Universiteit Leiden
The Netherlands

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

Dijk, B.M.A. van

Citation
Dijk, B. M. A. van. (2025, January 17). *Theory of mind in language, minds, and machines: a multidisciplinary approach*. Retrieved from https://hdl.handle.net/1887/4176419

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/4176419

Note: To cite this publication please use the final published version (if applicable).

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).

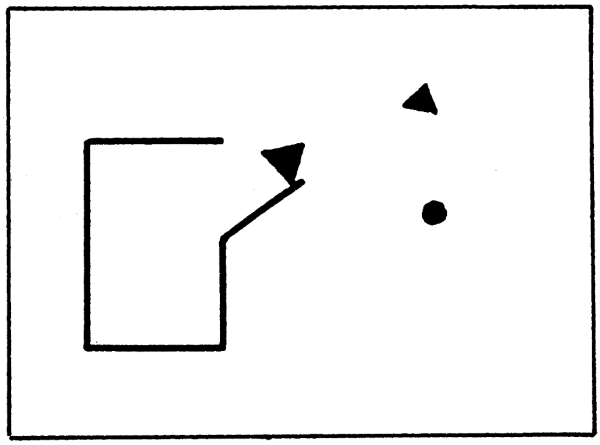


FIG. 1. EXPOSURE-OBJECTS DISPLAYED IN VARIOUS POSITIONS AND CONFIGURATIONS FROM THE MOVING FILM.
Large triangle, small triangle, disc and house.

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).

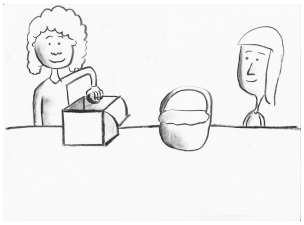
- **Statistical modelling –** Refers to statistical models such as linear models used for testing hypotheses regarding ToM (RQ1), CPR (RQ4), and language development (RQ2, RQ5).

- **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).

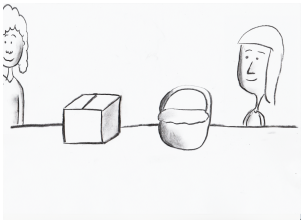
- **NLP –** Refers to experimental setups using vector models for inducing word semantics from text (RQ2), text classification based on linguistic features (RQ3), and a baseline of human ToM test performance against which LLMs are benchmarked (RQ6).

- **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.

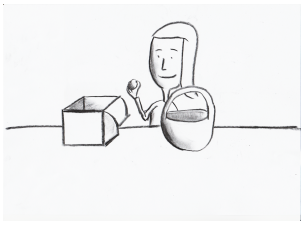
- **Theoretical analysis** – Refers to critical analysis of arguments in current debates on LLMs and their implications for studying human cognition and language understanding (RQ7). This also refers to the development of a pragmatic philosophical perspective on these debates.



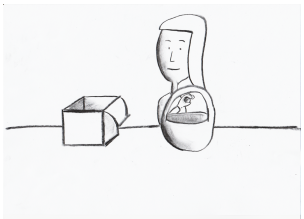
(a) This is Sally **(left)** and Anne **(right)**. They are playing. Sally has a box and Anne has a basket, and there is a ball. Sally puts the ball in her box...



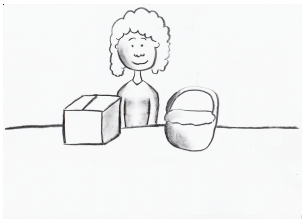
(b) Then Sally goes to play somewhere else.



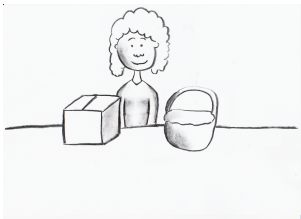
(c) Anne takes the ball from Sally's box...



(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.

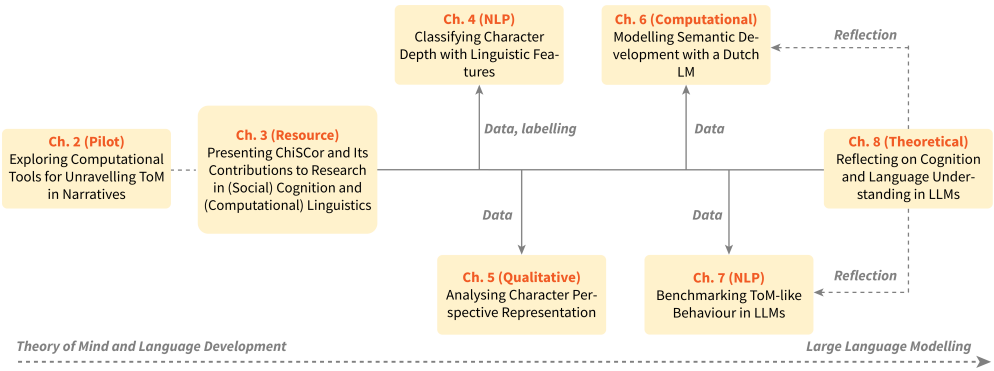


Figure 1.3: The structure of this dissertation as constituted by seven chapters and their themes.

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/`.