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

Dijk, B.M.A. van

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.