewhere else.



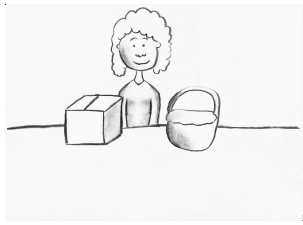
(c) Anne takes the ball from Sally's box...



(d) ...and she puts the ball in her basket. Anne also goes to play somewhere else for a while.



(e) Then Sally returns.



(f) Where will Sally look for the ball?

Figure 1.2: Illustration of a digital ToM experiment (here the Sally-Anne test originating from Baron-Cohen et al. (1985)) presented to children. Children see sub-figures (a) through (f) successively on a monitor or tablet, and at (f) answer an open question by typing their answers in a text box (not shown in the picture). Illustrations by Werner de Valk.

1.2.3 Datasets

Here we highlight the datasets employed throughout this dissertation. An overview of datasets used per RQ/chapter is given in Table 1.3. As can be seen, ChiSCor drives much of the work in this dissertation (RQ2 through RQ6). We also use the 'free essays' section of BasiScript (Tellings et al., 2018a), a 3.4M token corpus of freely written essays from thousands of children (7-12y) throughout The Netherlands (RQ2). Also, we use experimental data that result from carrying out various ToM tests with children from two different age groups (RQ6).

RQ (Chapter)	Dataset	Details
1 (2)	Pre-ChiSCor pilot set	51 stories from 51 children (4-10y)
2 (3)	Full ChiSCor	619 stories from 442 children (4-12y)
	BasiScript sample	Full 'free essays' section from BasiScript, ±33k essays from ±11k children (7-12y)
3 (4)	ChiSCor sample	442 first-told stories of 442 children (4-12y)
4 (5)	ChiSCor sample	150 stories in total from young (4-6y), middle (6-9y), and old (9-12y) age groups, 50 stories from 50 children per group
5 (6)	ChiSCor sample	90 stories selected from young (4-6y), middle (6-9y), and old (9-12y) age groups, 30 stories per group from 68 children in total
6 (7)	ChiSCor sample	Experimental results of ToM tests with 36 children from middle (7-8y) and 37 children from old (9-10y) age groups
7 (8)	NA	NA

Table 1.3: Overview of the datasets employed for each RQ/chapter.

1.2.4 Outline

This section is intended as a brief guide for the reader through the organisation of this dissertation, which is best read alongside the dissertation structure laid out in Figure 1.3.

Chapter 2 - Modelling Story Characters' Mental Depth

This *pilot* chapter preludes the key concepts and the data resources at issue in later chapters. It introduces narrative as a form of cognitive play at the intersection of ToM and language competence, Character Depth as a window on ToM competence in narrative, and extracting linguistic features from narratives using computational tools.

> **Published as:** Van Dijk, B.M.A. and Van Duijn, M.J. (2021). Modelling Characters' Mental Depth in Stories Told by Children Aged 4-10. In Fitch, T., Lamm, C., and Leber, H., editors, *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, pages 2384-2390. Cognitive Science Society.

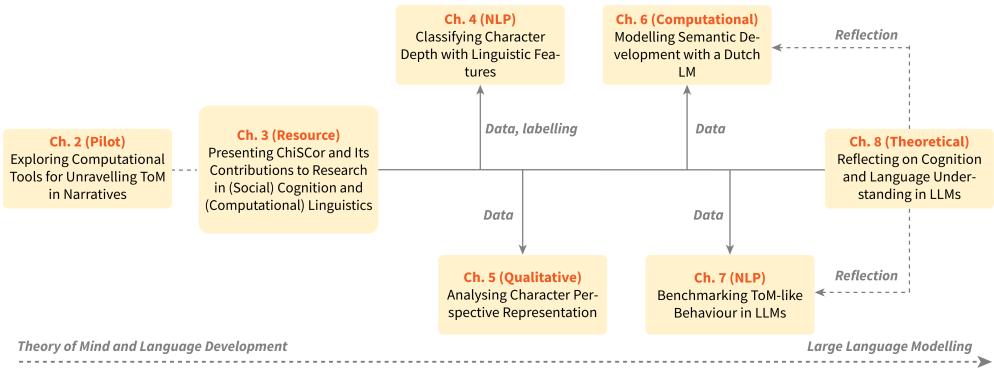


Figure 1.3: The structure of this dissertation as constituted by seven chapters and their themes.

Chapter 3 - ChiSCor: A Dutch Children's Story Corpus

This *resource* chapter details ChiSCor's compilation as a new resource for research in (social) cognition and (computational) linguistics. ChiSCor's larger scale enabled training a text classifier on a larger set of linguistic features (Chapter 4); looking at Character Perspective Representation in different age groups (Chapter 5); assessing language development in ChiSCor with a LLM (Chapter 6); and leveraging a child baseline for evaluation of LLM ToM performance (Chapter 7).

> **Published as:** Van Dijk, B.M.A.,* Van Duijn, M.J.,* Verberne, S., and Spruit, M.R. (2023). ChiSCor: A Corpus of Freely-Told Fantasy Stories by Dutch Children for Computational Linguistics and Cognitive Science. In Jiang, J., Reitter, D., and Deng, S., editors, *Proceedings of the 27th Conference on Computational Natural Language Learning*, pages 352-363. Association for Computational Linguistics. (* denotes equal contributions.)

Chapter 4 - Classifying Theory of Mind in Freely Told Stories

This *NLP* chapter extracts linguistic features from narratives using a text classifier at scale in the wake of Chapter 2, enabling a finer-grained and more robust perspective on the relation between language and ToM in narrative.

Published as: Van Dijk, B.M.A., Spruit, M.R., and Van Duijn, M.J. (2023). Theory of Mind in Freely-Told Children's Narratives: A Classification Approach. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics*, pages 12979-12993. Association for Computational Linguistics.

Chapter 5 - Character Perspective Representation in Freely Told Stories

This *qualitative* chapter draws on theory in stylistics to view ToM in narratives from a different angle compared to Chapter 2 and Chapter 4. The chapter illustrates how ChiSCor can accommodate different kinds of annotations and hence future work at the intersection of ToM, stylistics, and development.

Published as: Van Duijn, M.J., Van Dijk, B.M.A., and Spruit, M.R. (2022). Looking from the Inside: How Children Render Characters' Perspectives in Freely-told Fantasy Stories. In Clark, E., Brahman F., and Iyyer, M., editors, *Proceedings of the 4th Workshop on Narrative Understanding*, pages 66-76. Association for Computational Linguistics.

Chapter 6 - Analysing Semantic Development with a Language Model

This *computational* chapter spotlights language models which are central in the last three chapters. The chapter argues and demonstrates that they can be useful computational models in a developmental context, bearing on Chapter 7 and Chapter 8.

Published as: Van Dijk, B.M.A., Van Duijn, M.J., Kloostra, L., Spruit, M.R., and Beekhuizen, B.F. (2024). Using a Language Model to Unravel Semantic Development in Children's Use of a Dutch Perception Verb. In Zock, M., Chersoni, E., Hsu, Y., and De Deyne, S., editors, *Proceedings of the 8th Workshop on Cognitive Aspects of the Lexicon*, pages 98-106. European Language Resources Association.

Chapter 7 - Theory of Mind in Large Language Models

This *NLP* chapter benchmarks LLMs' ToM-like ability against children on three ToM tests. Instead of using LLMs as tools, this chapter shifts the perspective to using LLMs as psychological subjects. This use of LLMs constitutes a use case for a broader reflection on LLMs as models of human language and cognition in Chapter 8.

> **Published as:** Van Duijn, M.J.,* Van Dijk, B.M.A.,* Kouwenhoven, T.,* De Valk, W.M., Spruit, M.R., and Van Der Putten, P.W.H. (2023). Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art Models vs. Children Aged 7-10 on Advanced Tests. In Jiang, J., Reitter, D., and Deng, S., editors, *Proceedings of the 27th Conference on Computational Natural Language Learning*, pages 389-402. Association for Computational Linguistics. (* denotes equal contributions.)

Chapter 8 - Reflecting on Language and Cognition in Large Language Models

This *theoretical* chapter nuances strong negative claims on language understanding and cognition in LLMs, which relates to the topics of Chapter 6 and Chapter 7. This chapter also develops a pragmatic perspective on the attribution of 'real' language understanding and intentionality to humans and machines, which depends on the practical and social value such attribution has to us. By returning to the idea of ToM and language being foremost social tools as explained earlier in this introduction, this chapter goes full circle.

> **Published as:** Van Dijk, B.M.A., Kouwenhoven, T., Spruit, M.R., and Van Duijn, M.J. (2023). Large Language Models: The Need for Nuance in Current Debates and a Pragmatic Perspective on Understanding. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12641–12654. Association for Computational Linguistics.

Chapter 9 - Conclusions

The *conclusion* chapter presents answers to all research questions. It also provides discussion on the limitations of this dissertation and directions for future research.

1.3 Dissertation Contributions

Here we briefly summarise the materials and other publications resulting from this dissertation.

Materials

1. **Children's Story Corpus (ChiSCor) –** Refers to the Dutch corpus of children's freely told narratives. Although this dissertation draws mostly on the 619 Dutch

stories told by 442 children that constitute the bulk of ChiSCor, the corpus also includes a small subset (62) of English stories, high-quality .wav files of virtually all stories, demographic metadata of 202 Dutch and English-speaking children, annotation protocol and annotations, and automatically extracted linguistic features for Dutch and English stories. ChiSCor's data types and metadata are further explained in Chapter 3 and available at `https://doi.org/10.17026/SS/TGPDJF`.

2. **Experimental ToM data –** Refers to test results of a suite of classical and new ToM tests that were administered to 83 Dutch and English children (4-12y) in primary schools. Tests include, among others, the canonical Sally-Anne (Baron-Cohen et al., 1985; Wimmer and Perner, 1983) and Strange Stories tests (Happé, 1994), the Dutch version of the Reading the Mind in the Eyes Test tailored to children (a.k.a. RMET) (Van Der Meulen et al., 2017), and the Imposing Memory test, a hitherto unpublished test that originates from the work of Robin Dunbar and Anneke Haddad, that evaluates higher levels of recursive ToM (Van Duijn, 2016). These test results are included in ChiSCor's repository at `https://doi.org/10.17026/SS/TGPDJF`.

3. **ToM test set for LLMs** – Refers to a set of ToM tests used for benchmarking LLMs. The set includes the Sally-Anne and Strange Stories tests, and carefully made test deviations that stray away from original scenarios and thereby gauge LLMs' generalisation capabilities. In addition, the set includes the Imposing Memory test, which because of its unpublished nature has no deviations. These tests are further explained in Chapter 7 and included in the accompanying repository at `https://osf.io/426p9/`.

4. **Source code of all papers** – Every chapter comes with a link to an associated repository, from which the code and data used to extract features, train models, create figures and so on can be consulted.

Other publications

Besides the academic publications mentioned in Section 1.2.4, the following publications (in Dutch) intend to convey insights obtained in this dissertation to a broader audience:

1. Van Dijk, B.M.A. (2021). Inductiemachines. In *Filosofie - Tijdschrift* 31(4), pages 25-28. Available at `https://osf.io/jxaz6/`.
2. Van Dijk, B.M.A. (2022). Een kunstmatig intelligente spiegel. In *Filosofie - Tijdschrift* 32(6), pages 38-42. Available at `https://osf.io/u5mws/`.
3. Van Dijk, B.M.A. (2024). Serendipiteit in silico. In *Filosofie - Tijdschrift* 34(5), pages 14-17. Available at `https://osf.io/mrafk/`.

Chapter 2

Modelling Story Characters' Mental Depth

From age 3-4, children are generally capable of telling stories about a topic free of choice. Over the years their stories become more sophisticated in content and structure, reflecting various aspects of cognitive development. Here we focus on children's ability to construe characters with increasing levels of mental depth, arguably reflecting socio-cognitive capacities including Theory of Mind. Within our sample of 51 stories told by children aged 4-10, characters range from flat 'Actors' performing simple actions, to 'Agents' having basic perceptive, emotional, and intentional capacities, to fully-blown 'Persons' with complex inner lives. We argue for the underexplored potential of computationally extracted story-internal features (e.g. lexical/syntactic complexity) in explaining variance in Character Depth, as opposed to story-external features (e.g. age, socioeconomic status) on which existing work has focused. We show that especially lexical complexity explains variance in Character Depth, and this effect is larger than and not moderated by age.

This work was originally published as: Van Dijk, B.M.A. and Van Duijn, M.J. (2021). Modelling Characters' Mental Depth in Stories Told by Children Aged 4-10. In Fitch, T., Lamm, C., and Leber, H., editors, *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, pages 2384-2390. Cognitive Science Society.

2.1 Introduction

From early childhood children tell stories to themselves and others as part of their daily play activities (Cremin et al., 2017; Sutton-Smith, 2012). Such storytelling has been described as a kind of cognitive play that — besides being the source of a lot of fun — forms a natural crossroads of various key areas in child development (Bergen, 2002; Paley, 1990; Smith and Roopnarine, 2018). Telling stories involves language skills at the phonological, lexical, syntactic, and pragmatic levels (Southwood and Russell, 2004). It draws further on cognitive abilities such as memorising, planning, organising knowledge of the world (McKeough and Genereux, 2003), and empathising with others, in particular to work out how characters should behave, speak, feel, and think in ways that are relatable and interesting for an audience (Nicolopoulou, 1993; Van Duijn et al., 2015; Zunshine, 2006).

Here we are interested in the representation of mental activities of characters and the place this has in child development. Existing theoretical work has linked children's ability to render character minds to the mastery of socio-cognitive skills, in particular mindreading or Theory of Mind (ToM).[1] Empirical research has shown that the complexity of characters and their mental activities that children can deal with tends to increase with age (e.g. Nicolopoulou and Richner, 2007; Nicolopoulou and Ünlütabak, 2017).

For this chapter, we recorded and transcribed 51 oral stories elicited from Dutch children of different ages and backgrounds, during storytelling workshops integrated in their daily school and daycare environments. A total of 268 characters were represented in these stories, each of which we assessed in terms of its mental depth. To give two brief opening examples of what we looked at (excerpts translated from stories we collected earlier):

(1) *they sit neatly in a row but the other* [*puppy*] *always enters later* (child 4y;1m)

(2) *they sat down as always until he was not looking* [...] *then they went inside the school director's office and secretly took the key* (child 9y;11m)

Characters presented in excerpts (1) and (2) fall at the lower and higher ends of the scale of mental depth that we will introduce in more detail below respectively. Excerpt (1) introduces characters with arguably different perspectives on the staged set-

[1] For a general overview of literature on mindreading/ToM, i.e. the ability to take others' perspectives and reason about their behaviour in terms of emotional and intentional states, see Apperly (2012). For an overview of theoretical work linking ToM with children's stories see Nicolopoulou (2015) and Zunshine (2019).

ting: some are already inside, while another one enters later. However, there is no fleshing out of mental activity by any of these characters at all, and this is representative of the rest of the story, which revolves around movements (coming in, going out) and actions (eating) only. This is very different in (2), where the implied protagonists' awareness of what the school director does not see, and hence knows, is central in the story's plot.

In line with existing work, we observe an overall increase in mental Character Depth with the age of the children telling the stories in our sample. However, it is our aim to understand in more detail which factors drive children's ability to render more complex characters. To this end, we develop a framework using computational techniques and statistical modelling for mapping out relationships between, on one side, the mental depth of story characters and, on the other side, multiple story-external features (e.g. age, socioeconomics) and automatically parsed story-internal features (e.g. vocabulary, syntax).

Our results show that in particular the lexical complexity a story exhibits can be used as a reliable predictor of Character Depth: it explains a larger proportion of the variance compared to age and is not moderated by age. We discuss the role of lexical complexity and other variables in understanding children's ability to deal with characters of different levels of mental complexity, both within our current sample and in larger, more diverse samples in the future.

2.2 Background

Narrative plays a key role in human communication. On a daily basis adults and children alike use stories to share their perceptions and imaginations with others, typically in causally, temporally, and logically structured ways. Classic definitions of narrative often emphasise criteria such as goal-directedness, causality, or the unfolding of series of actions over time (Duinmeijer et al., 2012). However, in this research we cast the definitional net a bit wider and argue that children's stories could also be descriptions of situations, events, or characters in which goals, causal relations, or a clear temporal development are not immediately present. What we take as our central criterion here to demarcate stories from other speech phenomena is *mediatedness* or *transcendence*, marked by a departure from the actual speaker and its immediate here-and-now (cf. Nicolopoulou and Richner, 2007; Zeman, 2018). For example, children merely describing their situation during the storytelling workshop in which we collected our data would not be telling a story (e.g. *'I am sitting on a chair in the*

group circle...'), whereas children describing a real or imagined situation set elsewhere would be, even if that situation is not worked out any further with additional characters and events (e.g. *'Yesterday I had a silent disco...'*).

In this chapter we focus on two of the developmental trajectories that naturally intersect in stories that children tell, social cognition and language, against the background of their more general development, which we approximate via age and educational level of the parents/caregivers. Following a large body of research (for an overview see Milligan et al., 2007; Tompkins et al., 2019), we expect these trajectories to be interrelated and it is our longer-term aim to contribute to further understanding of this interrelatedness by studying stories that children tell. Here we develop a framework for mapping out features within such stories that we assume to be manifestations of developmental progression on the linguistic and socio-cognitive levels. Our hypotheses at this stage concern the co-occurrence of and relationships between these features within the stories; testing whether this is indeed indicative of the development of the children who tell them is outside the scope of this chapter.

Social cognition

Firstly, we are interested in socio-cognitive sophistication of the stories, which we operationalise as the mental depth that characters exhibit, in short, Character Depth (CD). Using a slightly adapted version of the typology introduced by Nicolopoulou and Richner (2007) we rate each character's mental activity on a nine-level scale. These levels fall under three main categories: `Actors` undergoing (level I) or performing (level II) simple actions, `Agents` having basic perceptive, expressive, emotional, and intentional capacities (levels III-V), and `Persons` capable of coordinating beliefs, desires, expectations, and so on, with different imagined realities (levels VI-VII) and/or other characters' cognitive states (levels VIII-IX; see Section 2.3 and Table 2.1 below for more details).

Language

Secondly, we are interested in the linguistic qualities of the stories, which we operationalise on two levels: vocabulary and syntax. As a measure of vocabulary sophistication (a.k.a. lexical complexity) we assessed the vocabulary of each story by computing the probability of the occurrence of each lemma that a child used approximated by frequencies in a benchmark lexicon. This metric builds on the idea that the difficulty of words from the perspective of a language learner is strongly negatively

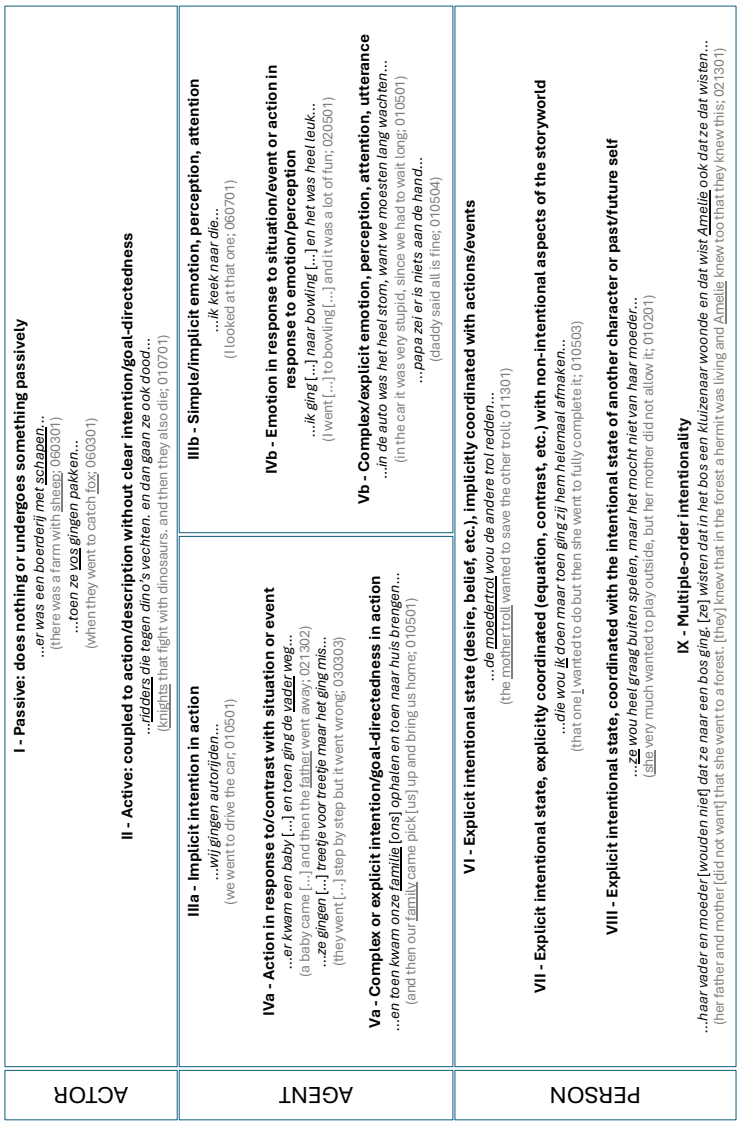


Table 2.1: Annotation scheme for CD. All examples are quotes from our dataset, followed by a somewhat liberal/idiomatic English gloss, followed by the unique ID of the story from which it was taken. Underlining indicates character to which the CD level applies in case of multiple characters in an example. Square brackets indicate elements of quotes that were reordered or omitted for purposes of readability.

correlated with how frequently it occurs (Vermeer, 2001). Thus, using less frequent words means using less probable words, and this we take to indicate a more complex vocabulary. The idea is that a more complex vocabulary functions as a communicative and mental toolbox that enables a child to render both the physical and social world better. This toolbox can be especially helpful when engaging in demanding tasks such as telling a story, where there is a sustained pressure for finding the right words to get the desired message across to an audience (Curenton and Justice, 2008).

As a measure of syntactic complexity, we calculated the average distance between syntactically dependent words. It is well-established that language structures which employ longer dependency distances between head words and dependent words are more difficult to process (Gibson, 1998; Gildea and Temperley, 2010). An example of this difference is given by King and Just (1991) in terms of subject-extracted relative clauses (3) and object-extracted relative clauses (4):

(3) The reporter who attacked the senator admitted the error.

(4) The reporter who the senator attacked admitted the error.

In both sentences the verb 'attacked' is dependent on the pronoun 'who'. In (4) these dependents are not adjacent, but have two words in between, which makes that part of the sentence more challenging to process. Average dependency distance seems to capture language skills more generally. For example, it can be used to distinguish English written by natives from that written by L2 learners (Oya, 2011) and speech from individuals with mild cognitive impairments from speech produced by typically developed speakers (Roark et al., 2007). Our idea here is that children capable of handling more complex syntactic structures, as indicated by their stories exhibiting higher average dependency distances, have more powerful formats available for representing events in the social and mental worlds, in discourse as well as in their own strands of reasoning (cf. De Villiers and De Villiers, 2014).

Hypotheses

Firstly, we hypothesise that stories exhibiting a more complex vocabulary contain characters with higher levels of mental depth. Secondly, we hypothesise that stories with larger syntactic dependency lengths contain characters with higher levels of mental depth.

Story-external features

Existing work has shown that the mental depth of characters in stories that children tell increases with their age (Nicolopoulou and Richner, 2007), which is why we include it in the model. Parent education functions as a proxy for socioeconomic status in our model; there is evidence that children from parents with a higher socioeconomic status perform better on ToM tasks (Shatz et al., 2003).

2.3 Methods

Dataset

For our data collection, we offered storytelling sessions to various institutions in the medium-sized Dutch cities Leiden, Tilburg and Utrecht. Three schools (two in Leiden, one in Utrecht), one daycare (Leiden) and one community centre (Tilburg) were willing to cooperate. Around 200 children in total participated in sessions held between September 2019 and June 2020. We were able to include 98 stories told by 54 children ($M_{age}(SD) = 6.81(1.66)$, range = 4.17-10.1; 30 females, 2 unknown) in our database after receiving consent forms from their parents. In order to maximise independence between observations we use only the first story told by each child, and due to missing information on the consent forms an additional 3 stories dropped out, resulting in a subset of 51 stories for this chapter. Our experiment and data management protocols were assessed and approved by the Ethical Committee of the Leiden University Faculty of Science (file no. 2020 – 002).

Our storytelling sessions were held in group circle settings. After briefly exploring some general features of stories interactively (e.g. 'What is a story?', 'What do we find in stories?') and narrating a short standard exemplary fantasy story, we invited children to tell a story about a topic free of choice. Voice recordings were made after informing the children about this. Afterwards, the recordings were pseudonymised and transcribed by the authors and research assistants twice: first orthographically (including 'noise' such as false starts, wrong conjugations, broken-off words, etc.), and second normalised, thus without noisy elements, to enhance compatibility with computational language processing tools. All transcripts were double-checked for consistency with the audio files. In addition to the story data, personal data such as age of the children and parental education levels were collected through consent forms. Transcripts, data, and code are available via `https://osf.io/k52e8/`.

Annotations

We loaded all pseudonymised transcriptions in the open online content annotation tool CATMA (version 6.1.3; (Horstmann, 2020)), where we created a tag set for CD. Tags within this set were based on the typology introduced by Nicolopoulou and Richner (2007). A few adaptations were made, however, in terms of the three main levels (`Actor`, `Agent`, and `Person`) our tag set remained compatible with the original typology. See Table 2.1 for descriptions and examples of the tags we have used to assign a CD level to each character. Our workflow included a first stage in which the authors of this chapter discussed the first 10 stories openly, followed by a second stage in which the remaining 41 stories were annotated by each of the authors independently. In the third stage, all tags that differed were discussed until consensus was reached. Finally, the annotations were considered fixed and downloaded from CATMA in TEI-XML format.

We extracted the maximum CD with a Python script. This feature represented the highest level of CD reached in a story on a scale from 0 to 9, corresponding with the levels in the topology set out in Table 2.1 when discarding subcategories indicated by letters (e.g. IVa and IVb both count as 4), where 0 indicates the theoretical option of no characters being presented in a story (which did not occur in our current dataset), value 1 corresponds with level I in Table 2.1, and so on.

Extracting Linguistic Features

Vocabulary Probability – Our approach was to take the textual vocabulary of a representative reference corpus, which consists of all the lemmas constituting the vocabulary of the corpus (Fengxiang et al., 2016). We use this benchmark to compute the probability of each story vocabulary, treating it as a subset of the textual vocabulary. Lemma probabilities were approximated by relative frequency counts in the reference corpus.

We obtained lemmas for each story by parsing normalised story transcripts with the memory-based Frog parser (Van Den Bosch et al., 2007). We used as reference corpus the 'free text' lexicon (FTL) of the BasiScript corpus (Tellings et al., 2018a), which consists of essays of primary school children with minimal teacher intervention, thus staying close to the free story paradigm. We removed punctuation marks and named entities from the FTL, which yielded a total number of token instances N of 3699822, and a vocabulary V of 46570 lemmas. The estimated probability of some lemma l_i

occurring in story S is given by

$$P(l_i) = \frac{(c_i + 1)\frac{N}{N+V}}{N}, \tag{2.1}$$

with c_i being the count of token instances of l_i in the FTL, adjusted for words not occurring in the FTL. This estimation is based on n-gram smoothing methods as outlined by Jurafsky and Martin (2024); we used Laplacian smoothing since the FTL includes many typical fantasy constructions such as 'trollensnot' (troll snot) with count 1, but not the similar construction 'eenhoornsnot' (unicorn snot) which occurs in our stories. We calculated the probability of the vocabulary of S with

$$L = \frac{1}{S_n} \sum_{i=1}^{n} P(l_i), \tag{2.2}$$

with the fraction being a normalising factor (S_n being the length of S), and converted them to per mille for convenient interpretation. The interpretation of L can be phrased as follows: if one draws a lemma from the FTL, how likely is it that it belongs to the story vocabulary? For complex vocabularies this probability will be lower.

Dependency Distance – We used the Alpino parser (Van Noord, 2006) to extract all dependencies per utterance. The dependency distance of the ith dependency relation DD_i is typically set to 1 for adjacent words, 2 if one extra word occurs in between the dependents, and so on. We follow Wang and Liu (2017) and compute overall dependency distance MD_{sent} for a sentence with n words by

$$MD_{sent} = \frac{1}{n-1} \sum_{i=1}^{n-1} |DD_i| \,. \tag{2.3}$$

Then, for a story consisting of multiple utterances,

$$MD_{story} = \frac{1}{u} \sum_{i=1}^{u} MD_i, \tag{2.4}$$

where u is the total number of utterances in a story.

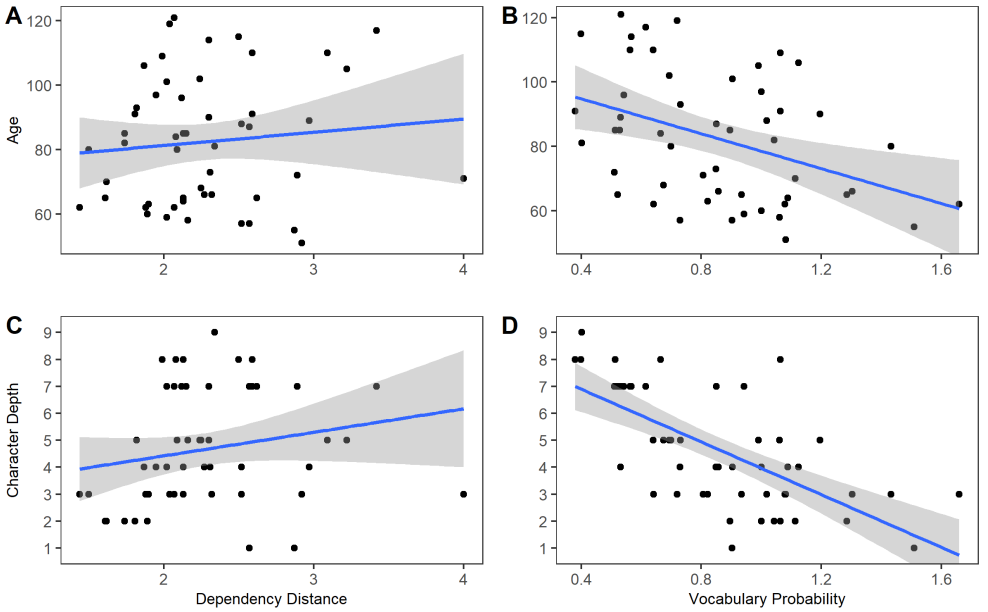


Figure 2.1: Correlation plots with Age in months, average Dependency Distance in number of words, average Vocabulary Probability in per mille, and Character Depth in levels.

2.4 Results

Bivariate explorations

Prior to constructing the linear model that we used for assessing our hypotheses, we explored various correlations between a subset of the features outlined above.

Firstly, it appears that Dependency Distance correlates weakly with Age (Figure 2.1 A, Pearson's $r = 0.104$) and Vocabulary Probability correlates moderately with Age (Figure 2.1B, Pearson's r = -0.403). It makes sense that as children grow older, both their vocabularies and syntax are becoming increasingly complex. Secondly, Dependency Distance correlates weakly with CD (Figure 2.1C, Pearson's r = 0.202) and Vocabulary Probability correlates strongly with CD (Figure 2.1D, Pearson's r = -0.670), indicating that the relationship between the parsed linguistic features and CD are in the expected directions, albeit in quite different gradations. In the next section we scrutinise these bivariate explorations using a linear regression model.

Predictor	β	*SE*	*t*	*p*	95% CI	
Intercept	4.716	.250	18.892	<.001	4.212	5.219
Vocabulary Probability	-1.117⋆	.289	-3.860	<.001	-1.701	-.534
Age	.582⋆	.265	2.193	.036	.047	1.117
Education Parents	.425	.265	1.602	.116	-.110	.961
Vocabulary Probability * Age	.161	.292	.551	.584	-.428	.750
Dependency Distance * Age	.146	.232	.628	.533	-.322	.614
Dependency Distance	-.016	.242	-.067	.947	-.110	.473

Table 2.2: Model estimates sorted on the magnitude of the standardised betas. Stars denote significance at the .05 level, two-tailed.

Hypothesis testing

We consider a linear multiple regression model most appropriate for the analysis; due to the limited number of observations per institution in our dataset, a mixed-effects model did not converge properly. Our model includes Dependency Distance, Vocabulary Probability, Age, Education Parents, and interactions between Vocabulary Probability and Age and between Dependency Distance and Age as predictors of CD. The model accounts for about 53% of the variance in CD $R^2 = .525, F_{6,44} = 8.132, p < .001$, with $M_{CD} = 4.667, SD_{CD} = 2.167$. Standardised coefficients sorted on magnitude are given in Table 2.2.

In line with our first hypothesis, we see that the simple effect of Vocabulary Probability has the largest negative and significant slope. This indicates that as the vocabulary of a story becomes less probable, i.e. the lexical complexity of that story goes up by our measure, characters tend to become more complex in terms of their mental depth, with other effects fixed at mean level. In addition, we observe in Table 2.2 a positive and significant simple effect of Age, which means that as children get older, the characters they use in their stories tend to get more complex in terms of mental depth, with other effects fixed at mean level. However, this effect is only a bit over half the magnitude of that of Vocabulary Probability ($\beta = .528$ versus $\beta = -1.117$).

We learn more about the relationship between Vocabulary Probability and Age by looking at the small non-significant interaction effect Vocabulary Probability * Age in Table 2.2. It indicates that the effect of vocabulary is not moderated by Age, in other words, is not significantly different for children of different ages. This is visible in Figure 2.2, where three lines indicate predictions of CD for various ages, but have similar slopes.

With respect to our second hypothesis, we observe in Table 2.2 that the simple

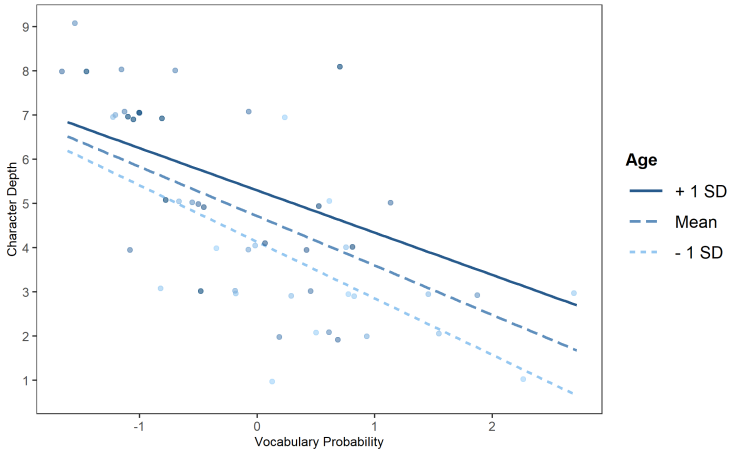


Figure 2.2: Interaction plot of Vocabulary Probability * Age with Age in months, average Vocabulary Probability (z-scored), and Character Depth in levels.

effect Dependency Distance and the interaction Dependency Distance * Age are both small and non-significant. Thus, contrary to our expectation, this model suggests that the distance between syntactically dependent words does not explain the observed variation in the levels of Character Depth, nor can we say that age plays a moderating role here. Finally, we can see in Table 2.2 that the main effect of Education Parents is positive and a bit smaller than age, but non-significant, suggesting that parental years of education do not reliably predict levels of characters' mental depth either.

Although we saw in the bivariate explorations that there is a moderate correlation between Vocabulary Probability and Age, (Pearson's r = -.403), we have no indications that these and other predictors pose multicollinearity issues for the estimates in our model, since all computed Variance Inflation Factors are below 1.44 (with a conservative threshold of 5). We thus find some tentative evidence for the idea that in our model, Vocabulary Probability and Age have independent effects.

2.5 Discussion

Our central finding is that lexical complexity is a key story-internal feature for predicting a story's socio-cognitive sophistication, as manifested in the mental depth of characters. This finding has multiple implications and possible interpretations. Firstly, it seems to follow that rich vocabularies are particularly helpful in organising and describing the storyworld, including its social and mental aspects. In theory, this

could be entirely independent of actual socio-cognitive skills possessed by the child telling the story: it could be merely a matter of being able or not to find the right words for fleshing out a character in terms of its emotional and intentional states.

However, with existing research in mind (e.g. De Villiers and De Villiers, 2014; Milligan et al., 2007) it appears more likely that our observed effect extends beyond the realm of the stories as such, and that possessing a more advanced vocabulary not only enhances a child's communication about the social world, but also supports its understanding of and ability to reason about socio-cognitive matters. Here it is particularly salient that the effect is larger than and not moderated by age. This adds a new perspective to the debate about the period in which children start to invoke others' mental states in their language (for an overview see Nicolopoulou, 2015).

Rather than disclosing a 'Rubicon' moment for ToM-language use in children, we propose a methodology that can show what it is about certain aspects of language development, such as having access to a more advanced lexicon, that engenders fleshing out mental activity in more detail, regardless of what age a child has. To substantiate such an interpretation, further research is needed focused on establishing firmer links between patterns observed inside stories and development as it takes place within the children that tell them. Here we see a role for collaborative work involving both (computational) linguists, narratologists and developmental psychologists.

For syntactic complexity the picture is quite different; we see no significant evidence for its contribution to Character Depth in our sample. Although in our bivariate explorations we saw a hint of the relation we hypothesised, in our model it was probably trumped by other effects. A reason for this could be that speech employs overall lower dependency distances compared to written text, which for children may even be stronger the case. If dependency effects are thus generally smaller, we must revisit this prediction with more data and maybe also compare and evaluate different metrics of syntactic complexity, such as clause length and words per finite verb.

A general remark about our methodology is that the use of computational language processing tools makes operationalising 'narrative sophistication', as we have done (and as is also proposed by Nicolopoulou (2016)), a lot easier, more reproducible, and more scalable. With larger datasets we might in the future be able to use story-internal variables to approximate children's narratological and linguistic capacities, as well as related cognitive skills, when no external information about the storytellers is available, or when collection and storage of sensitive data from children or parents is to be minimised.

In addition to (and to provide a more solid foundation for) such computational

approaches, we see multiple directions in which research may go that aims to deepen our understanding of the relationships between socio-cognitive development and narrative/linguistic competence. A possibility would be to include stories from a more diverse population, for example by involving atypically developing children, and/or collect additional data about each storyteller's performance on relevant standardised tasks (e.g. those used by Wellman and Liu (2004)). Another exciting possibility would be to compare our sample to story corpora in other languages, ideally differing substantially from Dutch in their syntactic and semantic structuring. Such extensions could help to further bootstrap patterns within the stories on trends in individual development, and shed light on directionality and causality of the interactions.

Finally, insufficient returned consent forms and other factors diminished the number of children per session we could include, which constrained this study to a fixed-effects model. Using more advanced random effect modelling we could most likely make better estimates of the relevant relationships, since such models would be able to take session-bound dependencies between for instance vocabularies into account. With this perspective in mind, we emphasise that a first improvement for our future research will be to focus on more participants per workshop session. Currently, the prospects for our story corpus are looking good: recent data collections in Spring 2021 yielded about 200 additional stories to be analysed. The goal for the rest of this year is to compile a corpus of at least 500 stories, consisting of around 8 hours of high-quality child speech recordings and 50000 tokens, that is open to researchers with all kinds of backgrounds and interests. A huge bonus so far is that children love our storytelling workshop, and are happy to see us come each time.

Chapter 3

ChiSCor: A Dutch Children's Story Corpus

In this resource chapter we release ChiSCor, a new corpus containing 619 fantasy stories, told freely by 442 Dutch children aged 4-12. ChiSCor was compiled for studying how children render character perspectives, and unravelling language and cognition in development, with computational tools. Unlike existing resources, ChiSCor's stories were produced in natural contexts, in line with recent calls for more ecologically valid datasets. ChiSCor hosts text, audio, and annotations for Character Depth and linguistic complexity. Additional metadata (e.g. education of caregivers) is available for one-third of the Dutch children. ChiSCor also includes a small set of 62 English stories. This chapter details how ChiSCor was compiled and shows its potential for future work with three brief case studies: i) we show that the syntactic complexity of stories is strikingly stable across children's ages; ii) we extend work on Zipfian distributions in free speech and show that ChiSCor obeys Zipf's law closely, reflecting its social context; iii) we show that even though ChiSCor is relatively small, the corpus is rich enough to train informative lemma vectors that allow us to analyse children's language use. We end with a reflection on the value of narrative datasets in Computational Linguistics.

This work was originally published as: Van Dijk, B.M.A.,* Van Duijn, M.J.,* Verberne, S., and Spruit, M.R. (2023). ChiSCor: A Corpus of Freely-Told Fantasy Stories by Dutch Children for Computational Linguistics and Cognitive Science. In Jiang, J., Reitter, D., and Deng, S., editors, *Proceedings of the 27th Conference on Computational Natural Language Learning*, pages 352-363. Association for Computational Linguistics. (* denotes equal contribution.)

3.1 Introduction

All of us tell stories on a daily basis: to share experiences, contextualise emotions, exchange jokes, and so on. There is a rich tradition of research into how such storytelling develops during infancy, and its relations with various aspects of children's linguistic and cognitive development (for an overview see Cremin et al., 2017). ChiSCor (Children's Story Corpus) was compiled to give a unique impulse to this tradition: it allows for (computationally) studying how children render character perspectives such as perceptions, emotions, and mental states throughout their cognitive and linguistic development.

Existing research connecting language and cognition has largely relied on standardised tests (for an overview see Milligan et al., 2007). Yet, recently researchers across fields have urged for data reflecting phenomena they study in their natural context. For instance, computational linguists call for better curated and more representative language datasets (Bender et al., 2021; Paullada et al., 2021), language pathologists question whether standardised linguistic tests capture children's actual linguistic skills (Ebert and Scott, 2014), and cognitive scientists call for more naturalistic measures of socio-cognitive competences (Beauchamp, 2017; Nicolopoulou and Ünlütabak, 2017; Rubio-Fernández, 2021). Following these considerations, ChiSCor has three key features: it contains fantasy stories that were told *freely*, within children's *social* classroom environments, and stories are supplemented with relevant *metadata*. As such, ChiSCor documents a low-resource language phenomenon, i.e. freely produced and socially embedded child language.

This chapter makes the following contributions. First, we release ChiSCor and describe its compilation, data, and annotations in detail (Section 3.2 and Section 3.3). Second, we show how ChiSCor fuels future work on the intersection of language, cognition, and computation, with three brief case studies (Section 3.4). We explore the Dependency Length Minimisation hypothesis (Futrell et al., 2015) with ChiSCor's language features and show that the syntactic complexity in children's stories is strikingly stable across different age groups. Also, we extend emerging work on Zipf's law in speech (e.g. Lavi-Rotbain and Arnon, 2023; Linders and Louwerse, 2023) and find that ChiSCor's token distribution approximates Zipf better than a reference corpus consisting of language written by children, which we explain by appealing to the Principle of Least Effort. Furthermore, we show that ChiSCor as a small corpus is rich enough to be used with Natural Language Processing (NLP) tools traditionally thought to require large datasets. We train informative lemma vectors with ChiSCor,

that can be used to analyse how coherently children use specific lemmas of interest, and potential bias in their language use.

Together, our case studies demonstrate that even though storytelling is a cognitively challenging task, the language children employ is no less sophisticated. And although corpora of narratives are often smaller, we show that we can (and should) leverage NLP-tools to unravel linguistic and cognitive mechanisms at work in children's language use. As discussed in Section 3.5, we see this as an important stepping stone towards building more ecologically valid language models.

3.2 Background

Various resources of Dutch child language exist. Before the 2000s, corpora typically consisted of child speech gathered in unstructured home settings involving smaller numbers of younger children (e.g. Schlichting, 1996; Wijnen and Verrips, 1998). Later, more structured language elicitation (e.g. with picture books) from larger samples of children was more common (e.g. Kuijper et al., 2015), and recently we have seen large corpora documenting thousands of essays in school settings (Tellings et al., 2018a), and many hours of speech recordings in human-machine interaction contexts (Cucchiarini and Van hamme, 2013).

Although these resources are valuable, what is currently lacking is a corpus of speech samples that are 1) produced freely in natural social settings, while being 2) sufficiently independent or 'decontextualised' to be a good reflection of children's capacities, and 3) accompanied by metadata about children's backgrounds. The rest of this section will discuss these three characteristics, on the basis of which ChiSCor was compiled.

1. The stories in ChiSCor were collected on a large scale in natural settings, because language as a social phenomenon is highly context-sensitive. The corpora mentioned above that include such settings are often limited in scale, whereas the newer corpora are large-scale, but cover language produced for a machine interface or in a school assignment context, thus are not socially embedded.

2. The stories in ChiSCor concern a special form of *decontextualised* language use, in which children cannot draw on cues (like picture books), feedback from interlocutors (as they could in a conversation), or much shared background knowledge with the audience (that hears a new fantasy story). Thus, the cognitive demands in producing decontextualised language are high, since children have

to simultaneously plan the story, monitor their language use, and make sure the audience can follow the plot (Nicolopoulou, 2019). As such, eliciting freely told narratives is an acknowledged method for sampling an individual child's language skills on phonological, lexical, syntactic, and pragmatic levels (Ebert and Scott, 2014; Nicolopoulou et al., 2015; Southwood and Russell, 2004), as well as for assessing cognitive abilities, including memorising, planning, organising world knowledge (McKeough and Genereux, 2003), and Theory of Mind (Nicolopoulou, 1993). Furthermore, proficiency in decontextualised language is known to be a good predictor of literacy and academic achievement (Snow and Dickinson, 1991). As far as we know, no larger-scale corpora of decontextualised Dutch child speech exist, and in the international context such corpora are also rare.

3. Existing resources often contain data on children's age and gender, but not on their backgrounds such as the educational levels of parents, which ChiSCor does contain (see Section 3.3). Metadata on subjects included in datasets becomes increasingly important, e.g. for gauging how representative language samples are (Bender et al., 2021), but also for follow-up work where e.g. partitioning the dataset is desired.

3.3 Methods

Story collection

We contacted primary schools, a day care and a community centre in the South and South-West of The Netherlands to offer storytelling workshops, in the period 2020-2023. Workshops generally consisted of three stages: first, we openly brainstormed with children about what stories are, without enforcing our ideas (e.g. what is a story, where can you find stories, what do you like about stories); second, we invited children to freely fill in the details of a fantasy story initiated by us as experimenters (e.g. filling in names, settings, events in a variation on the King Midas avarice myth); third and most importantly, we challenged children to individually make up and tell a fantasy story to their class peers, which we recorded.

Our storytelling workshop was inspired by the Story Telling Story Acting (STSA) paradigm, originally developed by Paley (1990) and used as a framework in empirical studies by Nicolopoulou and Richner (2007), and Nicolopoulou et al. (2015, 2022). Work by Nicolopoulou generally targets younger children using a longitudi-

Type	Quantity	Details
Audio	±11.5 hours	619 44.1kHz .wav files
Text	619 stories	±74k words, verbatim and normalised .txt files
Metadata	All 442 children	School age group
Extra metadata	147 children	Exact age, reading time, education caregivers, number of siblings, gender, language disorder (y/n), home language Dutch (y/n)
Linguistic features	All 619 stories	E.g. vocabulary perplexity, vocabulary diversity, syntactic tree depth, words before root verb, syntactic dependency distance
Annotations	All 619 stories	Character Depth (see Section 3.3)

Table 3.1: Details on ChiSCor's data. Besides the Dutch stories, ChiSCor also features an additional set of 62 English stories, for which audio, text, (extra) metadata, linguistic features and annotations are also available.

nal research practice integrated in the school curriculum, which involves both telling stories and acting them out. Our approach differed in that we included all primary school age groups (4-12y), but focused on storytelling only. Like in the STSA paradigm, children told stories live to an audience of peers, which comes close to narration in everyday social life: children explored themes like friendship and conflict, excitement over real and imagined events, and storytelling was interactive in the sense that their class peers reacted with laughter, disbelief, and so on.

High-quality recordings were made with a Zoom H5 recorder. Recordings were manually transcribed into verbatim and normalised versions. In the normalised stories employed in the case studies (Section 3.4), noise such as false starts and broken-off words was manually corrected with as little impact on semantics and syntax as possible. Our project was approved by the Leiden University Science Ethics Committee (ref. 2021-18). Caregivers were informed beforehand and could optionally provide additional metadata, which ±33% (147) did. Our corpus, metadata, and code are available at DANS.[1] See for more details on the data Table 3.1 and for sample stories Table 3.2.

Metadata

Here we highlight two variables from the metadata we collected: children's age and the educational levels of caregivers. Most ages are well-represented (Figure 3.1), but older children (ages 10-12) are underrepresented; fewer teachers from older age

[1] https://doi.org/10.17026/SS/TGPDJF.

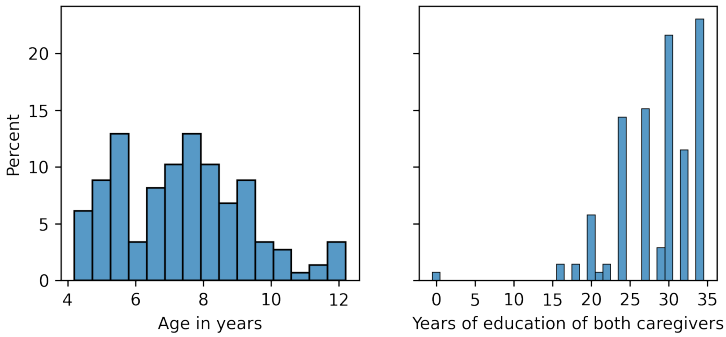


Figure 3.1: Ages of 147 children and educational levels of their caregivers. Bars in each plot stack up to 100%.

groups signed up for the workshop. For educational levels, we see that ±53% of the children has two highly educated caregivers (in the Dutch system, a higher degree equals a minimum of 15 years of education), while ±24% has caregivers with two vocational (or lower) degrees (a vocational degree equals a maximum of 12 years of education, see e.g. Van Elk et al. (2012)). Thus, in the part of our sample for which extra metadata is available, children from caregivers with higher socioeconomic status (SES) are overrepresented. Yet, selection bias is higher in the metadata than in the language samples in ChiSCor as a whole: while we were able to include stories told by children from schools in more challenged neighbourhoods in ChiSCor, metadata depended on caregivers filling out forms, which caregivers with higher SES did more often.

Annotations

Here we highlight two types of annotations available in ChiSCor: socio-cognitive annotations in the form of Character Depth (CD) annotations, and linguistic annotations in the form of automatically extracted features.

Regarding **social cognition**, ChiSCor provides Character Depth annotations that involve one label per story indicating the 'depth' of the most complex character encountered in a story (examples in Table 3.2). CD can be used as a window onto the socio-cognitive skills of storytellers and was adapted from Nicolopoulou and Richner (2007) and Nicolopoulou (2016). The scale ranges from 'flat' `Actors` merely undergoing or performing simple actions, to `Agents` having basic perceptive, emotional, and intentional capacities, possibly in response to their environments, to 'fully-blown'

Level	Example	ID
`Actor`	*Once upon a time there was a castle.* *There stood a throne in the castle and a princess sat on the throne.* *And the princess had a unicorn.*	093101
`Agent`	*Once upon a time there was a prince and he saw a villain.* *And then he called the police.* *And then the police came.* *And then he was caught. The end.*	023101
`Person`	*Once upon a time there was a girl.* *She really wanted to play outside. Her mother did not allow it.* *She went outside anyway and her mother asked where are you going?* *And the girl said I am going outside. The end.*	010101

Table 3.2: Translated stories from ChiSCor, traceable with ID. Underscoring shows the character the label is based on.

`Persons` with (complex) intentional states that are explicitly coordinated with the storyworld. Labelling was done with CATMA 6 (Horstmann, 2020) and in-text annotations are available on DANS. Labelling Character Depth requires expert annotation, given that children's stories often progress in non-obvious ways. Interrater agreement was obtained in two rounds. Two experts A and B first labelled a random subset of 8% of stories, yielding moderate agreement (Cohen's κ = .62). After calibration (discussing disagreements to consensus), A labelled the rest of the corpus, and B labelled another random 8% as a second check, for which Cohen's κ = .84 was obtained, indicating almost perfect agreement (Landis and Koch, 1977).

Regarding **linguistic features**, we extracted mean dependency distance between syntactic heads and dependents as measure of syntactic complexity with spaCy 3.5 (Honnibal and Johnson, 2015). We followed Liu (2008) and Liu et al. (2017) and calculated mean dependency distance with

$$DD(S) = \frac{1}{n-s} \sum_{i=1}^{n} |DD_i| \,, \tag{3.1}$$

where DD_i is the absolute distance in the number of words for the i-th dependency link, s the number of sentences and n the number of words in a story. Language employing larger dependency distances is more demanding for working memory, thus harder to process (Futrell et al., 2015; Grodner and Gibson, 2005). We further elaborate on dependency distance in a case study in Section 3.4.

We emphasise that many more linguistic features are included on DANS than we

can discuss here, e.g. lexical perplexity and syntactic tree depth as common measures of linguistic proficiency and development (see e.g. Kyle, 2016; McNamara et al., 2014).

3.4 Results

We conduct three small case studies to illustrate ChiSCor's potential. Since we aim to show ChiSCor's versatility to the broader research community, we draw in Study 1 on ChiSCor's own linguistic annotations and metadata; in Study 2 use ChiSCor in a corpus linguistics-style analysis on Zipf's law in child speech, and in Study 3 show the feasibility of using ChiSCor with NLP-tools that are traditionally thought to require larger train corpora.

Case study I: syntactic complexity

The Dependency Length Minimisation (DLM) hypothesis states that languages have evolved to keep syntactically related heads and dependents close together (such as an article modifying a noun), so that anticipation of a noun after an article is not stretched over many intervening words, which increases cognitive load and/or working memory costs (Futrell et al., 2015). Although DLM has been observed for various languages in various studies (e.g. Futrell et al., 2015; Gildea and Temperley, 2010), as far as we know, DLM for child speech has not been explored. ChiSCor concerns live storytelling, which is known to be a cognitively intense language phenomenon (see Section 3.2), which makes the DLM interesting to explore in ChiSCor's context. It is intuitive to expect that children employ smaller dependency distances to reduce cognitive load. We leverage ChiSCor's linguistic features (dependency distance as explained in Section 3.3) and metadata (age groups) to analyse the developmental trend under the DLM. Especially for younger children (e.g. 4-6y), DLM could be expected to be more pronounced, given that they are arguably less proficient language users with little formal language training in school.

Our modelling approach was as follows. In a linear model we included contrast-coded predictors such that each predictor indicated the mean dependency distance difference with the previous grade (i.e. backwards difference coding), to model a trend over age groups. Dependency distance conditioned on age is plotted in Figure 3.2 for 442 stories of 442 children, and coefficients of the model are given in Table 3.3. Note that for those children who told multiple stories, we included only the first story to maximise independence of observations.

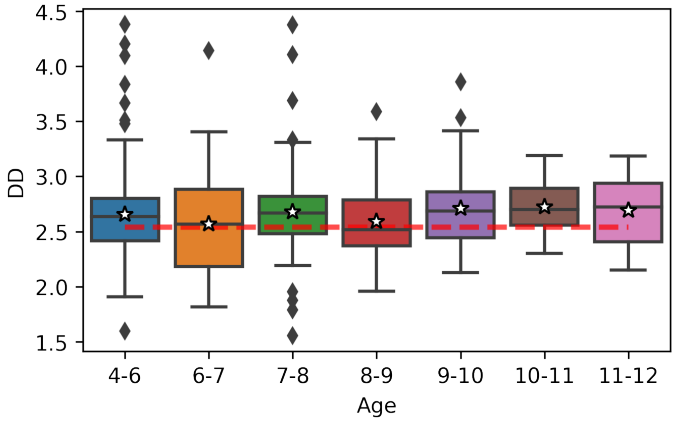


Figure 3.2: Dependency distance conditioned on age groups (in years) as customary in Dutch primary education. Dashed line indicates mean dependency distance reported by Liu (2008). Stars indicate means.

Predictor	β	*SE*	p
Intercept	2.66	.02	.00
Diff. 6-7/4-6	-.09	.07	.20
Diff. 7-8/6-7	.11	.07	.13
Diff. 8-9/7-8	-.09	.06	.16
Diff. 9-10/8-9	.12	.07	.08
Diff. 10-11/9-10	.01	.10	.91
Diff. 11-12/10-11	-.03	.12	.81

Table 3.3: Coefficients of the linear model. Each predictor indicates the difference in dependency distance with the previous age group.

Dependency distance appeared to be surprisingly stable across age groups: no single predictor significantly predicted dependency distance (Table 3.3, all $p > .05$), nor did all predictors together ($F_{6,435} = 1.078, p = .38, R^2_{adj} < .01$). Contrary to expectations, it was not the case that younger children, as less proficient language users, employ shorter dependency distances, nor do children employ significantly longer dependency distances as they grow older. Interestingly, in backwards difference coding, the intercept is the grand mean of dependency distance of all groups (2.66), which is close to the mean dependency distance of 2.52 found for Dutch written by adults and reported by Liu (2008).

We make a start with trying to explain why in storytelling for younger children (4-6y), we find higher dependency distances than expected. Manual examination of narratives from this group showed that children often use syntactically complex

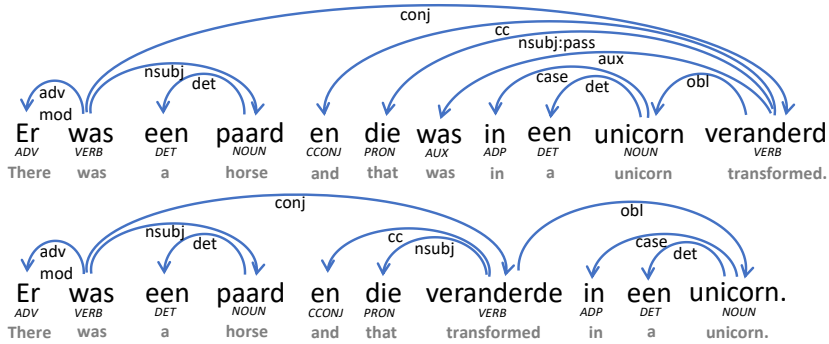


Figure 3.3: Top: original utterance from story 033201 in PaP with a mean dependency distance of 3.2. Bottom: paraphrase in SP (bottom) with a mean dependency distance of 2. The dependency labels shown were not used in this case study, but are based on Universal Dependencies as developed by Nivre et al. (2017).

constructions to refer to past events, even when simpler alternatives are available or preferred. The typical tense for narrative contexts is the Simple Past (SP) for many languages (Zeman, 2016), and SP can be used for completed and ongoing events in the past (Boogaart, 1999) in the story world. SP is syntactically simple; it requires only a single inflected verb. Young children, however, often use Present/Past Perfect (PrP/PaP) and Past Progressive (PP) constructions. These forms are used to indicate ongoing (PrP/PP) and completed (PaP) events in the past, and are syntactically similar in that they all involve an auxiliary depending on a (past) participle (PrP/PaP) or infinitive (PP) that is typically at utterance-final position, thus creating complex syntax. Figure 3.3 provides an illustration from ChiSCor of a child narrating a completed past event in PaP, which pushes dependency distance well beyond the average reported by Liu (2008), although the more efficient option would be SP.

Although it is known that young children in experimental contexts also refer to past events with PrP and PP constructions instead of SP (Schaerlaekens and Gillis, 1993; Van Koert et al., 2010), in the context of decontextualised language use and the DLM our finding was unexpected. We find a possible explanation in the work by Van Koert et al. (2010): separating tense (auxiliary) from lexical information (verb) yields more complex syntax on the one hand, but makes processing easier for an audience on the other hand. After all, the audience does not have to decode different types of information packed in a single inflected verb. The trade-off between syntactic simplicity and ease of processing could indeed explain why ChiSCor's spoken narratives, produced live in front of an audience of peers, contain relatively high pro-

portions of PrP and PP. Follow-up work would be needed to further substantiate this idea.

Case study II: Zipf distributions

Zipf distributions, where token frequencies are proportional to their rank r according to

$$f(r) \propto \frac{1}{r^{\alpha}}, \tag{3.2}$$

with $\alpha = 1$ (Zipf, 1932) were found for many language samples (Ferrer i Cancho, 2005; Lavi-Rotbain and Arnon, 2023; Smith, 2007; Tellings et al., 2014; Xiao, 2008; Yu et al., 2018), but are also subject to debate (for a review see Piantadosi, 2014); is Zipf a trivial mathematical artefact or a fundamental property of human cognition and language? As Linders and Louwerse (2023) note, to answer this question we should analyse Zipf in more natural forms of communication, such as speech instead of written language, and invoke cognitive mechanisms underlying Zipf, such as the Principle of Least Effort (PLE). The PLE assumes that senders prefer efficient communication using infrequent, hence often shorter and ambiguous words, whereas receivers prefer larger vocabularies of longer, infrequent words to more easily decode messages. Zipf distributions are considered the balanced trade-off between sender and receiver needs (Cancho and Solé, 2003).

The PLE is salient in ChiSCor's context: since live storytelling is a cognitively intense form of decontextualised language use (Section 3.2), this could lead to a bias in storytellers towards frequent tokens to alleviate cognitive load, a prediction made by Linders and Louwerse (2023). Yet, at the same time, if receiver needs are neglected, they cannot follow along; receivers cannot ask for clarification during storytelling as would be possible in e.g. normal conversations, which is something senders take into account to prevent losing their audience, which equals losing the point of storytelling. This balance is arguably less pronounced in written discourse, where there is opportunity to reconsider earlier parts, and no immediate interaction, thus less pressing receiver needs. Here we pit the token distribution of ChiSCor against that of BasiScript, a corpus of written child language (subsection 'free essays', $\pm$3.7M tokens from thousands of Dutch children aged 7-12y (Tellings et al., 2018a)), to compare Zipfian distributions in speech to the written domain.

We followed Piantadosi (2014) in performing a binomial split on the observed frequency of each token to avoid estimating frequency and rank on the same sample. We used Zipf's original formula introduced above rather than derivations to model to-

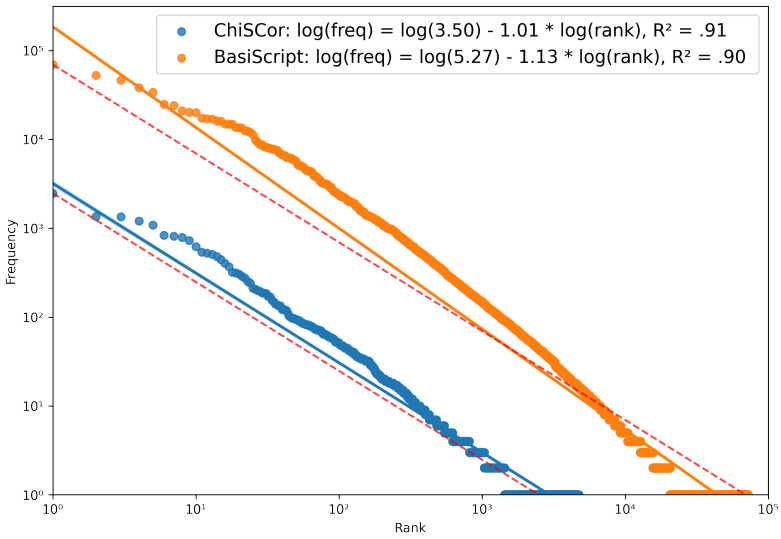


Figure 3.4: Rank-frequency plots of ChiSCor and BasiScript. Dashed red lines indicate Zipf's law with $\alpha = 1$, solid blue and orange lines indicate model fits.

ken distributions, following Linders and Louwerse (2023). We log-transformed (base 10) token rank and frequency to model Zipf linearly with

$$log(frequency) = log(intercept) + slope * log(rank). \tag{3.3}$$

We see in Figure 3.4 that both corpora approximate the plotted Zipf lines with good model fits ($R^2 \geq .90$). Yet, ChiSCor approximates the Zipf line more closely than BasiScript, with a slope closer to -1, supporting the idea that in live storytelling balancing sender *and* receiver needs is more pressing than in written language, even though in live storytelling a bias towards frequent tokens seems intuitive.

The larger negative slope (-1.13) fitted for BasiScript indicates that senders rely more on frequent tokens and employ less infrequent tokens, which confirms the prediction that in written discourse, receiver needs are less pressing. Senders apparently prefer a smaller vocabulary of more frequent hence ambiguous tokens, contra receivers who prefer a vocabulary comprising more infrequent terms that are easier to decode. Follow-up work could investigate Zipf distributions in both corpora beyond tokens, e.g. on parts-of-speech or utterance segments (Lavi-Rotbain and Arnon, 2023; Linders and Louwerse, 2023).

Case study III: lexical semantics with Word2Vec

The third case study demonstrates the usability of ChiSCor as a relatively small corpus with common NLP-tools. We use a Word2Vec model (Mikolov et al., 2013) to visualize lexico-semantic differences in children's language use in ChiSCor and BasiScript. It is commonly assumed that training high-quality word vectors requires large corpora ($>$ 100 million tokens) (Altszyler et al., 2017; Mikolov et al., 2013); ChiSCor and BasiScript are much smaller with $\pm$74k and $\pm$3.7M tokens respectively. Still, it is worthwhile to see how well ChiSCor allows a computer to infer lexico-semantic information, since vector representations are the starting point for many downstream NLP tasks, and studies in computational and cognitive linguistics (e.g. Beekhuizen et al., 2021; Samir et al., 2021).

We obtained lemma vectors from both ChiSCor and BasiScript with Word2Vec as implemented in Gensim 4.1.2 (Řehůřek and Sojka, 2010). For ChiSCor, the CBOW algorithm yielded the best result, for BasiScript this was Skip-gram. Vector quality was evaluated visually during training with reduced-dimensionality plots of a set of 35 common nouns, verbs, connectives, etc. that occur proportionally in both corpora. The results are given in Figure 3.5. Here we see that overall vectors from both corpora allow intuitive syntactic groupings (e.g. conjunctions *but*/*because* and verbs *to think*/*to know*) and semantic groupings (e.g. *mommy*/*daddy*, *not*/*none*). To verify this quantitatively, we computed cosine similarities between the 595 possible pairs of the 35 lemmas plotted in Figure 3.5 with

$$\cos(\mathbf{v}, \mathbf{w}) = \frac{\mathbf{v} \cdot \mathbf{w}}{\|\mathbf{v}\|\|\mathbf{w}\|}, \tag{3.4}$$

where $\mathbf{v}$ and $\mathbf{w}$ are two vectors representing lemmas from one corpus, and computed the overlap between the corpora. We found a fair correlation $\rho(595) = .45, p < .01$ (Akoglu, 2018), which is salient: it shows that from ChiSCor as relatively small corpus, rich lexico-semantic information can be learned as effectively as from BasiScript, which is 46 times larger.

Lemma vectors also allow us to analyse how children use particular lemmas of interest. There is some nuance in the groupings in Figure 3.5: for ChiSCor in Figure 3.5a, especially the verbs referring to cognitive states (*to think*, *to know*, *to wish*, *to want*) and perceptual states (*to hear*, *to see*) are more clearly grouped and positioned compared to BasiScript in Figure 3.5b (where e.g. *to wish*, *to see*, and *to want* have less obvious positions). Since these lemmas have about equal relative frequencies in both corpora, it is likely that for these verbs, the lemma *context* is semantically more

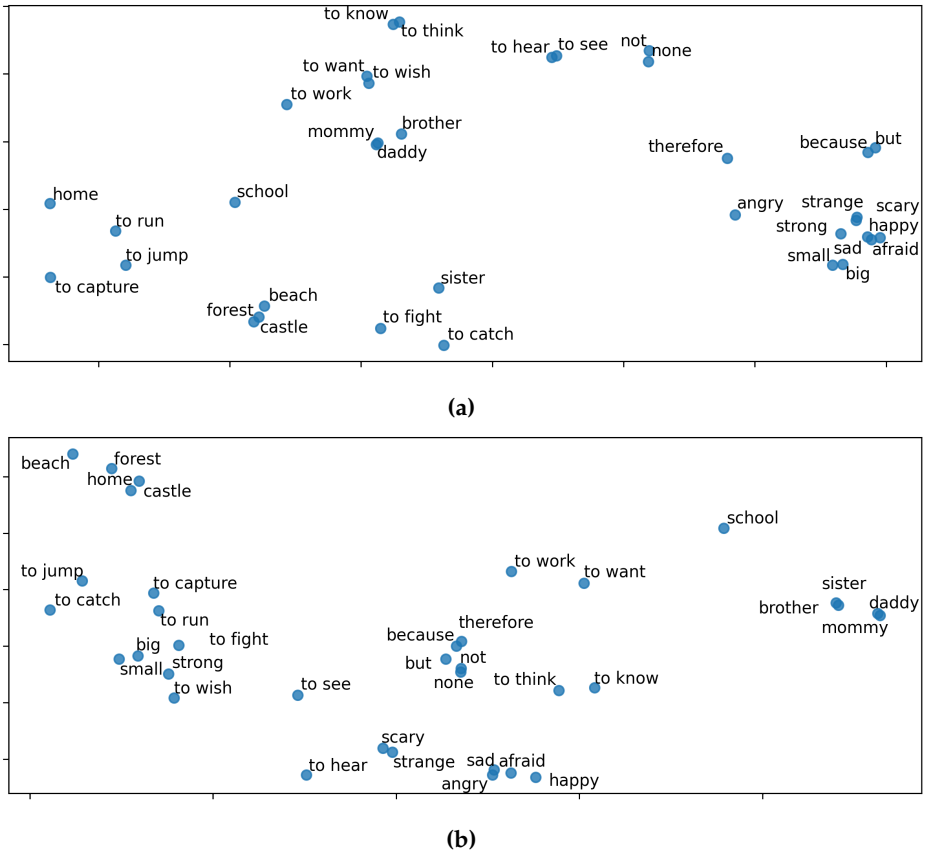


Figure 3.5: t-SNE projections (Maaten and Hinton, 2008) of the latent spaces of 100-dimensional lemma vectors of ChiSCor (a) and BasiScript (b). Lemma positions should not be compared between but within plots, as the axes of the plots have no explicit interpretation.

clear and coherent in ChiSCor compared to BasiScript. On the other hand, conjunctions (*but*, *because*, *therefore*) are more coherently grouped in BasiScript compared to ChiSCor (where *therefore* has a less obvious position).

Apparently, children use verbs referring to cognitive/perceptual states more coherently in ChiSCor, while conjunctions are more coherently used in BasiScript. In live storytelling, communicating clearly and coherently what was thought and/or perceived seems more critical than in written storytelling, as the audience cannot access earlier information as they could in a written story, and this information is critical for understanding and relating to narratives more generally (Zunshine, 2006). On the other hand, in written stories, children have more time to reflect on, and, if

necessary, correct their use of conjunctions to link clauses, making the context more clear and coherent. This example shows that ChiSCor is usable with common NLP tools to unravel children's language use in detail, even though it is relatively small.

Lemma vectors can also reveal bias in children's speech. A well-known gender bias in language is the woman-home/man-work stereotype (Bolukbasi et al., 2016; Wevers, 2019), which in ChiSCor and BasiScript can be investigated with gendered categories *mommy*, and *daddy*, and attributes *home* and *to work*. As we see in Figure 3.5, *mommy* and *daddy* occupy similar positions, so initially we do not expect much difference in their cosine similarity with *home* and *to work*. A standard approach to verify this, is to compute the difference in cosine similarity of an attribute with one category versus another, e.g. *home* and *mommy* vs. *daddy*. For ChiSCor, difference scores were small: for *home* and *mommy* vs. *daddy* .031, for *to work* and *mommy* vs. *daddy* .076. The difference scores were comparably small for BasiScript: .049 and .001 respectively. These smaller scores indicate that neither gender is more strongly associated with one attribute than the other, suggesting little gender bias in the corpora, contra earlier work on bias in child language (e.g. Charlesworth et al., 2021). Still, future work should leverage ChiSCor and incorporate more gendered categories (e.g. *she*, *he*), more attributes (e.g. *baby*, *office*), average these vectors and apply more advanced vector arithmetic to put this initially surprising result to the test.

3.5 Discussion

Storytelling datasets are scarce, which is a shortcoming in existing resources, given that live storytelling challenges children to leverage linguistic, cognitive, and social competences to tell a story that engages an audience. These competences can be analysed through stories, manually or with computational tools, to learn more about child development. We demonstrated that ChiSCor has properties that other established language samples also have, such as a Zipfian token distribution. Moreover, ChiSCor's close fit to the Zipfian curve testifies to the *social context* of the language contained in it and the Principle of Least Effort that is likely at work there.

In addition, even though storytelling is a cognitively demanding task, we demonstrated that the stories in ChiSCor are syntactically surprisingly complex, and we offered a tentative explanation why especially younger children may employ complex syntax, which could be related to ChiSCor's context of live storytelling in front of an audience. Lastly, we have shown that ChiSCor can be used to learn a semantic vector space that is as intuitive as the semantic space of a much larger reference corpus. This

opens up possibilities for using ChiSCor with tools that are traditionally deemed fit only for much larger corpora, to assess the coherence of contexts in which children use particular words of interest. For example, we found that words detailing cognitive and perceptual states were more clearly differentiated in ChiSCor compared to BasiScript as corpus of written child language. Such words concern information that is critical to understand a plot that cannot be consulted again in live storytelling, possibly leading children to use these words more carefully and coherently.

The social context of ChiSCor's narratives and its influence on language production invite us to reflect on a more general issue: the dominance of written (web) text in Computational Linguistics and NLP. Researchers increasingly question scraping together ever larger, uncurated and undocumented resources (Bender et al., 2021; Paullada et al., 2021), that is, datasets without metadata, and it is subject to debate how helpful such large-scale written datasets are in understanding language acquisition and modelling cognition (e.g. Mahowald et al., 2024; Warstadt and Bowman, 2022). Indeed, spoken language is different from written language in many ways, as Linders and Louwerse (2023) note: it is mainly acquired naturally (unlike writing) and predates writing in both the evolutionary and developmental sense. Most critically, speech is typically situated in a social setting with other language users, evanescent, spontaneous, and grounded in a particular context, to mention just a few out of many defining characteristics.

Still, with Large Language Models (LLMs) as prime current example of the reliance on large written datasets, such models have helped disclose what is *in principle* learnable from word co-occurrence statistics and a simple word prediction training objective, such as the capacity to represent language input hierarchically (Manning et al., 2020). Although we should take LLMs serious as the current best yet data-hungry distributional learners we have (Contreras Kallens et al., 2023), the next challenge is to achieve the same performance with more ecologically valid, smaller datasets and smaller neural architectures; here, corpora like ChiSCor could be part of the solution. Since ChiSCor has information on the age groups of the children who produced the language, future work could, for example, partition ChiSCor to employ train and/or test sets that more realistically model children's language use at different stages of their development. And since ChiSCor covers language from the speech domain, it provides an interesting opportunity to explore training language models on language with a different nature. Still, we do not mean to claim that ChiSCor solves all issues regarding LLMs and training data, but we hope to contribute a dataset that can be a part of the move towards better datasets for Computational

Linguistics, a dataset that, in the words of Bender et al. (2021), 'is only as large as can be sufficiently documented'.

Lastly, we like to emphasise that since ChiSCor features high-quality audio besides text, it naturally opens directions for multi-modal research. For example, research on detecting characters' emotions will benefit from adding information on prosody. Also, research aimed at improving speech-to-text models will benefit from the voices of 442 unique children of different ages, and accompanying transcripts, that can be used for fine-tuning existing speech-to-text models.

3.6 Conclusion

This chapter introduced ChiSCor as a versatile resource for computational work on the intersection of child language and cognition. ChiSCor is a new corpus of Dutch fantasy stories told freely by children aged 4-12 years, containing high-quality language samples that reflect the social settings in which they were recorded in many details. We provided three case studies as examples of how ChiSCor can fuel future work: studying language development with ChiSCor's out-of-the-box age metadata and linguistic features, modelling Zipf distributions with ChiSCor, and linking ChiSCor to common NLP-tools to study children's language use in action. Besides verbatim and normalised texts, ChiSCor comes with 619 high-quality audio samples of 442 children, metadata on the backgrounds of 147 children, annotations of Character Depth, and extracted linguistic features that will be useful for a variety of researchers. In addition to Dutch stories, ChiSCor comes with a small additional set of 62 English stories with the same additional metadata and annotations as the Dutch stories.

Four years have passed since we started compiling ChiSCor. We look back on many great moments with the children who were happy to share their fantasies and cleverly constructed plots with us. We encourage readers of this chapter to have a look at the corpus — both for research purposes and for fun.

3.7 Limitations

Within the subset of our corpus that contains extra metadata (Section 3.3) older children and children from lower socioeconomic backgrounds are underrepresented. This may limit the generalisability of future work done with ChiSCor. This is partly due to a bias resulting from the way our metadata was obtained; the larger set of

619 stories is likely more balanced. A second limitation concerns Character Depth annotations: a large part of Character Depth labels depends on one expert. A third limitation is that for BasiScript, a license has to be signed before one can use it. Thus, we cannot provide its lexicon or the corpus on DANS, which makes parts of our study less directly reproducible.

3.8 Ethics Statement

In compiling this corpus, the researchers were frequently in touch with school principals, teachers, children and parents to find an appropriate way to collect, store and analyse the stories and metadata. Our study was reviewed and approved by the Leiden University Science Ethics Committee (ref. 2021-18). Regarding model efficiency, the spaCy models used to extract linguistic information are pre-trained, easy to use, and extraction of lexical and syntactic information did not take more than a couple of minutes. Further, the Gensim models used to train word vectors are also lightweight, easy to use, and equally efficient qua training time.

Chapter 4

Classifying Theory of Mind in Freely Told Stories

Children are the focal point for studying the link between Theory of Mind (ToM) and language competence. ToM and language are often studied with younger children and standardised tests, but as both are social competences, data and methods with higher ecological validity are critical. We leverage a corpus of 442 freely told stories by Dutch children aged 4-12y, recorded in their everyday classroom environments, to study ToM and language with Natural Language Processing tools. We labelled stories according to the mental depth of story characters children create (Character Depth), as a proxy for their ToM competence 'in action', and built a classifier with features encoding linguistic competences identified in existing work as predictive of ToM. We obtain good and fairly robust results (F1-macro = .71), relative to the complexity of the task for humans. Our results are explainable in that we link specific linguistic features such as lexical complexity and sentential complementation, that are relatively independent of children's age, to higher levels of Character Depth. This confirms and extends earlier work, as our study includes older children and socially embedded data from a different domain. Overall, our results support the idea that language and ToM are strongly interlinked, and that in narratives the former can scaffold the latter.

This work was originally published as: Van Dijk, B.M.A., Spruit, M.R., and Van Duijn, M.J. (2023). Theory of Mind in Freely-Told Children's Narratives: A Classification Approach. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics*, pages 12979-12993. Association for Computational Linguistics.

4.1 Introduction

One key reason language is critical to us humans is that it allows us to communicate and manipulate others' mental states (Clark, 1996; Dor, 2015). Anticipating what others feel, believe, and intend, is key to navigating the social world and having meaningful interactions, and language evolved as an essential tool to achieve that (see e.g. Tomasello, 2003, 2014; Verhagen, 2005). Thus, there is a strong link between language competence on the one hand, and the competence to reason about and understand others' mental states on the other; the latter is known as Theory of Mind (ToM) (Apperly, 2012; Baron-Cohen, 2001).

There is a long tradition of research in child development to understand how emerging competence in language and ToM interact, typically with standardised tests, carried out in lab settings with younger children, often below age 7 (for reviews see Beaudoin et al., 2020; Milligan et al., 2007). Yet, researchers in child development and cognition call for more ecologically valid data to study language and ToM as social phenomena; ToM displayed in experimental settings may look different from ToM used in navigating the real social world and daily activities such as pretend play and storytelling (Beauchamp, 2017; Beaudoin et al., 2020; Nicolopoulou and Ünlütabak, 2017; Rączaszek-Leonardi et al., 2018; Rubio-Fernández et al., 2019; Rubio-Fernández, 2021). In addition, especially for ToM, researchers call to also include older subjects (Apperly et al., 2009) and methods that capture a wider variety of ToM skills (Ensink and Mayes, 2010).

We argue that children's stories are a natural choice to study language and ToM competence in a social context. In narrating, children draw on various linguistic skills in producing a story, for example, structuring clauses with temporal and causal connectives (Nicolopoulou, 2016). Furthermore, narratives are typically rich in the feelings, beliefs and intentions of story characters, that resonate well with our own (Zunshine, 2006), thus inviting children to leverage their ToM skills in rendering these character minds. We employ 442 freely told narratives by 442 Dutch children aged 4-12 in a classification task, that we approach with features encoding the linguistic skills identified as predictive for ToM performance in earlier empirical work. Doing so, we evaluate and extend existing work on the links between language and ToM in a natural social context and for a larger age range.

We employ an adapted version of Character Depth (CD), originating from Nicolopoulou and Richner (2007), as window onto children's ToM competence. For labelling, CD indicates the mental complexity of characters, from flat characters with-

out inner lives, to characters with basic intentionality, actions and emotions, to fully-blown characters with (complex) desires, beliefs, and intentions. Our approach meets the 'intensional requirement' of any Natural Language Processing (NLP) task defined by Schlangen (2021), which is having *a theory* on the relation between input (story) and output (CD label), next to the extensional requirement, which is simply the set of stories and labels. If the aim is to model humans' cognitive abilities with NLP-tools, then drawing on established work in other fields for meeting the intensional requirement is key.

The work in Chapter 2 has suggested that linguistic features (e.g. vocabulary complexity) play a key role, besides age, in predicting ToM in natural language data, but was limited in scale; here we approach language and ToM in narratives at scale from a NLP perspective. Our logistic regression classifier performs well (F1-macro = .71) drawing on purely linguistic features that are relatively independent of children's age. We are able to link specific features to specific CD levels: stories employing higher CD also employ, for example, more pragmatic markers, more complex words, and more sentential complementation. Our results support the idea that language and ToM are intertwined, and that language can scaffold children's reasoning about the social world.

This chapter proceeds as follows. In Section 4.2 we reflect on relevant work. In Section 4.3 we elaborate on our dataset, labelling, feature engineering and classifier setup. We present results in Section 4.4, and contextualise them in Section 4.5.

4.2 Background

Few have used NLP tools on child language to study ToM, but Kovatchev et al. (2020) pioneered classifying children's ToM competence on two standardised ToM tests, the Strange Stories Task (Happé, 1994) and Silent Film Task (Devine and Hughes, 2013). In such tests, children are typically presented a vignette containing a social situation (verbally and/or visually) and are asked to explain why a character is behaving in a certain way (e.g. being ironic), thus inviting children to refer to characters' mental states. Kovatchev et al. (2020) labelled $\pm$11k answers on questions as either incorrect, partially correct, or correct, depending on how appropriately children referred to characters' mental states, and obtained good performance (F1-macro = .91) with a DistilBERT Transformer. Indeed, accurate automatic scoring is valuable for processing standardised ToM tests. It can reduce the need for resource-intensive human evaluation of answers at larger scale (for example, Kovatchev et al. (2020) processed tests

conducted with ±1k children), and explaining how models learn to identify correct answers can further our understanding of the relation between ToM and language.

Kovatchev et al. (2020) however did not focus on the language children use to reason about ToM, although their error analysis suggests that this is worthwhile to do. One source of confusion identified for DistilBERT, is that children's answers sometimes explicate what characters would say or think. This evidences a child shifting to a different *perspective*, which is a precursor to ToM competence (De Mulder, 2011; Rubio-Fernández, 2021). A syntactic device to achieve such shifts is sentential complementation: *Character X thinks/sees/said that it is raining*, and its mastery predicts children's understanding of false beliefs (De Villiers, 2005, 2007; Lohmann and Tomasello, 2003).

Yet, since it is debated whether the role of sentential complementation holds beyond the false-belief context (De Mulder, 2011; Slade and Ruffman, 2005), it would be interesting to see whether complementation can be linked to ToM in children's *natural language productions* where reasoning about characters' mental states is natural, like narratives. As shown in the example above, complementation does not exclusively scaffold reasoning about mental states, but also communication and perception, which arguably provide less direct access to mental states (see Chapter 5). With modern NLP-tools, complementation in natural language can be efficiently extracted and linked to children's ToM performance, as Rabkina et al. (2019) have demonstrated, and we argue that this is also worthwhile for other linguistic competences.

In our view, narratives are natural devices to study language and ToM. In children's narratives, increasingly complex ways to represent characters' inner states can be found (Nicolopoulou and Richner, 2007), which is why we look beyond standardised tests, and draw on a Character Depth typology established in developmental work for labelling stories (see Section 4.3). Narrative elicitation is an established way of sampling children's language skills at lexical, syntactic, phonological and pragmatic levels (Ebert and Scott, 2014; Nicolopoulou et al., 2015; Southwood and Russell, 2004), but also for examining cognitive abilities, including memorising, planning, organising world knowledge (McKeough and Genereux, 2003), and ToM (Nicolopoulou, 1993). The narratives central in this chapter result from children's free storytelling for a live audience of peers (see Section 4.3), which yields a window on children's language and ToM competence that is more ecologically valid.

Like Kovatchev et al. (2020), we classify child language, though not test answers but a smaller set of narratives, that are linguistically speaking likely more varied. We rely on logistic regression and custom features that encode earlier findings on lan-

guage competence and ToM to obtain explainable performance. With Shapley values we compute feature importance in the game-theoretic fashion defined by Lundberg and Lee (2017). Shapley values encode the contribution a specific feature makes to a model's prediction. If a model is a function $v(x)$ that consists of a 'team' of N features $\{1, 2, ...n\}$, then $S \subseteq N$ denotes a possible subset of features. The marginal contribution of feature f is the difference between the model's output on a given input with f included i.e. $v(S \cup \{f\})$, and $v(S)$, where f is not included. The average marginal contribution (Shapley value) is this difference computed over all possible subsets of features without f, i.e. $S \subseteq N \setminus \{f\}$:

$$\varphi(f) = \frac{1}{N} \sum_{S \subseteq N \setminus \{f\}} \binom{n-1}{|S|}^{-1} v(S \cup \{f\}) - v(S). \tag{4.1}$$

Shapley values are calculated for each feature and each class in multiclass classification, and are additive, that is, they sum up to the difference between the expected value and the model prediction with all features present.

4.3 Methods

Dataset

We collected 442 stories at various Dutch primary schools, a day care, and a community centre, from 442 children aged 4-12y. Story collection was embedded in a workshop, which consisted of three stages. In the first stage, we brainstormed about stories openly with the children without providing our own opinions, for example on what stories are, where you can find stories, what is engaging about stories, etc., to introduce the theme. In the second stage, children were free to draw on their imagination to fill in the details of a fantasy story told by the experimenter. For the group until age 10-11, this was a variation on the King Midas avarice myth, and details children could fill in were e.g. about where the king lives, what his possessions were, what things he turned into gold, etc. Older children had a different story template but the same approach. This second stage served as preparation for the final and for this study critical stage, where children were invited to individually make up and tell their own fantasy stories to their class peers.

Our workshop was inspired by the Story Telling Story Acting (STSA) practice, originally developed by Paley (1990) and further employed in empirical studies by Nicolopoulou et al. (2015); Nicolopoulou and Richner (2007); Nicolopoulou et al.

(2022). The storytelling children do in this paradigm is thoroughly social: they speak live to an audience of peers, that can provide feedback in the form of expressions of disbelief, laughter, etc., and children's storytelling explores common themes like friendship, conflict, and so on.

The stories were recorded with a Zoom H5 recorder. Our project was approved by the Leiden University Science Ethics Committee (ref. no. 2021-18), and parents were informed before classroom visits. Recordings were manually transcribed into verbatim and normalised versions. In the normalised stories central in this chapter, false starts, broken-off words, wrong verb conjugations and other errors were corrected with minimal impact on semantics and syntax. With regard to story lengths in words, there is positive skew ($\bar{x} = 128, \sigma = 176.40, Q_1 = 40, Mdn = 87, Q_3 = 164$); in longer stories linguistic properties are likely more reliably estimated. Our data, annotations and code are available on OSF.[1]

Labelling

Nicolopoulou and Richner (2007) and Nicolopoulou (2016) were among the first to study CD in children's freely told narratives. The idea, also employed in Chapter 2, is that CD is a window onto children's ToM competence. For example, if a child adequately constructs a story character that tries to convince another character to go ice skating, then it is safely assumed that it can coordinate multiple mental states (two desires). However, this does not necessarily give a complete view on an individual child's ToM competence; a narrative with only flat characters, may or may not imply a narrator with lower ToM competence. Here we rather disclose the linguistic contexts tied to ToM competence given by different CD levels, thus ToM 'in action'. In a similar vein, stories do not necessarily yield a full view on individual children's linguistic competence.[2] We employ an adapted version of the three-level character typology developed by Nicolopoulou and Richner (2007):

- **Actors** are non-psychological characters, often physically described. They lack clear intentionality and goal-directedness. They typically don't act but are acted upon. If they act it is without clear intention or goal;

[1] `https://osf.io/2es6w/`.

[2] We note that similar issues regarding the validity of standardised ToM tests come currently increasingly to the fore; they may be confounded by lower-level skills (e.g. emotion recognition), or the third-person perspective in which vignettes are presented (Quesque and Rossetti, 2020), or even by superficial aspects such as familiarity with the test materials, the use of real humans or figurines in testing, and phrasing differences in the test questions (Beaudoin et al., 2020).

Level	Example	ID
`Actor`	*Once upon a time there was a castle.* *There stood a throne in the castle and a princess sat on the throne.* *And the princess had a unicorn.*	093101
`Agent`	*Once upon a time there was a prince and he saw a villain.* *And then he called the police.* *And then the police came.* *And then he was caught. The end.*	023101
`Person`	*Once upon a time there was a girl.* *She really wanted to play outside. Her mother did not allow it.* *She went outside anyway and her mother asked where are you going?* *And the girl said I am going outside. The end.*	010101

Table 4.1: Translated stories from ChiSCor, traceable with ID. Underscoring shows the character the label is based on.

- **Agents** exhibit implicit intentions-in-action, emotions and perceptions. Agents' actions are goal-directed and they can respond to events in the story world verbally or with actions and emotions;
- **Persons** display explicit mental states and intentional reasoning: they want, believe, and intend things, in relation to events in the story world, other characters' mental states, or their own (future/past) mental states.

Following work in developmental psychology we give one CD label per story, indicating the 'deepest' level achieved by any character in the story (Nicolopoulou and Richner, 2007; Nicolopoulou, 2016). Labelling CD is a form of expert annotation, as children's story plots are not always obvious. To establish interrater agreement we proceeded as follows. First, two experts A and B labelled a random subset of 8% of stories, resulting in moderate agreement (Cohen's κ = .62). After discussing disagreements to consensus (i.e. calibration), A labelled the rest of the corpus, and as second verification, B labelled another random 8%, for which Cohen's κ = .84 was obtained, which indicates almost perfect agreement (Landis and Koch, 1977). See Table 4.1 for examples of CD levels and Table 4.2 for level distribution; `Actor` stories are underrepresented, which challenges inducing characteristics of this level. As we are dealing with pure language samples of children, we considered oversampling or data augmentation not appropriate.

Nicolopoulou and Richner (2007) showed CD development over age: as young children (4-6y) grow older they tell relatively more `Person` and less `Actor` stories. For older children this has not been explored, but we can see in Figure 4.1 that in our

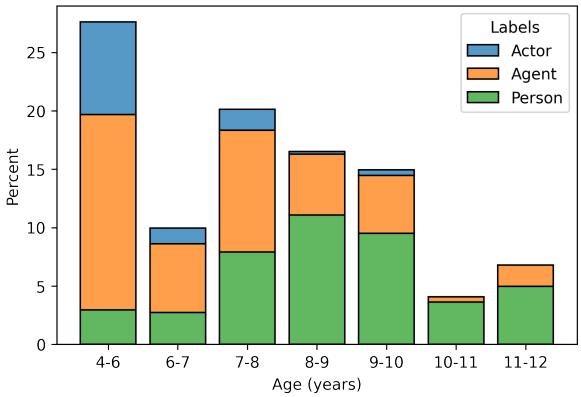


Figure 4.1: Character Depth levels by the age groups standard in Dutch primary education. Bars stack to 100%.

Actor	Agent	Person	Total
52 (12%)	201 (45%)	189 (43%)	442 (100%)

Table 4.2: Character Depth label distribution in our full dataset.

data, children also tell relatively more `Person` and less `Agent` and `Actor` stories as they grow older. Our CD labelling thus tracks meaningful variation in ToM competence over the 4-12 year age range. Age is a strong story-external predictor of CD (see Chapter 2); yet, here we do not include it in our classifier. We think it is valuable to try to label CD purely from textual variables, anticipating collecting data without needing to store sensitive background information of children, or leveraging text datasets where such information is unavailable. Also, from a more general perspective, CD levels indicate the kind of socio-cognitive information present in texts. In advanced applications such as conversational agents, memorising socio-cognitive information is important for making interactions successful. Knowing the linguistic properties of socio-cognitive information (`Person` stories), could be helpful information to add to multi-modal conversational agents that draw on gaze and speaker activity (e.g. Tsfasman et al., 2022).

Feature engineering

Here we describe the engineering of features that encode language competences predictive of ToM competence in children.

- **Lexical Complexity (LC).** We calculated the perplexity PP of the story vocabulary V as set of lemmas $\{l_1, l_2 ... l_n\}$ with

$$PP(V) = \sqrt[n]{\frac{1}{P(l_1, l_2, ... l_n)}}. \tag{4.2}$$

  Lemma probabilities were approximated with relative frequencies from the BasiScript lexicon, a Dutch corpus of written child essays (Tellings et al., 2018a). Lemma frequency estimates lemma complexity (Vermeer, 2001): infrequent lemmas yield higher perplexity relative to the lexicon. A more complex vocabulary has been found to predict ToM competence and CD (De Mulder, 2011, see also Chapter 2). The idea here is that a more complex vocabulary works as a toolbox, enabling the representation of more complex aspects of reality, including the social realm.

- **Lexical Diversity (LD).** We modelled the lexical diversity of stories with the Measure of Textual Lexical Diversity (MTLD). MTLD calculates the average length of word sequences for which a type-token ratio of at least 0.72 is maintained; MTLD is robust to texts of differing lengths (McCarthy and Jarvis, 2010). Since LD ignores word complexity, it is a proxy for vocabulary size (but not complexity), which is found to predict performance on various ToM tasks (Milligan et al., 2007; Slade and Ruffman, 2005).

- **Dependency Distance (DD).** As measure of syntactic skills we extracted dependency distance DD between syntactic heads and dependents with spaCy version 3.2.0 (Honnibal and Johnson, 2015). Following Liu (2008) we calculated mean DD with

$$DD(S) = \frac{1}{n-s} \sum_{i=1}^{n} |DD_i| , \tag{4.3}$$

  where DD_i is the absolute distance in number of words for the i-th dependency link, s the number of sentences, and n the number of words in story S. Language employing larger DD is more demanding for working memory and thus harder to process (Futrell et al., 2015; Grodner and Gibson, 2005). Here DD is a measure of children's general syntactic proficiency, which has been linked to ToM competence on standardised tests (Astington and Jenkins, 1999; Milligan et al., 2007; Slade and Ruffman, 2005).

- **Clausal Complementation (CC).** We extracted the average number of clausal complements per utterance with spaCy. Mastering CC has been linked to performance

on several false belief tasks (De Villiers, 2005, 2007; Hale and Tager-Flusberg, 2003; Lohmann and Tomasello, 2003); here we examine its predictive power in the narrative domain. Complementation syntactically scaffolds reasoning about beliefs, desires, speech and perception (see Section 4.2).

- **Pragmatic Markers (PM).** We compute the average use per utterance of *pragmatic markers*: words used to indicate deixis and common ground (Rubio-Fernández, 2021). As markers of deixis we include demonstratives 'this' (*deze*), 'that' (*dat, die*), 'here' (*hier*), and 'there' (*daar*). As marker of common ground we use the definite article 'the' (*de/het*). These markers all invoke a character's *perspective* in space or time (e.g. 'Come here!'), or shared knowledge (e.g. 'I saw the key' vs. 'I saw a key'); children's competence in these more basic forms of handling others' perspectives is argued to be a precursor to ToM competence (De Mulder, 2011; Rubio-Fernández, 2021).

- **Social Words (SOC)**. Linguistic Inquiry and Word Count (LIWC) is a tool that extracts words belonging to specific categories (Tausczik and Pennebaker, 2010). The 'social' category indicates family, friends, social interactions and personal pronouns (e.g. 'mother', 'to invite', 'she'). The social content children employ is here taken to reflect the finding that ToM competence depends on frequent social interactions (Nelson, 2005), and that family size and sibling relation quality contribute to ToM competence (Hughes and Leekam, 2004; McAlister and Peterson, 2007). Thus, we expect that stories with more social content have higher CD.

- **Lemmas.** With spaCy we obtained binarised bag-of-words vector representations of stories to retrieve lemmas typical for specific CD levels. Lemmas occurring in less than 5% of stories were excluded. Some lemmas more clearly fit specific CD levels than others; for example, 'to think' has mental state content, thus fits `Person` level, but this is less obvious for e.g. temporal ('then'), and causal ('because') connectives. Mastery of/exposure to mental state verbs like 'to think' has been linked to performance on various standard ToM tasks (Lohmann and Tomasello, 2003; San Juan and Astington, 2017); by transforming stories into bag-of-words vectors, we are able to automate lexical analysis of narratives that in developmental work often relied on hand-coding (Nicolopoulou et al., 2022).

We had 205 features in total (6 custom features + 199 lemmas). Since the aim is to predict CD purely from textual features, our custom features must be relatively independent of age (to prevent predicting CD from age through language) and from

	Precision	Recall	F1
Actor	.71 (.55)	.50 (.52)	.59 (.52)
Agent	.76 (.74)	.68 (.70)	.72 (.72)
Person	.76 (.79)	.89 (.85)	.82 (.82)
Average	.74 (.69)	.69 (.69)	**.71 (.69)**

Table 4.3: Performance metrics on an initial test set, and on 100 different train-test splits (averages in parentheses).

one another. We computed Variance Inflation Factors (VIF) for custom features and dummy-coded age groups, with the youngest group (4-6y) as reference. We adopted a threshold of 5 as indicating problematic multicollinearity (James et al., 2013); all VIF were low ≤ 1.54, indicating that features are relatively independent.

4.4 Results

Our analysis was implemented with scikit-learn version 1.0.1 (Pedregosa et al., 2011) and proceeded as follows. First, we obtained an initial random 80%-20% train-test split. We chose logistic regression, since unlike generative classifiers like Naive Bayes, logistic regression is more robust regarding correlated features. In addition, we preferred logistic regression as probabilistic classifier to geometrically motivated classifiers like Support Vector Machines. To curb overfitting, we tuned regularisation type and strength of our logistic classifier with 5-fold cross-validation, which suggested L2 regularisation and higher regularisation strength ($\alpha = .075$). Overfitting is a threat as validation and test stories can differ from training examples. We then did a full training, and with Shapley values considered the linguistic information associated with different CD levels. We gauged robustness of the model by re-training it with the same settings on 100 different train-test splits. In all splits, the label distribution visible in Table 4.2 was maintained. In training, class weights were computed based on Table 4.2 that during training, induce a larger penalty on errors made for the infrequent class (`Actor`).

Performance metrics are given in Table 4.3. For the initial split, performance is reasonably good with a F1-macro of .71, given task complexity for humans (Section 4.3), and against the background of a majority vote baseline which always decides `Agent` and is accurate 45% of the time, but performance is a bit lower for `Actor` stories. The model seems robust on `Agent` and `Person` stories, as performance on the additional splits is comparable, but less robust for `Actor` stories. Overall, higher CD

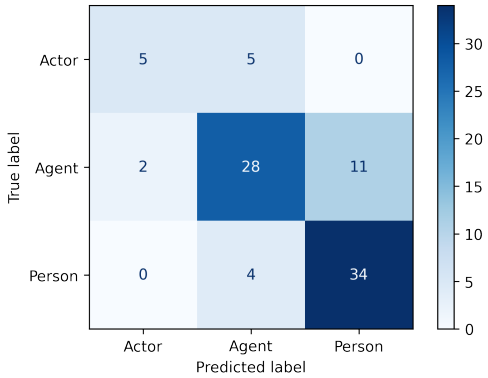


Figure 4.2: Confusion matrix for initial test set.

levels coincide with better performance. In Figure 4.2 we see that the most dissimilar CD levels (`Actors` and `Persons`) are never confused, which is intuitive.

Feature importance

We now disclose the linguistic information the model associated with specific CD levels during training with feature importance as given in Figure 4.3.

For `Actor` stories, we see that lexical complexity (LC), complementation (CC), pragmatic markers (PM), and dependency distance (DD) are all negative indicators. Thus, `Actor` stories are overall linguistically less complex. We also see other negative indicators that indeed fit other levels better: verbs 'to see', 'to go', 'to say', 'to come' for the `Agent` level, as they indicate action and perception, and 'to want' for `Person` level, which is explicitly intentional. Connectives 'not' and 'but' are also negative indicators, suggesting that clauses and utterances in `Actor` stories are less explicitly linked. The only positive indicator is adverb 'than' (*dan* in Dutch), which is in `Actor` stories often used for (quasi-temporally) stringing together events.

For `Agent` stories we see as positive indicators use of pragmatic markers (PM) and larger dependency distance (DD), next to the verb 'to go' and preposition 'to', which fit `Agent` as action-centred CD level. For the rest we see features that were also negative indicators for the `Actor` level, such as the intentional verb 'to want', and connectives 'not', 'but', and 'thus', likely for the same reasons as mentioned above. Also, we see pronoun 'he' as negative indicator, useful for shifting a story to a third-person perspective, which is natural in narratives. Overall `Agent` stories appear to be linguistically more complex than `Actor` stories.

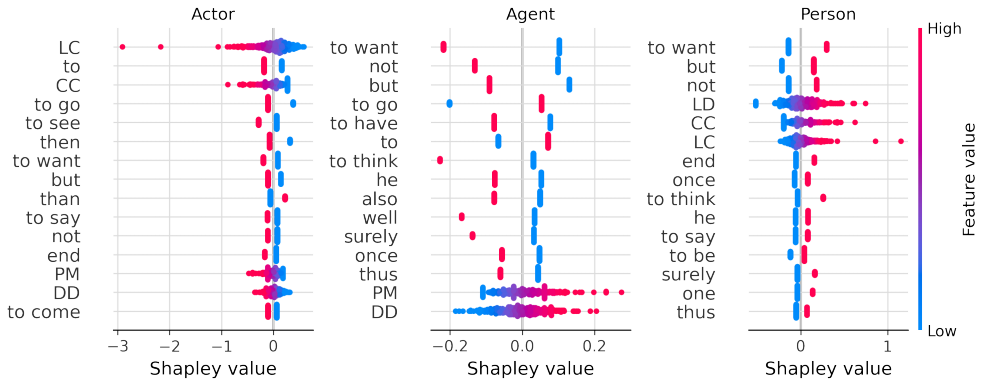


Figure 4.3: Shapley values for the 15 most important features per label. Value size (X-axes) quantifies feature importance; value sign whether the feature is a positive/negative indicator of a particular label; colour indicates for which values of that feature. For example, for clausal complementation (CC), red positive Shapley values under the `Person` label indicate that more clausal complementation makes a `Person` label more likely; blue negative values indicate that less clausal complementation makes a `Person` label less likely.

`Person` stories are linguistically most complex. They employ higher lexical diversity (LD), lexical complexity (LC), and more complementation (CC). Verbs with intentional content ('to want', 'to think') are clear and intuitive indicators. All connectives that negatively indicated `Actor` and `Agent` levels, positively indicate `Person` stories ('but', 'not', 'thus'), suggesting that `Person` stories have more explicitly linked clauses. In addition, the pronoun 'he' suggests that a third-person perspective is more often employed in `Person` stories. Further, in `Person` stories communication also seems to play a key role ('to say').

Error analysis

Here we briefly discuss two prediction errors in `Actor` recall (`Actor` stories mistaken for `Agent`), the metric with lowest values in Table 4.3. For story 083101, we see in Figure 4.4 that many linguistic features (e.g. DD, CC, PM) indicating less linguistic complexity, push the decision line towards the correct prediction; the same applies to the absence of various lemmas (e.g. 'to want', 'not') identified in Section 4.4. Yet, this story is an outlier as it employs some highly unusual words (driving up LC), which sharply reduces the probability of deciding `Actor`; Actor stories are overall lexically less complex.

For story 010601 we see that linguistic features (e.g. CC, LC) indicating less lin-

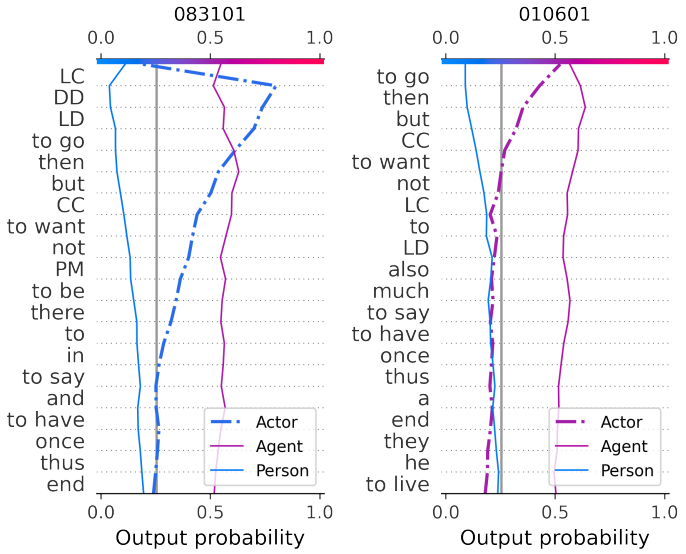


Figure 4.4: Decision plots for two recall errors (story IDs 083101 and 010601), that show the impact of features on the vertical decision lines given for each label. These plots are best read from bottom to top. Each decision line plots the probability the classifier assigns to a specific label and it may increase (push right) or decrease (push left) based on the features given on the rows. These plots presuppose knowledge of feature importances as discussed in the previous section and Figure 4.3. For example, given that we already know that `Actor` stories less often employ PM and CC, we can take these features for story 083101 to indicate absence of PM and CC, since they push the decision line to the right thus increase the probability of `Actor`.

guistic complexity, plus the absence of particular lemmas ('to go', 'but'), push the decision towards `Actor`. The issue here is probably that the features of which absence has a large impact on the decision for `Actor`, also favour `Agent` ('to want', 'not', 'but'), making the levels less distinguishable (their lines have partly similar trajectories). Thus, `Actor` and `Agent` labels would benefit from having more unequivocal indicators. Importantly, besides exposing wrong decisions, Figure 4.4 also illustrates that multiple custom linguistic features and lemma features shape the classifier's decisions.

4.5 Discussion

We employed a logistic classifier on labelling Character Depth for 442 freely told stories. Feature engineering was used to encode key linguistic competences identified in empirical work as predictive of ToM performance. The goal was to see how these

features are reflected in ToM as manifested by CD. CD was predicted from linguistic features only, which were relatively independent of age. We now discuss the link between specific features and CD levels in the broader context of ToM, and further reflect on language and ToM competence in narratives as context-dependent phenomena.

We saw that stories with flat characters (`Actors`) are identified by the model as employing less complex words, less complementation, less pragmatic markers, and lower dependency distance. In addition, the clauses and utterances in these stories seem less explicitly linked with connectives. Thus, stories without clear ToM competence 'in action', are also stories in which we see less advanced language competence 'in action'. Our results here mostly confirm and extend existing work on ToM and language, but we saw no role for social words or lexical diversity. Stories in which children do not provide insight in character minds, thus where the texts concerns mostly physical descriptions, apparently solicit less complex linguistic scaffolds. A caveat for `Actor` stories is that our results were less robust compared to other CD levels (Table 4.3).

In `Agent` stories, ToM competence 'in action' starts to take off with characters exhibiting implicit intentions, intentions-in-action, emotions and perceptions. In the example in Table 4.1, that the prince calls the police after perceiving a villain *implicitly* suggests a goal or intention with the action. As developmental work cited in Section 4.2 shows, this is a precursor to explicitly spelling out the character's mental states, that then further contextualises actions and events (as the girl's desire in the `Person` example does in Table 4.1). In this light, it is interesting that in `Agent` stories the use of pragmatic markers emerges, another precursor to ToM, that involves handling deixis, which constitutes basic character perspective management (Section 4.3). Another tentative indicator that a full third-person perspective shift, natural to narratives, is not typical for `Agent` stories, is the pronoun 'he' as negative indicator, although this perspective can also be construed with other third-person pronouns. Regarding other features, `Agent` stories exhibit larger dependency distance, thus syntactically more complex utterances; yet, the fact that various connectives are negative indicators also suggests children add less explicit coherence between clauses and utterances. We see no indications that the lexical properties of stories or social words are tied to the `Agent` level. Thus, our results partly confirm and extend earlier work especially regarding pragmatic markers and syntax, and this result seems robust (Table 4.3).

`Person` stories exhibit the highest level of ToM competence in that characters

show explicit (complex) intentional states, related to events, actions or other characters' mental states in the story world. Complementation indicates `Person` stories and thus seems to scaffold ToM beyond the false belief context (De Villiers, 2000), likely to convey desires, beliefs, and speech, as evidenced by the lemmas indicative for this class (Section 4.4). `Person` stories are lexically more diverse and complex, in line with other work on predicting ToM in narratives (see Chapter 2): a larger *and* more complex vocabulary could provide better tools to grasp and represent the social world. `Person` stories are not distinctively associated with pragmatic markers, social words, or syntactic complexity as represented in our model. Yet, regarding syntax, various connectives as positive indicators suggest that `Person` stories have more explicitly structured clauses and utterances. Thus, our result partly confirms and extends earlier research (Section 4.3), and seems robust (Table 4.3). Stories in which children provide most insight in character minds, thus texts in which (complex) socio-cognitive information is explicitly present, apparently solicit more complex language scaffolds regarding the lexical domain, which is traditionally strongly linked to a host of ToM-related skills (see Section 4.3).

We conclude with a reflection on language and ToM competence in narratives as context-sensitive, yet *natural* language data. Some reviewers remarked that ToM in narratives needs a separate accompanying measure, to make sure we are really talking about a child's ToM ability when we are talking about CD. There are strong reasons to think that ToM is a complex, multi-faceted ability, given the many definitions of ToM that exist (Quesque and Rossetti, 2020; Schlinger, 2009), and the many different standardised tests that have been designed and employed (Milligan et al., 2007; Wellman, 2018). As stated in footnote 2 (Section 4.3), these tests have their own limitations; benchmarking CD with an existing standardised measure yields no simple answer to the question whether we are now talking about children's actual ToM. That does not make standardised tests uninformative, but contextualises their merit: if we agree that ToM (and language) are social competences, we should also test them in social contexts, not to claim superiority over but rather to complement work done in controlled settings.

Our classroom context has as advantages regarding ToM, that children feel more motivated to do a fun task, engage with narratives as natural finding place for mental state content, have freedom to explore the (social) scenario they want, and that their language use has a social goal: immersing the audience in their narratives as possible worlds. This social context may stimulate children more to challenge their language skills. To entice their audience, children may leverage their vocabulary skills to re-

fer to rare settings, uncommon objects, unorthodox characters, and peculiar social situations which is not possible in standardised language tests like the Peabody Picture Vocabulary test (Dunn and Dunn, 1997). Additionally, children may also recycle complex linguistic structures and plots from prior exposure to narratives in their own narratives, to entice their audience. Thus, the influence of the social context could result in more complex language use than one would expect based on age, which makes the direct relation between age and language competence in narratives less obvious.

Overall, our results support the link between more complex language and ToM. That said, not all ToM-related content requires complex language. Explicating character thought could linguistically also be represented without complement, e.g. with Free Direct Thought as in 'Was she angry with him?' (Leech and Short, 2007). Moreover, the words used in this thought are not complex, nor is the syntax. This example serves to illustrate the point that in our approach, our classifier makes no assumptions at the outset about the linguistic complexity of ToM-related content.

4.6 Conclusion

This chapter aimed to disclose the relation between language competence and Theory of Mind in children's freely told narratives. Language competence was encoded in custom linguistic features; the mental depth of story characters was a proxy for Theory of Mind competence 'in action'. We linked specific linguistic contexts to lower and higher levels of Theory of Mind in narratives. Overall, we found that stories with flat, mentally undeveloped characters (`Actors`) are linguistically less complex, compared to stories employing characters displaying intention-in-action, emotion, and perception (`Agents`), which in turn are linguistically less complex compared to fully-blown characters with explicit intentionality (`Persons`). We classified Character Depth without drawing on children's age and obtained good performance on an initial train-test split, relative to the complexity of the task for humans (F1-macro = .71). This result was fairly robust on 100 different splits, but to a smaller extent for `Actor` stories. Overall our results support the hypothesis that in children as focal point for studying language and ToM development, language and ToM are intertwined and reinforce each other, using data from older children obtained in social settings.

4.7 Limitations

One limitation concerns the annotations: although there were two independent expert annotators that together annotated 16% of the stories, the rest of the annotations depended on a single expert. A second limitation is that in retraining and testing models on different splits, feature importance can vary a bit, since for example outliers (an example is given in Figure 4.4) are sometimes part of the train set, and sometimes not. Third, especially for the `Actor` level, the model was less robust, so results regarding the linguistic properties of `Actor` stories may generalise less well to other research contexts, but this remains to be seen; we can for example imagine a comparable analysis of ToM and language competence in *written* Dutch essays by school children, as provided by the BasiScript corpus (Tellings et al., 2018a). Lastly, the BasiScript lexicon used for calculating lexical complexity (Section 4.3) is free, but a license must be signed before use, which can be obtained from the hosting institution. Also, LIWC as used for extracting the social words feature (Section 4.3) is a proprietary tool. Thus, features for lexical complexity and social words cannot be reproduced from scratch, although the results of using these tools are included in our data csv files. Another limitation is that in this study we cannot differentiate between language and ToM competence of neurotypical and neurodivergent children, as we collect no such medical data.

4.8 Ethics Statement

This study was approved by the Leiden University Science Ethics Committee (ref. no. 2021-18). The story corpus employed in this chapter was compiled in close consultation with school teachers, principals, parents, and children. We used lightweight classifiers that for our research purposes required little compute. By offering all children in a classroom the opportunity to freely tell a story and participate, and by including schools in a variety of areas and environments across the South and South-West of The Netherlands, we aimed to be as inclusive in our data collection as possible.

Chapter 5

Character Perspective Representation in Freely Told Stories

Story characters not only perform actions, they typically also perceive, feel, think, and communicate. Here we are interested in how children render characters' perspectives when freely telling a fantasy story. Drawing on a sample of 150 narratives elicited from Dutch children aged 4-12y, we provide an inventory of 750 instances of Character Perspective Representation (CPR), distinguishing fourteen different types. We observe first of all that character perspectives are ubiquitous in freely told children's stories and take more varied forms than traditional frameworks can accommodate. Second, we discuss variation in the use of different types of CPR across age groups, finding that character perspectives are being fleshed out in more advanced and diverse ways as children grow older. Thirdly, we explore whether such variation can be meaningfully linked to automatically extracted linguistic features, thereby probing the potential for using automated tools from Natural Language Processing to extract and classify character perspectives in children's stories.

This work was originally published as: Van Duijn, M.J., Van Dijk, B.M.A., and Spruit, M.R. (2022). Looking from the Inside: How Children Render Characters' Perspectives in Freely-told Fantasy Stories. In Clark, E., Brahman F., and Iyyer, M., editors, *Proceedings of the 4th Workshop on Narrative Understanding*, pages 66-76. Association for Computational Linguistics.

5.1 Introduction

Story characters are everywhere around us: we meet them in the books we read, the TV series we get caught up in, or in a gossipy tale we tell each other during everyday social gatherings. Some characters may be modelled on real people, whereas others exist only in the imagined worlds of fantasy and fiction.

In its most basic form, a story character is an entity involved in some kind of action or description. Yet typically we also get to share in some of its *perspectives* on the story world and the objects, events, and other characters within it. There are long-standing traditions in linguistics and literary studies, especially within the subfields of stylistics and narratology, studying the ways in which such character perspectives can be rendered (e.g. Banfield, 1973; Leech and Short, 2007; Vandelanotte, 2009). Three main types commonly distinguished in studies of speech and thought representation are *direct*, *indirect*, and *free (in)direct* speech or thought (see Table 5.2 for examples). While most attention has been paid to literary texts, scholars have also identified such types in cinema (Verstraten, 2009), theatre (McConachie and Hart, 2006), and other domains such as news articles (Sanders, 2010), everyday conversations between parents and young infants (Köder, 2016), or speech from individuals with psycho-pathological conditions (Van Schuppen et al., 2020).

It is largely an open question as of yet how children render characters' intentions, perceptions, emotions, speech, and thought when asked to freely tell a fantasy story. This is worthwhile exploring for a variety of reasons. It has been widely argued that representing different perspectives reflects a central function of language usage (e.g. Dancygier et al., 2016): human interaction is characterised by 'polyphony', meaning that we rarely only express our own perspective. Instead, the default is that we use language to *orchestrate* multiple perspectives.

Even though this pervades all speech domains, stories are a key finding place for linguistic and narratological patterns supporting this function (Fludernik, 1996), and arguably also the 'sandbox' where both children and adults test and refine their perspective orchestration skills (Vermeule, 2009). Mapping how children of different ages render character perspectives is as such of interest to language acquisition research, but also to cognitive psychology as it provides insight into how children learn to understand the social world and others' minds, and the role narratives can play herein. Tools from Natural Language Processing (NLP) can fuel all such research, for example by automatically identifying contextual information associated with different character perspectives. NLP researchers, in turn, can learn about phe-

nomena relevant for embarking on tasks involving more complex classification or extraction of perspectivised content.

In the current contribution we draw on a sample of 150 stories, told by children aged 4-12y as part of storytelling workshops we offered across The Netherlands. Our sample features 750 instances of Character Perspective Representation (CPR), which we categorise into fourteen different types based on manual qualitative analysis. As discussed below, the type categories and analytical framework we adopt are primarily inspired on 'classic' speech and thought representation literature (mainly Leech and Short, 2007). However, we complement our framework with additional types based on research into children's development as storytellers and relevant insights from cognitive linguistics, allowing for a more refined and inclusive way of mapping character perspectives.

The best way to introduce our approach in concrete terms is to discuss the analysis of an example story (given in Table 5.1). Doing so will also make clear how we position this chapter: as an effort to build a bridge between *qualitative* analysis of narrative material as traditionally done in the humanities, and *quantitative* analysis, driven by the automatic extraction of linguistic information, as customary in computational approaches. However, in Section 5.3 below, we will first provide more details on our language sample and annotations, introduce two automatically extracted linguistic features, and then discuss an example story (given in Table 5.1) and our CPR typology. Thereafter we discuss in Section 5.4 the relation between the occurrence of different CPR types and the age of the storytellers, as well as the relation with lexical and syntactic characteristics of the utterances in which the CPR types occur. We end with a reflection on our findings in Section 5.5.

5.2 Background

Children tell stories to themselves and others as part of their daily play activities (Cremin et al., 2017; Sutton-Smith, 1986). While being the source of a lot of fun in the first place, such storytelling has been analysed as a form of *cognitive play* that is essential for child development in various key areas, including the acquisition and refinement of communicative skills (Southwood and Russell, 2004), organising knowledge of the (social) world (McKeough and Genereux, 2003), and empathising with others and understanding their motives and intentions (Gallagher and Hutto, 2008; Nicolopoulou, 2018; Zunshine, 2019). Phenomena of CPR are situated at a natural crossroads of these key developmental areas: their absence or presence in freely

told stories arguably reflects children's communicative abilities, but also their understanding of the social world and capacity to imagine others' inner workings. Here we explore the occurrence of different patterns of CPR across different age groups, and we believe that our contribution can ultimately fuel research in developmental psychology and language acquisition research. However, it is important to note that claims about whether the patterns we find in our stories are indicative of a specific child's development are outside the scope of this chapter.

5.3 Methods

Data

The storytelling workshops for the creation of our database were held between 2019-2021 at seven elementary schools, a daycare, and a community centre located in various areas across in The Netherlands. Each session was held in a classroom setting involving 5-30 children at a time, of varying ages between 4-12. Sessions started by discussing some general characteristics of stories (e.g. 'Where can you find stories?', 'What kind of stories do you like?') and interactively narrating an exemplary fantasy story with the participating children. Next, we invited children to take the floor and tell a fantasy story about a topic free of choice. After informing children about this, voice recordings were made, which were pseudonymised and transcribed afterwards by the authors and research assistants. Transcripts were double-checked for consistency with the audio files.

As of now, we have collected over 600 stories in our database called ChiSCor (Children's Story Corpus).[1] Our data collection and data management protocols were assessed and approved by the Leiden University Faculty of Science Ethics Committee (file no. 2020 – 002).

For the current research we drew a sample from ChiSCor according to the following steps:

1. We included only the first story told by each child (many children told multiple stories), which reduces dependence between stories. This yielded a subset of 350 stories.

2. We selected stories with a length (in number of words, $\bar{x} = 108.64, \sigma = 99.62$) falling in the interquartile range (IQR), i.e. 50% around the median ($min =$

[1] The stories used in this chapter, along with our current annotations and scripts, are available via the Open Science Framework (OSF): `https://osf.io/9q32v/`

$4, Q1 = 35, Med = 75, Q3 = 151, max = 626$), to prevent over- or underrepresentation of data from children with exceptionally long or short stories.

3. We then defined three age categories, 'Young', 'Middle' and 'Older', in line with the division common in Dutch primary education into 'Onderbouw', 'Middenbouw', and 'Bovenbouw'. Young corresponds to Onderbouw which involves ages 4-6; Middle corresponds to Middenbouw which involves ages 6-9; Older corresponds to Bovenbouw which involves ages 9-12.

4. We included 150 stories in total (12879 words), 50 for each group. For the Young and Middle groups these were randomly drawn out of 60 and 78 stories falling within the IQR, respectively. The older group had only 39 stories within the IQR; here we added 11 stories closest to $Q1$ and $Q3$ to balance the age groups.

Annotations

The 150 stories were put into a large table in random order and without showing additional information to avoid (unconscious) interference with decisions in the annotation process.[2] Existing line breaks, introduced during transcription of the audio recordings according to a standardised protocol, were used to chunk each story into smaller units, henceforth referred to as 'utterances'. We identified 568 unique characters that in total made 1472 appearances (the same character can obviously appear in multiple utterances within the same story), 722 of which involved only descriptions or simple actions without insight being offered into the character's perspective. The remaining 750 appearances were given one of fourteen different labels representing our types of CPR. In rare cases where multiple types applied, the most 'advanced' label was chosen in terms of the stages introduced below.

One author of this chapter who has a background in grammar and narratological theory, took the lead in the annotation process, while regularly discussing categorical distinctions as well as individual utterances with the second author. In some specific cases, expertise was gathered from external experts. While we can see how this procedure may be problematic from the perspective of current standards in NLP, two considerations should be added with regard to our approach in this chapter. Firstly, we point out that we base our annotations on long-standing traditions of textual analysis within cognitive linguistics, narratology, and stylistics, known to support high degrees of intersubjective agreement and reproducibility between researchers within

[2] E.g. the age or school of the storyteller. Note that such interference could only be avoided to a certain degree; after all, we were ourselves involved in recording the stories.

these fields (for a broader discussion of a 'grounded theory' approach, see Charmaz, 2006). Secondly, it is important to note that the statistical analyses in Section 5.4 are based on *merged* categories. While discussion is sometimes possible about the most appropriate type of label for specific utterances (e.g. deciding between direct and indirect speech on grammatical grounds; see also Köder, 2016), such discussions would rarely affect the overarching merged category under which this utterance falls.[3] Nevertheless, we consider it an important next step within our larger project to gather CPR annotations from at least one additional, independent annotator.

We discuss our full typology of CPR further below, along with the example story and inventory of the occurrence of each type in our sample. However, it is important to single out one type beforehand: `ego-narration`. We see this as a 'preliminary stage' of the fuller mastery of CPR that is characteristic of the other thirteen types. We marked cases as `ego-narration` if there was no (or an unclear) distinction between the child narrating the story and a referent indicated with first-person pronouns ('I', 'me', 'we', 'us') within a story.

Consider the following example from story with ID 022501 in ChiSCor:

(1) [...] *and I do a lot of horse riding / and ride a lot of horses / and we have a lot of very sweet horses in the stables* [...]

Example (1) counts as `ego-narration`, since the 'I' who regularly does a lot of horse riding refers to the child in the immediate situation of telling the story. This is different in the following example from story 082601:

(2) [...] *and then came well myself in fact who came with a gun / and I said why are you fighting Batman and Superman* [...]

In example (2), the 'I' is making an appearance in a story world clearly detached from the here-and-now of telling the story.[4]

The rationale for singling out `ego-narration` as a preliminary phenomenon is that it evidences a lack of 'transcendence' (Zeman, 2020), marking a departure from the actual speaker and its immediate here-and-now, which we consider a key feature of storytelling. Such transcendence is warranted by a distinction between the child telling the story (ego), the narrator seen as a theoretical entity or 'role', and characters

[3] An exception is found in line 7 of the example story presented in Table 5.1.

[4] The full Dutch stories can be found in our OSF repository (footnote 1). Utterances are separated with forward slashes. English translations are our own and were made only for the purpose of discussing them here; CPR annotations within this chapter are based on the Dutch stories.

within the story.[5]

What the remaining thirteen types of character representation have in common is that they exhibit storytelling in this sense, i.e. a specific form of communication in which a narrator-entity provides all kinds of linguistic cues inviting listeners (or readers) to imagine a story world including objects, characters, actions, events, etc. (Dancygier, 2011). In this way it is possible for narrators to tell a story entirely from the 'outside perspective', without directly cuing listeners to imagine what the story world would look like from any character's point of view; this is what we observed in utterances containing only character appearances consisting of descriptions or simple actions, plus in utterances containing no character appearances at all. In each of the remaining utterances we found essentially a mix of narrator and character perspectives. The way in which, and degree to which these character perspectives were explicitly fleshed out and/or separated from that of the narrator, determine which of the thirteen CPR types applies.

Linguistic features

There is evidence that socio-cognitive skills, in particular the capacity to understand and reason about others' mental states known as Theory of Mind (Apperly, 2012), are positively correlated to lexical and syntactic proficiency in children. For example, children possessing a larger vocabulary, or mastering clausal complementation, perform better in reasoning about others' mental states in standardised clinical tasks (for an overview see Milligan et al., 2007).

We see overlap between children's development of socio-cognitive capacities and their ability to flesh out characters' perspectives in a narrative. Therefore we include lexical and syntactic complexity here as two theoretically motivated features, that can potentially provide us with information about the linguistic context in which different CPR types occur, and connect this to age groups of the storytellers in our sample. Doing so, we might also anticipate linguistic information encoded in (the middle layers of deeper) neural networks, that could be helpful for automatically extracting and/or classifying perspectivised information in children's narratives in the future (Jawahar et al., 2019).

To calculate Lexical Complexity (LC), we approximated for each word in utterance U featuring CPR its lemma probability $P(l)$ by its relative frequency count in the BasiScript lexicon, a large benchmark corpus of written child output (Tellings et al.,

[5]We refrain from going into the widely debated narratological concept of the narrator here and refer to Zeman (2020) for a to-the-point overview.

	Utterance
1.	*a girl went to the zoo and she saw a huge lot of tigers and other animals* [...]
2.	*and she went home all alone*
3.	*but her little brother was left behind he was sitting on the monkey*
4.	*then said the sister of the little boy where is my little brother now*
5.	*she went back again to the zoo*
6.	*then she saw that the little brother was sitting on the monkey*
7.	*oh little brother where are you now*
8.	*the end*

Table 5.1: Example story with ID 072201.

2018a). The perplexity of the utterance $PP(U)$ is then given by the set of lemmas $U = \{l_1, l_2...l_N\}$ and its probabilities with

$$PP(U) = \sqrt[N]{\frac{1}{P(l_1, l_2, ...l_N)}}. \tag{5.1}$$

Utterances with more infrequent lemmas show higher perplexity with respect to the lexicon. Lemma frequency has been argued to be a good measure of lemma complexity given that infrequent lemmas are overall harder to learn (Vermeer, 2001).

To calculate Syntactic Complexity (SC), for each utterance U featuring CPR we extracted a dependency tree, a directed graph $G = (V, A)$ with V as the set of words and A as the set of arcs indicating dependency relations between words. We extracted the maximum number of arcs between the root node and a leaf node in G. This measure is also known as tree depth and is a common measure of syntactic complexity: utterances employing longer paths are syntactically more complex (Dell'Orletta et al., 2011).

CPR types in an example story

In order to illustrate our approach in more detail, we will now discuss the analysis of a story excerpt given in Table 5.1, featuring five types of CPR found throughout our sample. Afterwards, the remaining types will be briefly introduced along with a complete overview of examples and counts in Table 5.2.

First of all, we can observe that this is a story narrated in third person, past tense. For a large part it consists of narrator descriptions of actions and situations ('went to the zoo', utterance 1; 'went home all alone', utterance 2; 'her little brother was left behind' and 'sitting on a monkey', utterance 3; etc.); however, as listeners/readers

we also get a few glimpses into the perspective of one character: the 'girl'.

In utterance 1 we learn about the animals she 'saw'. It could be defended that this is still entirely the narrator's voice telling us 'from the outside' what the girl would have been seeing at the zoo. Yet, as discussed in Section 5.2 above, and in line with what cognitive linguists have argued in recent years (e.g. Van Duijn and Verhagen, 2018), we suggest that perspectivisation of content in narratives can be seen on a cline, ranging from pure narrator view on the one extreme, to full character view with minimal narrator mediation on the other extreme. Following this approach, the report of what the girl 'saw' in utterance 1 implies a modest but certain invitation for listeners or readers of the story to imagine the girl's perspective on objects within the story world: 'a huge lot of tigers and other animals'. This is a case of Perception (PER) in our system of types. Another instance is found in utterance 6.

What is more, we note a difference between how the situation of the 'little brother' is described ('was left behind', 'sitting on a monkey', utterance 3) and some of the descriptions of actions performed by the 'girl' (e.g. 'went home all alone', utterance 2; 'went back again to the zoo', utterance 5). Following developmental psychologists and children's story researchers Nicolopoulou and Richner (2007) we classify the latter as cases of Intention-in-action (IIA), i.e. actions coupled to a clear goal or result within the immediate story context. As these authors argue, such actions are not yet fully explicit intentional states that would disclose characters' perspectives for the audience, but hint at them implicitly. So, compared to PER and other forms of CPR discussed below, IIA represents the lowest degree of inviting a shift from the narrator's to a character's perspective. Yet mere descriptions of a character's situation, appearance, attributes, or actions without an immediately specified result or goal do not invite such a shift at all, or to an even lesser degree. This is why we see IIA as the most basic type in our staging of CPR.

In utterance 4 we find a case of Direct Speech (DS) with an inquit formula ('said the sister of the little boy') and a reported clause ('where is my little brother now'.[6]) The reported clause has three features supporting our classification as DS. Firstly, a shift to the present tense can be observed ('is' as opposed to 'said' in the inquit formula). Secondly, there is a shift from the third to the first person as expressed by the pronoun 'my'. And thirdly, the addition of 'now' ('nou' in the original Dutch story) can be seen as an idiomatic exclamation, expressing a degree of wonder or confusion (which is not satisfactorily covered by the English translation 'now'). This wonder or confusion

[6]The absence of a question mark after the reported clause is due to standardised transcription of the recorded oral stories.

is clearly to be interpreted as part of the 'girl'-character's experience, and not of the narrator's, just as 'my little brother' from the character's point of view indicates the same referent as 'the little boy' from the narrator's point of view. The present tense is congruent with the girl-character's experience at the moment of speaking within the story plot.

Finally, utterance 7 features Free Direct Speech (FDS). Here we see the same shift to present tense ('are') and the same exclamation ('nou' in the Dutch original), complemented with another exclamation at the beginning of the sentence ('oh'). The absence of an inquit formula makes it a case of FDS rather than DS. Or, a different possible interpretation of utterance 7 is that we are looking at a form of 'monologue intérieur' in which the girl-character produces this utterance for herself, rendering it a case of Free Direct Thought (FDT) rather than speech. The context does not resolve this ambiguity. One can argue that she is addressing the boy, given that she has just found him in the preceding utterance, but one can equally well argue that utterance 7 should be read as an internal expression of her surprise, given that he is sitting on a monkey.

CPR types in the rest of the stories

In Table 5.2 it can be seen that ego-narration (EGO-NARR), the preliminary stage of CPR we distinguished earlier, occurs 47 times in our sample. IIA, which we consider to be CPR in its most basic form, is with 350 occurrences by far the most frequently observed type. Usage of IIA entails that the narrator reports what a character is doing, and to what end. Similarly, with PER, of which we recorded 53 instances, it is the narrator who reports what a character is perceiving. Both happen without the narrator intruding into the character's mental world: rather, a description is given that invites the listener to imagine what a character intends or perceives, thereby effectively getting to share in the character's perspective on the story world to some degree. Narrative Reports of Speech Acts (NRSA) and cases of (Free) Indirect Speech (*(F)IS*), relate what a character says or said primarily in the words of the narrator, while (Free) Direct Speech (*(F)DS*) is to be taken as the literal rendition of a character's words.

Still, what all these forms of speech reporting have in common is that they do not imply that the narrator has direct insight into characters' minds. Here too it is strictly speaking the listener who is cued to draw conclusions about a character's perspective based on the report of what they say or said. This contrasts with thought

Type	Example	Counts	ID
Ego-narration (EGO−NARR)	*'I love music'*	47	061401
Intention-in-action (IIA)	*'she went back again to the zoo'*	350	072201
Perception (PER)	*'she saw a huge lot of tigers and other animals'*	53	072201
Narrative Report of Speech Act (NRSA)	*'she did not ask the teacher about it'*	15	033401
Direct Speech (DS)	*'then said the sister* [...] *where is my little brother now'*	74	072201
Free Direct Speech (FDS)	*'oh little brother where are you now'*	14	072201
Indirect Speech (IS)	*'she said that they had to stop swimming'*	5	114201
Free Indirect Speech (FIS)	NA	NA	NA
Narrative Report of Mental State (NRMS)	*'he did not like that'*	98	061401
Viewpoint Package (VP)	*'because he entered secretly'*	44	101901
Direct Thought (DT)	*'then he thought I want to protect her'*	17	052901
Free Direct Thought (FDT)	*'shall I make some invitations for her friends'*	1	052901
Indirect Thought (IT)	*'the family thought that they were safe'*	17	112301
Free Indirect Thought (FIT)	*'he could wish for everything that he now wants'*	15	014901

Table 5.2: Fourteen types of CPR and their frequencies as found in our sample.

representation in its different forms, where access to a character's mind is relied on by default.[7] This goes for Direct Thought (DT) and Indirect Thought (IT) alike, even though in the latter case the contents of the character's thoughts are rendered in the narrator's words (see also the examples in Table 5.2). Narrative Report of Mental State (*NRMS*) is an ambiguous type in this respect; it can sometimes imply access to a character's mind, but in other cases reflect the narrator's reading of a mental state 'from the outside' (viz. characterising someone as 'happy' can be based on their behaviour as well as on narratorial access to their inner life).

[7] In classic narrative theory this is referred to as narrator omniscience; cf. Margolin (2014). Furthermore, for an extensive discussion of FIS and FIT as forms mixing elements of direct and indirect representation, see Vandelanotte (2009).

Looking at frequencies in the representation of speech and thought, it is apparent that DS is the most used type of speech representation (74 occurrences), whereas the much more indirect NRMS is most frequent (98) in representing thought. Finally, the type Viewpoint Package (VP), recorded 44 times, is introduced by us based on recent work by Van Duijn and Verhagen (2018) that we found useful in our children's story context. In short, Viewpoint Packages are single words implying a mental state contrasting with a state of affairs or with another mental state. For example, if a character does something 'secretly', this implies that there is a perspective from which this is *not noticed* and a perspective from it is indeed *desired* that it remains unnoticed.

We follow Nicolopoulou and Richner (2007) in their analysis suggesting that, for a storyteller, IIA and PER require less advanced efforts on a cognitive level, compared to handling character speech representation. Dealing with character thought, in turn, is argued to be more advanced on a cognitive level than handling speech, for exactly the reason discussed in the preceding paragraph: thought representation requires the narrator to intrude into character minds, whereas speech representation does not. Following this analysis, plus our own analysis of ego-narration, the order in which we present the fourteen types in Table 5.2 can be seen as indicating different stages, ranging from preliminary (EGO−NARR), to basic (IIA, PER), to intermediate (NRSA, (F)DS, (F)IS), to advanced (NRMS, VP, (F)DT, (F)IT).

Hypotheses

First it is our aim to explore variation in the use of CPR types within our sample as a whole. Second, we hypothesise that the occurrence of these types is not uniformly distributed over age groups. From the idea that some CPR types can be seen as more advanced than others, as we discussed in the previous section, we predict that preliminary and basic types of CPR occur more often at younger ages, while intermediate and advanced types are more often found in older children. Third, we aim to explore links between CPR types and linguistic information extracted using NLP tools. We predict that more advanced types of CPR are more likely to co-occur with utterances exhibiting higher lexical and syntactic complexity.

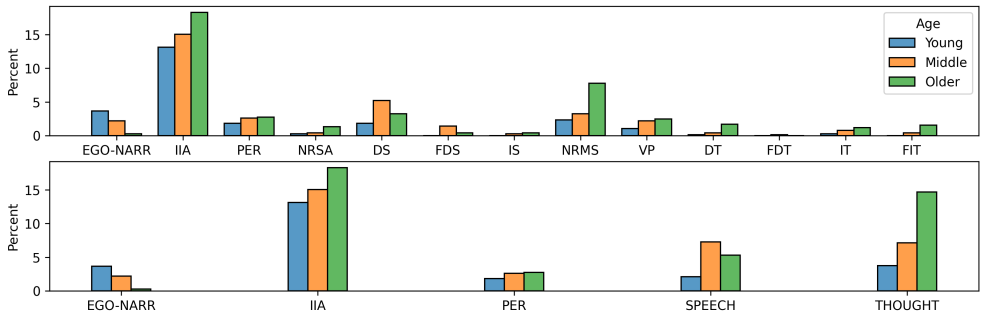


Figure 5.1: Occurrence of the original (top) and merged (bottom) CPR types in stories by children from three age groups, in percentages. Young: 4-6y, Middle: 6-9y, and Old:9-12y. Bars stack to 100%.

Type	O_{young}	O_{middle}	O_{old}	E	χ^2	p
EGO−NARR	28	17	2	15.67	21.74	<.001*
IIA	96	114	140	116.67	8.39	.015
PER	12	20	21	17.67	2.75	.252
SPEECH	12	55	41	36	26.72	<.001*
THOUGHT	28	53	111	64	56.66	<.001*

Table 5.3: Observed frequencies (O), expected values (E), and χ^2 statistics with $df = 2$ for all merged CPR types. Since we run 5 separate tests on the same variable, α was set to $.05/5 = .01$. Asterisks indicate $p < \alpha$.

5.4 Results

Development

For statistical analyses of the observed counts we merged CPR types that are theoretically closely related. In line with the stages discussed above, NRSA, DS, FDS, IS, and FIS were grouped as SPEECH, and NRMS, VP, DT, FDT, IT, and FIT as THOUGHT. CPR as found in our sample is plotted for both the five merged and thirteen original types in Figure 5.1. We conducted several χ^2 goodness-of-fit tests to probe whether observed frequencies for a given CPR type differed significantly from a uniform distribution among the three age groups. Test statistics and p-values are given in Table 5.3, with

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}, df = k - 1. \tag{5.2}$$

We see that younger children use a lot more EGO−NARR, but older children a lot less compared to the expected value E; the distribution is significantly different from

Lexical Complexity				Syntactic Complexity			
Age	$\bar{x}$	σ	N	Age	$\bar{x}$	σ	N
Young	5.72	.63	176	Young	2.69	.85	176
Middle	5.96	.65	259	Middle	2.70	.97	259
Older	5.99	.63	315	Older	2.75	.85	315

Table 5.4: Descriptive statistics for lexical complexity and syntactic complexity, for a total of 750 utterances featuring CPR from 150 stories, 50 stories per age group.

uniform. This suggests children 'outgrow' ego-narration, which we argued is a preliminary stage of CPR, and as hypothesised it seems to disappear from children's storytelling as they get older. For both IIA and PER, which we called basic types of CPR, the distributions do not differ significantly from uniform. Thus, there are no age-specific preferences among children for either IIA or PER, contra our hypothesis that these basic types occur mainly at young age.

With regard to SPEECH, the distribution among age groups is significantly different from uniform. We see little use among young children compared to the expected value E, but a peak in use at middle age and then a slight decrease in use for the older group. This supports our hypothesis that SPEECH, which we argued is an intermediate type of CPR, is increasingly used at a later age. THOUGHT is significantly different from uniformly distributed and seems to take off rather late. The younger and middle groups use less THOUGHT compared to the expected value E, whereas the older group uses it a lot more. This pattern observed regarding THOUGHT offers clearest support for our prediction that advanced types of CPR are increasingly employed at a later age.

In summary, children of all ages in our sample tell stories in which character perspectives are represented. As children grow older, perspectives of their characters tend to be fleshed out in more diverse and advanced ways. For the middle group we observe that characters more often speak and have various kinds of thoughts and other mental states. The older group relies even more often on forms of thought representation, and slightly less on character speech; possibly using the first partly instead of the latter.

Linguistic contexts

Next we examine whether more complex types of CPR co-occur with utterances that are lexically and syntactically more complex. We automatically extracted Lexical Complexity (LC) and Syntactic Complexity (SC) for each utterance.

		Lexical Complexity		Syntactic Complexity	
	Effect	Estimate	*SE*	Estimate	*SE*
Simple	(Intercept)	5.759*	0.138	2.363*	.182
	IIA	.211	.151	.327	.204
	PER	-.028	.228	.733*	.318
	SPEECH	.331	.238	.002	.329
	THOUGHT	.344	.186	.419	.255
	Middle	.055	.239	-.051	.306
	Older	1.114*	.470	.602	.673
Interaction	IIA x Middle	.091	.254	.198	.331
	PER x Middle	.216	.328	-.201	.448
	SPEECH x Middle	.172	.323	.248	.431
	THOUGHT x Middle	.049	.285	-.014	.377
	IIA x Older	-.896	.474	-.433	.682
	PER x Older	-.575	.521	-.689	.748
	SPEECH x Older	-1.022*	.515	-.313	.739
	THOUGHT x Older	-1.136*	.488	-.770	.700
Random	Child (intercept)	.103	.012	.076	.010
	Residual	.329	.021	.744	.031

Table 5.5: Terms for two linear mixed models with by-child varying intercepts. EGO−NARR and Young age group are the reference classes, i.e. the intercept is the average perplexity/tree depth for an utterance of a young child with the Ego-narrator type. Asterisks indicate $p < .05$.

For LC, we first lemmatised utterances with the spaCy parser (Honnibal and Johnson, 2015), and calculated the lexical perplexity; for SC, we also used the spaCy parser to extract the maximum depth of the parsed tree, as described in Section 5.3. Means and standard deviations of the extracted features are given in Table 5.4. As can be seen, average differences for both lexical and syntactic complexity are small across the three age groups. Next, we employed LC and SC as dependent variables in two linear mixed models. We included our five merged types of CPR as categorical predictors and included interactions with our three age groups, to find out whether different CPR types have significantly different mean LC and SC values, while taking potential age differences into account. Coefficients are given in Table 5.5. Our overall finding is that the link between lexical and syntactic complexity and specific types of CPR is not as we anticipated.

We first discuss the results for LC. Here we see that the only significant simple effect is Older, which means that with respect to the young EGO−NARR reference class, older children use ego-narration in utterances that are lexically more complex than young children do. Further, we see two significant negative interactions with SPEECH and THOUGHT, indicating that as we ascend from our reference class to older children

that use these intermediate and advanced forms of CPR, the lexical complexity of the utterances decreases, which is contrary to what we hypothesised with respect to LC. We do not see evidence for our hypothesis that average LC for more complex types of CPR is higher compared to ego-narration, while taking age differences into account.

Next we elaborate on our results for SC. Here we see no evidence for our hypothesis that more complex forms of CPR co-occur in utterances that have higher average syntactic complexity, while taking age differences into account. Main and interaction effects are all insignificant, except PER as simple effect, which implies that with respect to our young EGO−NARR reference class, average SC is higher when young children employ PER as type of CPR. This is contrary to what we hypothesised, as PER is a basic CPR type which we expected to co-occur with less complex syntax.

Our results are not in line with earlier work suggesting that children's more advanced lexical and syntactical skills co-occur with better socio-cognitive skills (as reviewed in e.g. Milligan et al., 2007). One possibility is that the way we looked at lexical and syntactic information in utterances here, provides a too limited view on the contexts in which different types of CPR occur. Given that other studies demonstrate that lexical complexity on the level of the entire stories children tell, predicts the occurrence of more sophisticated story characters (see Chapter 2 and Chapter 4) we suggest that automatically extracted information on the story level (as opposed to the utterance level only) could be more helpful for modelling CPR occurrence in the future.

5.5 Discussion

Our inventory shows that CPR is ubiquitous in freely told children's stories and that it takes many different forms. We discussed that classification of perspective phenomena into a system of CPR types requires knowledge of linguistic and narratological theory, and that it is regularly dependent on thorough analysis of utterance context within a story. Reliance on a single annotator is a weakness of this study; however, we believe to have met the goal of building a meaningful (foundation of a) bridge between long-standing research traditions in the humanities and current approaches in the computational sciences.

Regarding ego-narration we have identified cases exhibiting a problematic mixing between children's own perspective and the narrator's or characters' perspectives in the story, and argued for seeing these as a preliminary stage of CPR. Also, building on existing work from developmental psychology and cognitive linguistics, we have

introduced the types IIA, PER, and VP in our analysis, covering perspectives implied in actions, perceptions, and single lexical units such as 'secretly'. This was particularly useful for getting a grasp on the more basic stages of perspective coordination as present in our sample of children's stories. Although we did not see occurrence of these basic stages peak at younger ages, as we expected, we presented evidence that indeed more complex types are implemented more frequently at later ages.

Furthermore, our aim was to link automatically extracted linguistic information to the occurrence of different types of CPR, while also taking age differences into account. The picture that emerged for lexical and syntactic complexity was more complicated than we anticipated. By taking into account dependencies between utterances coming from the same speaker by using random intercepts, and by including interactions with age in our statistical models, we tried to describe as much variation as possible in the language children use when rendering character perspectives. As we saw, overall average differences in lexical and syntactic complexity between ages were small at the outset, and we were not able to link higher linguistic complexity to advanced types of CPR. Here the overall sparse occurrence of several of the individual types likely calls for exploiting a larger part of our story database in the future. We also learned that using perplexity and tree depth to describe the immediate (utterance-level) contexts in which CPR types occur, is challenging, suggesting that additional types of linguistic information from wider (story-level) contexts could be needed.

All in all, these findings and lessons encourage us to pursue the line of inquiry set out in this chapter. This will also require refining our framework, models, and automatically extracted information in interaction with linguistic and narratological theory, for which additional interdisciplinary cooperation is indispensable.

Chapter 6

Analysing Semantic Development with a Language Model

In this chapter we employ a Language Model (LM) to gain insight into how complex semantics of Dutch Perception Verb (PV) *zien* ('to see') emerge in children. Using a Dutch LM as representation of mature language use, we find that for ages 4-12y 1) the LM accurately predicts PV use in children's freely told narratives; 2) children's PV use is close to mature use; and 3) complex PV meanings with attentional and cognitive aspects can be found. Our approach illustrates how LMs can be meaningfully employed in studying language development, hence takes a constructive position in the debate on the relevance of LMs in this context.

This work was originally published as: Van Dijk, B.M.A., Van Duijn, M.J., Kloostra, L., Spruit, M.R., and Beekhuizen, B.F. (2024). Using a Language Model to Unravel Semantic Development in Children's Use of a Dutch Perception Verb. In Zock, M., Chersoni, E., Hsu, Y., and De Deyne, S., editors, *Proceedings of the 8th Workshop on Cognitive Aspects of the Lexicon*, pages 98-106. European Language Resources Association.

6.1 Introduction

Recent Language Models (LMs) based on Transformer architectures (Vaswani et al., 2017) reflect semantic knowledge present in a language community. BERT vectors (Devlin et al., 2019), for example, are able to distinguish different senses of the same word (Rogers et al., 2020; Vulić et al., 2020; Wiedemann et al., 2019). These LMs implement the distributional hypothesis that words with similar meanings tend to occur in similar contexts, and they represent both word type and word token meanings with real-valued vectors (Lenci and Sahlgren, 2023). The latter allows LMs to encode polysemy and different usages of words.

Despite this, LMs' relevance in the context of language development is disputed: their architecture and volume of training input have been argued to make them incomparable to children (e.g. Bunzeck and Zarrieß, 2023; Prystawski et al., 2022; Warstadt and Bowman, 2022). Yet, others argue that LMs can show which linguistic phenomena are *in principle* learnable from distributional information, bearing on learnability debates (Contreras Kallens et al., 2023; Piantadosi, 2023; Wilcox et al., 2023).

Here we leverage LMs' rich semantic information to gain insight in children's semantic and pragmatic development. Addressing the question whether children's pragmatic use of lexical items develops over time or, conversely, is adult-like from the start, we use a Dutch LM as representation of mature language use and study the Dutch Perception Verb (PV) *zien* ('to see'). We find that children's use of *see* is close to mature use across the 4-12y age range, and that for all ages the familiar mature usage patterns of the verb can be identified. As such, this chapter further illustrates the relevance of LMs in studying language development, by reflecting on LMs as representations of mature language use and setting up appropriate tasks and metrics.

6.2 Background

Little empirical work employs modern LMs in language development, the exception being work comparing word acquisition in children and LMs (Chang and Bergen, 2022; Laverghetta Jr and Licato, 2021). This is understandable given the debate on the validity of LMs in the child context: LMs and children differ in key respects including word exposure (Warstadt and Bowman, 2022) and learning mechanisms (Bunzeck and Zarrieß, 2023).

Still, LMs are arguably useful representations of *mature language use* by being

trained on corpora of adult language, and are therefore of value in modelling language understanding. LMs can be viewed as an incremental methodological step compared to earlier corpus studies comparing children's verb use to mature use, that relied on manual annotation or feature engineering to identify different senses of mature verb use (e.g. Adricula and Narasimhan, 2009; Parisien and Stevenson, 2009), but different senses, as we will show, can also be conveniently retrieved from LMs. These and other considerations have led to increasing acknowledgement of LMs' relevance for analysing language development (Contreras Kallens et al., 2023; Lappin, 2023), and efforts to make LMs more comparable to the child context (Warstadt et al., 2023).

Here we address the relevance of LMs in the developmental context by analysing children's lexical semantic development with LMs. We target children's use of Dutch PV *zien* ('to see') as a case study, which has been frequently analysed in language development (e.g. Davis, 2020; Davis and Landau, 2021). Studies of perception verbs across languages have shown that visual perception verbs have extended meanings beyond their denotational meaning 'entity X visually perceives object or event Y', that involve additional aspects of e.g. *attention* ("Let's see if I can find the keys") and *cognition* ('I see what you mean') (San Roque et al., 2018; San Roque and Schieffelin, 2019). Such meaning extensions are salient for children with a limited lexicon, where meaning extension of known words allows children to express new meanings efficiently (Nerlich and Clarke, 1999). In addition, since visual perception is argued to have strong metaphorical mappings to knowledge and understanding (e.g. Johnson, 1999), *see* can be a window onto how children learn to represent (socio-)cognitive content with language (Sweetser, 1990).

This work addresses the question of when meaning extension occurs. Some argue that literal understandings of PVs emerge first in young children (e.g. Davis, 2020; Davis and Landau, 2021; Elli et al., 2021; Landau and Gleitman, 2009), while others argue pragmatic meanings are likely present early due to the social situatedness of language learning (e.g. Enfield, 2023; San Roque and Schieffelin, 2019). In the latter case, the discursive relation between the visual perception event and the events surrounding it may be more salient for a language learner than the encoding of visual perception per se. For example, a young child's utterance *see ball* may be followed by the caregiver showing the ball, or focusing its attention on the ball — further attentional aspects that are likely relevant components of the message for the child beyond the denotational content of visual perception having taken place. While focusing on a single verb may seem limited, we believe as a case study, visual perception verbs are well-chosen as a starting point for generalising the proposed approach,

since their acquisitional pathway and pragmatic usages (as described above) are well understood.

We focus on children's use of *see* in ChiSCor, a corpus of freely told stories by Dutch children (4-12y) in classroom settings (for details see Chapter 3), since complex PV meanings can be especially relevant in the narrative domain. For example, that character X *sees* entity Y may not only imply that X literally perceives Y, but also that X *evaluates* Y or *discovers* Y. Such information, which may be crucial for the 'tellability' of the story (Labov and Waletzky, 1967), can be efficiently transmitted through PVs. Narratives are 'natural' sandboxes for children to challenge their language competence in various ways (Frizelle et al., 2018), including the development of lexical pragmatics.

6.3 Methods

Language data

We extracted all 308 occurrences of *see* from 619 stories of 442 children (4-12y) in ChiSCor. We manually inspected these occurrences and removed unintelligible usages (mainly transcription errors) as well as stories exceeding a context window larger than 512 tokens, resulting in 210 occurrences. We assigned occurrences to a Young (4-6), Middle (6-9) or Old (9-12) age group, following the age binning in Dutch primary education, and included only PV occurrences from one story per child, resulting in 30 Young, 82 Middle and 42 Old PV occurrences. To balance the sample across age groups, we randomly sampled 30 occurrences from the Middle and Old age group.

A known problem with LMs is that data contamination can lead them to solve tasks by memorisation (Deng et al., 2024). ChiSCor is likely not in the train data of recent LMs, as the corpus is recent and 'hidden' behind view-only links in research papers. Further, ChiSCor's free storytelling is unlike other available Dutch corpora that involve language elicitation and as such constitutes language that tests LMs' generalisation capabilities.

LMs as benchmark models

Using LMs as representation of mature language use requires evidence that the LM models the linguistic phenomenon and domain at issue reliably. We draw on findings that word representations in BERT encode rich semantic information about word polysemy (Garí Soler and Apidianaki, 2021; Wiedemann et al., 2019), although not per-

fectly. Also, Dutch LMs are for a large part trained on narrative texts (e.g. De Vries et al., 2019; Delobelle et al., 2020), and LMs in general have been shown to model coherence in written narratives well (Laban et al., 2021). In sum, earlier work supports the idea that LMs encode mature PV use in narratives.

Choice of LMs

For reasons of computational efficiency, validity with respect to the child context, and reproducibility, we chose `RobBERT-2023-dutch-large`, a Dutch BERT-like LM (Delobelle et al., 2020). RobBERT has 455M parameters trained on 19.5B tokens and is more in line with the 100M token training input a 10-year-old has seen (Warstadt and Bowman, 2022), compared to often employed larger LMs like GPT-3 (175B parameters, 500B tokens (Brown et al., 2020)).[1] RobBERT is accessible through the `HuggingFace Transformers` ecosystem (Wolf et al., 2019).

Recent work on LM relevancy to human language acquisition in the BabyLM challenge (Warstadt et al., 2023), highlighted smaller LMs with optimised architectures and train objectives, and curated train data for training developmentally plausible models (Samuel et al., 2023). However, such Dutch LMs are not yet available and training models from scratch is generally not feasible for researchers studying language acquisition. RobBERT was a fitting resource as it is optimised compared to BERT and has a simpler training objective (masked language modelling only) (Liu et al., 2019). These aspects go some way towards the findings of the BabyLM challenge (Samuel et al., 2023; Warstadt et al., 2023).

Task design and metrics

To use LMs as representations of mature language use, zero-shot evaluation settings as described by Laban et al. (2021) are preferred. This means using LMs of-the-shelf without further pre-training on the target domain or fine-tuning to stay close to the mature language use encoded in the LM, similar to how factual knowledge can be retrieved from LMs without fine-tuning (Petroni et al., 2019). We use various possibilities available through LMs to assess whether and how children's use of *see* differs from mature use.

Our first task consists of predicting *see* in children's narratives. We present RobBERT with stories containing a masked instance of *see*, as in the (translated) excerpt

[1] In the context of this chapter scale differences between BERT and GPT-3 are most salient, but we acknowledging that GPT-3 as unidirectional decoder-only model is also qualitatively different from BERT-like models like RobBERT.

in (1):

(1) [...] one time robot was travelling. and all of a sudden he <**mask**> a wolf. and he ran away quickly. [...] (Story ID 052301)

In our experiment we provided full stories as context to RobBERT, which varied in number of words ($\bar{x} = 187, \sigma = 108$). If children's usage differs from adults, the LM might have difficulty predicting the PV correctly.

As a second measure, we compute the negative log-likelihood NLL or surprisal for a prediction for a masked token w_m with

$$NLL(w_m) = -\log p(w_m | w_{1\ldots m-1}, w_{m+1\ldots n}) \tag{6.1}$$

with the `fill-mask` pipeline from `HuggingFace Transformers`. This measure provides further context to the predictive accuracy measure presented above: lower NLL implies that the predicted token is less surprising i.e. closer to mature use as encoded in the LM, and more generally indicates how well a given context supports a specific token on the masked position (PV or other).

Lastly, we use the tokens in RobBERT's top-5 predictions for masked instances of *see* as 'near neighbours' that can reveal the additional discursive meanings that the usage of PVs supports. Our data and notebooks are available at `https://osf.io/8eyvf/`.

6.4 Results

Predictive accuracy

First, we assessed RobBERT's overall performance in predicting *see* at masked positions in all 90 PV occurrences. Accuracy is overall high (.83, Table 6.1), and although lower for Young (.70) we found no significant difference in accuracy between ages with an ANOVA ($F_{2,87} = 2.974, p = .056$).

This shows that RobBERT models children's PV use in the narrative domain well. The 15 errors were mainly in Young and showed confusion of *seeing* with 'finding', 'having', 'looking' and 'getting', meaning that contexts underconstrained the use of *see*. Although these other verbs can be valid tokens on masked positions (e.g. 'found' in (1)), here our aim was to see if RobBERT adequately models that *see* can subsume such other possible meanings in narratives.

Metric	Young	Middle	Old	Overall
Accuracy	.70 (30)	.90 (30)	.90 (30)	.83 (90)
Surprisal	.40 (21)	.23 (27)	.32 (27)	.31 (75)
Top-5	1.00 (30)	1.00 (30)	.97 (30)	.99 (90)

Table 6.1: Metrics for RobBERT. Accuracy: percentage that *see* was predicted. Surprisal: NLL computed for predictions of *see*. Top-5: proportion that *see* was in top-5 predictions. Number of PV occurrences (i.e. observations) in parentheses.

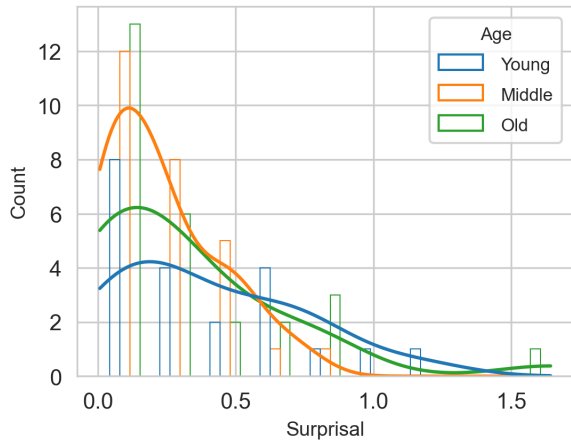


Figure 6.1: Surprisal distributions.

Surprisal

Second, we analysed potential age effects in mean surprisal for 75 correct predictions of *see*. For example, RobBERT may be less surprised by PV use for Old compared to Young or Middle, indicating that PV use of Old children is closer to mature use than for Young. Interestingly, surprisal distributions tend to 0 for all ages (Figure 6.1), suggesting that use of *see* is overall close to mature use irrespective of age. And although mean surprisal between Young, Middle, and Old differs (Table 6.1), pairwise comparisons with Tukey's HSD (Tukey, 1949) revealed no significant age effects. This indicates that PV use by children of all ages is about equally close to mature use as approximated by RobBERT.

Top-5 alternative predictions

For virtually all age groups, *see* is in the top-5 predictions (Table 6.1), which supports the idea that by examining top-5s we get insight into extended meanings of *see*. For

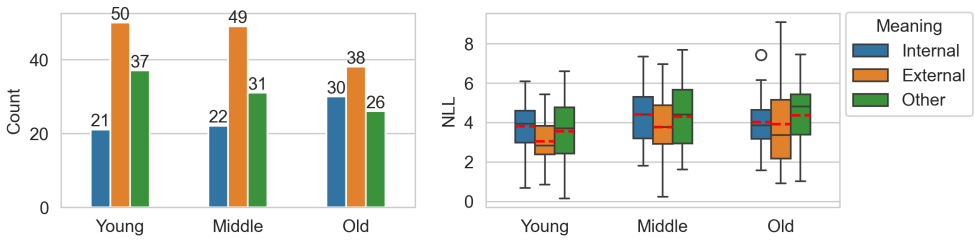


Figure 6.2: Frequencies (left) and surprisal dist. (right) of internal, external, and other meanings of 304 top-5 lemmas. Bars (left) stack to 100%; dashed red lines (right) indicate means.

90 PV occurrences and their top-5s (450 tokens) we lemmatised tokens and removed *see* and lemmas that were not verbs (e.g. 'many', 'and', 'at'), resulting in 304 lemmas. We then took the set and classified 65 lemmas as having roughly `external`, `internal`, or `other` meaning. `External` implies a meaning pertaining to plain action (e.g. 'to go', 'to come', 'to carry', 'to throw'); `internal` a meaning pertaining to an attentional (e.g. 'to notice', 'to meet') or cognitive state (e.g. 'to think', 'to know'). `Other` pertains to auxiliary verbs and PVs not the focus of the current study (e.g. 'to have', 'to hear').

The idea is that top-5 lemmas indicate what possible meanings PV contexts support, even if these lemmas are not necessarily intuitive substitutions. For example, substituting 'threw' for <**mask**> in (1) renders the excerpt less intuitive. Yet, this immediate context as a sequence of *external* actions better supports understanding *seeing* also as a causal part of a sequence of external actions, than as *seeing* as part of narrative components reflecting a character's attentional or cognitive *internal* states (cf. examples in Table 6.2).

We assessed frequencies of `external`, `internal` and `other` meanings, and their mean surprisal over age groups to identify potential age differences in occurrence and closeness to mature use. Regarding frequency, although `external` and `other` meanings decrease over age while `internal` meanings increase over age (Figure 6.1, left), we found no significant age effects with a χ^2 test of independence $\chi^2(4, N = 304) = 5.044, p = .283$, suggesting that all the different meanings are about equally frequent in Young, Middle and Old groups. Regarding surprisal (Figure 6.2, right), distributions for `external`, `internal` and `other` meanings are relatively similar both within and between age groups. Pairwise comparisons with Tukey's HSD found only a significant difference at the $p < .05$ level between mean surprisal for `external` meanings for Young and Old.

Age	Ex.	PV context
Young	**(2)**	*.. and when he returned. then he saw/knew that the princess was gone. and they lived happily ever after.* (102901)
	(3)	*.. and then they were lost again. and then they saw/searched the castle. and then they went in the castle.* (122901)
	(4)	*.. but then the teacher came and then she was already too late. the teacher had seen/caught them. and then you get a punishment from the teacher.* (033401)
Middle	**(5)**	*.. but then they lost each other all of a sudden. and then Wergje saw/met another rabbit. and it asked how are you called.* (072301)
	(6&7)	*.. because when he was home. then he saw/noticed/discovered that he had the other scales. but then he went to fly on it and he wanted to find his own dragon again.* (022301)
Old	**(8)**	*.. once arrived at the cave Puta completely forgot that you were not allowed to touch the big diamond. Puta saw/checked out the diamond and found it so beautiful. and he touched it accidentally.* (034801)
	(9)	*.. so then the fat little king went on his fat broom to the cry for help. and what did he see/think. the cry came from a little fat guinea pig that looked very much like the king.* (023801)
	(10)	*.. and he ever wanted one time to try it with his eyes closed. to see/test can I grab that donut well with my eyes closed.* (034501)

Table 6.2: Translated PV contexts with top-5 internal lemmas (underlined) with lowest surprisal. Story IDs given in parentheses. All excerpts were translated from the original stories.

We illustrate complex meanings of *see* present in all age groups, by providing the three `internal` meanings that were closest to mature use (i.e. with lowest surprisal) and their PV contexts in Table 6.2. We make three observations. First, `internal` meanings with attentional and cognitive aspects can be but are not exclusively cued by surface linguistic frames such as complementation that RobBERT simply picks up, as example (4) and (9) show. In (4) 'caught' implies that the teacher knows what the 'she' character is up to; in (9) 'think' renders the realisation where the cry of help is coming from a representation in the mind of the king. Second, `internal` meanings are varied: from more purely attentional where characters simply become aware of something or find something out as in (6&7), to more social (5), and evaluative attentional aspects (8). Third, although `internal` meanings with cognitive aspects have the most abstract lemmas ('think', 'know') that are argued to be harder to master (Barak et al., 2012), cognitive meanings were found in both Young (2), (4) and Old (9) children.

6.5 Discussion

Our results show that complex meanings of the Dutch perception verb *zien* ('to see') are about equally frequent in all age groups and that children's use of the PV is overall not significantly different from mature use. This contrasts with earlier work that has argued that children initially acquire more literal meanings of PVs (Adricula and Narasimhan, 2009; Davis, 2020; Davis and Landau, 2021; Elli et al., 2021; Landau and Gleitman, 2009) (Section 6.2), although we note that children in our sample are older (four years and older) than children in earlier studies (typically between two and four years).

Our result aligns with the idea that it is the social context that cues various complex senses of *see* in children (e.g. Enfield, 2023; San Roque and Schieffelin, 2019), and with the idea that (young) children may employ PVs like *see* as linguistic devices for learning to represent cognitive and attentional states (Johnson, 1999; Sweetser, 1990). We argue that our finding can be explained by the social context provided by live storytelling. PVs like *see* are linguistic devices for efficiently communicating about characters' attentional and cognitive states that are key to understanding the story, as PVs can compress redundant information that would make the story tedious. An earlier chapter has shown that in children's live storytelling, contexts of PVs like *hear* and *see* are coherent and clear, as evidenced by the rich PV vectors that can be trained from limited amounts of narrative data (Chapter 3).

Narrative language data may explain the contrast between our and earlier findings, as storytelling has been argued to solicit 'maximal behaviour' in that it challenges children's linguistic competence (Frizelle et al., 2018; Southwood and Russell, 2004), more than the speech produced by children in child-caregiver interactions would do, which typically take place in mundane contexts. Some earlier work contrasting with our results relied on language data from such child-caregiver interactions (e.g. Adricula and Narasimhan, 2009; Davis and Landau, 2021). The latter work also employed smaller sample sizes with less unique children and more PV use per child compared to the current study, which may compress the variation in complex semantics we find in our analysis.

Interestingly, RobBERT accurately predicted *see* in narratives of children of all ages; we argue that this is not a mere frequency effect (i.e. *see* being more frequent in train data than alternatives), given that top-5 predictions often reveal RobBERT's correct mapping of the nuanced senses of PVs. Also, RobBERT's aptitude in handling PV use in narratives is interesting insofar children's stories are not obvious regarding

wording, characters and themes. One issue pointed out by a reviewer is whether LMs with Transformer architectures are the best fit for representing linguistic knowledge of a mature Dutch language user, or whether other models should be used, e.g. from the BabyLM challenge (Warstadt et al., 2023). The best-performing LMs in this challenge employed Transformer architectures that are essentially optimised versions of vanilla BERT models regarding training objective, architecture and dataset (Samuel et al., 2023). With our choice for RobBERT we aimed to make the comparison to the human case as valid as possible with an existing resource (see Section 6.3).

In any case, from the BabyLM challenge we learn that the Transformer architecture is also in more modest training setups a powerful encoder of linguistic information. Our claim is not that Transformers are therefore good (cognitive) models of human language users, which is still debated (see e.g. Paape, 2023, and Chapter 8). Rather, when it comes to specific linguistic aspects such as mature semantic and pragmatic knowledge, LMs as sophisticated distributional learners represent this information in a convenient fashion. For using such computational models as representations of mature language use, the primary question is if their *behaviour* for a specific linguistic phenomenon is sufficiently complex, which for many modern BERT-like models seems the case. But representations of mature use could also be created in other ways, e.g. by clustering different verb senses with features based on verb argument structure in a large corpus of mature language use. Thus, LMs are more of an analytical tool here than direct models of humans. That said, it is still worthwhile and necessary to make LMs more similar to the human context.

6.6 Conclusion

This chapter provided a case study on Dutch children's (4-12y) use of *zien* ('to see') and the emergence of complex semantics in the use of this perception verb. We showed that 1) a recent Dutch LM can predict use of *see* in narratives produced by children of different ages reliably; 2) children's use of *see* is close to mature use for all ages; and 3) complex meanings of *see* with attentional and cognitive aspects can be found across all ages. Our results align with work that argues that meaning extension occurs early in children and with the idea that via perception verbs, children may learn to represent socio-cognitive content.

We also showed how LMs can be meaningfully leveraged in developmental contexts. We hope to provide future researchers with useful reflection on how to proceed when using LMs as representations of mature language use, choosing models, and

setting up tasks and metrics.

6.7 Limitations

A limitation of this study is that we provided the whole story as context for predicting a masked occurrence of *zien* ('to see'), but for space limitations we could only discuss complex meanings with smaller story excerpts as in Table 6.2. This may suggest that complex PV meanings can be determined from small pieces of narrative after all. Yet, when doing the same task with smaller PV contexts as in Table 6.2, i.e. a sentence before and after the sentence featuring an occurrence of *see*, RobBERT's overall accuracy drops from .83 to .57 and overall surprisal increases from .31 to .59 (see Table 6.1), which suggests that RobBERT needs to take the whole story into account to model PV use adequately. This means that there is more relevant information in the context beyond what we show in the immediate PV context that renders RobBERT's predictions of masked tokens accurate and supports additional meanings of *see*.

Another limitation is that we had to translate story excerpts into English, as also providing Dutch excerpts required too much space. Some awkwardness in translations could not be avoided. For example, Dutch has a verb 'betrappen' that always has a cognitive meaning similar to 'catching somebody red-handed', whereas 'catching' in English can also have a more obvious action-related meaning. 'Betrappen' was a token prediction in RobBERT's top-5 with low surprisal that we had to translate as 'caught' in example (4) in Table 6.2.

6.8 Ethics Statement

In this study we used the ChiSCor story corpus and we refer to Chapter 3 for further details regarding ethical considerations and approval that was obtained for collecting language data from children. Regarding computational efficiency, we chose a relatively small, open and free to use Large Language Model that can also be employed with limited computational resources.

Chapter 7

Theory of Mind in Large Language Models

To what degree should we ascribe cognitive capacities to Large Language Models (LLMs), such as the ability to reason about intentions and beliefs known as Theory of Mind (ToM)? Here we add to this emerging debate by (i) testing 11 base and instruction-tuned LLMs on capabilities relevant to ToM beyond the dominant false-belief paradigm, including non-literal language usage and recursive intentionality; (ii) using newly rewritten versions of standardised tests to gauge LLMs' robustness; (iii) prompting and scoring for open besides closed questions; and (iv) benchmarking LLM performance against that of children aged 7-10y on the same tasks. We find that instruction-tuned LLMs from the GPT family outperform other models, and often also children. Base-LLMs are mostly unable to solve ToM tasks, even with specialised prompting. We suggest that the interlinked evolution and development of language and ToM may help explain what instruction-tuning adds: rewarding cooperative communication that takes into account interlocutor and context. We conclude by arguing for a nuanced perspective on ToM in LLMs.

This work was originally published as: Van Duijn, M.J.,* Van Dijk, B.M.A.,* Kouwenhoven, T.,* De Valk, W.M., Spruit, M.R., and Van Der Putten, P.W.H. (2023). Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art Models vs. Children Aged 7-10 on Advanced Tests. In Jiang, J., Reitter, D., and Deng, S., editors, *Proceedings of the 27th Conference on Computational Natural Language Learning*, pages 389-402. Association for Computational Linguistics. (* denotes equal contribution.)

7.1 Introduction

Machines that can think like us have always triggered our imagination. Contemplation of such machines can be traced as far back as antiquity (Liveley and Thomas, 2020), and peaked with the advent of all kinds of 'automata' in the early days of the Industrial Revolution (Voskuhl, 2019) before settling in computer science from the 1950s (Turing, 1950). Currently people around the world can interact with powerful chatbots driven by Large Language Models (LLMs), such as OpenAI's ChatGPT (Achiam et al., 2024), and wonder to what degree such systems are capable of thought.

LLMs are large-scale deep neural networks, trained on massive amounts of text from the web. They are vastly complex systems: even if all details about their architecture, training data, and optional fine-tuning procedures are known (which is currently not the case for the most competitive models), it is very difficult to oversee their capabilities and predict how they will perform on a variety of tasks. Researchers from linguistics (Manning et al., 2020), psychology (Binz and Schulz, 2023; Kosinski, 2024; Webb et al., 2023), psychiatry (Kjell et al., 2023), epistemology (Sileo and Lernould, 2023), logic (Creswell et al., 2023), and other fields, have therefore started to study LLMs as new, 'alien' entities, with their own sort of intelligence, that needs to be probed with experiments, an endeavour recently described as 'machine psychology' (Hagendorff, 2023). This not only yields knowledge about what LLMs are capable of, but also provides a unique opportunity to shed new light on questions surrounding our own intelligence (Binz and Schulz, 2024; Dillion et al., 2023).

Here we focus on attempts to determine to what degree LLMs demonstrate a capacity for Theory of Mind (ToM), defined as the ability to work with beliefs, intentions, desires, and other mental states, to anticipate and explain behaviour in social settings (Apperly, 2012). We first address the question **how LLMs perform on standardised, language-based tasks used to assess ToM capabilities in humans.** We extend existing work in this area, surveyed in Section 7.2, in four ways:

1. By testing 11 LLMs (see Table 7.1) for a broader suite of capabilities relevant to ToM beyond just the dominant false-belief paradigm, including non-literal language understanding and recursive intentionality (A *wants* B to *believe* that C *intends*...);

2. By using newly written versions of standardised tests with varying degrees of deviation from the originals;

3. By including open questions besides closed ones;

4. By benchmarking LLM performance against that of children aged 7-8 (n=37) and 9-10 (n=36) on the same tasks.

Section 7.3 contains details of our test procedures for both children and LLMs. After reporting the results in Section 7.4, we turn to the question **how variation in performance of the LLMs we tested can be explained** in Section 7.5. We conclude by placing our findings in the broader context of strong links between language and ToM in human development and evolution, and tentatively interpret what it means for a LLM to pass (or fail) ToM tests.

We are aware of issues regarding LLM training and deployment, for example regarding the biases they inherit (Bender et al., 2021; Lucy and Bamman, 2021), problems for educators (Sparrow, 2022), and ethical concerns in obtaining human feedback (Perrigo, 2023). Ongoing reflection on the use of LLMs is necessary, but outside the scope of this chapter.

7.2 Background

Large language models

The field of Natural Language Processing (NLP) has been revolutionised by the advent of the Transformer architecture (Devlin et al., 2019; Vaswani et al., 2017) in deep neural networks that can induce language structures through self-supervised learning. During training, such models iteratively predict masked words from context in large sets of natural language data. They improve at this task by building representations of the many morphological, lexical, and syntactic rules governing human language production and understanding (Grand et al., 2022; Manning et al., 2020; Rogers et al., 2020). Models exclusively trained through such self-supervision constitute what we refer to as 'base-LLMs' in this chapter.

Base-LLMs can generate natural language when they are prompted with completion queries ('A mouse is an ...'). They can also be leveraged successfully for an array of other challenges, such as question-answering and translation, which often requires task-specific fine-tuning or prompting with specific examples, known as few-shot-learning (Brown et al., 2020). This makes them different from a new generation of LLMs that we refer to as 'instruct-LLMs' in this chapter, and to which the currently most competitive models belong. In instruction-tuning, various forms of human

feedback are collected, such as ranking most suitable responses, which then forms the reward signal for further aligning these models to human preferences through reinforcement learning (Ouyang et al., 2022). The resulting LLMs can be prompted with natural language in the form of instructions to perform a wide variety of tasks directly.

A key realisation is thus that LLMs are given either no explicitly labelled data at all, or, in the case of instruct-LLMs, data with human labels pertaining to relatively general aspects of communicative interaction. As such they are part of a completely different paradigm than earlier language models that were trained on, for example, datasets of human-annotated language structures (e.g. Nivre et al., 2016). This means that when LLMs are capable of such tasks as solving co-reference relationships or identifying word classes (Manning et al., 2020), this arises as an *emergent* property of the model's architecture and training on different objectives. Given that such emergent linguistic capabilities have been observed (Grand et al., 2022; Reif et al., 2019), it is a legitimate empirical question which other capacities LLMs may have acquired as 'by-catch'.

Theory of Mind in humans and LLMs

ToM, also known as 'mindreading', is classically defined as the capacity to attribute mental states to others (and oneself), in order to explain and anticipate behaviour. The concept goes back to research in ethology in which Premack and Woodruff (1978) famously studied chimpanzees' abilities to anticipate behaviour of caretakers. When focus shifted to ToM in humans, tests were developed that present a scenario in which a character behaves according to its *false beliefs* about a situation, and not according to the reality of the situation itself — which a successful participant, having the benefit of spectator-sight, can work out.

Initial consensus that children could pass versions of this test from the age of 4 was followed by scepticism about additional abilities it presumed, including language skills and executive functioning, which led to the development of simplified false-belief tests based on eye gaze that even 15 month old children were found to 'pass' (Onishi and Baillargeon, 2005). While this line of research also met important criticism (for a review see Barone et al., 2019), it highlights two key distinctions in debate from the past decades: implicit-behavioural versus explicit-representational and innate versus learned components of ToM. Some researchers see results from eye-gaze paradigms as evidence for a native or very early developing capacity for

belief-attribution in humans (Carruthers, 2013) and hold that performance on more complex tests is initially 'masked' by a lack of expressive skills (cf. also Fodor, 1992). Others have attempted to explain eye gaze results in terms of lower-level cognitive mechanisms (Heyes, 2014) and argued that the capacity for belief attribution itself develops gradually in interaction with more general social, linguistic, and narrative competencies (Heyes and Frith, 2014; Hutto, 2008; Milligan et al., 2007). Two-systems approaches (Apperly, 2012) essentially reconcile both sides by positing that our ToM capacity encompasses both a basic, fast, and early developing component and a more advanced and flexible component that develops later.

In computational cognitive research, a variety of approaches to modelling ToM has been proposed (e.g. Arslan et al., 2017; Baker et al., 2011). More recently neural agents (Rabinowitz et al., 2018b) have been implemented, along with an increasing number of deep learning paradigms aimed at testing first- and second-order ToM via question-answering. Initially this was done with recurrent memory networks (Grant et al., 2017; Nematzadeh et al., 2018) using datasets of classic false-belief tests from psychology, but after issues surfaced with simple heuristics for solving such tasks, scenarios were made more varied and challenging (Le et al., 2019). From the inception of BERT as one of the first language models (Devlin et al., 2019), we have seen roughly two approaches for testing ToM in LLMs: many different ToM scenarios integrated in large benchmark suites (e.g. Ma et al., 2023a; Sap et al., 2022; Shapira et al., 2024; Sileo and Lernould, 2023; Srivastava et al., 2023), and studies that modified standardised ToM tests as used in developmental and clinical research for prompting LLMs (e.g. Brunet-Gouet et al., 2023; Bubeck et al., 2023; Chowdhery et al., 2022; Kosinski, 2024; Marchetti et al., 2023; Moghaddam and Honey, 2023; Ullman, 2023). This chapter adds to the latter tradition in four respects, as explained in the introduction.

7.3 Methods

Here we describe our tasks and procedures for testing LLMs and children.[1]

Theory of Mind tests

Sally-Anne test, first-order (SA1) – The Sally-Anne test (Baron-Cohen et al., 1985; Wimmer and Perner, 1983) is a classic first-order false belief test. It relies on a narrative in which Sally and Anne stand behind a table with a box and a basket on it.

[1] All code, materials, and data are available on OSF: `https://osf.io/426p9/`.

When Anne is still present, Sally puts a ball in her box. When Sally leaves, Anne retrieves the ball from the box and puts it in her own basket. The story ends when Sally returns and the participant is asked the experimental question 'Where will Sally look for the ball?' The correct answer is that she will look in her box. We followed up by asking a motivation question, 'Why?', to prompt an explanation to the effect of 'she (falsely) believes the object is where she left it'.

Sally-Anne test, second-order (SA2) – While SA1 targets the participant's judgement of what a character *believes* about the location of an unexpectedly displaced object, in SA2 the participant needs to judge what a character *believes* that *another character believes* about the location of an ice cream truck (Perner and Wimmer, 1985). Sally and Anne are in a park this time, where an ice cream man is positioned next to the fountain. Anne runs home to get her wallet just while the ice cream man decides to move his truck to the swings. He tells Sally about this, but unknown to her, he meets Anne on the way and tells her too. Sally then runs after Anne, and finds her mother at home, who says that Anne picked up the wallet and went to buy ice cream. The experimental question now is 'Where does Sally think Anne went to buy ice cream?', with as correct answer 'to the fountain', also followed up with 'Why?', to prompt an explanation to the effect of 'Sally doesn't know that the ice cream man told Anne that he was moving to the swings'.

Strange Stories test (SS) – The Strange Stories test (Happé, 1994; Kaland et al., 2005) depicts seven social situations with non-literal language use that can easily be misinterpreted, but cause no problems to typically developed adults. To understand the situations, subjects must infer the characters' intentions, applying ToM. For example, in one of the test items a girl wants a rabbit for Christmas. When she opens her present, wrapped in a big enough box, it turns out that she received a pile of books. She says that she is really happy with her gift, after which subjects are asked the experimental question 'Is what the girl says true?', with correct answer 'No'. They can motivate their answer after the question 'Why does she say this?', with as correct answer "to avoid her parents' feelings being hurt". Items increase in difficulty and cover a lie, pretend play scenario, practical joke, white lie (example above), misunderstanding, sarcasm, and double bluff.

Imposing Memory test (IM) – The Imposing Memory test was originally developed by Kinderman et al. (1998), but the test has been revised several times; we rely on an unpublished version created by Anneke Haddad and Robin Dunbar (Van Duijn, 2016), originally for adolescents, which we adapted thoroughly to make it suitable for children aged 7-10y. Our version features two different stories, fol-

lowed by true/false questions, 10 of which are 'intentionality' and 12 of which are 'memory' questions. For instance, in one story Sam has just moved to a new town. He asks one of his new classmates, Helen, where he can buy post stamps for a birthday card for his granny. When Helen initially sends him to the wrong location, Sam wonders whether she was playing a prank on him or just got confused about the whereabouts of the shop herself. He asks another classmate, Pete, for help. As in the original IM, the intentionality questions involve reasoning about different levels of recursively embedded mental states (e.g. at third-level: 'Helen *thought* Sam *did not believe* that she *knew* the location of the store that sells post stamps'), whereas the memory questions require just remembering facts presented in the story (to match third-level intentionality questions, three elements from the story are combined, e.g. 'Sam was looking for a store where they sell post stamps. He told Pete that he had asked Helen about this').

Testing procedures

Scoring – For both children and LLMs test scores were determined in the following way. For each of the SA1 and SA2 items, as well as for the seven SS items, a correct answer to the experimental question yielded 1 point. These answers were discrete and thus easy to assess ('box', 'fountain', 'no', etc.). For the motivation question a consensus score was obtained from two expert raters, on a range from 0-2 with 0 meaning a missing, irrelevant, or wrong motivation, 1 meaning a partly appropriate motivation, and 2 meaning a completely appropriate motivation that fully explained why the character in each scenario did or said something, or had a mental or emotional mental state. Thus, the maximum score for the SA1, SA2, and SS was 3 points per item, which were averaged to obtain a score between 0 and 1. For each correct answer to a true/false question in the IM, 1 point was given, and IM scores were averaged over its items as well. All scores and ratings can be found on OSF.

Deviations – We tested the LLMs on the original SA and SS scenarios, but also on manually created deviations that increasingly stray from their original formulations, to prevent LLMs from leveraging heuristics and memorising relevant patterns from the training data. Thus, deviations probe the degree to which performance on ToM tests in LLMs generalises. Deviation 0 was always the original test scenario (likely present in the training data); deviation 1 was a superficial variation on the original, e.g. with only objects and names changed (similar to Kosinski (2024)), whereas deviation 2 was a completely new scenario where only the ToM-phenomenon at issue

	LLMs	Source	Size
Base	Falcon	Penedo et al. (2023)	7B
	LLaMA	Touvron et al. (2023)	30B
	GPT-davinci	Brown et al. (2020)	175B
	BLOOM	Scao et al. (2022)	176B
Instruct	Falcon-instruct	Penedo et al. (2023)	7B
	Flan-T5	Chung et al. (2024)	11B
	GPT-3 (text-davinci-003)	Ouyang et al. (2022)	175B
	GPT-3.5-turbo	Ouyang et al. (2022)	175B
	PaLM2	Anil et al. (2023)	175-340B
	PaLM2-chat	Anil et al. (2023)	175-340B
	GPT-4	Achiam et al. (2024)	>340B

Table 7.1: LLMs used in this study. Model sizes are undisclosed for GPT-4 and for PaLM2 and PaLM2-chat, thus we base ourselves on secondary sources for estimations; Knight (2023) and Elias (2023), respectively.

was kept constant (e.g. 'second-order false belief' or 'irony'). Since our adaptation of the IM test has hitherto not been used or published, we did not include deviations for this test.

Testing LLMs

We leveraged 11 state-of-the-art LLMs: 4 base-LLMs and 7 instruct-LLMs (see Table 7.1). Inference parameters were set such that their output was as deterministic as possible (i.e. a temperature $\approxeq$ zero or zero where possible) improving reproducibility. Each inference was done independently to avoid in-context learning or memory leakage between questions. This means that for each question, the prompt repeated the following general structure: [`instruction`] + [`test scenario`] + [`question`].

Instruct-LLMs were prompted in a question-answering format that stayed as close as possible to the questionnaires given to children, without further custom prompting or provision of examples. Instructions were also similar to those given to children (e.g. 'You will be asked a question. Please respond to it as accurately as possible without using many words.'). The 'Why'-questions in SA1 and SA2 were created by inserting the experimental question and answer the LLM gave into the prompt: [`instruction`] + [`test scenario`] + [`question`] + [`LLM answer`] + [`'Why?'`]. This was not necessary for SS, given that experimental and motivation questions could be answered independently.

For base-LLMs, known to continue prompts rather than follow instructions, staying this close to the children's questionnaires was not feasible. For the SA and SS we

therefore fed base-LLMs the scenario as described before, but formulated the questions as text-completion exercises (e.g. 'Sally will look for the ball in the '). Additionally, when creating the motivation questions for SA1 and SA2, we inserted the *correct* answer to the experimental question, instead of the LLM's answer. This was because base-LLMs so often derailed in their output that the method described for instruct-LLMs did not yield sensible prompts. Base-LLMs thus had an advantage here over children and instruct-LLMs, who were potentially providing a motivation following up on an incorrect answer they gave to the experimental question.

For the closed questions in the IM we attempted to streamline the output of base-LLMs by including two example continuations in the desired answer format. These examples were based on trivial information we added to the scenarios, unrelated to the actual experimental questions. For example: 'Helen: I wear a blue jumper today. This is [incorrect]', where it was added in the story that Helen wears a green jumper. This pushed nearly all base-LLM responses towards starting with '[correct]' or '[incorrect]', which we then assessed as answers to the true/false questions. We considered a similar prompt structure for SA and SS, amounting to adopting few-shot learning for base-LLMs throughout (Brown et al., 2020), but given that reformulating questions as text-completion exercises was by itself effective to get the desired output format, we refrained from inserting further differences from how instruct-LLMs were prompted. It is important to note that our prompts were in general not optimised for maximal test performance, but rather designed to stay as uniform and close to the way children were tested as possible, enabling a fair comparison among LLMs and with child performance.

Testing children

Children were recruited from one Dutch and one international school in the South-West of the Netherlands: 37 children in the younger group (7-8y) and 36 children in the older group (9-10y). Children were administered digital versions of the SA and SS for the younger group, and of the IM for the older group, which they completed individually on tablets or PCs equipped with a touch screen. Test scenarios and questions were presented in a self-paced text format and all SA and SS questions were followed by an open text field in which they had to type their answer. As the IM features long scenarios, voice-overs of the text were included to alleviate reading fatigue. Here children had to answer by pressing yes/no after each question. To reduce memory bottlenecks, accompanying drawings were inserted (see Figure 1.2 in

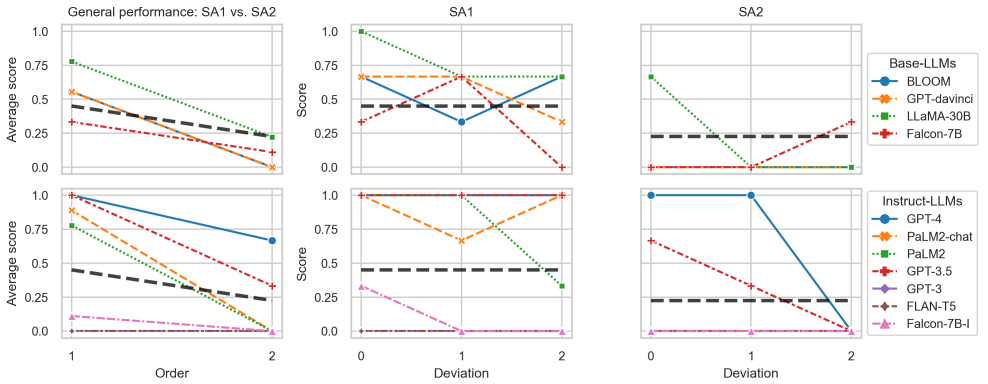


Figure 7.1: Performance on Sally-Anne tests for base-LLMs (top row) and instruct-LLMs (bottom row). Left column depicts performance on first- and second-order ToM (i.e. SA1 vs. SA2), averaged over the original and rewritten test versions (deviations). Middle and right columns depict performance for SA1 and SA2 over levels of deviation from the original test (0, 1, and 2 as explained in Section 7.3). Dashed black lines indicate average child performance (n=37, age 7-8 years).

the Introduction) and navigating back and forth throughout the tests was enabled. Informed consent for each child was obtained from caregivers, and the study was approved by the Leiden University Science Ethics Committee (ref. no. 2021-18). Test answers were evaluated and scored parallel to the approach for LLMs.

7.4 Results

Sally-Anne Test

Overall performance on SA1 versus SA2 is given in Figure 7.1, left column. Most base-LLMs perform above child level on first-order ToM (BLOOM, Davinci, LLaMA-30B) but fall at or below child level on second-order ToM. A similar pattern is visible for instruct-LLMs: most models perform well above child level on first-order (GPT-4, GPT-3.5, PaLM2-chat, PaLM2), but not on second-order ToM. Exceptions are GPT-4 and GPT-3.5: while degrading on second-order, they remain above child level. For both base- and instruct-LLMs, smaller models tend to perform worse (Falcon-7B, Falcon-7B-I, FLAN-T5) with GPT-3's structurally low scores as striking exception. This is inconsistent with results reported by Kosinski (2024) for GPT-3, which is probably due to the fact that Kosinski applied a text-completion approach whereas we prompted GPT-3 with open questions.

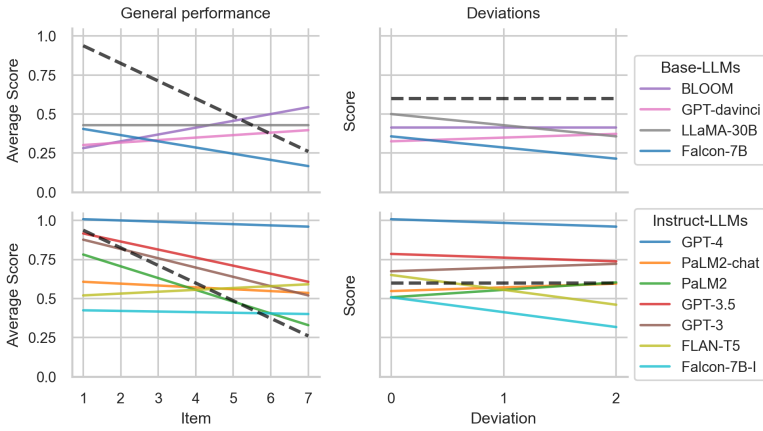


Figure 7.2: Performance on Strange Stories for base-LLMs (top row) and instruct-LLMs (bottom row). Left column shows overall performance, averaged over deviations from the original test. Right column shows performance over deviations, averaged over items. Dashed black lines indicate average child performance (n=37, 7-8y).

When we consider the performance on SA1 and SA2 over deviations (middle and right columns in Figure 7.1), we see once more that almost all LLMs struggle with second-order ToM, since performance decreases already on deviation 0 (i.e. the original test scenario), except for GPT-3.5 and GPT-4. Yet, it is the *combination* of second-order ToM and deviation 2 that pushes also GPT-3.5 and GPT-4 substantially below child levels, except for Falcon-7B, although the instruction-tuned version of this model (Falcon-7B-I) fails on all second-order questions.

Strange Stories Test

General performance on SS is given in Figure 7.2, left column. Whereas child performance declines as items become more complex (from 1 to 7; see Section 7.3), this is overall less the case for LLM performance. As a result, all models surpass child level at some point, except for the smallest model, Falcon-7B. All base-LLMs score below child level on most items but perform above child level on the most difficult ones, except Falcon-7B. For instruct-LLMs, we see that GPT-4 approaches perfect scores throughout. GPT-3 and GPT-3.5 perform at or close to child level on item 1, after which their performance declines somewhat, while staying well above child level. Other instruct-LLMs show a mixed picture: PaLM2-chat and FLAN-T5 surpass child

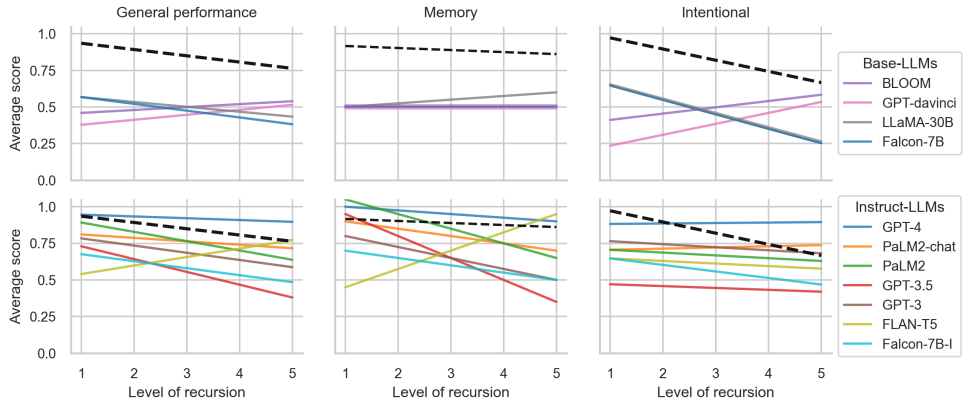


Figure 7.3: Performance on Imposing Memory test for base-LLMs (top row) and instruct-LLMs (bottom row). Left column depicts overall performance over five levels of recursion, averaged over deviations. Middle and right columns depict performance for Memory and Intentional questions. Dashed lines indicate average child performance (n=36, 9-10y).

level earlier than PaLM2. Interestingly, smaller FLAN-T5 outperforms larger PaLM2 and PaLM2-chat on more difficult items. Falcon-7B-I, as smallest instruct-LLM, performs overall worst.

If performance is plotted over deviations (right column in Figure 7.2) we see little impact on most base-LLMs. For instruct-LLMs, it is striking that deviation levels have almost no effect on the larger models (GPT-4, PaLM2, PaLM2-chat, GPT-3, GPT-3.5), but do more dramatically lower performance of smaller models (FLAN-T5, Falcon-7B-I). In sum, base-LLMs perform below child level, except for the most complex items. Several large instruct-LLMs match or surpass child level throughout, others only for more complex items. Unlike for the SA test, deviation levels seem to have little negative impact for SS.

Imposing Memory Test

The classical finding for the IM test is that error rates go up significantly for questions involving higher levels of recursive intentionality, but not for memory questions on matched levels of complexity, suggesting a limit to the capacity for recursive ToM specifically (Stiller and Dunbar, 2007).[2]

[2]While there is consensus in the literature that higher levels of intentionality are significantly harder for participants than lower levels, by various measures, there is debate about the difference with memory

We verified this for our child data (n=36) with two mixed linear models for memory and intentional questions with random intercepts. We included five predictors that were contrast-coded such that each predictor indicated the difference in average IM performance with the previous level. For intentional questions, only the difference between level two and one was significant ($\beta = -0.222, p < .05$), marking a cut-off point after which performance remained consistently low. For memory questions, performance remained high across all levels ($>.85$), except for level four, where scores were significantly lower than at level three ($\beta = -0.292, p < .00$), but went up again at level five ($\beta = 0.208, p < .00$). Thus, in line with earlier work, we find a cut-off point after which scores on intentionality questions remained consistently low, compared to scores on matched memory questions. We have no clear explanation for the dip in performance on memory questions at level four, but observe that it is driven by low scores on only one specific question out of a total of four for this level, which children may have found confusing.

In Figure 7.3 we see that all base-LLMs perform below child level, in general and on both intentionality and memory questions, and there is little variation in performance, except that larger base-LLMs (BLOOM, GPT-davinci) improve on higher levels of recursion. Regarding instruct-LLMs, we see largely the same picture, as they almost all perform below child level, in general and on both types of questions. The exception is GPT-4, which performs consistently well on all levels and stays above child level after first-order intentionality. For the difference between memory and intentional questions, instruct-LLMs perform better on easier memory questions, and drop towards the end, while on intentional questions, they already start lower and stay relatively constant. Lastly, it is remarkable that FLAN-T5, as one of the smallest instruct-LLMs, overall increases performance as recursion levels go up, and ends at child level. For GPT-3.5, which performs worst of all instruct-LLMs on this task, we see the exact opposite.

Notes on child performance

It can be observed that performance for SA was overall low compared to what could be expected from children aged 7-8 years: $\bar{x} = 0.45$ for SA1 and $\bar{x} = 0.23$ for SA2. We have two complementary explanations for this. Firstly, as discussed in Section 7.3, children had to read the tests on a screen, after which they had to type answers in open text fields. This is a challenging task by itself that relies on additional skills in-

questions; see e.g. Lewis et al. (2017). For a critical discussion of measuring recursive intentionality in general, see Wilson et al. (2023).

cluding language proficiency, conscientiousness, digital literacy, and more. Secondly, whereas 'passing' originally only means that a child can work out where Sally will look (for the ball, or for Anne on her way to buy ice cream), we also asked for a motivation, which makes the test more demanding. For the SS test, completed by the same group of children, we see the expected pattern that scores show a downward tendency as test items become increasingly difficult. The older group, aged 9-10, completed the IM. As discussed above, IM scores resonate with earlier work. Given that we see child performance not as the central phenomenon under observation in this chapter, but rather as a reference for LLM performance, further discussion is outside our scope.

7.5 Discussion

Summing up the results for the Sally-Anne tests, while it is less surprising that base-LLMs and smaller instruct-LLMs struggle with increasing test complexity and deviations, it is striking that second-order ToM immediately perturbs some large instruct-LLMs (e.g. PaLM2-chat), and that adding deviations from the original test formulations pushed down performance of even the most competitive models (e.g. GPT-4, GPT-3.5). This initially suggests that performance on ToM tasks does not generalise well beyond a few standard contexts in LLMs, in line with earlier work (Sap et al., 2022; Shapira et al., 2024; Ullman, 2023).

For the Strange Stories test we saw that base-LLMs perform generally below child level. Most instruct-LLMs perform close to or above child level, particularly as items become more complex, and child performance drops much more dramatically than LLM performance. Levels of deviation from the original test formulation seem to have made almost no impact for the SS test, suggesting that the capacity to deal with non-literal language targeted by the Strange Stories*does* generalise to novel contexts. We conclude that instruct-LLMs are quite capable at interpreting non-literal language, a skill that in humans involves ToM.

Since the training data of LLMs includes numerous books and fora, which are typically rich in irony, misunderstanding, jokes, sarcasm, and similar figures of speech, we tentatively suggest that LLMs are in general well-equipped to handle the sort of scenarios covered in the Strange Stories. This should in theory include base-LLMs, but it could be that their knowledge does not surface due to the test format, even after specialised prompting. Going one step further, we hypothesise that Sally-Anne is generally harder for LLMs given that this test relies less on a very specific sort of

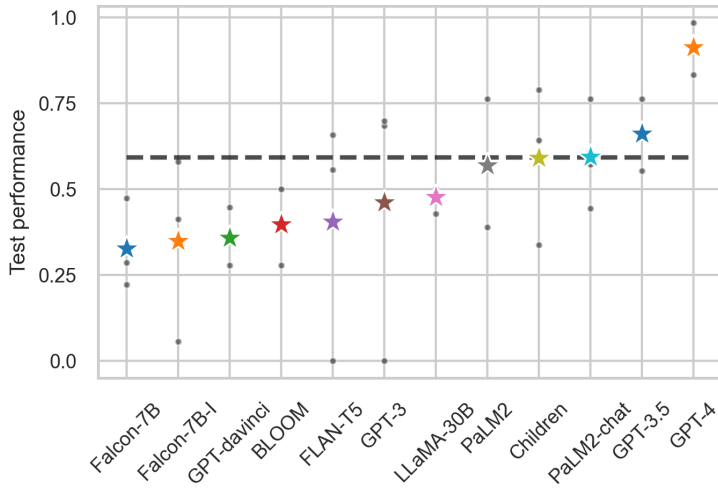


Figure 7.4: Grand mean performance (stars) of all mean test scores (dots) for children and LLMs.

advanced language ability, but more on a type of behaviourally-situated reasoning that LLMs have limited access to during training (see also Mahowald et al., 2024). The Imposing Memory test was the most challenging for both base- and instruct-LLMs. Since our version of this test was never published before, it constitutes another robustness test, which only GPT-4 as largest instruct-LLM seems to pass well.

The gap between base- and instruct-LLMs is best summarised in Figure 7.4. Here we see that no base-LLM achieves child level: all LLMs approaching or exceeding child performance are larger instruct-LLMs. Our adapted prompts and insertion of correct answers for motivation questions for the SA test did not make a difference. We suggest that another issue for base-LLMs, besides the prompt format, was prompt length. This was highest for IM, which can explain why they struggled most with this test. Prompt length, in relation to the models' varying context window sizes and ability to engage in what Hagendorff et al. (2023) call chain-of-thought reasoning, merits further research (see also Liu et al., 2023). We tested whether there was a difference between model performance on closed versus open questions across all three tasks, but found no signal: the models that struggled with closed questions were also those that performed low on open questions (for more details see OSF).

Evidence is emerging that most LLM capacities are learned in the self-supervised pre-training phase (Gudibande et al., 2023; Ye et al., 2023), which suggests that base-

LLMs are essentially 'complete' models. Instruction-tuning, however, even in small amounts adds adherence to the desired interaction format and teaches LLMs, as it were, to apply their knowledge appropriately (Zhou et al., 2023a). We see a parallel between instruction-tuning and the role for *rewarding cooperative communication* in human evolution and development. It has been argued extensively that human communication is fundamentally cooperative in that it relies on a basic ability and willingness to engage in mental coordination (e.g Grice, 1975; Verhagen, 2015). It is a key characteristic of the socio-cultural niche in which we evolved that, when growing up, we are constantly being rewarded for showing such willingness and cooperating with others to achieve successful communicative interactions (Tomasello, 2008). Reversely, if we do not, we are being punished, explicitly or implicitly via increasing social exclusion (David-Barrett and Dunbar, 2016). This brings us back to our context: instruction-tuning essentially rewards similar cooperative principles, but punishes the opposite, which may amount to an enhanced capacity for *coordinating with an interaction partner's perspective*, in humans and LLMs alike. This is reflected in performance on ToM tasks, which are banking on this capacity too.

7.6 Conclusion

We have shown that the majority of recent LLMs operate below performance of children aged 7-10y on three standardised tests relevant to ToM. Yet, those that are largest in terms of parameters, and most heavily instruction-tuned, surpass children, with GPT-4 well above all other models, including more recent competitors like PaLM2-chat and PaLM2. We have interpreted these findings by drawing a parallel between instruction-tuning and rewarding cooperative interaction in human evolution. We conclude that researching the degree to which LLMs are capable of anything like thought in the human sense has only just begun, which leaves the field with exciting challenges ahead.

Chapter 8

Reflecting on Language and Cognition in Large Language Models

Current Large Language Models (LLMs) are unparalleled in their ability to generate grammatically correct, fluent text. LLMs are appearing rapidly, and debates on LLM capacities have taken off, but reflection is lagging behind. Thus, in this theoretical chapter, we first zoom in on the debate and critically assess three points recurring in critiques of LLM capacities: i) that LLMs only parrot statistical patterns in the training data; ii) that LLMs master formal but not functional language competence; and iii) that language learning in LLMs cannot inform human language learning. Drawing on empirical and theoretical arguments, we show that these points need more nuance. Second, we outline a pragmatic perspective on the issue of 'real' understanding and intentionality in LLMs. Understanding and intentionality pertain to unobservable mental states we *attribute* to other humans because they have *pragmatic value*: they allow us to abstract away from complex underlying mechanics and predict behaviour effectively. We reflect on the circumstances under which it would make sense for humans to similarly attribute mental states to LLMs, thereby outlining a pragmatic philosophical context for LLMs as an increasingly prominent technology in society.

This work was originally published as: Van Dijk, B.M.A., Kouwenhoven, T., Spruit, M.R., and Van Duijn, M.J. (2023). Large Language Models: The Need for Nuance in Current Debates and a Pragmatic Perspective on Understanding. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12641–12654. Association for Computational Linguistics.

8.1 Introduction

The performance of Large Language Models (LLMs) has recently reached high levels (see e.g. Bommasani et al., 2021; Mahowald et al., 2024). LLMs are deep neural networks with a Transformer architecture (Vaswani et al., 2017), trained to predict masked words from context, using massive text datasets.[1] During training, LLMs learn to represent input syntactically in hierarchical form, and they also learn semantic relations (Rogers et al., 2020), which are useful features in summarising, question-answering, and translating text. Examples of recent LLMs are LLaMA (Touvron et al., 2023), GPT-3 & 4 (Achiam et al., 2024; Brown et al., 2020), and PaLM (Chowdhery et al., 2022).

LLMs have sparked a lot of debate, inside and outside academia, around the question what their successes and failures say about linguistic capacities in AI systems, but also in humans. In the first part of this chapter, we scrutinise three key points recurring in arguments by experts critical of LLM capacities (e.g. Bender and Koller, 2020; Bender et al., 2021; Bisk et al., 2020; Browning and LeCun, 2022; Floridi, 2023; Mahowald et al., 2024; Marcus, 2022; Mitchell and Krakauer, 2023; Shanahan, 2024), although there are also experts who are more optimistic on this matter (e.g. Agüera y Arcas, 2022a,b; Berger and Packard, 2022; Cerullo, 2022; Sejnowski, 2022; Piantadosi and Hill, 2022; Piantadosi, 2023; Sahlgren and Carlsson, 2021). These points are:

1. that all LLMs can do is predict next words;
2. that LLMs can only master formal as opposed to functional language competence;
3. that language learning in LLMs cannot inform human language learning.

In the first part of this chapter, we aim to nuance these points and show that they are hard to maintain in the face of empirical work on LLMs and theoretical arguments. In the second part, this leads us to develop a pragmatist perspective on LLMs, for which we draw on work by Daniel Dennett, Richard Rorty, and others. 'Real' language understanding and intentionality consist of *attributions* of unobservable mental states, that humans make on the basis of observable behaviour. We do so because this has *pragmatic value*: it simplifies complex underlying biophysical processes and allows us

[1] Some LLMs additionally benefit from further fine-tuning, e.g. reinforcement learning from human feedback (Christiano et al., 2017). Since evidence is emerging that most of LLMs' capabilities are learned during pre-training (Gudibande et al., 2023; Ye et al., 2023; Zhou et al., 2023a), we abstract away from this aspect in this chapter.

to predict future behaviour. Instead of asking whether LLMs have 'real' understanding and intentionality, we ask under what circumstances regarding LLM behaviour and their role in society, it is reasonable for humans to make mental models of LLMs that include capacities like understanding and intentionality.

In sum, our aim is to contribute to a more realistic framework for understanding LLMs within academia and beyond, which is better grounded in empirical and philosophical work. Since LLMs' impact on research and society likely increases in the future, properly understanding them is key. Still, although we defuse various critiques of LLMs in this chapter, it is not our purpose to advocate their deployment without ongoing reflection on their implications. Examples include their environmental impact (Bender et al., 2021), biases (Lucy and Bamman, 2021), problems for educators (Sparrow, 2022), and ethical issues in adding human feedback (Perrigo, 2023), but these issues go beyond the scope of this chapter.

8.2 Theoretical Analysis

In this section, we qualify three key points from the debate regarding LLM capacities by drawing on theoretical and empirical work.

I. LLMs only predict next words

Shanahan (2024) claims that whenever we prompt a LLM, for example with

(1) *The capital city of the Netherlands is ...*

we actually ask

(2) *Given the statistical patterns you learned from dataset Y during training, what word is most likely to follow the provided sequence?*

and the answer to (1) is likely 'Amsterdam'. According to this assumption, this is all there is to it; we should not speak about the model's potential topographical knowledge, nor should we say that the model understands the question in any way comparable to how humans understand it. We can see similar claims in Bender et al. (2021), Chomsky (2023), Floridi (2023), Marcus (2022), and in weaker form in Mitchell and Krakauer (2023).

Although it is true that the good performance of LLMs on many tasks stems from a simple training objective, which is predicting masked words from context, we argue

that this point overlooks the complex ways in which LLMs are able to represent information. During training, LLMs induce various semantic and syntactic features that the model uses internally to *represent* the input in a manner that can be extracted, for example, by analysing model weights or patterns of neuronal activation. An example regarding syntax is that LLMs are able to hierarchically represent input (Hewitt and Manning, 2019; Mahowald et al., 2024; Manning et al., 2020; Rogers et al., 2020). That is, they are capable of internally parsing the example into syntactic chunks, such that the prepositional phrase ('of The Netherlands') provides information about the noun phrase ('The capital city'), which constitutes a clue to the answer ('Amsterdam'). In addition, regarding semantics, the vector representations of words that neural networks induce are shown to be context-sensitive and rich enough to capture conceptual relations in line with human judgements (Grand et al., 2022; Piantadosi and Hill, 2022; Reif et al., 2019). Moreover, in our example, 'The Netherlands' and 'Amsterdam' are likely geometrically related in the vector space of a LLM, which provides a further clue regarding the answer.

Our syntactic and semantic examples here are not necessarily the way LLMs represent the relevant linguistic information; it is not trivial to extract representations from LLMs (Rogers et al., 2020). The point is rather that LLMs are capable of further *representing* input in various ways that are *not reducible* to either their training data or objective (Piantadosi and Hill, 2022). While they are not explicitly trained to represent input hierarchically, or represent semantic relations, such properties emerge while becoming better at their relatively simple stochastic training objective (Manning et al., 2020). On second thought, this should not be too surprising, given that there is a lot of linguistic information 'hidden' in the web text used to train LLMs, which LLMs (partially) reconstruct. Note that this argument does not require LLMs to 'really' understand or know their inputs or representations.

The assumption that LLMs can only echo statistical regularities is important to qualify. So-called 'underclaiming' (downplaying what LLMs do and learn), resonates more broadly in the academic sphere, which could hinder studying how LLMs work in detail (Bowman, 2022), and exploring whether they are useful for studying questions about human language usage. The assumption likely stems from the idea that probabilistic modelling of language may successfully simulate or approximate linguistic facts (i.e. generate coherent language), but is of such different nature that it does not provide any further insight into human language (Norvig, 2012; Piantadosi, 2023). Although it may seem on the surface that LLMs are just language simulation machines, this overlooks all the linguistic complexity that is stored in the weights

that are updated during training, with word prediction providing a powerful supervision signal. Beyond the potential impact of 'underclaiming' in academic debates (Bowman, 2022), this assumption could reinforce simplistic views of what LLMs are and why they are useful in society at large.

II. LLMs master the form, not the function of language

The distinction between formal and functional language competence we draw on here stems from Mahowald et al. (2024): formal language competence concerns employing information about linguistic rules and patterns in producing coherent output, whereas functional language competence draws on further cognitive capacities, such as formal reasoning, intentional reasoning, and situation modelling. Mahowald et al. (2024) paraphrase this difference as the difference between being good at *language* and being good at *thought*; in their view, LLMs master language but not thought. They motivate this distinction with the finding that the two competences recruit independent brain circuits, and discuss persons with aphasia as a concrete example: they can have limited formal linguistic competence, yet still be able to compose music, solve logic puzzles, reason about other persons' mental states, thus, leverage thought independently of language. Whether one buys into its neural grounding or not, the distinction between formal and functional language competence as such is a useful one to make, and we see similar oppositions in Bender and Koller (2020), where the distinction is made between LLMs' mastery of linguistic form as opposed to extra-linguistic meaning, and in Bisk et al. (2020), Browning and LeCun (2022), and Floridi (2023).

Disentangling language and thought – Here we do not claim that LLMs have thought that can be meaningfully separated from language, as we are agnostic on this matter, but we question some of the methods currently used to disclose thought.

First of all, whereas persons with aphasia can be tested on their capacity to employ thought in a way that is clearly independent of language (e.g. composing music), for LLMs this is not possible. For example, for the common sense reasoning and intentional reasoning (a.k.a. Theory of Mind/ToM) humans do, two vital functional capacities, benchmarks inevitably rely on presenting a particular (social) situation using linguistic prompts (e.g. Binz and Schulz, 2023; Borji, 2023; Collins et al., 2022; Creswell et al., 2023; Kosinski, 2024; Sap et al., 2022; Ullman, 2023). Implicitly or explicitly, such works draw on the assumption that in LLMs, there must be a distinction between thought as internal symbolic system representing abstract relations, and lan-

guage as a mapping between these representations and their outward expression in text. Although this is not unreasonable to think, given that LLMs have many emergent capacities for which they were not explicitly trained (Section 8.1), currently we do not know how much formal linguistic information LLMs leverage when performing such tasks. LLM output could, for example, be the result of specific semantic or syntactic relations with the input, while it seems unlikely that a human would approach such tasks in the same way. Thus, in assessing thought in LLMs, language and thought are confounded.

This issue has an analogy in testing thought in children. When for instance testing ToM, confounding factors are always present, as the myriad tests of ToM that exist and the different modalities they solicit (vision, speech, text) illustrate (Quesque and Rossetti, 2020). General language and memory abilities of children are typically controlled with additional tasks (Milligan et al., 2007); few tests exist that rely on language alone (Beaudoin et al., 2020). Still, many seemingly superficial aspects shape performance on such tests, such as how questions are phrased (Beaudoin et al., 2020; Siegal and Beattie, 1991). The influence of superficial linguistic artefacts of tests can be controlled to some extent when conducting ToM tests with LLMs, for example, preventing memorisation by rewriting tests such that they are not in the training data (e.g. Kosinski, 2024; Shapira et al., 2024), but this is only the beginning of disentangling language and thought in LLM output.

Moreover, disentangling language and thought is difficult, because for many cognitive test we have an idea of what the test operationalises, but not when they are used in the context of LLMs. For the ToM context, one hallmark test is the 'unexpected contents' test (Perner et al., 1987), where a Smarties box with unexpected contents (e.g. a pencil) is shown to a child. The child is asked what a friend, unfamiliar with the box, would think its contents are, thereby asking it to manage two conflicting beliefs: the false belief imputed to the friend, and its own belief about the box' contents. This conflict management, as instantiation of ToM ability, is arguably what the test operationalises, and something humans know from subjective experience, which makes it easier to understand what was measured when such tests are used on humans. Yet, this is far less clear when using such tests on LLMs.

Thought is a continuum – Here we assume, for the sake of argument, that we can separate thought from language ability in LLMs with cognitive tests (which we questioned above). We further reflect on how various tests are currently being used to deny or affirm thought in LLMs. Again, we do not argue that LLMs have or lack thought here, but rather that we should suspend conclusions, based on the issues

raised below.

Quite some cognitive tests currently employed with LLMs make assumptions about thought that need qualification. Many are designed so that a LLM can only fail or succeed on them (e.g. (part of) tests employed by Borji (2023), Bubeck et al. (2023), Kosinski (2024), Sap et al. (2022), Shapira et al. (2024), and Ullman (2023)). Yet, thought capacity is better understood as a *continuum* (Beaudoin et al., 2020; Sahlgren and Carlsson, 2021). That is, thought capacity is unequally distributed in humans; people who excel in logic, may also be horrible composers, or may struggle to recognise other persons' intentional states.

In addition, we are typically much more lenient towards failure in exercising thought. In daily contexts, humans are generally susceptible to a host of misplaced heuristics and formal errors (Dasgupta et al., 2022; Haselton et al., 2015), but we generally do not conclude from this that humans do not have unique thought capacity. Sahlgren and Carlsson (2021) make a similar claim for language understanding: different language users are good at different things, at different times, in different situations. Thus, it is unsurprising that disagreement exists among NLP scholars about proper operationalisations of various tasks in Natural Language Understanding, such as Natural Language Inference (Subramonian et al., 2023), a disagreement that can also be anticipated for other cognitive tests.

To make our evaluations of thought in LLMs as compelling as possible, we can employ more sophisticated measures, and aim for more nuance in the interpretations of results, before we can deny (or affirm) thought in LLMs. Instead of focusing merely on (average) success or failure, knowing that a LLM has, for example, 51% confidence in a wrong answer is already more informative than just knowing the LLM erred. Fortunately, work on more nuanced evaluations of thought in LLMs is emerging (e.g. Binz and Schulz, 2023; Collins et al., 2022). Alternatively, we could evaluate the model's intermediate reasoning steps in solving a complex reasoning task, besides only the answer, by employing Chain-of-Thought prompting (Wei et al., 2022). Evidence is emerging that in the context of ToM tests, more sophisticated prompting improves performance (Moghaddam and Honey, 2023).

From a methodological perspective, testing thought in LLMs and humans differs a lot because the entities at issue differ a lot. Yet, we can improve the comparability of testing. A single output of a LLM on a single test item likely does not yield a good estimate of its capacities; in testing humans, we typically ask multiple humans to do the same item. Thus, we could for example initialise LLMs with slightly higher temperature values multiple times on the same test item, to get a fuller view

on LLMs' abilities, and to obtain a larger sample of responses on which statistical tests are possible. Although we know that low temperature settings make models deterministic, not much evaluation of slightly higher temperature settings in relation to performance has been done, with the exception of Moghaddam and Honey (2023). In addition, we know from work on knowledge extraction in LLMs, that paraphrasing a particular input improves model performance in retrieving knowledge and relations (Jiang et al., 2020). Paraphrasing test items is not only a way to increase model performance, but could also provide a way to increase our confidence in our estimates of thought capacity as indicated by LLM performance on multiple paraphrased items.

Lastly, from a more general perspective, failure of a LLM on a cognitive test does not imply that the system does not have the mappings required to do the test; the tests could also be less suited to retrieve them (Bommasani et al., 2021).

III. LLMs are irrelevant to human language acquisition

Bisk et al. (2020) argue that, since children cannot acquire a language by merely listening to the radio, it is likewise wrong to expect that LLMs can acquire language by purely ingesting text from the internet; similar claims are offered in Bender and Koller (2020) and Chomsky (2023). This point relies on the presumed poverty of 'extralinguistic' information in the train data of LLMs. In language acquisition, children draw not only on linguistic input but also on sense perception (e.g. seeing and touching the world), motor experience (e.g. moving objects), and interaction with caretakers (e.g. feedback). Also, children receive far less language input compared to LLMs (Warstadt and Bowman, 2022). Thus, if language acquisition in children and LLMs is so different from the start, it seems language acquisition in LLMs cannot inform language acquisition in humans.

LLMs as useful distributional models – LLMs and children evidently differ a lot. Yet, as Sahlgren and Carlsson (2021) formulate, LLMs are theoretically and practically our current 'best bet' for machines to acquire language understanding, given the empirical work documenting LLM proficiency in many language tasks (see e.g. Bommasani et al., 2021; Wei et al., 2023). Like any scientific model they are wrong in some respects, but they seem our current best *distributional* models to study *specific* aspects of language acquisition.

For example, Chang and Bergen (2022) use BERT and GPT-2 as distributional agents that exclusively learn from word co-occurrence statistics. They employ among other things word frequency, lexical class and word length, as known predictors of

word acquisition in children, and predict word acquisition in LLMs and children to gauge the extent to which these known effects in children can be accounted for by statistical learning mechanisms. Chang and Bergen (2022) show that language acquisition in language models and children differs in key respects (language models are more frequency-driven), but is also similar (learning in both takes longer for words embedded in longer utterances). As Chang and Bergen (2022) note, distributional models can be used in similar fashion to explore the extent to which the acquisition of semantics or syntax in children can be accounted for by statistical learning.

Cevoli et al. (2023) provide another example by unravelling lexical ambiguity with BERT. The authors show that psychological theories positing complex mechanisms for representing ambiguity, are not necessary to explain how such representations are acquired, since they can be decoded from distributional information in text. This illustrates another key role language models can play in language acquisition: as Warstadt and Bowman (2022) note, language models in ablation studies can show whether target linguistic knowledge (e.g. verb-subject agreement in triply embedded clauses) is learnable in an ablated environment (e.g. without triply embedded clauses in the training data). Such studies are helpful in identifying sufficient conditions for obtaining specific linguistic knowledge in language acquisition. Thus, the examples mentioned above echo the broader point made by various scholars that we should see LLMs as distributional learners that show what linguistic phenomena are *in principle* learnable from statistical information in text (Contreras Kallens et al., 2023; Evanson et al., 2023; Wilcox et al., 2023).

Warstadt and Bowman (2022) note that in language acquisition contexts, LLMs need to be less advantaged to humans in one key aspect: the amount of training data. That is, they need to be made more ecologically valid. The latter is indeed important, and fortunately, work is emerging which shows that it is possible to train LLMs with more realistic amounts of data, that at the same time perform equally well in predicting human neural and behavioural patterns as models trained with large datasets (e.g. Hosseini et al., 2022; Wilcox et al., 2023). Yet, it is equally remarkable that LLMs have become so successful, despite being very *disadvantaged* as well (no multimodal input, no feedback in learning, no sensorimotor input). We argue that it is not obvious how we should weigh such disadvantages and advantages in LLMs' language learning. LLMs make wrong assumptions about language acquisition in key respects, but all scientific models do this (Baayen, 2008; Box, 1976), while this does not render such models useless: they can provide a lower bound on what linguistic phenomena are learnable in principle from distributional information.

How poor is training data? – Here we discuss the assumption that data used to train LLMs lacks extralinguistic information required in language acquisition, by considering what extralinguistic information LLMs learn to represent during training. Since humans use language to do a variety of things (Sahlgren and Carlsson, 2021), such as providing explanations, describing all sorts of objects and processes, and entertaining and convincing others, it is natural to assume that LLMs can recover some of the knowledge about e.g. properties of objects in the world, communicative intents, and users' mental states. Recent work shows that LLMs are able to represent conceptual schemes for worlds (e.g. for direction) they have never observed (Patel and Pavlick, 2022), thus it seems that LLMs have a sufficiently rich conceptual structure to decode at least some of the extralinguistic information present in text, as a surrogate grounding. Similarly, Abdou et al. (2021) show that the internal representations of language models show a topology of colours that corresponds to human perceptual topology. In addition, evidence emerges that LLMs are able to represent communicative intents behind texts (Andreas, 2022), and the ways LLMs represent semantic features of various object concepts align with humans (Hansen and Hebart, 2022). Moreover, studies in which LLMs are trained and tested on synthetic tasks, provide an even stricter scenario for testing whether LLMs are able to decode emergent properties from simple input. LLMs trained on simple input such as lists of player moves in a board game, are able to recover emergent properties such as game rules, valid future moves, and board states (Li et al., 2022). For additional examples of extralinguistic grounding, see Bowman (2023).

8.3 A Pragmatic Perspective on LLMs

This section sketches a more general, pragmatist philosophical context for LLMs. Although LLMs are prominent in academia and society, philosophical reflection is lagging behind. This is lamentable, given that LLMs and the way they are deployed raise pressing philosophical questions. Here we develop a pragmatist view on LLMs with the following claims that we will motivate with reference to philosophical pragmatism:

1. All three key points from the debate about the capacities of LLMs as discussed above, ultimately revolve around the issue of 'real' understanding and intentionality, but fail to address what that means;

2. Once we try to explain what 'real' understanding and intentionality are, we

find that these (and mental states more generally) are not accessible in others we interact with, irrespective of whether they are humans or other kinds of systems;

3. Attributing mental states to others has foremost pragmatic value, in that they help us to abstract underlying complexity away, predict behaviour, and obtain goals in the world;

4. Given the increasing prominence of LLMs, interacting with them in terms of mental state attribution will likely become more common, yet lacks a comprehensive theory;

5. This practice is fully explainable from a pragmatist perspective, although in different communities, such as the scientific community, different pragmatist values may play a role, that makes this practice less acceptable for this community.

Invoking 'real' understanding

Various scholars have claimed that LLMs are incapable of 'really' understanding language and using intentionality like humans (e.g. Bender and Koller, 2020; Bishop, 2021; Browning and LeCun, 2022; Floridi, 2023; Mahowald et al., 2024). Indeed, it seems that the critiques of LLMs as autocomplete systems that do not know how language functions in the world, or as language learners that cannot learn by drawing on such functions, implicitly invoke this claim. Still, in such critical works it is seldom made explicit what 'real' understanding or intentionality amounts to. Such works often revolve around John Searle's 'Chinese Room' thought experiment (Searle, 1980). We illustrate this with the following quote from Bender and Koller (2020):

(3) *This means we must be extra careful in devising evaluations for machine understanding, as Searle (1980) elaborates with his Chinese Room experiment: he develops the metaphor of a "system" in which a person who does not speak Chinese answers Chinese questions by consulting a library of Chinese books according to predefined rules. From the outside, the system seems like it "understands" Chinese, although in reality no actual understanding happens anywhere inside the system.* (p. 5188)

The argument presented in (3) is that, for any system, being able to deliver the expected output on a range of inputs is insufficient for having 'actual understanding',

where it is important to note that the perspective of anyone interacting with the system is 'from the outside'. The issue with this thought experiment is that, although the idea of 'real' understanding is implied, it is not explained, which makes the argument incomplete.

Explaining 'real' understanding

The Chinese Room argument appeals to a situation where we *would* grant that the system understands Chinese: if the human in the system understands Chinese. That is, this human would need to have a set of *mental states* involving knowledge, beliefs, and intentions, such that in producing output, the human does not draw on predefined rules, but rather on its *knowledge* of Chinese, *beliefs* about the desired output, and further communicative *intent*. This would constitute an example of what is meant by a human having 'real' understanding. Nonetheless, this explanation cannot save the thought experiment as presented above, since it makes no difference for anyone interacting with the system if we would replace the rule-abiding human with the human as full mental agent, since from the outside, there is only the system's behaviour to observe, which does not change.

This distinction between what is observable, e.g. behaviour, and what is inaccessible or unobservable, e.g. mental states, is a distinction known in the philosophy of science (see e.g. Churchland, 1985; Fodor, 1987; Van Fraassen, 1980), but it is also at work in the empirical domain. For example, as Rabinowitz et al. (2018b) note with reference to Dennett (1991), we make mental models of others' internal states that are *inaccessible* from the outside, and that make 'little to no reference' to the underlying mechanisms of the agent that produce the observed behaviour. Our point here is that we are *always* confined to the observable behaviours of other agents, regardless of whether they are humans or machines. Whenever we claim that 'real' understanding and intentionality are lacking in some other agent, we make a claim about states that are in principle inaccessible from the outside.

Pragmatic value

We are nevertheless fully entitled to make 'mental models' of other humans, that is, attribute beliefs, desires, and intentions to them, because this is useful in everyday interaction: it has clear *pragmatic value*. This point is perhaps best known in the form worked out by Daniel Dennett as the 'intentional stance': by attributing mental states to other humans, we abstract away from their underlying biophysical complexity,

while still having a ground for anticipating future behaviour (Dennett, 1989). If we see a person running towards a bus stop, attributing the desire to catch the bus makes the behaviour intelligible and allows us to predict further behaviour, e.g. waving to the bus driver. Similarly, attributing the set of mental states that constitutes 'real' language understanding to other humans, makes their behaviour intelligible, and smooths our social interactions. This pragmatic perspective is closely related to the idea that mental states are key concepts for humans that have a strong *social* justification (Rorty, 2009); they help a community to achieve its goals in the world, and that is all the justification we need to use them. Exactly because attributing mental states to others has such clear pragmatic value, we have sufficient reason to take them seriously. From this perspective, it is counterproductive to adopt a behaviourist (i.e. denying their importance or existence) or essentialist (i.e. accepting them only if there is evidence that they are 'real') attitude towards mental states.

Pragmatic value and LLMs

We can make a similar claim for LLMs, even though humans and LLMs differ. With regard to observable behaviour, humans can deploy more subtle and multimodal observable behaviours compared to LLMs, like tone of voice, facial expressions, gestures, even unconsciously. So the observable behaviour that underlies our mental models of humans is arguably much richer, which gives us more details to work with when attributing mental states to others. At the same time, we should acknowledge that the way we interact with LLMs is strikingly different from the way we interacted with artificially intelligent systems before.

LLMs' language output is grammatically correct, fluent, and critically, increasingly well adapted to context, user, and input. This is starting to challenge assumptions about what it is to be human and what it is to be a machine, and what it is to communicate as a human with an intelligent system that communicates in many ways like a human would do (Guzman and Lewis, 2020). The increasing sophistication of interaction has led to humans viewing such systems as distinct, social entities, and as a consequence, humans are triggered even more to attribute mental states to such systems (see e.g. Guzman and Lewis, 2020; Stuart and Kneer, 2021).

In the context of LLMs, attributing mental states to LLMs has often been addressed as oversensitive anthropomorphisation, with our mental models being illusions 'in the eye of the beholder' (Bender et al., 2021). Such critiques overlook that making mental models of intelligent systems can have clear pragmatic value, in that

they abstract away from the underlying complexities of LLMs, and at the same time help us to predict and explain their behaviour, and achieve goals in the world. Obviously there are complex systems for which making mental models makes less sense, e.g. for a Mars rover, where our goal of landing it on Mars is better served by physical models. On the other hand, our interaction with LLMs as complex systems, as we show with examples below, is often best served by attributing mental states to them 'as if' they were socially intelligent in the way we think other humans are.

Our mental model about what an LLM 'knows' or 'wants', can allow us, among other things, to communicate our requests succinctly ('Do you *know* how to do X' in prompting), explain errors ('The system *confuses* X for Y'), formulate a next step in interaction ('The system now *expects* input X'), or gauge reliability of output ('How strong is your *belief* that X?'). And this need not apply only to our own interaction with LLMs, but is also relevant for explaining LLM behaviour to other humans. LLMs optimised for dialogue, (e.g. ChatGPT, PaLM2-chat), increasingly enable this form of interaction that involves mental state language.

This development should not surprise us, given that language is a tool that has evolved for communicating and manipulating mental states to achieve goals in the world (Clark, 1996; Tomasello, 2014), for example resolving conflict and working together. Similarly, in child development, language competence and the ability to reason about mental states strongly overlap (for an overview see Milligan et al., 2007). Furthermore, scholars argue that children in learning word meanings (for example for verbs of perception like 'to look') do not just learn abstract sign-object mappings (that interlocutor X literally perceives object Y), but foremost their pragmatic effects, which is for children typically directing an interlocutor's attention to various concrete objects (Enfield, 2023; San Roque and Schieffelin, 2019), and such forms of joint attention are a precursor to ToM (Tomasello et al., 1995). In a similar vein, current research that focuses on ways to have LLMs 'reflect' on their 'confidence' in their assertions, or on uncertainty in their input, can be understood in terms of the pragmatic value this has for LLM users, that in the normal world also deal with uncertainty in information (Kadavath et al., 2022; Zhou et al., 2023b).

Given the increasing prominence of LLMs in society, we can expect that making mental models of LLMs will become more common.[2] We can already see some examples where LLMs are specifically used to impersonate individuals, for example, helpdesk service agents (Brynjolfsson et al., 2023), influencers (Lorenz, 2023),

[2]Note that our account is not intended to be normative, in that we are not claiming that humans should make mental models of LLMs.

deceased beloved ones (Pearcy, 2023), virtual friends (Marr, 2023), and personal assistants (Chen, 2023). These may strike one as rather worrying examples, but such developments could have pragmatic value for humans in that LLMs can give them a sense of relationship or consolation. This is *not* to say that we advocate such deployments of LLMs, as they have many unaddressed ethical implications, but rather that there are conditions imaginable, in which mental state attribution to LLMs is explainable, justified, and has pragmatic value. Here we rather want to stress that the larger role LLMs (in whatever future form) will likely play in society, demands a theory of our interactions with them that does not simplify our behaviour to 'anthropomorphisation'.

Pragmatic value and science

We want to emphasise that pragmatic values can be different in different communities, since they may have different goals in the world. In a scientific community that attempts to describe/explain LLMs and their purported cognitive capacities in more detail than is typically required in daily life, researchers may balk at attributing mental states to LLMs. Yet, they should not do so because LLMs do not have any 'real' understanding and intentionality, as we saw that this claim misses the point. Mental states are not intended as literal accounts of the underlying complexity of humans or machines, and the subjective experience associated with 'real' understanding is not something we can access from the outside, and therefore deny outright in other entities.

A better reason, grounded in the pragmatist perspective we offer here, seems to be that mental models made in everyday interaction may allow us to explain and predict behaviour, but lack other pragmatic values critical in the scientific community. If mental states are to play a role in a scientific description or explanation of LLMs, then they must, for example, *also* cohere with other currently accepted theories or models; offer an elegant or simple explanation, have a large scope, etc. (Van Fraassen, 1980). Such values are pragmatic, because they do not primarily depend on the relation a theory has with the observable world; there is, for example, no reason to think that the world must be elegant or simple because our theories are, or that phenomena in the world cohere because our theories cohere.

The upshot is that attributing mental states to LLMs may not cohere well with empirical work on mental states in other fields that map them to patterns of neuronal activity in the brain, for which neurons in LLMs currently constitute at best only a

loose analogy. Or it may not cohere with more theoretical work that holds that the possibility of mental states in machines entails a category mistake (as introduced by Ryle (1950)), as mental states are properties of beings such as humans, which fundamentally differ from machines. By considering pragmatic values at play in different communities, we are able to explain why, on a general level, attributing mental models to LLMs can be explainable and justifiable, but at the same time could be less acceptable in the scientific community that has different pragmatic values.

8.4 Discussion

Although in Section 8.3 we discussed LLM capacities and mental states mostly at a fundamental level, our arguments are also relevant for engineers working on concrete systems that employ LLMs. Such systems will always require reflection on understanding and mental states in humans and machines, which our pragmatic outlook can inform. Our arguments are agnostic about the explicit taxonomies and frameworks of the mental, which engineers may develop and employ in such systems, as it is the system's behaviour that the pragmatist is typically most interested in, and it can be realised in various ways. In the design of LLMs, no such taxonomies or frameworks exist (Kosinski, 2024; Trott et al., 2023), but it is possible that systems that do have them manifest equally complex behaviour. In a similar vein, we can imagine training scenarios that include visual (or other multi-modal) input as a proxy for grounding denotations of words in the world, which would also make LLM behaviour more sophisticated, as disambiguating input is arguably simpler with an additional information channel. Evidence is emerging that enriching LLMs with vision modules as surrogate grounding allows such models to learn new words more efficiently (Ma et al., 2023b).

A related point is that, although a pragmatic account of mental states in humans abstracts away from their complex underlying biophysical correlates, pragmatism does not entail that there is no point in trying to disclose such correlates scientifically, with the aim of opening the black box. A biophysical account of mental states may have pragmatic values for a community of scientists as explained above, and also broader pragmatic value for society in that it can help us to, for example, treat dysfunctional mental states better. This biophysical account resides at a different level of explanation, and does not necessarily conflict with pragmatic accounts of the mental in general. In the case of LLMs, the pragmatist has similarly no principal issues with trying to find out what patterns of (artificial) neuronal activation are correlated with

mental state content in LLM input and output.

8.5 Conclusion

The goal of this chapter was to provide further reflection on LLMs in two ways. First, we scrutinised three key points surfacing in recurring critiques on LLMs, and found that on empirical and theoretical grounds, these points need more nuance. Our conclusions are that LLMs are more than exploiters of statistical patterns; that we need better measures for evaluating thought competence in LLMs before we can draw conclusions; and that LLMs have a role to play in language acquisition, as our current best distributional models.

Second, we provided a philosophical context for LLMs from a pragmatist perspective. An unresolved question underlying various critiques of LLMs, is whether they have something like 'real' language understanding and intentionality. We argued that whether we attribute unobservable mental states to other entities, including the set that would constitute 'real' language understanding and intentionality, depends on how much pragmatic value this has to us, not on whether mental states are actual properties of the entities at issue.

LLMs (in whatever future form) will become more prominent in the years to come. We hope to have contributed to a better understanding of what LLMs can(not) do, as well as to a philosophically informed understanding of our interaction with LLMs that is more than a story of mere anthropomorphisation.

8.6 Limitations

In this chapter, we addressed LLMs as the set of large Transformer-based deep neural networks that are trained with cloze tasks, using large text datasets. Still, there is some variation in this set, as LLMs can have different sizes, different architectures, training datasets, methods for further fine-tuning, and so on. Up to this point, it is typically the case that larger LLMs trained with more data obtain the best performance on a variety of tasks, which also makes that such larger LLMs are overrepresented in evaluations of general LLM capacities. OpenAI's flagship models like GPT-3 and ChatGPT are LLMs that frequently recur in tests (although the work of e.g. Shapira et al. (2024) is an exception).

In addition, new models are appearing at a fast pace, such as LLaMA (Touvron

et al., 2023), Falcon (Penedo et al., 2023), and PaLM2 (Anil et al., 2023). It remains to be seen how these new models fare on various tests, such as those for cognition, but they are fairly similar regarding their neural network architecture, training data, and training objective. At the same time, signs are emerging that OpenAI's flagship models may be slowly deteriorating with respect to their performance in writing code and doing basic math (Chen et al., 2024).

All these developments together challenge the idea that there is something like 'the' LLM, which is a simplification we made in this chapter that is not doing complete justice to the large zoo of LLMs that currently exists. In addition, the continuing updates they are undergoing to make them derail less quickly, safer, less bias-driven, more efficient, and so on, also imply that they are a moving target in many discussions. These fast developments may also limit the import of the arguments into the more distant future, as it is hard to foresee for example developments in different neural architectures and training regimes.

Chapter 9

Conclusions

This dissertation aimed to deepen our understanding of the relation between Theory of Mind (ToM) and language by combining computational, qualitative, and experimental methods. It proposed to study ToM and language through a new language resource consisting of children's freely told narratives, and offered empirical studies on ToM and language in both humans and computational models of language and cognition. This dissertation's empirical work was accompanied by broader reflection on how we can understand ToM and language in the context of modern Artificial Intelligence.

In this concluding chapter, we answer the Main Research Question (MRQ) in Section 9.1 by discussing answers to all Research Questions (RQs) as presented in Section 1.2.1 and placing them in a broader context. We end by providing a reflection on this dissertation in Section 9.2, which includes a discussion of its limitations (Section 9.2.1), and an outlook on future work (Section 9.2.2).

9.1 Answers to Research Questions

RQ1 (Chapter 2)

> *How can we predict the mental complexity of story characters with computational tools?*

To answer this question, this chapter introduced and explored the use of narratives freely told by children (4-10y) of different ages as language data. A set of 51 stories was collected and annotated for Character Depth (CD) as proxy for ToM. The chapter

explored the extraction of linguistic features from narratives to predict CD. We found that higher lexical complexity of narratives predicted the occurrence of mentally more developed story characters, besides age as another story-external predictor, and the effect of lexical complexity was the same for both younger and older children. We argued that a more complex vocabulary allows a child to better organise and represent mental aspects of the story world. Syntax, however, was not associated with CD.

From a broader perspective, children's vocabularies play a key role in development: studying children's vocabularies is studying how they represent the world (Alexander Pan, 2011). Developmental differences are more pronounced in children's vocabulary compared to syntax, as vocabulary is more sensitive to language exposure and experience, hence varies more (Hoff, 2006; Alexander Pan, 2011). Vocabulary has also been robustly linked to other variables such as academic achievement and reading comprehension (Griffin et al., 1998), but this is less obvious for syntax, except that mastery of specific syntactic structures has been shown to drive socio-cognitive development (Lohmann and Tomasello, 2003).

RQ2 (Chapter 3)

> *What is the contribution of narrative language data to research in (social) cognition and (computational) linguistics?*

In this chapter we presented ChiSCor (Children's Story Corpus) as a new natural language resource consisting of 619 fantasy stories told by 442 children aged 4-12y in social contexts. In addition, three case studies showcased ChiSCor's potential for future work and they together answer RQ2.

The first case study analysed syntactic complexity in stories. We found that overall dependency distance in stories was similar to that reported for adult language use, and stable over the primary school age range (4-12y), which is surprising given that storytelling is cognitively demanding. Our finding aligns with work that shows that children master syntax rapidly (McNeill, 1966), and with the idea that storytelling solicits 'maximal behaviour' in challenging children's ToM, linguistic and cognitive competences (Frizelle et al., 2018; Southwood and Russell, 2004).

The second case study benchmarked ChiSCor's token frequencies as approximately Zipfian, and closer to Zipf's law than BasiScript as reference corpus of Dutch children's written essays (Tellings et al., 2018a). We argued that this finding testifies to the needs of speaker and receiver that are balanced in live, oral storytelling. This case study fuels further work on cognitive pressures in free speech, which hith-

erto received less attention than written language, but may also impact word length, syllable duration, and syntactic structures, which can all be studied with ChiSCor's narratives (De Palma et al., 2021; Regier et al., 2015).

The third case study showed that with ChiSCor as relatively small dataset (±74k tokens), lemma vectors can be trained that are as informative as vectors trained on reference corpus BasiScript which is 46 times larger (Tellings et al., 2018a). Although Word2Vec is a stepping stone towards LLMs and not a language model itself, our finding aligns with work that shows that with smaller, curated datasets, LLMs that perform well can be trained (Samuel et al., 2023). There is also evidence that narratives form particularly effective training data for LLMs (Eldan and Li, 2023), and we argue that this is because narratives are often self-contained worlds, populated with (fictional) characters and their experiences, that may act as 'surrogate groundings' for learning to model different types of factual and social information.

RQ3 (Chapter 4)

> *How can a text classification task complement existing experimental work on the relation between Theory of Mind and language in children?*

We found in 442 narratives from 442 children (4-12y) in ChiSCor overlap between ToM and linguistic complexity. By drawing on theory-motivated feature engineering, manual labelling of Character Depth as proxy for ToM, and a text classifier, our overall finding was that stories with mentally more sophisticated characters were also more linguistically complex. This finding is mostly in line with Chapter 2 regarding the importance of lexical features, although the effects of syntax are much more pronounced in this chapter compared to Chapter 2, possibly due to the larger sample and variation in syntactic features included. Our finding is relevant to work in NLP that classifies different levels of ToM or perspective representation in natural language, but does so in less explainable ways (Lee et al., 2021; Kovatchev et al., 2020; Sharma et al., 2020).

Our finding supports the idea that narratives challenge children to show 'maximal behaviour' regarding their language use in creating mentally complex characters. We are however aware that children's narratives provide a window on, but not necessarily a full image of, ToM and language competence. Interestingly a similar argument runs for controlled tests of ToM (e.g. Beaudoin et al., 2020; Bloom and German, 2000) and language (e.g. Dockrell and Marshall, 2015). Hence, we see our approach in this chapter as *complementing* experimental work, as ToM and language are multi-faceted

phenomena that should be studied in both experimental and social contexts.

RQ4 (Chapter 5)

> *What different types of Character Perspective Representation occur in ChiSCor's narratives and what is their relation to children's age and language use?*

We annotated different types of Character Perspective Representation (CPR) in a sample of 150 stories from 150 children (4-12y) in ChiSCor, and found that children employ almost the full CPR typology provided by Leech and Short (2007). We also found a relation with age in that CPR types that provide more direct access to character minds, are more often found in older children, although we also found that children of all ages draw about equally on basic types of CPR that provide less direct access to character minds.

Contrary to our expectations, we found no clear overlap between more advanced types of CPR and lexically and syntactically more complex contexts of CPR use, which would have been in line with findings of Chapter 2 and Chapter 4. Possibly, both the relatively sparse occurrences of some CPR types and the fact that contexts were single utterances (i.e. limited contexts), barred reliable estimation of linguistic properties.

RQ5 (Chapter 6)

> *In what way can we meaningfully employ Language Models in studying children's language development?*

Meaningfully employing LLMs in the context of developmental research involves several steps. First, taking measures to prevent data contamination as much as possible. Second, for using a LLM as a representation of mature language use as typical use case in developmental contexts, verifying that the LLM has a decent amount of knowledge of the domain at issue. Third, choosing a LLM that is modest regarding volume of training data and size, which will improve generalisability of findings (Warstadt and Bowman, 2022). Fourth, using models zero-shot, so that the linguistic knowledge of the LLM can be directly employed, and using straightforward metrics like surprisal.

Following these steps, we employed a Dutch LM to predict masked instances of the perception verb *see* in a sample of 90 narratives from 68 children (4-12y) in ChiSCor and unravelled children's semantic development. We found that children's

use of *see* is close to mature use and manifests various complex attentive and cognitive meanings. This finding supports the idea that perception verbs can be linguistic devices for children to learn to represent information about characters' attentional and cognitive states (Johnson, 1999; Sweetser, 1990), bearing on the relation between ToM and language. Our finding contrasts with work arguing that in children's language use, denotational meanings of PVs are initially dominant (Adricula and Narasimhan, 2009; Davis, 2020; Davis and Landau, 2021; Elli et al., 2021; Landau and Gleitman, 2009, *inter alia*), but aligns with work arguing that complex meanings may be present early already because of the social situatedness of language (e.g. Enfield, 2023; San Roque and Schieffelin, 2019).

RQ6 (Chapter 7)

> *To what extent do Large Language Models show behaviour that is consistent with having Theory of Mind-like competence?*

To answer this question, we employed ToM tests from developmental psychology that present narratives in which characters' beliefs, desires and intentions are crucial for understanding the story. Both children (n=73, 7-12y) and LLMs (n=11, base and instruction-tuned) answered narrative comprehension questions to test their ability to reason about character mind states, i.e. ToM. We found that overall, only the largest commercial LLMs (GPT-4, GPT-3.5, PaLM2-chat) performed at or above child level, but we also found generalisation issues, in that these LLMs excel in figurative language use but struggle with more advanced situation modelling, and are sometimes sensitive to reformulations of original ToM tests. LLMs fine-tuned to follow instructions outperformed their counterparts that were only pre-trained.

We explained our findings by linking fine-tuning LLMs to follow instructions to the reward for cooperative communication in human language evolution and development. Successful cooperation depends for a large part on coordinating with an interaction partner's perspective (Clark, 1996; Grice, 1975; Tomasello, 2008), which aligns with ToM and is what instruction-tuning rewards.

Overall, ToM-like ability differs in humans and LLMs in that in LLMs it seems less robust. Yet, we noted that ToM tests are in the child context foremost *behavioural* probes that correlate with various other linguistic, social and academic skills. For LLMs these further correlations are hitherto unexplored, which leaves LLM success or failure on ToM tests an open question.

The discussion regarding ToM in LLMs has parallels to animal cognition, where

the issue is whether from observational evidence of primate behaviour, we should make the inference that primates represent beliefs, or adopt simpler explanations that primates simply respond to behavioural regularities which is not ToM (Trott et al., 2023). This point also concerns the internal validity of ToM tests and the way they operationalise ToM through behaviour. However, validity as concept has multiple dimensions, including one that relates to practical value (Van Haastrecht et al., 2023), which recurs in our answer on the next research question.

RQ7 (Chapter 8)

> *What are the implications of Large Language Models' complex behaviour for studying human language understanding and cognition?*

We identified and critically discussed three claims that often recur in debates about LLM behaviour in relation to language understanding and cognition.

Regarding claim 1) that LLMs are mere autocomplete devices, we argued that this overlooks much of the complexity of their internal representations learned during training. With respect to claim 2) that LLMs have formal but not functional language competence, we pointed out flaws in the current methodologies of assessing functional language competence in LLMs: among other things, the reliance on simplistic tasks, and double standards in evaluating reliance on heuristics in humans and LLMs. We concluded that at this point we cannot jump to conclusions about LLMs lacking or having functional competence. Concerning claim 3) that LLMs are not relevant in understanding human language acquisition, we argued that LLMs can be useful distributional models that show which linguistic and cognitive phenomena are *in principle* learnable from word co-occurrence statistics. This could be ToM (Kosinski, 2024; Trott et al., 2023), specific grammatical phenomena (Wilcox et al., 2023), or cognitive heuristics and biases (Suri et al., 2024).

Building forth on the position we developed in response to the claims mentioned above, we concluded by setting out a pragmatic philosophical framework regarding language understanding and intentionality in LLMs. We argued that the often recurring critique that humans tend to 'anthropomorphise' LLMs, overlooks the *pragmatic value* of ascribing mental states to LLMs. Pragmatic value means being able to efficiently predict and explain behaviour in a way that abstracts away from underlying complex mechanisms, which is essentially what we do for other humans all the time.

MRQ

How can we unravel the relation between Theory of Mind and language using computational methods and narrative?

Turning to our main research question, we have disclosed various *patterns* connecting ToM and language. Stories employing mentally more complex story characters have specific linguistic properties, some straightforward like the presence of mental state verbs, others more subtle such as the lexical complexity of a story or the presence of pragmatic markers. Statistical modelling, feature engineering and text classification as *computational* methods proved fruitful in extracting these patterns from children's narratives. We believe that the finding that narratives that manifest specific ToM levels have specific linguistic profiles, links children's ToM to their language/storytelling competences in a more direct and informative way than prediction through age or socioeconomic story-external features only, which was the focus of earlier work. It suggests that the way children use language in a context *that is relevant to them*, carries relevant information about their socio-cognitive ability. This opens up a richer set of methods and activities to study and engage children's ToM that likely aligns better with children's experiences and interests, and with ToM as a social tool to communicate and coordinate other humans' intentions and behaviour: from pretend play, to sharing experiences in group circle settings, to more focused storytelling in spoken or written form. We deem all such activities worthwhile for children's development, even though they do not yield immediate quantifiable results for a teacher. Though narratives have linguistic profiles that can be linked to ToM, this did not apply to the linguistic contexts of Character Perspective Representation (CPR), as related form of ToM.

We presented ChiSCor as new resource of children's *narratives* for research in (social) cognition and (computational) linguistics. We have shown that narrating in an interactive setting leaves social 'fingerprints' in how speaker and receiver needs are balanced in children's story language. In doing so, we attempted to lift the veil of other possible interesting properties of children's natural language use, which is underexplored in developmental research. It is true that experimental contexts allow focusing on specific (socio-)cognitive and linguistic phenomena and some degree of control. At the same time, the salience of a story that is narrated live by a peer may engage children's competences differently compared to the same story that is presented in a controlled setting. Early in language ontogeny already, children learn that language is for doing things in the world: directing attention, surprising, joking,

etc. – effects which may not be obvious for children in controlled settings. Hence, it is unsurprising that we found that the words that are key for the audience to understand and follow the story, are used clearly and coherently by children, as from their contexts of use we trained rich lexico-semantic vectors. For if the audience loses the gist of the narrative, it cannot respond as expected to the planned suspense and surprisal, and the effects of the story language are lost. We believe there is opportunity for future work in computational linguistics that draws on (smaller) socially embedded, curated datasets. Such datasets will become more relevant for modelling language development where LLMs can function as powerful statistical learners.

We used narratives as windows on the linguistic and (socio-)cognitive competences of the narrators. A Dutch language model was employed to demonstrate the early onset of complex attentive and cognitive meanings in children's use of perception verb *zien* ('to see') in ChiSCor, which is relevant to children's developing ToM. Narratives typically also contain rich factual and social information, which is why we gauged LLMs' ToM-like ability in reasoning about social information presented in narrative format, which only the largest, commercial instruction-tuned LLMs can do at or above child level.

In this dissertation, LLMs as computational models provided a novel and unorthodox perspective on ToM and language. This perspective is not without critique, as LLMs are argued to lack (socio-)cognitive capacities, 'real' language understanding and intentionality like humans. We showed that such views are too simplistic and offered a pragmatic framework for understanding the relation between complex observable behaviour (of human and machine) on the one hand, and our (warranted) attribution of mental states on the other.

All in all, this dissertation aimed to provide new perspectives on ToM and language, drawing on computational methods and narratives, in a classic developmental context but also that of modern artificial intelligence. Throughout this dissertation, computational methods were complemented with theories and methods from other fields such as narratology, developmental psychology, and philosophy. We hope that the reader considers this work as rich, multifaceted, and captivating as ToM and language are themselves.

9.2 Reflection

This dissertation constitutes yet another experiment regarding ToM and language with both the writer and reader as subject. During the process of putting thoughts

into words, the writer was constantly trying to picture the mind of the reader and possible understandings and beliefs concerning the dissertation's contents therein. For the reader at this point, questions concerning the writer's intentions with this dissertation and its beliefs regarding ToM and language are hopefully sufficiently addressed.

This work would however not be science if it did not have elements that warrant further reflection. Thus, in the remainder of this chapter, we reflect on the limitations of this dissertation and sketch opportunities for future work.

9.2.1 Limitations

We first reflect on our reliance on manually labelling ToM competence in narratives. We then discuss parsing accuracy which plays a key role in extracting information from text. We continue with discussing our view on stories as windows on development and their relation with controlled tests. We conclude with elaborating on recent insights regarding issues with prompting in LLMs.

Labelling Theory of Mind

In Chapter 2 and Chapter 4, we employed an adapted version of the Character Depth (CD) labelling originating from Nicolopoulou and Richner (2007), given in full in Table 2.1. For the mentally most developed type of character, `Person`, we can see that the focus is on children's *explicit* representation of intentional states (levels VI-VIII). However, a conclusion from Chapter 6 is that children also frequently represent information about characters' (complex) attentional and cognitive states via the use of perception verbs, hence in a more *implicit* fashion. Example (4) in Table 6.2 illustrates this: here being 'seen' implies being 'caught in the act', i.e. *seeing* condenses a complex coordination of multiple characters' mental states (*knowing* about something of which someone else *desires* that it remains unknown). Thus, this example stages a mentally sophisticated character, that according to Table 2.1 is an `Agent` since we are dealing with mere perceptions (levels IIIb and Vb) where mental states are not explicit. So, as our research progressed, we found out that by initially requiring indicators of advanced ToM in storytelling to be explicit, we risked overlooking the subtleties in language with which children may represent socio-cognitive content.

Another point is that having a single CD label per story may obscure the (complex) constellations of characters in stories, which is relevant to the work done in Chapter 2 and Chapter 4. For example, it is not obvious what a `Person` label says

regarding stories with different amounts of additional Person or Agent characters. Because this complexity is hard to bin, we opted for a single label per story. Future work could tie in with more general work on automatic character extraction and character density (e.g. Vala et al., 2015) to unravel networks of character types.

Lastly, another worry may be that the CD labelling and the linguistic features used to predict it are not independent. That is, the description of e.g. Persons as in Table 2.1 explicates various lexical cues sufficient for a Person label, such as mental state verbs (e.g. 'to think', 'to want') that the classifier in Chapter 4 simply picks up on, as displayed by the feature importances in Figure 4.3. While such cues were indeed found to be typical for Person stories, there are reasons to assume they are not related to the linguistic features in a problematic way. First, although 'to want' and 'to think' are typically found with complement structures, complementation can be used with many other verbs expressing communication and perception, thus is not necessarily an indicator of Person as opposed to Agent stories. Second, these mental state verbs are among the 80 most frequent lemmas in the BasiScript lexicon (Tellings et al., 2018a), meaning they are not very infrequent which would render them drivers of lexical complexity. Third, character thought can also be explicated linguistically without these lexical cues, infrequent (complex) lemmas or complement structures at all, for example with Free Direct/Indirect Thought, which we found to be occurring in stories in Chapter 5. Though it may be debatable whether such statements constitute explicit intentional states, we regarded them as such in our CD labelling. Lastly, from a more general perspective, we can make a distinction between lower-level features that motivate label choice for an annotator, e.g. an absence of action and mental state verbs (Actor), presence of verbs of perception (Agent) or mental state verbs (Person), and higher-level linguistic features that likely do not play into label choice such as lexical diversity, lexical complexity, dependency distance, and the occurrence of pragmatic markers.

Parsing accuracy

In various chapters we relied on NLP-parsers for lemmatising story vocabularies and parsing syntactic dependencies. This introduced a margin of error as such parsers are not perfect; in the case of ChiSCor they were used outside the language domains on which they were trained, which is typically written or spoken adult text. On the other hand, manually lemmatising and syntactic parsing of hundreds of stories is not feasible. This is likely why large Dutch language resources like JASMIN-CGN,

Metric	Frog	Alpino	spaCy	Overall
Lemmatisation	.93 (942)	-	.95 (590)	.94 (1532)
Dependency parsing	-	.80 (207)	.83 (218)	.81 (425)

Table 9.1: Performance of various parsers in extracting linguistic features. The metric for lemmatisation is accuracy of lemma and associated POS-tag, for dependency parsing this is the Unlabelled Attachment Score (accuracy of dependency links), as we do not use dependency labels in this dissertation. Numbers of parsed lemmas/dependencies in parentheses.

BasiLex and BasiScript offer layers of automatically extracted annotations, such as Part-of-Speech (POS) tagging and lemmatisation that were not manually verified (e.g. Cucchiarini et al., 2008; Tellings et al., 2014, 2018a), and why studies employ (part of) these annotations as-is (Harmsen et al., 2021; Monster et al., 2022; Strik et al., 2010; Tellings et al., 2018b, *inter alia*).

Here we provide an indication of parser accuracy by manual verification of Frog (Van Den Bosch et al., 2007) and Alpino (Van Noord, 2006), the parsers used in Chapter 2, for lemmatisation and dependency parsing respectively.[1] We also verified spaCy (Honnibal and Johnson, 2015), a NLP-pipeline used in Chapter 3 through Chapter 5 for both lemmatisation and dependency parsing. For verifying lemmatisation, we drew 10 random stories and for verifying dependency parsing we drew 20 random sentences from the datasets used in the respective chapters: ChiSCor's pilot set or full ChiSCor. For lemmatisation, associated POS-tags were also evaluated as they are needed to disambiguate lemmas. For example, in Dutch *leven* ('to live') can be both verb and noun.

As can be seen in Table 9.1, overall performance is high for lemmatisation, but lower while still acceptable for dependency parsing. Frog and Alpino perform slightly worse compared to spaCy. Part of the reason performance is overall acceptable could be that we used parsers on *normalised* transcriptions where some of the noise in storytelling was manually corrected: false starts, broken-off words, wrong conjugations, and so on were removed, while keeping the impact on syntax and semantics of the utterances as minimal as possible.[2] We also tried to disambiguate utterance boundaries by carefully attending to pause and pitch in the recordings, and formatted transcription files as having one utterance per line. Yet, utterance boundaries are not always obvious for children, especially for younger children, where pauses and pitch differences are harder to discern. So, the metrics provided in Table 9.1 give an indication of

[1] Notebooks and data are available at `https://osf.io/25ead/`.

[2] ChiSCor provides the original recordings, raw transcripts and normalised transcripts so that normalisations can be consulted.

parser performance and associated margin of error in feature extraction at scale, but do not provide a complete view.

Stories as windows on development

Throughout this dissertation we assumed that ChiSCor's freely told stories can be used to unravel children's ToM and language development. Although we find various links between children's ToM and their language use in stories, this dissertation does not link ToM or language to *external*, validated measures of these competences such as the Strange Stories test (Happé, 1994) or the Peabody Picture Vocabulary test (Dunn and Dunn, 1997). So from a developmental perspective, one could argue that stories are rather fleeting windows on children's competences where we have little control over what children actually produce in a story regarding ToM or language. This limits the scope of the dissertation's conclusions.

However, both experimental and storytelling setups use similar approaches in that both tend to link *constructs* that are not observable like ToM and language competence, to observable behaviour. Both setups face issues like auxiliary task demands such as processing and memorising (Bloom and German, 2000; Hu and Frank, 2024), which makes it hard to maintain that one setting has a privileged view over the other. So, we emphasise here and in other chapters that we see our approach as *complementing* existing experimental work.

Reliance on prompting

Recently, prompting (as introduced by Brown et al. (2020)) as a technique to straightforwardly test LLMs' linguistic and cognitive capacities via natural language, is under scrutiny as it asks LLMs to engage in 'meta-linguistic judgements'. For example, in obtaining sentence acceptability judgements, a direct approach would compute and evaluate the surprisal of a grammatical against an ungrammatical sentence (following the line of Chapter 6). However, the prompting approach would be to feed an instruction in natural language to the model, something along the line of 'Here are two sentences, A and B. Which one is grammatical and which one is not?', besides the actual sentence pair to evaluate. The additional meta-linguistic judgement a LLM has to engage in in the prompting approach, may negatively impact performance as LLMs may struggle with this auxiliary task (Hu and Levy, 2023), especially smaller models.

We relied on prompting in Chapter 7, so our finding that smaller LLMs and LLMs

without further instruction-tuning perform below child level may be different if LLM behaviour on ToM tests is considered through more direct metrics like surprisal in token predictions. The issue regarding LLMs' sensitivity to prompting is an ongoing debate, which is reminiscent of the open issue in how the mode of presentation of ToM tests – linguistically or other – influences performance in children (Bloom and German, 2000; Quesque and Rossetti, 2020). We incorporated new insights on more direct measures of LLM capabilities in Chapter 6, which appeared as research paper later than Chapter 7.

9.2.2 Future Work

Here we elaborate on directions for future work following up on the studies constituting this dissertation.

1. **Narrative and computation –** Computational narratology as a field is not new (cf. Mani, 2014; Santana et al., 2023), but LLMs could make it easier to extract information from narratives, which is deemed a challenging task as narratives are rich in information but vary a lot (DeBuse and Warnick, 2024). An example is character network extraction from the perspective that characters and their (inter)actions drive the narrative plot (Vala et al., 2015). Another example is the extraction of (development in) plot structure in children's narratives, that was previously identified manually (Botvin and Sutton-Smith, 1977). For both examples, LLMs can extract narrative information by prompting with narrative comprehension questions, following the line of Chapter 7. As in both examples LLMs are mainly used as tools for information extraction (and not as models of cognition), LLMs' sensitivity to prompting as discussed in the previous section seems less of an issue. Such approaches have been applied successfully already for event segmentation in narratives with LLMs (Michelmann et al., 2023).

2. **Leveraging LLMs in developmental contexts –** Although the role of using LLMs as fully-blown language learners is contested, they can still be useful tools in developmental contexts. An example is using LLMs as representations of mature language use, continuing the line of Chapter 6. LLMs can be used to estimate the distance between children's and adults' contexts of use of any word of interest, and in that way model word acquisition. They can also estimate the coherence between sets of sentences or utterances via sentence order prediction tasks and thereby model narrative skills. This use of LLMs is interesting in combination with longitudinal language data of children, as it allows

mapping language development in a novel way. Distance and coherence are essentially surprisal metrics, which have links with cognitive processes such as learning (Saffran, 2020).

3. **Child-specific LLMs –** Recent work shows that smaller, curated datasets can be used to train well-performing LLMs (Eldan and Li, 2023; Samuel et al., 2023). Following up on Chapter 3 we see potential in training a LLM on smaller sets of children's natural language exclusively. This would allow exploring whether the same learned representations concerning formal linguistic and other cognitive properties (as mentioned in Chapter 8) found in existing LLMs, also occur in LLMs trained on child data, and how they differ.

Bibliography

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., and Søgaard, A. (2021). Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online. Association for Computational Linguistics.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., et al. (2024). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Adricula, N. and Narasimhan, B. (2009). 'Understanding is Understanding by Seeing': Visual Perception Verbs in Child Language. In *Proceedings of the 44th Boston University Conference on Language Development*, pages 18–27, Boston. Somerville, MA: Cascadilla Press.

Agüera y Arcas, B. (2022a). Artificial neural networks are making strides towards consciousness. `https://www.economist.com/by-invitation/2022/09/02/artificial-neural-networks-are-making-strides-towards-consciousness-according-to-blaise-aguera-y-arcas`. *The Economist* (Jun 11th). Accessed on: 2023-01-28.

Agüera y Arcas, B. (2022b). Do Large Language Models Understand Us? *Daedalus*, 151(2):183–197.

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3):91–93.

Alabdulkarim, A., Li, S., and Peng, X. (2021). Automatic story generation: Challenges and attempts. In Akoury, N., Brahman, F., Chaturvedi, S., Clark, E., Iyyer, M., and Martin, L. J., editors, *Proceedings of the Third Workshop on Narrative Understanding*, pages 72–83, Virtual. Association for Computational Linguistics.

Alexander Pan, B. (2011). Assessing vocabulary skills. In Hoff, E., editor, *Research Methods in Child Language*, pages 100–112. John Wiley and Sons.

Altszyler, E., Ribeiro, S., Sigman, M., and Fernández Slezak, D. (2017). Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. In *XVIII Simposio Argentino de Inteligencia Artificial (ASAI)-JAIIO 46 (Córdoba, 2017)*.

Andreas, J. (2022). Language models as agent models. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., Shafey, L. E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., et al. (2023). PaLM 2 Technical Report. *arXiv preprint arXiv:2305.10403v3*.

Apperly, I. (2012). *Mindreaders: the Cognitive Basis of "Theory of Mind"*. Psychology Press.

Apperly, I. A., Samson, D., and Humphreys, G. W. (2009). Studies of adults can inform accounts of theory of mind development. *Developmental psychology*, 45(1):190.

Arslan, B., Taatgen, N. A., and Verbrugge, R. (2017). Five-year-olds' systematic errors in second-order false belief tasks are due to first-order theory of mind strategy selection: A computational modeling study. *Frontiers in psychology*, 8:275.

Astington, J. W. and Jenkins, J. M. (1999). A longitudinal study of the relation between language and theory-of-mind development. *Developmental psychology*, 35(5):1311.

Baayen, H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Baker, C., Saxe, R., and Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the Cognitive Science Society*, volume 33.

Banfield, A. (1973). Narrative style and the grammar of direct and indirect speech. *Foundations of language*, 10(1):1–39.

Barak, L., Fazly, A., and Stevenson, S. (2012). Modeling the Acquisition of Mental State Verbs. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 1–10, Montréal, Canada. Association for Computational Linguistics.

Baron-Cohen, S. (2001). Theory of mind in normal development and autism. *Prisme*, 34(1):74–183.

Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

Barone, P., Corradi, G., and Gomila, A. (2019). Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis. *Infant Behavior and Development*, 57:101350.

Beauchamp, M. H. (2017). Neuropsychology's social landscape: Common ground with social neuroscience. *Neuropsychology*, 31(8):981–1002.

Beaudoin, C., Leblanc, E., Gagner, C., and Beauchamp, M. H. (2020). Systematic Review and Inventory of Theory of Mind Measures for Young Children. *Frontiers in Psychology*, 10.

Beekhuizen, B., Armstrong, B. C., and Stevenson, S. (2021). Probing Lexical Ambiguity: Word Vectors Encode Number and Relatedness of Senses. *Cognitive Science*, 45(5):e12943.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Bender, E. M. and Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Bergen, D. (2002). The role of pretend play in children's cognitive development. *Early Childhood Research & Practice*, 4(1).

Berger, J. and Packard, G. (2022). Using natural language processing to understand people and culture. *American Psychologist*, 77(4):525.

Bian, N., Han, X., Lin, H., Lu, Y., He, B., and Sun, L. (2024). Rule or Story, Which is a Better Commonsense Expression for Talking with Large Language Models? *arXiv preprint arXiv:2402.14355*.

Binz, M. and Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Binz, M. and Schulz, E. (2024). Turning large language models into cognitive models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Bishop, M. J. (2021). Artificial intelligence is stupid and causal reasoning will not fix it. *Frontiers in Psychology*, 11:2603.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., and Turian, J. (2020). Experience Grounds Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Blank, I. A. (2023). What are large language models supposed to model? *Trends in Cognitive Sciences*, 27(11):987–989.

Bloom, P. and German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, 77(1):B25–B31.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., et al. (2021). On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*.

Boogaart, R. (1999). *Aspect and temporal ordering. A contrastive analysis of Dutch and English*. PhD thesis, Vrije Universiteit Amsterdam.

Borji, A. (2023). A Categorical Archive of ChatGPT Failures. *arXiv preprint arXiv:2302.03494*.

Botting, N. (2002). Narrative as a tool for the assessment of linguistic and pragmatic impairments. *Child language teaching and therapy*, 18(1):1–21.

Botvin, G. J. and Sutton-Smith, B. (1977). The development of structural complexity in children's fantasy narratives. *Developmental Psychology*, 13(4):377.

Bowman, S. R. (2022). The Dangers of Underclaiming: Reasons for Caution When Reporting How NLP Systems Fail. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7484–7499, Dublin, Ireland. Association for Computational Linguistics.

Bowman, S. R. (2023). Eight Things to Know about Large Language Models. *arXiv preprint arXiv:2304.00612*.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Brahman, F., Huang, M., Tafjord, O., Zhao, C., Sachan, M., and Chaturvedi, S. (2021). "let your characters tell their story": A dataset for character-centric narrative understanding. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W., editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1734–1752, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen,

M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Browning, J. and LeCun, Y. (2022). AI and The Limits Of Language. `https://www.noemamag.com/ai-and-the-limits-of-language`. *Noema Magazine* (Aug 23rd). Berggruen Institute. Accessed on: 2023-01-28.

Bruner, J. (1990). Culture and human development: A new look. *Human development*, 33(6):344–355.

Brunet-Gouet, E., Vidal, N., and Roux, P. (2023). Can a Conversational Agent Pass Theory-of-Mind Tasks? A Case Study of ChatGPT with the Hinting, False Beliefs, and Strange Stories Paradigms. In *International Conference on Human and Artificial Rationalities*, pages 107–126. Springer.

Brunner, A. (2013). Automatic recognition of speech, thought, and writing representation in german narrative texts. *Literary and linguistic computing*, 28(4):563–575.

Brynjolfsson, E., Li, D., and Raymond, L. R. (2023). Generative AI at Work. *Working Paper Series*, (31161). Cambridge, MA: National Bureau of Economic Research.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Bunzeck, B. and Zarrieß, S. (2023). GPT-wee: How Small Can a Small Language Model Really Get? In Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., and Cotterell, R., editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 35–46, Singapore. Association for Computational Linguistics.

Cancho, R. F. I. and Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3):788–791.

Carruthers, P. (2013). Mindreading in infancy. *Mind & Language*, 28(2):141–172.

Cerullo, M. (2022). In Defense of Blake Lemoine and the Possibility of Machine Sentience in Lamda. `https://philarchive.org/rec/CERIDO`. *PhilPapers*. Accessed on: 2023-02-20.

Cevoli, B., Watkins, C., Gao, Y., and Rastle, K. (2023). Shades of meaning: Uncovering the geometry of ambiguous word representations through contextualised language models. *arXiv preprint arXiv:2304.13597*.

Chang, T. A. and Bergen, B. K. (2022). Word Acquisition in Neural Language Models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., and Banaji, M. R. (2021). Gender Stereotypes in Natural Language: Word Embeddings Show Robust Consistency Across Child and Adult Language Corpora of More than 65 Million Words. *Psychological Science*, 32(2):218–240.

Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. Sage Publications.

Chen, B. X. (2023). How ChatGPT and Bard Performed as My Executive Assistants. `https://www.nytimes.com/2023/03/29/technology/personaltech/ai-chatgpt-google-bard-assistant.html`. *New York Times* (May 5th). Accessed on: 2023-10-15.

Chen, L., Zaharia, M., and Zou, J. (2024). How Is ChatGPT's Behavior Changing Over Time? *Harvard Data Science Review*, (2).

Chomsky, N. (2023). The False Promise of Chat-GPT. `https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html`. *New York Times* (June 6th). Accessed on: 2023-01-30.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep Reinforcement Learning from Human Preferences. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, et al. (2024). Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, 25(70):1–53.

Churchland, P. M. (1985). The ontological status of observables: in praise of the superempirical virtues. *Images of science*, pages 35–47.

Clark, H. H. (1996). *Using Language*. Cambridge University Press.

Clement, M. (1991). Present—preterite: Tense and narrative point of view. *Linguistics in the Netherlands*, 8(1):11–20.

Collins, K. M., Wong, C., Feng, J., Wei, M., and Tenenbaum, J. (2022). Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Contreras Kallens, P., Kristensen-McLachlan, R. D., and Christiansen, M. H. (2023). Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, 47(3):e13256.

Cremin, T., Flewitt, R., Mardell, B., and Swann, J. (2017). *Storytelling in Early Childhood. Enriching language, literacy and classroom culture*. Routledge Abingdon.

Creswell, A., Shanahan, M., and Higgins, I. (2023). Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations*.

Cucchiarini, C., Driesen, J., Sanders, E., et al. (2008). Recording Speech of Children, Non-Natives and Elderly People for HLT Applications: the JASMIN-CGN Corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Cucchiarini, C. and Van hamme, H. (2013). *The JASMIN Speech Corpus: Recordings of Children, Non-natives and Elderly People*, pages 43–59. Springer Berlin Heidelberg, Berlin, Heidelberg.

Curenton, S. M. and Justice, L. M. (2008). Children's preliteracy skills: Influence of mothers' education and beliefs about shared-reading interactions. *Early Education and Development*, 19(2):261–283.

Cuzzolin, F., Morelli, A., Cirstea, B., and Sahakian, B. J. (2020). Knowing me, knowing you: theory of mind in AI. *Psychological medicine*, 50(7):1057–1061.

Dancygier, B. (2011). *The language of stories: A cognitive approach*. Cambridge University Press.

Dancygier, B., Lu, W., and Verhagen, A. (2016). *Viewpoint and the Fabric of Meaning: Form and Use of Viewpoint Tools across Languages and Modalities*. De Gruyter Mouton.

Dasgupta, I., Lampinen, A. K., Chan, S. C., Creswell, A., Kumaran, D., McClelland, J. L., and Hill, F. (2022). Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*.

David-Barrett, T. and Dunbar, R. I. M. (2016). Language as a coordination tool evolves slowly. *R. Soc. open sci.*, 3:160259.

Davis, E. E. (2020). *Does seeing mean believing? The development of children's semantic representations for perception verbs*. PhD thesis, The Johns Hopkins University.

Davis, E. E. and Landau, B. (2021). Seeing and believing: the relationship between perception and mental verbs in acquisition. *Language Learning and Development*, 17(1):26–47.

De Mulder, H. N. (2011). *Putting the pieces together: The development of theory of mind and (mental) language*. PhD thesis, Utrecht University.

De Palma, P., Garcia-Camargo, L. A., Kilfoyle, J., Vandam, M., and Stover, J. (2021). Speech tested for zipfian fit using rigorous statistical techniques. *Proceedings of the Linguistic Society of America*, 6(1):394–402.

De Villiers, J. G. (2000). Language and theory of mind: What are the developmental relationships? In Baron-Cohen, S., Tager-Flusberg, H., and Cohen, D. J., editors, *Understanding other minds: Perspectives from developmental cognitive neuroscience*, pages 83–123. Oxford University Press.

De Villiers, J. G. (2005). Can Language Acquisition Give Children a Point of View. In Astington, J., editor, *Why language matters for theory of mind*, pages 186–219. Oxford University Press.

De Villiers, J. G. (2007). The interface of language and Theory of Mind. *Lingua: an International Review of General Linguistics.*, 117(11):1858–1878.

De Villiers, J. G. and De Villiers, P. A. (2014). The role of language in theory of mind development. *Topics in Language Disorders*, 34(4):313–328.

De Vries, W., Van Cranenburgh, A., Bisazza, A., Caselli, T., Van Noord, G., and Nissim, M. (2019). BERTje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.

DeBuse, M. A. and Warnick, S. (2024). Plot extraction and the visualization of narrative flow. *Natural Language Engineering*, 30(3):480–524.

Dell'Orletta, F., Montemagni, S., and Venturi, G. (2011). READ–IT: Assessing Readability of Italian Texts with a View to Text Simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83.

Delobelle, P., Winters, T., and Berendt, B. (2020). RobBERT: a Dutch RoBERTa-based Language Model. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Deng, C., Zhao, Y., Tang, X., Gerstein, M., and Cohan, A. (2024). Investigating Data Contamination in Modern Benchmarks for Large Language Models. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8706–8719, Mexico City, Mexico. Association for Computational Linguistics.

Dennett, D. (1989). *The Intentional Stance*. MIT Press.

Dennett, D. C. (1991). Two Contrasts: Folk Craft vs. Folk Science and Belief vs. Opinion. In Greenwood, J. D., editor, *The Future of Folk Psychology: Intentionality and Cognitive Science*, pages 135–48. Cambridge University Press.

Devine, R. T. and Hughes, C. (2013). Silent films and strange stories: Theory of mind, gender, and social experiences in middle childhood. *Child development*, 84(3):989–1003.

Devine, R. T., Kovatchev, V., Grumley Traynor, I., Smith, P., and Lee, M. (2023). Machine learning and deep learning systems for automated measurement of "advanced" theory of mind: Reliability and validity in children and adolescents. *Psychological Assessment*, 35(2):165.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dillion, D., Tandon, N., Gu, Y., and Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Dirkson, A., Verberne, S., Van Oortmerssen, G., Gelderblom, H., and Kraaij, W. (2023). How do others cope? extracting coping strategies for adverse drug events from social media. *Journal of Biomedical Informatics*, 139:104228.

Dockrell, J. E. and Marshall, C. R. (2015). Measurement issues: Assessing language skills in young children. *Child and Adolescent Mental Health*, 20(2):116–125.

Dor, D. (2015). *The Instruction of Imagination. Language as a Social Communication Technology*. Oxford University Press.

Duinmeijer, I., de Jong, J., and Scheper, A. (2012). Narrative abilities, memory and attention in children with a specific language impairment. *International Journal of Language & Communication Disorders*, 47(5):542–555.

Dunn, L. M. and Dunn, L. M. (1997). *PPVT-III: Peabody Picture Vocabulary Test*. Circle Pines, MN: American Guidance Services.

Ebert, K. D. and Scott, C. M. (2014). Relationships between narrative language samples and norm-referenced test scores in language assessments of school-age children. *Language, Speech, and Hearing Services in Schools*, 45(4):337–350.

Eekhof, L. S. (2024). *Reading the Mind. The Relationship Between Social Cognition and Narrative Processing*. PhD thesis, Radboud University Nijmegen Nijmegen.

Eekhof, L. S., Van Krieken, K., and Sanders, J. (2020). VPIP: A lexical identification procedure for perceptual, cognitive, and emotional viewpoint in narrative discourse. *Open Library of Humanities*, 6(1).

Eldan, R. and Li, Y. (2023). TinyStories: How Small Can Language Models Be and Still Speak Coherent English? *arXiv preprint arXiv:2305.07759*.

Elias, J. (2023). Google's newest A.I. model uses nearly five times more text data for training than its predecessor. `https://www.cnbc.com/2023/05/16/googles-palm-2-uses-nearly-five-times-more-text-data-than-predecessor.html`. *CNBC* (May 16th). Accessed on: 2023-05-30.

Elli, G. V., Bedny, M., and Landau, B. (2021). How does a blind person see? Developmental change in applying visual verbs to agents with disabilities. *Cognition*, 212:104683.

Enfield, N. (2023). Linguistic concepts are self-generating choice architectures. *Philosophical Transactions of the Royal Society B*, 378(1870).

Ensink, K. and Mayes, L. C. (2010). The development of mentalisation in children from a theory of mind perspective. *Psychoanalytic Inquiry*, 30(4):301–337.

Evanson, L., Lakretz, Y., and King, J. R. (2023). Language acquisition: do children and language models follow similar learning stages? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12205–12218, Toronto, Canada. Association for Computational Linguistics.

Fengxiang, F., Yang, Y., and Yaqin, W. (2016). The probability distribution of textual vocabulary in the English language. *Journal of Quantitative Linguistics*, 23(1):49–70.

Fernández, C. (2013). Mindful storytellers: Emerging pragmatics and theory of mind development. *First Language*, 33(1):20–46.

Ferrer i Cancho, R. (2005). The variation of Zipf's law in human language. *The European Physical Journal B-Condensed Matter and Complex Systems*, 44(2):249–257.

Flobbe, L., Verbrugge, R., Hendriks, P., and Krämer, I. (2008). Children's application of theory of mind in reasoning and language. *Journal of Logic, Language and Information*, 17:417–442.

Floridi, L. (2023). AI as Agency without Intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology*, 36(1):15.

Fludernik, M. (1996). *Towards a 'natural' narratology*. Routledge.

Fodor, J. (1992). A theory of the child's theory of mind. *Cognition*, 44(3):283–296.

Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. MIT press.

Frank, M. C. (2023). Large language models as models of human cognition. *PsyArXiv preprint psyarXiv:wxt69*.

Frizelle, P., Thompson, P. A., McDonald, D., and Bishop, D. V. (2018). Growth in syntactic complexity between four years and adulthood: Evidence from a narrative task. *Journal of Child Language*, 45(5):1174–1197.

Futrell, R., Mahowald, K., and Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Gallagher, S. and Hutto, D. D. (2008). Understanding Others Through Primary Interaction and Narrative Practice. In Zlatev, J., Racine, T., Sinha, C., and Itkonen, E., editors, *The Shared Mind: Perspectives on Intersubjectivity.*, pages 17–38. John Benjamins.

Ganti, A., Wilson, S., Ma, Z., Zhao, X., and Ma, R. (2022). Narrative detection and feature analysis in online health communities. In Clark, E., Brahman, F., and Iyyer, M., editors, *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 57–65, Seattle, United States. Association for Computational Linguistics.

Garí Soler, A. and Apidianaki, M. (2021). Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Gildea, D. and Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.

Grand, G., Blank, I. A., Pereira, F., and Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*, 6(7):975–987.

Grant, E., Nematzadeh, A., and Griffiths, T. L. (2017). How can memory-augmented neural networks pass a false-belief task? In *Proceedings of the Annual Meeting of the Cognitive Science Society, 39*, pages 472–432.

Grice, P. (1975). Logic and conversation. In Cole, P. and Morgan, J., editors, *Syntax and semantics. Vol. 3: Speech acts*, pages 41–58. Academic Press, New York.

Griffin, P., Burns, M. S., and Snow, C. E. (1998). *Preventing reading difficulties in young children*. National Academies Press.

Grodner, D. and Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive science*, 29(2):261–290.

Gudibande, A., Wallace, E., Snell, C., Geng, X., Liu, H., Abbeel, P., Levine, S., and Song, D. (2023). The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.

Gurney, N., Marsella, S., Ustun, V., and Pynadath, D. V. (2021). Operationalizing theories of theory of mind: a survey. In *AAAI Fall Symposium*, pages 3–20. Springer.

Guzman, A. L. and Lewis, S. C. (2020). Artificial intelligence and communication: A human–machine communication research agenda. *New Media & Society*, 22(1):70–86.

Hagendorff, T. (2023). Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods. *arXiv preprint arXiv:2303.13988*.

Hagendorff, T., Fabi, S., and Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838.

Hale, C. M. and Tager-Flusberg, H. (2003). The Influence of Language on Theory of Mind: A Training Study. *Developmental science*, 6(3):346–359.

Hansen, H. and Hebart, M. N. (2022). Semantic features of object concepts generated with GPT-3. *arXiv preprint arXiv:2202.03753*.

Happé, F. G. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, 24(2):129–154.

Harmsen, W. N., Cucchiarini, C., and Strik, H. (2021). Automatic Detection and Annotation of Spelling Errors and Orthographic Properties in the Dutch BasiScript Corpus. *Computational Linguistics in the Netherlands Journal*, 11:281–306.

Haselton, M. G., Nettle, D., and Andrews, P. W. (2015). The evolution of cognitive bias. *The handbook of evolutionary psychology*, pages 724–746.

Heath, S. B. (1986). Taking a cross-cultural look at narratives. *Topics in language disorders*, 7(1):84.

Heider, F. and Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, 57(2):243–259.

Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word representations. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Heyes, C. (2014). False belief in infancy: a fresh look. *Developmental Science*, 17(5):647–659.

Heyes, C. M. and Frith, C. D. (2014). The cultural evolution of mind reading. *Science*, 344(6190):1243091.

Hoeksema, J., de Glopper, K., and Van Noord, G. (2022). Syntactic profiles in secondary school writing using PaQu and SPOD. In Fišer, D. and Witt, A., editors, *CLARIN: The Infrastructure for Language Resources*. Berlin: De Gruyter.

Hoff, E. (2006). How social contexts support and shape language development. *Developmental review*, 26(1):55–88.

Honnibal, M. and Johnson, M. (2015). An Improved Non-monotonic Transition System for Dependency Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Horstmann, J. (2020). Undogmatic Literary Annotation with CATMA. In Nantke, J. and Schlupkothen, F., editors, *Annotations in Scholarly Editions and Research*, pages 157–176. De Gruyter Berlin and Boston.

Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., and Fedorenko, E. (2022). Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. *bioRxiv preprint bioRxiv:2022.10.04.510681*.

Hu, J. and Frank, M. (2024). Auxiliary task demands mask the capabilities of smaller language models. In *First Conference on Language Modeling*.

Hu, J. and Levy, R. (2023). Prompting is not a substitute for probability measurements in large language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Hughes, C. and Leekam, S. (2004). What are the Links Between Theory of Mind and Social Relations? Review, Reflections and New Directions for Studies of Typical and Atypical Development. *Social development*, 13(4):590–619.

Hutto, D. D. (2008). *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*. The MIT Press.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*, volume 112. Springer.

Jawahar, G., Sagot, B., and Seddah, D. (2019). What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Jiang, Z., Xu, F. F., Araki, J., and Neubig, G. (2020). How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Johnson, C. (1999). Metaphor vs. Conflation in the Acquisition of Polysemy: The Case of SEE. In Hiraga, M. K., Wilcox, S., and Sinha, C., editors, *Cultural, Psychological and Typological Issues in Cognitive Linguistics: Selected papers of the bi-annual ICLA meeting in Albuquerque*, pages 155–169.

Jurafsky, D. and Martin, J. H. (2024). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. Stanford University. Online resource, 3rd edition.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Dodds, Z. H., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., et al. (2022). Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*.

Kaland, N., Møller-Nielsen, A., Smith, L., Mortensen, E. L., Callesen, K., and Gottlieb, D. (2005). The Strange Stories test - a replication study of children and adolescents with Asperger syndrome. *European child & adolescent psychiatry*, 14(2):73–82.

Karlsen, J., Hjetland, H. N., Hagtvet, B. E., Braeken, J., and Melby-Lervåg, M. (2021). The concurrent and longitudinal relationship between narrative skills and other language skills in children. *First Language*, 41(5):555–572.

Karsdorp, F., Kranenburg, P. V., Meder, T., and Van Den Bosch, A. (2012). Casting a spell: Identification and ranking of actors in folktales. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, pages 39–50. Lisbon, Portugal: Edições Colibri.

Kidd, D., Ongis, M., and Castano, E. (2016). On literary fiction and its effects on theory of mind. *Scientific Study of Literature*, 6(1):42–58.

Kinderman, P., Dunbar, R., and Bentall, R. P. (1998). Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, (2):191–204.

King, J. and Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of memory and language*, 30(5):580–602.

Kjell, O. N., Kjell, K., and Schwartz, H. A. (2023). Beyond rating scales: With targeted evaluation, language models are poised for psychological assessment. *Psychiatry Research*, page 115667.

Knight, W. (2023). A New Chip Cluster Will Make Massive AI Models Possible. `https://www.wired.com/story/cerebras-chip-cluster-neural-networks-ai/`. *WIRED magazine* (Aug 24th). Accessed on: 2023-05-30.

Köder, F. M. (2016). *Between direct and indirect speech: The acquisition of pronouns in reported speech*. PhD thesis, University of Groningen.

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Kouwenhoven, T., Verhoef, T., De Kleijn, R., and Raaijmakers, S. (2022). Emerging grounded shared vocabularies between human and machine, inspired by human language evolution. *Frontiers in Artificial Intelligence*, 5:886349.

Kovatchev, V., Smith, P., Lee, M., Grumley Traynor, I., Luque Aguilera, I., and Devine, R. (2020). "What is on your mind?" Automated Scoring of Mindreading in Childhood and Early Adolescence. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6217–6228, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kuijper, S. J. M., Hartman, C. A., and Hendriks, P. (2015). Who Is He? Children with ASD and ADHD Take the Listener into Account in Their Production of Ambiguous Pronouns. *PloS one*, 10(7):e0132408.

Kyle, K. (2016). *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication*. PhD thesis, Georgia State University.

Laban, P., Dai, L., Bandarkar, L., and Hearst, M. A. (2021). Can Transformer Models Measure Coherence In Text: Re-Thinking the Shuffle Test. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1058–1064, Online. Association for Computational Linguistics.

Labov, W. and Waletzky, J. (1967). Narrative analysis: Oral versions of personal experience. In Holm, J., editor, *Essays on the Verbal and Visual Arts*, pages 12–44. University of Washington Press.

Landau, B. and Gleitman, L. R. (2009). *Language and experience: Evidence from the blind child*. Harvard University Press.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Lappin, S. (2023). Assessing the Strengths and Weaknesses of Large Language Models. *Journal of Logic, Language and Information*, pages 1–12.

Laverghetta Jr, A. and Licato, J. (2021). Modeling age of acquisition norms using transformer networks. In *The International FLAIRS Conference Proceedings*, volume 34.

Lavi-Rotbain, O. and Arnon, I. (2023). Zipfian Distributions in Child-Directed Speech. *Open Mind*, 7:1–30.

Le, M., Boureau, Y.-L., and Nickel, M. (2019). "revisiting the evaluation of theory of mind through question answering". In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Lee, Y. K., Lee, I., Park, J. E., Jung, Y., Kim, J., and Hahn, S. (2021). A Computational Approach to Measure Empathy and Theory-of-Mind from Written Texts. *arXiv preprint arXiv:2108.11810*.

Leech, G. N. and Short, M. (2007). *Style in fiction: A linguistic introduction to English fictional prose*. Pearson Education.

Lenci, A. and Sahlgren, M. (2023). *Distributional semantics*. Cambridge University Press.

Lewis, P. A., Birch, A., Hall, A., and Dunbar, R. I. M. (2017). Higher order intentionality tasks are cognitively more demanding. *Social Cognitive and Affective Neuroscience*, 12(7):1063–1071.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. (2022). Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. In *The Eleventh International Conference on Learning Representations*.

Lin, B. Y., Ravichander, A., Lu, X., Dziri, N., Sclar, M., Chandu, K., Bhagavatula, C., and Choi, Y. (2024). The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*.

Linders, G. M. and Louwerse, M. M. (2023). Zipf's law revisited: Spoken dialog, linguistic units, parameters, and the principle of least effort. *Psychonomic Bulletin & Review*, 30(1):77–101.

Liu, H. (2008). Dependency Distance as a Metric of Language Comprehension Difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Liu, H., Xu, C., and Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of life reviews*, 21:171–193.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2023). Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Roustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Liveley, G. and Thomas, S. (2020). Homer's intelligent machines. In Cave, S., Dihal, K., and Dillon, S., editors, *AI narratives. A history of imaginative thinking about intelligent machines*, pages 25–48. Oxford University Press.

Lohmann, H. and Tomasello, M. (2003). The role of language in the development of false belief understanding: A training study. *Child Development*, 74(4):1130–1144.

Lorenz, T. (2023). An influencer's AI clone will be your girlfriend for $1 a minute. `https://www.washingtonpost.com/technology/2023/05/13/caryn-ai-technology-gpt-4/?utm_campaign=wp_main&utm_medium=social&utm_source=twitter`. *Washington Post* (May 13th). Accessed on: 2023-05-30.

Lucy, L. and Bamman, D. (2021). Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Lundberg, S. M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Ma, X., Gao, L., and Xu, Q. (2023a). ToMChallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. In Jiang, J., Reitter, D., and Deng, S., editors, *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 15–26, Singapore. Association for Computational Linguistics.

Ma, Z., Pan, J., and Chai, J. (2023b). World-to-words: Grounded open vocabulary acquisition through fast mapping in vision-language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 524–544, Toronto, Canada. Association for Computational Linguistics.

Ma, Z., Sansom, J., Peng, R., and Chai, J. (2023c). Towards a holistic landscape of situated theory of mind in large language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1011–1031, Singapore. Association for Computational Linguistics.

Maaten, L. V. D. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

Mani, I. (2014). Computational narratology. *Handbook of narratology*, pages 84–92.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., and Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Marchetti, A., Di Dio, C., Cangelosi, A., Manzi, F., and Massaro, D. (2023). Developing ChatGPT's theory of mind. *Frontiers in Robotics and AI*, 10:1189525.

Marcus, G. F. (2022). Nonsense on Stilts. `https://garymarcus.substack.com/p/nonsense-on-stilts`. *Blog post* (Jun 6th). Accessed on: 2023-01-30.

Margolin, U. (2014). *Narrator*. Hamburg University.

Marr, B. (2023). Artificial Intimacy: How Generative AI Can Now Create Your Dream Girlfriend. `https://www.forbes.com/sites/bernardmarr/2023/09/28/artificial-intimacy-how-generative-ai-can-now-create-your-dream-girlfriend/?sh=7a1a0f72464a`. *Forbes Magazine* (May 13rd). Accessed on: 2023-10-10.

McAlister, A. and Peterson, C. (2007). A longitudinal study of child siblings and theory of mind development. *Cognitive Development*, 22(2):258–270.

McCarthy, P. M. and Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

McConachie, B. and Hart, F. E. (2006). *Performance and Cognition. Theatre Studies and the Cognitive Turn*. Routledge.

McKeough, A. and Genereux, R. (2003). Transformation in narrative thought during adolescence: The structure and content of story compositions. *Journal of Educational Psychology*, 95(3):537.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., and Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

McNeill, D. (1966). The creation of language by children. In Lyons, J. and Wales, R. J., editors, *Psycholinguistics papers: The proceedings of the 1966 Edinburgh conference*. Edinburgh University Press.

Michelmann, S., Kumar, M., Norman, K. A., and Toneva, M. (2023). Large language models can segment narrative events similarly to humans. *arXiv preprint arXiv:2301.10297*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In Bengio, Y. and LeCun, Y., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Miller, J. F. (1991). Quantifying productive language disorders. *Research on child language disorders: A decade of progress*, pages 211–220.

Milligan, K., Astington, J. W., and Dack, L. A. (2007). Language and Theory of Mind: Meta-Analysis of the Relation Between Language Ability and False-Belief Understanding. *Child Development*, 78(2):622–646.

Mitchell, M. and Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

Moghaddam, S. R. and Honey, C. J. (2023). Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*.

Monster, I., Tellings, A., Burk, W. J., Keuning, J., Segers, E., and Verhoeven, L. (2022). Word properties predicting children's word recognition. *Scientific Studies of Reading*, 26(5):373–389.

Nelson, K. (2005). Language Pathways into the Community of Minds. In Astington, J., editor, *Why language matters for theory of mind*, pages 26–49. Oxford University Press.

Nematzadeh, A., Burns, K., Grant, E., Gopnik, A., and Griffiths, T. (2018). Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400, Brussels, Belgium. Association for Computational Linguistics.

Nerlich, B. and Clarke, D. D. (1999). Elements for an integral theory of semantic change and semantic development. In *Meaning Change–Meaning Variation. Workshop held at Konstanz*, volume 1, pages 123–134.

Nicolopoulou, A. (1993). Play, cognitive development, and the social world: Piaget, Vygotsky, and beyond. *Human development*, 36(1):1–23.

Nicolopoulou, A. (2007). The interplay of play and narrative in children's development: Theoretical reflections and concrete examples. In Goncu, A. and Gaskins, S., editors, *Play and development*, pages 249–275. Psychology Press.

Nicolopoulou, A. (2015). Young children's pretend play and storytelling as modes of narrative activity: From complementarity to cross-fertilization? In Douglas, S. and Stirling, L., editors, *Children's play, pretense, and story*, pages 7–28. Routledge.

Nicolopoulou, A. (2016). Promoting oral narrative skills in low-income preschoolers through storytelling and story acting. In Cremin, T., Flewitt, R., Mardell, B., and Swann, J., editors, *Storytelling in Early Childhood*, pages 63–80. Routledge.

Nicolopoulou, A. (2018). Pretend and social pretend play: Complexities, continuities, and controversies of a research field. In Smith, P. K. and Roopnarine, J. L., editors, *The Cambridge Handbook of Play: Developmental and Disciplinary Perspectives*, page 183–199. Cambridge University Press.

Nicolopoulou, A. (2019). Using a storytelling/story-acting practice to promote narrative and other decontextualized language skills in disadvantaged children. In Veneziano, E. and Nicolopoulou, A., editors, *Narrative, literacy and other skills: Studies in intervention*, pages 263–284. John Benjamins Publishing Company.

Nicolopoulou, A., Cortina, K. S., Ilgaz, H., Cates, C. B., and de Sá, A. B. (2015). Using a narrative-and play-based activity to promote low-income preschoolers' oral language, emergent literacy, and social competence. *Early childhood research quarterly*, 31:147–162.

Nicolopoulou, A., Ilgaz, H., Shiro, M., and Hsin, L. B. (2022). "And they had a big, big, very long fight:" The development of evaluative language in preschoolers' oral fictional stories told in a peer-group context. *Journal of Child Language*, 49(3):522–551.

Nicolopoulou, A. and Richner, E. S. (2007). From Actors to Agents to Persons: The Development of Character Representation in Young Children's Narratives. *Child development*, 78(2):412–429.

Nicolopoulou, A. and Ünlütabak, B. (2017). Narrativity and Mindreading Revisited: Children's Understanding of Theory of Mind in a Storybook and in Standard False Belief Tasks. In Ketrez, F. N., Özyürek, A., Özçalişkan, Ş., and Küntay, A. C., editors, *Social Environment and Cognition in Language Development*, pages 151–166. John Benjamins.

Niles, J. D. (1999). *Homo narrans: The poetics and anthropology of oral literature*. University of Pennsylvania Press.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Nivre, J., Zeman, D., Ginter, F., and Tyers, F. (2017). Universal Dependencies. In Klementiev, A. and Specia, L., editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Norvig, P. (2012). Colorless green ideas learn furiously: Chomsky and the two cultures of statistical learning. *Significance*, 9(4):30–33.

Onishi, K. H. and Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308(5719):255–258.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Oya, M. (2011). Syntactic dependency distance as sentence complexity measure. In *Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics*, volume 1.

Paape, D. (2023). When Transformer models are more compositional than humans: The case of the depth charge illusion. *Experiments in Linguistic Meaning*, 2:202–218.

Paley, V. G. (1990). *The Boy Who Would Be a Helicopter: The Uses of Storytelling in the Kindergarten*. Cambridge, MA: Harvard University Press.

Parisien, C. and Stevenson, S. (2009). Modelling the acquisition of verb polysemy in children. In *Proceedings of the CogSci2009 Workshop on Distributional Semantics beyond Concrete Concepts*, pages 17–22, Austin, Texas. Cognitive Science Society.

Patel, R. and Pavlick, E. (2022). Mapping Language Models to Grounded Conceptual Spaces. In *International Conference on Learning Representations*.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., and Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

Pearcy, A. (2023). 'It was as if my father were actually texting me': grief in the age of AI. `https://www.theguardian.com/technology/2023/jul/18/ai-chatbots-grief-chatgpt`. *The Guardian* (October 15th). Accessed on: 2023-10-15.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., and Launay, J. (2023). The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Perner, J., Leekam, S. R., and Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British journal of developmental psychology*, 5(2):125–137.

Perner, J. and Wimmer, H. (1985). "John thinks that Mary thinks that..." attribution of second-order beliefs by 5-to 10-year-old children. *Journal of experimental child psychology*, 39(3):437–471.

Perrigo, B. (2023). Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. `https://time.com/6247678/openai-chatgpt-kenya-workers/`. *Time Magazine* (Jan 18th). Accessed on: 2023-01-25.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language Models as Knowledge Bases? In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Piantadosi, S. (2023). Modern language models refute Chomsky's approach to language. In Gibson, E. and Poliak, M., editors, *From fieldwork to linguistic theory: A tribute to Dan Everett*, pages 353–414. Language Science Press.

Piantadosi, S. and Hill, F. (2022). Meaning without reference in large language models. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Propp, V. (1968). *Morphology of the Folktale*. University of Texas Press.

Prystawski, B., Grant, E., Nematzadeh, A., Lee, S. W., Stevenson, S., and Xu, Y. (2022). The emergence of gender associations in child language development. *Cognitive Science*, 46(6):e13146.

Quesque, F. and Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science*, 15(2):384–396.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., and Botvinick, M. (2018a). Machine theory of mind. In *International conference on machine learning*, pages 4218–4227. PMLR.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. M. A., and Botvinick, M. (2018b). Machine theory of mind. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4218–4227. PMLR.

Rabkina, I., Nakos, C., and Forbus, K. (2019). Children's Sentential Complement Use Leads the Theory of Mind Development Period: Evidence from the CHILDES Corpus. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, pages 2634–2639. The Cognitive Science Society.

Regier, T., Kemp, C., and Kay, P. (2015). Word meanings across languages support efficient communication. In MacWhinney, B. and O'Grady, W., editors, *The handbook of language emergence*, pages 237–263. Wiley Online Library.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.

Reif, E., Yuan, A., Wattenberg, M., Viegas, F. B., Coenen, A., Pearce, A., and Kim, B. (2019). Visualizing and measuring the geometry of BERT. *Advances in Neural Information Processing Systems*, 32.

Roark, B., Mitchell, M., and Hollingshead, K. (2007). Syntactic complexity measures for detecting mild cognitive impairment. In *Biological, translational, and clinical language processing*, pages 1–8.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Rorty, R. (2009). *Philosophy and the Mirror of Nature*. Princeton University Press.

Rubio-Fernández, P. (2021). Pragmatic markers: the missing link between language and Theory of Mind. *Synthese*, 199(1):1125–1158.

Rubio-Fernández, P., Mollica, F., Oraa Ali, M., and Gibson, E. (2019). How do you know that? Automatic belief inferences in passing conversation. *Cognition*, 193:104011.

Ryle, G. (1950). *"The Concept of Mind"*. The University of Chicago Press.

Rączaszek-Leonardi, J., Nomikou, I., Rohlfing, K. J., and Deacon, T. W. (2018). Language development from an ecological perspective: Ecologically valid ways to abstract symbols. *Ecological Psychology*, 30(1):39–73.

Saffran, J. R. (2020). Statistical language learning in infancy. *Child development perspectives*, 14(1):49–54.

Sahlgren, M. and Carlsson, F. (2021). The Singleton Fallacy: Why Current Critiques of Language Models Miss the Point. *Frontiers in Artificial Intelligence*, 4:682578.

Samir, F., Beekhuizen, B., and Stevenson, S. (2021). A formidable ability: Detecting adjectival extremeness with DSMs. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4112–4125, Online. Association for Computational Linguistics.

Samuel, D., Kutuzov, A., Øvrelid, L., and Velldal, E. (2023). Trained on 100 million words and still in shape: BERT meets British National Corpus. In Vlachos, A. and Augenstein, I., editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.

San Juan, V. and Astington, J. W. (2017). Does language matter for implicit theory of mind? The effects of epistemic verb training on implicit and explicit false-belief understanding. *Cognitive Development*, 41:19–32.

San Roque, L., Kendrick, K. H., Norcliffe, E., and Majid, A. (2018). Universal meaning extensions of perception verbs are grounded in interaction. *Cognitive Linguistics*, 29(3):371–406.

San Roque, L. and Schieffelin, B. B. (2019). Perception verbs in context: Perspectives from Kaluli (Bosavi) child-caregiver interaction. In O'Meara, C., Speed, L. J., San Roque, L., and Majid, A., editors, *Perception Metaphors*, pages 347–368. John Benjamins.

Sanders, J. (2010). Intertwined voices: Journalists' modes of representing source information in journalistic subgenres. *English Text Construction*, 3(2):226–249.

Santana, B., Campos, R., Amorim, E., Jorge, A., Silvano, P., and Nunes, S. (2023). A survey on narrative extraction from textual data. *Artificial Intelligence Review*, 56(8):8393–8435.

Sap, M., Le Bras, R., Fried, D., and Choi, Y. (2022). Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. (2022). BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Schaerlaekens, A. M. and Gillis, S. (1993). *De taalverwerving van het kind: een hernieuwde oriëntatie in het Nederlandstalig onderzoek*. Wolter-Nordhoff.

Schank, R. C. (1995). *Tell me a story: Narrative and intelligence*. Northwestern University Press.

Schlangen, D. (2021). Targeting the Benchmark: On Methodology in Current Natural Language Processing Research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online. Association for Computational Linguistics.

Schlichting, J. E. P. T. (1996). *Discovering syntax: An empirical study in Dutch language acquisition*. Nijmegen: Nijmegen University Press.

Schlinger, H. D. (2009). Theory of mind: An overview and behavioral perspective. *The Psychological Record*, 59:435–448.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424.

Sejnowski, T. (2022). Large language models and the reverse turing test. *arXiv preprint arXiv:2207.14382*.

Sellars, W. (1956). *Empiricism and the Philosophy of Mind*. University of Minnesota Press, Minneapolis.

Shanahan, M. (2024). Talking about large language models. *Commun. ACM*, 67(2):68–79.

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., and Shwartz, V. (2024). Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. In Graham, Y. and Purver, M., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian's, Malta. Association for Computational Linguistics.

Shapiro, L. R. and Hudson, J. A. (1991). Tell me a make-believe story: Coherence and cohesion in young children's picture-elicited narratives. *Developmental psychology*, 27(6):960.

Sharma, A., Miner, A., Atkins, D., and Althoff, T. (2020). A computational approach to understanding empathy expressed in text-based mental health support. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Shatz, M., Diesendruck, G., Martinez-Beck, I., and Akar, D. (2003). The influence of language and socioeconomic status on children's understanding of false belief. *Developmental Psychology*, 39(4):717.

Siegal, M. and Beattie, K. (1991). Where to look first for children's knowledge of false beliefs. *Cognition*, 38(1):1–12.

Sileo, D. and Lernould, A. (2023). MindGames: Targeting theory of mind in large language models with dynamic epistemic modal logic. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4570–4577, Singapore. Association for Computational Linguistics.

Slade, L. and Ruffman, T. (2005). How language does (and does not) relate to theory of mind: A longitudinal study of syntax, semantics, working memory and false belief. *British Journal of Developmental Psychology*, 23(1):117–141.

Smith, P. K. and Roopnarine, J. L. (2018). *The Cambridge Handbook of Play: Developmental and Disciplinary Perspectives*. Cambridge University Press.

Smith, R. D. (2007). Investigation of the Zipf-plot of the extinct Meroitic language. *Glottometrics*, 15:53–61.

Snow, C. E. and Dickinson, D. K. (1991). *Some skills that aren't basic in a new conception of literacy*, pages 179–191. State University of New York Press.

Southwood, F. and Russell, A. F. (2004). Comparison of Conversation, Freeplay, and Story Generation as Methods of Language Sample Elicitation. *Journal of Speech, Language and Hearing Research*, 47(2):366–376.

Sparrow, J. (2022). 'Full-on robot writing': the artificial intelligence challenge facing universities. `https://www.theguardian.com/australia-news/2022/nov/19/full-on-robot-writing-the-artificial-intelligence-challenge-facing-universities`. *The Guardian* (November 18th). Accessed on: 2023-01-25.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., et al. (2023). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Stammbach, D., Antoniak, M., and Ash, E. (2022). Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data. In Clark, E., Brahman, F., and Iyyer, M., editors, *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 47–56, Seattle, United States. Association for Computational Linguistics.

Stiller, J. and Dunbar, R. I. (2007). Perspective-taking and memory capacity predict social network size. *Social Networks*, 29(1):93–104.

Strik, H., Loo, J. V. D., Doremalen, J. V., and Cucchiarini, C. (2010). Practicing syntax in spoken interaction: automatic detection of syntactical errors in non-native utterances. In *Proc. Second Language Studies: Acquisition, Learning, Education and Technology (L2WS 2010)*.

Stuart, M. T. and Kneer, M. (2021). Guilty artificial minds: Folk attributions of mens rea and culpability to artificially intelligent agents. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–27.

Subramonian, A., Yuan, X., Daumé III, H., and Blodgett, S. L. (2023). It Takes Two to Tango: Navigating Conceptualizations of NLP Tasks and Measurements of Performance. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3234–3279, Toronto, Canada. Association for Computational Linguistics.

Suri, G., Slater, L. R., Ziaee, A., and Nguyen, M. (2024). Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *Journal of Experimental Psychology: General*, 153(4):1066–1075.

Sutton-Smith, B. (1986). Children's fiction making. In Crites, S. and Sarbin, T. R., editors, *Narrative psychology: The storied nature of human conduct*, pages 67–90. Praeger Publishers/Greenwood Publishing Group.

Sutton-Smith, B. (2012). *The folkstories of children*. University of Pennsylvania Press.

Sweetser, E. (1990). *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*, volume 54. Cambridge University Press.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of language and social psychology*, 29(1):24–54.

Tellings, A., Hulsbosch, M., Vermeer, A., and Van den Bosch, A. (2014). BasiLex: An 11.5 million words corpus of Dutch texts written for children. *Computational Linguistics in the Netherlands Journal*, 4:191–208.

Tellings, A., Oostdijk, N., Monster, I., Grootjen, F., and Van Den Bosch, A. (2018a). BasiScript: A corpus of contemporary Dutch texts written by primary school children. *International Journal of Corpus Linguistics*, 23(4):494–508.

Tellings, A., Oostdijk, N., Monster, I., Grootjen, F., and Van Den Bosch, A. (2018b). Spelling errors of 24 cohorts of children across primary school 2012-2015: a BasiScript corpus study. *Computational Linguistics in the Netherlands Journal*, 8:83–98.

Tomasello, M. (2003). The key is social cognition. *Language in mind: Advances in the study of language and thought*, pages 44–57.

Tomasello, M. (2008). *Origins of Human Communication*. MIT Press, Cambridge, MA.

Tomasello, M. (2014). The ultra-social animal. *European Journal of Social Psychology*, 44(3):187–194.

Tomasello, M. et al. (1995). Joint attention as social cognition. *Joint attention: Its origins and role in development*, 103130:103–130.

Tompkins, V., Farrar, M. J., and Montgomery, D. E. (2019). Speaking Your Mind: Language and Narrative in Young Children's Theory of Mind Development. *Advances in Child Development and Behavior*, 56:109–140.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Trott, S., Jones, C., Chang, T., Michaelov, J., and Bergen, B. (2023). Do Large Language Models know what humans know? *Cognitive Science*, 47(7):e13309.

Tsfasman, M., Fenech, K., Tarvirdians, M., Lorincz, A., Jonker, C. M., and Oertel, C. (2022). Towards creating a conversational memory for long-term meeting support: predicting memorable moments in multi-party conversations through eye-gaze. In *ICMI 2022-Proceedings of the 2022 International Conference on Multimodal Interaction*. Association for Computing Machinery (ACM).

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, LIX:433–460.

Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Vala, H., Jurgens, D., Piper, A., and Ruths, D. (2015). Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On The Difficulty of Detecting Characters in Literary Texts. In Màrquez, L., Callison-Burch, C., and Su, J., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 769–774, Lisbon, Portugal. Association for Computational Linguistics.

Van Den Bosch, A., Busser, B., Canisius, S., and Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In *Selected papers of the 17th computational linguistics in the Netherlands meeting*, pages 99–114. CLIN Leuven, BE.

Van Der Meulen, A., Roerig, S., de Ruyter, D., van Lier, P., and Krabbendam, L. (2017). A comparison of children's ability to read children's and adults' mental states in an adaptation of the reading the mind in the eyes task. *Frontiers in psychology*, 8:594.

Van Duijn, M. and Verhagen, A. (2018). Beyond triadic communication: A three-dimensional conceptual space for modelling intersubjectivity. *Pragmatics & Cognition*, 25(2):384–416.

Van Duijn, M. J. (2016). *The lazy mindreader. A humanities perspective on mindreading and multiple-order intentionality*. PhD thesis, Leiden University.

Van Duijn, M. J., Sluiter, I., and Verhagen, A. (2015). When narrative takes over: The representation of embedded mindstates in Shakespeare's Othello. *Language and Literature*, 24(2):148–166.

Van Eecke, P., Verheyen, L., Willaert, T., and Beuls, K. (2023). The candide model: How narratives emerge where observations meet beliefs. In Akoury, N., Clark, E., Iyyer, M., Chaturvedi, S., Brahman, F., and Chandu, K., editors, *Proceedings of the The 5th Workshop on Narrative Understanding*, pages 48–57, Toronto, Canada. Association for Computational Linguistics.

Van Elk, R., Lanser, D., and Gerritsen, S. (2012). *Relatie Opleidingsniveau en Arbeidsaanbod*. Netherlands Bureau for Economic Policy Analysis (CPB).

Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

Van Haastrecht, M., Brinkhuis, M., Peichl, J., Remmele, B., and Spruit, M. (2023). Embracing trustworthiness and authenticity in the validation of learning analytics systems. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 552–558.

Van Koert, M., Hollebrandse, B., and Van Hout, A. (2010). Gaan 'go' as dummy auxiliary in Dutch children's tense production. *GAGL: Groninger Arbeiten zur germanistischen Linguistik*, (51):43–54.

Van Noord, G. (2006). At Last Parsing Is Now Operational. In *Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles. Conférences invitées*, pages 20–42.

Van Schuppen, L., Van Krieken, K., and Sanders, J. (2020). Variations in Viewpoint Presentation: The 'Pear Story' as Told by People with a Schizophrenia Diagnosis. *Open Library of Humanities*, 6(2):2.

Vandelanotte, L. (2009). *Speech and thought representation in English: A cognitive-functional approach*. Mouton de Gruyter.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Verhagen, A. (2005). *Constructions of Intersubjectivity. Discourse, Syntax, and Cognition*. Oxford University Press.

Verhagen, A. (2015). Grammar and cooperative communication. In Dabrowska, E. and Divjak, D., editors, *Handbook of Cognitive Linguistics*, pages 232–252. De Gruyter Mouton, Berlin, München, Boston.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied psycholinguistics*, 22(2):217–234.

Vermeule, B. (2009). *Why Do We Care About Literary Characters?* John Hopkins University Press.

Verstraten, P. (2009). *Film Narratology*. University of Toronto Press.

Voskuhl, A. (2019). *Androids in the enlightenment: Mechanics, artisans, and cultures of the self*. University of Chicago Press.

Vulić, I., Ponti, E. M., Litschko, R., Glavaš, G., and Korhonen, A. (2020). Probing Pretrained Language Models for Lexical Semantics. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Wang, Y. and Liu, H. (2017). The effects of genre on dependency distance and dependency direction. *Language Sciences*, 59:135–147.

Wardetzky, K. (1990). The structure and interpretation of fairy tales composed by children. *Journal of American Folklore*, pages 157–176.

Warstadt, A. and Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In Lappin, S. and Bernardy, J.-P., editors, *Algebraic Structures in Natural Language*, pages 17–60. CRC Press.

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., and Cotterell, R. (2023). Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., and Cotterell, R., editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Webb, T., Holyoak, K. J., and Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Wei, C., Wang, Y.-C., Wang, B., and Kuo, C.-C. J. (2023). An overview on language models: Recent developments and outlook. *arXiv preprint arXiv:2303.05759*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E. H., Le, Q., and Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. *CoRR*, abs/2201.11903.

Wellman, H. M. (2018). Theory of mind: The state of the art. *European Journal of Developmental Psychology*, 15(6):728–755.

Wellman, H. M. and Liu, D. (2004). Scaling of theory-of-mind tasks. *Child development*, 75(2):523–541.

Wevers, M. (2019). Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990. In Tahmasebi, N., Borin, L., Jatowt, A., and Xu, Y., editors, *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 92–97, Florence, Italy. Association for Computational Linguistics.

Wiedemann, G., Remus, S., Chawla, A., and Biemann, C. (2019). Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*.

Wijnen, F. and Verrips, M. (1998). The Acquisition of Dutch Syntax. In Wijnen, F. and Verrips, M., editors, *The Acquisition of Dutch*, pages 223–300. John Benjamins Publishing Company.

Wilcox, E. G., Futrell, R., and Levy, R. (2023). Using computational models to test syntactic learnability. *Linguistic Inquiry*, pages 1–44.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., and Levy, R. (2022). Learning syntactic structures from string input. *Algebraic Structures in Natural Language*, pages 113–138.

Wilson, R., Hruby, A., Perez-Zapata, D., Van Der Kleij, S. W., and Apperly, I. A. (2023). Is recursive "mindreading" really an exception to limitations on recursive thinking? *Journal of Experimental Psychology: General*, 152(5):1454–1468.

Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xiao, H. (2008). On the Applicability of Zipf's Law in Chinese Word Frequency Distribution. *Journal of Chinese Language and Computing*, 18(1):33–46.

Yamshchikov, I. and Tikhonov, A. (2023). What is wrong with language models that can not tell a story? In Akoury, N., Clark, E., Iyyer, M., Chaturvedi, S., Brahman, F., and Chandu, K., editors, *Proceedings of the The 5th Workshop on Narrative Understanding*, pages 58–64, Toronto, Canada. Association for Computational Linguistics.

Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y., et al. (2023). A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. *arXiv preprint arXiv:2303.10420*.

Yu, S., Xu, C., and Liu, H. (2018). Zipf's law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation. *CoRR*, abs/1807.01855.

Zeman, S. (2016). Perspectivization as a link between narrative micro-and macrostructure. In Igl, N. and Zeman, S., editors, *Perspectives on Narrativity and Narrative Perspectivization*, pages 17–42. John Benjamins Amsterdam.

Zeman, S. (2018). What is a narration–and why does it matter. In Hübl, A. and Steinbach, M., editors, *Linguistic foundations of narration in spoken and sign language*, pages 173–206. Amsterdam/Philadelphia: Benjamins.

Zeman, S. (2020). Parameters of narrative perspectivization: The narrator. *Open Library of Humanities*, 6(2):28.

Zhao, X., Wang, T., Osborn, S., and Rios, A. (2023). BabyStories: Can reinforcement learning teach baby language models to write better stories? In Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., and Cotterell, R., editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 186–197, Singapore. Association for Computational Linguistics.

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., YU, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., and Levy, O. (2023a). LIMA: Less Is More for Alignment. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.

Zhou, K., Jurafsky, D., and Hashimoto, T. (2023b). Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language*. Harvard University Press.

Zunshine, L. (2006). *Why We Read Fiction: Theory of the Mind and the Novel*. Ohio State University Press.

Zunshine, L. (2019). What Mary Poppins knew: Theory of Mind, children's literature, history. *Narrative*, 27(1):1–29.

Summary

Typical human beings assume the perspective of other human beings frequently and with little effort. Doing so can be described as having a theory about what these other humans believe, desire, and intend, that is, as having a *Theory of Mind*. Having a Theory of Mind is an extremely useful tool for coordinating other humans' behaviours and navigating the social world. And just like this summary can be thought of as the linguistic code compressing the author's mind, so do humans encode and decode Theory of Mind often (but not exclusively) with spoken and written language.

Humans refine their Theory of Mind and language competences from a very young age. Since the two are linked, it is worthwhile to examine their intersections, particularly in *storytelling* as a phenomenon that engages both. Stories generally invite an audience to get immersed in a story world populated by story characters that have their own mental lives. Hence, stories are natural loci for trying to answer questions about Theory of Mind and language. For example, what do the character minds children create and their linguistic representations look like? What are more generally the 'linguistic fingerprints' of stories that deal with character minds of varying complexities? And, given that storytelling involves a Theory of Mind about the audience beyond the narrative, what are the properties of the storytelling language? This dissertation addresses these questions through the compilation of a Dutch children's story corpus on which various research methods are employed, varying from manual annotation to information extraction stemming from computer science.

Drawing on computer science here is no coincidence. Current artificially intelligent systems, in particular *large language models*, have become fluent language users to a degree of sophistication that was hard to foresee a decade ago. Hence, it is natural to ask in what way we can use these computational models to unravel the story language of children in novel ways. And to ask how adequately these large language models handle character minds, beliefs, desires and intentions in stories themselves

in comparison to children. And to ask what a philosophically sound way to better understand large language models' understanding can be. Though this dissertation started with employing techniques from computer science to extract information from stories, the latter questions also shift the role played by computer science - in particular language models - in this dissertation: from toolkit, to representation of mature language use, to subject in Theory of Mind experiments, and ultimately as focal point of a philosophical discussion revolving around better understanding machine understanding.

This dissertation has found various answers to these two different but complementary strands of questions. Regarding children, language and Theory of Mind, we found story characters with different kinds of mental lives. Story characters range from 'flat' in the sense of being merely staged or performing actions without clear goals, to more developed in featuring goal-directed behaviour and perceptions and emotions, to fully blown 'round' characters manifesting explicit and (complex) mental states. As characters become mentally more complex, stories display more complex linguistic profiles: they feature more complex syntax, more frequent use of pragmatic markers, a more complex and diverse vocabulary, and more explicitly linked clauses, among other things. Further, the language of live storytelling reflects the optimal balance between speaker and listener needs more closely than a comparable corpus of written stories, testifying to the social nature of live storytelling.

Regarding Theory of Mind and language in the context of modern artificial intelligence, we found that a language model can be used to disclose complex attentional and cognitive semantics in the way children use Dutch perception verb *zien* ('to see') in their stories. Yet, as subjects in Theory of Mind experiments, most language models do not fare better than children in dealing with beliefs, desires and intentions in stories, except for GPT-4, GPT-3.5, and PaLM2-chat. And in trying to better understand the kind of understanding that language models possess, we argued for a pragmatic framework where the aptitude of using 'understanding', 'believing', and 'thinking' and similar language for encoding the mental realm, is acceptable if this has practical value to us as humans and does not depend on the properties of the system.

By analysing Theory of Mind and language from a multidisciplinary angle, this dissertation aimed to contribute to an ongoing line of research that hopefully continues to entice both the theories and minds of researchers and the broader public.

Samenvatting

De meeste mensen kunnen regelmatig en bijna moeiteloos het perspectief van anderen aannemen. Dit kunnen we beschrijven als het hebben van een idee of theorie over wat een ander denkt, wil, of van plan is, oftewel over diens geest, vandaar dat deze vaardigheid ook wel *Theory of Mind* genoemd wordt. Het hebben van een Theory of Mind is een nuttig hulpmiddel om het gedrag van anderen te begrijpen en te voorspellen om zo de sociale wereld te kunnen navigeren. En zoals deze samenvatting gezien kan worden als de talige code die de gedachten van de auteur comprimeert, zo coderen en decoderen mensen vaak (maar niet uitsluitend) hun Theory of Mind met gesproken en geschreven taal.

Mensen verfijnen hun Theory of Mind en taalvaardigheden al vanaf jonge leeftijd. Aangezien de twee met elkaar verbonden zijn, is het de moeite waard hun raakpunten te onderzoeken in het *vertellen van verhalen*, een fenomeen dat uit beide vaardigheden put. Verhalen nodigen een publiek uit om zich onder te dompelen in een verhaalwereld die bevolkt wordt door personages met hun eigen mentale levens. Verhalen zijn dus uitermate geschikt voor het beantwoorden van vragen over Theory of Mind en taal. Welke soorten verhaalpersonages creëren kinderen, en hoe krijgen deze vorm in taal? Wat zijn in bredere zin de 'taalkundige vingerafdrukken' van verhalen die verschillende soorten verhaalpersonages opvoeren? En, aangezien verhalen vertellen een activiteit is die een Theory of Mind over wat er omgaat in het publiek vereist, wat zijn de eigenschappen van dit taalgebruik? Dit proefschrift beantwoordt deze vragen met een nieuw Nederlands kinderverhalencorpus dat ontrafeld wordt met methoden variërend van handmatige annotatie tot informatie-extractie uit de informatica.

Dat de informatica hier om de hoek komt kijken is geen toeval. Moderne systemen in de kunstmatige intelligentie, in het bijzonder de *large language models* oftewel grote taalmodellen, zijn vloeiende taalgebruikers geworden op een niveau dat een decennium geleden nog moeilijk te voorzien was. Het ligt daarom voor de hand

om te vragen op welke manier we deze computationele modellen kunnen gebruiken om de taal van kinderverhalen te analyseren. En om te vragen hoe goed deze grote taalmodellen zelf in staat zijn om te gaan met de overtuigingen, wensen en intenties van verhaalpersonages, in vergelijking met kinderen. En om te onderzoeken op welke manier we het begrijpen van taalmodellen moeten begrijpen. Waar dit proefschrift begint met het gebruiken van technieken uit de informatica om informatie uit kinderverhalen te extraheren, verschuiven deze vragen langzaam de rol die de informatica speelt in dit proefschrift - in het bijzonder de grote taalmodellen: van gereedschapskist, naar afspiegeling van volwassen taalgebruikers, naar proefpersoon in Theory of Mind-experimenten, tot brandpunt van een filosofische discussie over het beter begrijpen van het begrijpen van machines.

Dit proefschrift biedt de volgende antwoorden op deze twee verschillende maar complementaire stromen van vragen. Met betrekking tot kinderen, taal en Theory of Mind, bestaan er verhaalpersonages met verschillende soorten mentale levens. Verhaalpersonages variëren van 'platte' personages slechts aanwezig of handelend zonder duidelijke doelen, tot verder ontwikkelde personages met doelgerichte acties en (complexe) waarnemingen en emoties, tot volwaardige 'ronde' personages met expliciete (complexe) mentale toestanden. Naarmate personages mentaal complexer worden, vertonen verhalen ook vaker complexere taalkundige profielen. Zo bevatten ze vaker voegwoorden, complexere zinsstructuren, pragmatische markeringen, en een complexer en diverser vocabulaire. Verder zijn in gesproken verhalen de behoeften van spreker en luisteraar beter in balans dan in geschreven verhalen, wat de sociale aard van 'live' verhalen vertellen in de klas benadrukt.

Met betrekking tot Theory of Mind, taal en moderne kunstmatige intelligentie, vonden we dat met een taalmodel kan worden geïllustreerd dat jonge kinderen het Nederlandse perceptiewerkwoord 'zien' al in de zin van 'begrijpen' of 'realiseren' kunnen gebruiken. Als proefpersonen in Theory of Mind-experimenten presteren de meeste taalmodellen echter slechter dan kinderen in het omgaan met informatie over overtuigingen en intenties in verhalen, met uitzondering van GPT-4, GPT-3.5 en PaLM2-chat. En in het beter begrijpen van het begrijpen van taalmodellen, wordt een pragmatisch kader geboden waarin het gebruik van mentale termen zoals 'begrijpen', 'geloven', en 'denken' acceptabel is zolang dit de mens praktisch dient; de fysieke eigenschappen van het onderliggende systeem spelen hier minder een rol.

Door deze multidisciplinaire benadering van Theory of Mind wil dit proefschrift bijdragen aan een onderzoekslijn die hopelijk zowel theorieën als geesten van onderzoekers en het bredere publiek blijft prikkelen.

Acknowledgements

I remember calling my partner Willeke, right after a job interview in Max van Duijn's office and mentioning I was sure that we had a match. My feeling was right, as from that moment we shared many unforgettable experiences that contributed to this dissertation in direct and indirect ways: from visiting chaotic classrooms to record children's stories to cycling London's empty roads in the middle of the night. Max, thank you for the freedom to develop my own line of research within our collaboration, for always being supportive, and for being open to talk about anything. We discussed many things larger than research for which I am grateful. The most remarkable moments were perhaps the ones we were not together: you presented our story corpus twice exactly when Willeke was giving labour to my two children. Not many supervisors can say the same thing.

Marco Spruit, perhaps you felt at first a bit puzzled about my topic in relation to your own work, but our collaboration turned out great. In fact, we are now collaborating in the medical domain - as your field of expertise - with the tools and insights obtained during my time as PhD candidate, which I consider a good plot twist. Thank you for always being straight to the point, for showing how in academia one can properly unwind, and for staying composed when reviewers did not seem to appreciate our work.

Tom Kouwenhoven, starting from our Leiden coffee strolls, you were my fellow office mate in Huygens 126. You helped me solve numerous issues with my code and always provided a constructive critical eye on ongoing projects. I enjoyed working on our joint papers and always felt motivated going to the office knowing I would see you. You have become a good friend, illustrated by the fact that you cooked for hours in my kitchen while not even being able to enjoy the food yourself.

Barend Beekhuizen, thank you for your hospitality at the University of Toronto, for the insightful discussions we had about children's language use, and for the op-

portunity to present my work. Toronto was a milestone in my time as PhD candidate that was beyond my imagination.

Ageliki Nicolopoulou, it was very special to meet the author of many papers seminal to my own work. Thank you for making our symposium a success; I feel privileged to have met you and to have discussed children's captivating story worlds.

Thanks to many research assistants for help with collecting and labelling data: Iris Jansen, Isabelle Blok, Li Kloostra, Lola Vandame, Nikita Ham, Werner de Valk, and Yasemin Tunbul. Special thanks to Werner for the great drawings accompanying our experiments as the Creative Intelligence Lab's (CIL) artist-in-residence, and to Li Kloostra for proofreading various papers.

Thanks to the CIL members for making this PhD journey a fun one, including (but not limited to) Dan Xu, Danica Mast, Giulio Barbero, Maarten Lamers, Marcello Gómez-Maureira, Marianne Bossema, Max Peeperkorn, Peter van der Putten, Ramira van der Meulen, Rob Saunders, Ross Towns, Tessa Verhoef, Tom Breedveld, and Zane Kripe. I was also happy to be part of Marco's Translational Data Science lab, and thank Alireza Shojaifar, Armel Lefebvre, Chaïm van Toledo, Hielke Muizelaar, Injy Sarhan, Jim Achterberg, Marcel Haas, Max van Haastrecht, and Samar Samir. I feel lucky to still collaborate with many of you. Special thanks to Max van Haastrecht as co-organiser of the PhD seminar and the fun social events we organised together. Peter Dekker, thanks for the fruitful exchanges we had about living the PhD life. Suzan Verberne, thank you for your advice at various stages of my project and the opportunity to present my work in your Text Mining and Retrieval lab. I also thank the Media Technology MSc programme staff for the opportunity to develop my teaching skills during my PhD.

Thanks dear Albert van Dijk and Marianne van Dijk - who we miss dearly - for your love and support. The unique Van Dijk thing I share with Daan and Tobias van Dijk I will always value. Thanks also to Leonor Villa Acosta and Cristhian Pinzon Villa for sharing many precious moments. Pieter Verduijn, having your artistic rendition of this dissertation's contents as the cover is an honour. Theresa Montenarello, thanks for helping with digitalising Pieter's work while still being in your baby bubble. Thanks also to Teun and Corrie Verduijn for caring for my children many times while I was working on this dissertation.

Lastly I would to like express my deep gratitude to Willeke Verduijn, not my best friend or buddy but life partner, who has seen it all the last four years. Please know I learn more from you every day about love, patience, kindness, and what truly matters in life, than any degree could offer.

List of publications

Asterisks denote equal contributions.

1. Van Dijk, B.M.A. and Van Duijn, M.J. (2021). Modelling Characters' Mental Depth in Stories Told by Children Aged 4-10. In Fitch, T., Lamm, C., and Leber, H., editors, *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, pages 2384-2390.

2. Van Dijk, B.M.A.,* Van Duijn, M.J.,* Verberne, S., and Spruit, M.R. (2023). ChiS-Cor: A Corpus of Freely-Told Fantasy Stories by Dutch Children for Computational Linguistics and Cognitive Science. In Jiang, J., Reitter, D., and Deng, S., editors, *Proceedings of the 27th Conference on Computational Natural Language Learning*, pages 352-363. Association for Computational Linguistics.

3. Van Dijk, B.M.A., Spruit, M.R., and Van Duijn, M.J. (2023). Theory of Mind in Freely-Told Children's Narratives: A Classification Approach. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics*, pages 12979-12993. Association for Computational Linguistics.

4. Van Duijn, M.J., Van Dijk, B.M.A., and Spruit, M.R. (2022). Looking from the Inside: How Children Render Character's Perspectives in Freely-told Fantasy Stories. In Clark, E., Brahman F., and Iyyer, M., editors, *Proceedings of the 4th Workshop on Narrative Understanding*, pages 66-76. Association for Computational Linguistics.

5. Van Dijk, B.M.A., Van Duijn, M.J., Kloostra, L., Spruit, M.R., and Beekhuizen, B.F. (2024). Using a Language Model to Unravel Semantic Development in Children's Use of a Dutch Perception Verb. In Zock, M., Chersoni, E., Hsu, Y., and

De Deyne, S., editors, *Proceedings of the 8th Workshop on Cognitive Aspects of the Lexicon*, pages 98-106. European Language Resources Association.

6. Van Duijn, M.J.,* Van Dijk, B.M.A.,* Kouwenhoven, T.,* De Valk, W.M., Spruit, M.R., and Van Der Putten, P.W.H. (2023). Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art Models vs. Children Aged 7-10 on Advanced Tests. In Jiang, J., Reitter, D., and Deng, S., editors, *Proceedings of the 27th Conference on Computational Natural Language Learning*, pages 389-402. Association for Computational Linguistics.

7. Van Dijk, B.M.A., Kouwenhoven, T., Spruit, M.R., and Van Duijn, M.J. (2023). Large Language Models: The Need for Nuance in Current Debates and a Pragmatic Perspective on Understanding. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12641–12654. Association for Computational Linguistics.

Curriculum Vitae

Bram van Dijk was born in Bogotá, Colombia in 1991. He obtained his bachelor's degree in Social Science in 2014 and his master's degree in History and Philosophy of Science in 2018, both from Utrecht University. After briefly working as a research assistant for various Dutch universities, he started in 2020 as a PhD candidate collaborating with dr. Max van Duijn's NWO-funded project 'A Telling Story', that was focused on unravelling Theory of Mind in children's stories. In 2023 their joint paper presenting the Dutch children's story corpus ChiSCor won the best paper award at the Conference for Computational Natural Language Learning in Singapore. In the same year, Bram was briefly a visiting researcher at the University of Toronto, collaborating with dr. Barend Beekhuizen on analysing the semantics of children's use of perception verbs. Bram completed courses in deep learning, text mining, and science communication, and co-taught the course 'Sciences and Humanities' in the Media Technology MSc programme as part of developing complementary academic skills in his PhD trajectory. Currently he is working as a postdoctoral researcher in prof. dr. Marco Spruit's Translational Data Science lab at the Leiden University Medical Center, where he focuses on the application of Computational Linguistics in the medical domain.