Theory of Mind in Freely-Told Children's Narratives: A Classification Approach

Bram van Dijk[1], **Marco Spruit**[1,2] and **Max van Duijn**[1]
[1]Leiden Institute of Advanced Computer Science
[2]Leiden University Medical Centre
{b.m.a.van.dijk, m.r.spruit, m.j.van.duijn}@liacs.leidenuniv.nl

Abstract

Children are the focal point for studying the link between language and Theory of Mind (ToM) competence. Language and ToM are often studied with younger children and standardized 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 4-12, recorded in their everyday classroom environments, to study language and ToM with NLP-tools. We labelled stories according to the mental depth of story characters children create, 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 ages, 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.

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 in navigating the social world and having meaningful interactions, and language evolved as an essential tool to achieve that (see e.g. Verhagen, 2005; Tomasello, 2003, 2014). 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) (Baron-Cohen, 2001; Apperly, 2012).

There is a long tradition of research in child development to understand how emerging competence in language and ToM interact, typically with standardized tests, carried out in lab settings with younger children, often below age 7 (for overviews see Milligan et al., 2007; Beaudoin et al., 2020). Yet, researchers in child development and cognition call for more ecologically valid data to study language and ToM as social phenomena (Beauchamp, 2017; Nicolopoulou and Ünlütabak, 2017; Rączaszek-Leonardi et al., 2018; Rubio-Fernández et al., 2019; Beaudoin et al., 2020; Rubio-Fernández, 2021). 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 mindreading competence. For labelling, CD indicates the *mental complexity* of characters, from flat characters without 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 re-

Findings of the Association for Computational Linguistics: ACL 2023, pages 12979–12993
July 9-14, 2023

quirement' of any 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.

Earlier work has suggested that linguistic features (e.g. vocabulary complexity) play a key role, besides age, in predicting ToM in natural language data (Van Dijk and Van Duijn, 2021), but was limited in scale; here we approach language and ToM in narratives at scale from a NLP-perspective. Our logistic 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 paper proceeds as follows. In Section 2 we reflect on relevant work, and in Section 3 we elaborate on our data and labelling. We explain feature engineering and classifier setup in Section 4, present results in Section 5, and contextualise results in Section 6.

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 standardized 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 and colleagues labelled ~11k answers on questions as either incorrect/partially correct/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 standardized 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 language and ToM.

Kovatchev et al. (2020) however do not focus on the *language children use* to reason about ToM, although their error analysis suggests that this is worthwhile to do. For one source of confusion identified for the Transformer, is that children's answers sometimes explicate what characters would *say or think*. This evidences a child shifting to a different *perspective* (van Duijn et al., 2022), 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 (Lohmann and Tomasello, 2003; De Villiers, 2005, 2007).

Yet, since it is debated whether the role of sentential complementation holds beyond the false-belief context (Slade and Ruffman, 2005; De Mulder, 2011), 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 (van Duijn et al., 2022). 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. Nicolopoulou and Richner (2007); Nicolopoulou (2016); Van Dijk and Van Duijn (2021); van Duijn et al. (2022) have found that in children's narratives, increasingly complex ways to represent characters' intentions, speech and thought can be found, which is why we look beyond standardized tests, and draw on a character depth typology established in developmental work (Nicolopoulou and Richner, 2007), for labelling stories (see Section 3). Narrative elicitation is an established way of sampling children's language skills at lexical, syntactic, phonological and pragmatic levels (Southwood and Russell, 2004;

Ebert and Scott, 2014; Nicolopoulou et al., 2015), but also on examining cognitive abilities, including memorizing, planning, organizing world knowledge (McKeough and Genereux, 2003), and ToM (Nicolopoulou, 1993). The narratives central in this paper result from children's free storytelling for a live audience of peers (see Section 3.1), 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 language 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 expected marginal 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 all possible subsets of features, including $\emptyset$. To find the marginal contribution of feature f we must compare the model's output on a given input with f included, i.e. $v(S \cup \{f\})$, against outputs of all models implementing all possible subsets of features without f, i.e. $S \subseteq N \setminus \{f\}$:

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

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

3 Data & annotation

3.1 Dataset: Stories

We collected 442 stories at various Dutch primary schools, a day care, and a community center, from 442 children aged 4-12. 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 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 story *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, 2022); Nicolopoulou and Richner (2007). 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 paper, 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]

3.2 Labelling: Character Depth

Nicolopoulou and Richner (2007); Nicolopoulou (2016) were among the first to study Character Depth (CD) in children's freely told narratives. The idea, also employed in Van Dijk and Van Duijn (2021), is that CD is a window on 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 gives not necessarily a complete view on an individual child's ToM competence; a narrative with only flat characters,

[1] https://osf.io/2es6w/?view_only=3c00905bd8fc40d9b269a6d4747e918c

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 as 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.*	023103
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 1: Translated stories from our data, traceable with ID. Underscoring shows the character the label is based on.

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 Nicolopoulou and Richner (2007)'s three-level character typology:

- **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;
- **Agents** exhibit implicit intentions-in-action, emotions and perceptions. Agents' actions are goal-directed and they can respond to events in the storyworld 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 storyworld, 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 1 for examples of CD levels and Table 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 (4-6) 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 1 that in our 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 age range. Age is a strong story-external predictor of CD (Van Dijk and Van Duijn, 2021); 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, memorizing 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

[2]We note that similar issues regarding the validity of standardized 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).

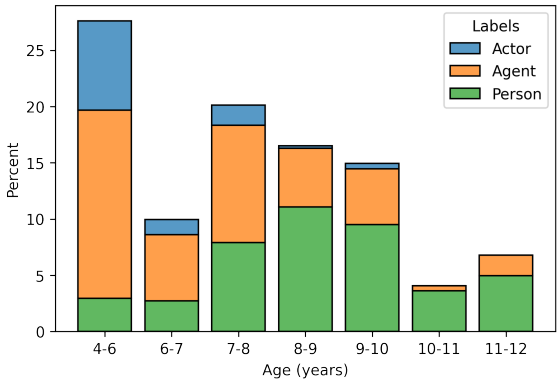

Figure 1: CD levels by the age groups standard in Dutch primary education. Bars stack to 100%.

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

Table 2: CD label distribution in our full dataset.

activity (e.g. Tsfasman et al., 2022).

4 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 $PP(V) = \sqrt[n]{\frac{1}{P(l_1,l_2,...l_n)}}$. Lemma probabilities were approximated with relative frequencies from the BasiScript lexicon, a Dutch corpus of written child essays (Tellings et al., 2018). 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; Van Dijk and Van Duijn, 2021). 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 (not complexity), which is found to predict performance on various mindreading tasks (Slade and Ruffman, 2005; Milligan et al., 2007).

- **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 $DD(S) = \frac{1}{n-s}\sum_{i=1}^{n}|DD_i|$, 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 standardized tests (Astington and Jenkins, 1999; Slade and Ruffman, 2005; Milligan et al., 2007).

- **Clausal Complementation (CC).** We extracted the average number of clausal complements per utterance with spaCy. Mastering CC has been linked to performance on a number of false belief tasks (Hale and Tager-Flusberg, 2003; Lohmann and Tomasello, 2003; De Villiers, 2005, 2007); here we examine its predictive power in the narrative domain. Complementation syntactically scaffolds reasoning about beliefs, desires, speech and perception (see Section 2).

- **Pragmatic Markers (PM).** We compute the average use per utterance of what Rubio-Fernández (2021) coined *pragmatic markers*: words used to indicate deixis and common ground. 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).** 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 binarized '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-word 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 + 199 lemmas). Since the aim is to predict CD purely from textual features, our custom features must be relatively independent from age (to not predict CD from age through language) and from one another. We computed Variance Inflation Factors (VIF) for custom features and dummy-coded age groups, with the youngest group (4-6) as reference. We adopted a threshold of 5 (James et al., 2013) as indicating problematic multicollinearity; all VIFs were low ≤ 1.54, indicating that features were relatively independent.

5 Analysis

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 regularization type and strength of our logistic classifier with 5-fold CV, which suggested L2 regularization and higher regularization 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 disclosed the linguistic information associated with 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 2 was maintained. In training, class weights were computed based on Table 2 that during training, induce a larger penalty on errors made for the infrequent class (Actor).

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

Performance metrics are given in Table 3. For the initial split, performance is reasonably good with a F1-macro of .71, given task complexity for humans (Section 3.2), 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 levels coincide with better performance. In Figure 3 we see that the most dissimilar CD levels (Actors and Persons) are never confused, which is intuitive.

5.1 Feature importance with Shapley values

We now disclose the linguistic information the model associated with specific CD levels during training with feature importance given in Figure 2.

For **Actors**, 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

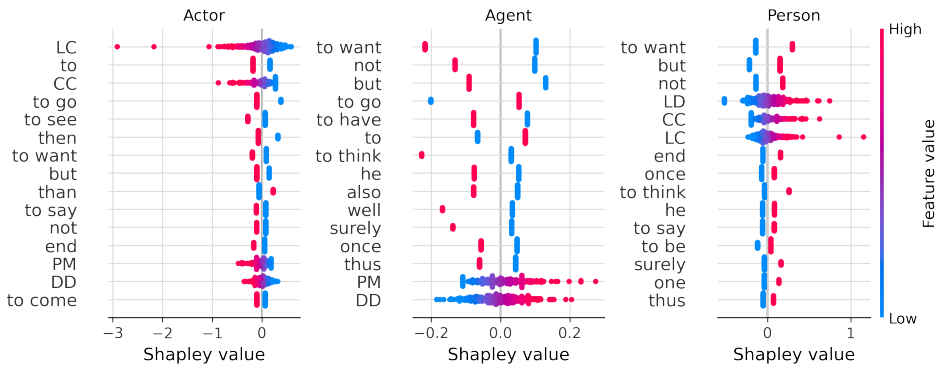

Figure 2: Shapley values for the 15 most important features per label. Value size (X-axes) quantifies importance; value sign whether the feature is a positive/negative indicator of a particular label; colour indicates for which values of that feature. For example, clausal complementation (CC) has mostly positive, red Shapley values under the Person label, indicating that more clausal complementation makes a Person label more likely; the blue negative values indicate that less clausal complementation makes a Person label less likely.

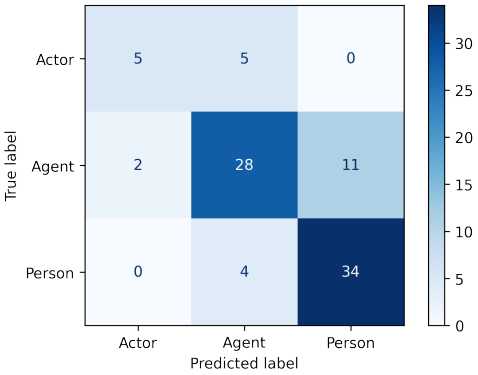

Figure 3: Confusion matrix for initial test set.

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 **Agents** we see as positive indicators use of pragmatic markers (PM) and dependency distance (DD), next to the verb 'to go' and preposition 'to', which fit Agent as action-centered 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 (van Duijn et al., 2022). Overall Agent stories appear to be linguistically more complex than Actor stories.

Person stories only have positive indicators. They are arguably linguistically most complex, as 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 Person levels, positively indicate Person stories ('but', 'not', 'thus'), suggesting that Person stories have more explicitly linked clauses. In addition, pronoun 'he' suggest 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').

5.2 Error analysis

Here we briefly discuss two prediction errors in Actor recall (Actor stories mistaken for Agent), the metric with lowest values in Table 3. For story 083101, we see in Figure 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 for the absence of various lemmas (e.g. 'to want', 'not') identified in Section 5.1. Yet, this story seems an outlier as it employs some highly unusual words (driving up LC), which sharply reduces the probability of deciding Actor; Actor stories are overall less lexically complex.

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

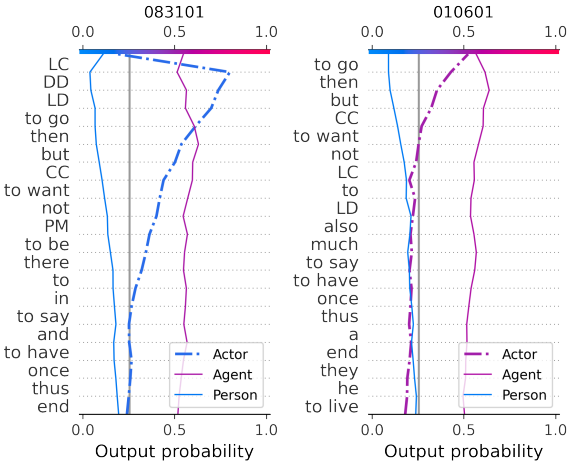

Figure 4: Decision plots for two recall errors (story IDs 083101 and 010601), that show the impact of features on the decision lines.

ity, 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 similar trajectories). Thus, Actor and Agent labels would benefit from having more unequivocal indicators. Importantly, besides exposing wrong decisions, Figure 4 also illustrates that multiple custom and lemma features shape the decision.

6 Discussion

We employed a logistic classifier on labelling Character Depth (CD) for 442 freely-told stories (Section 3). Feature engineering was used to encode key linguistic competences identified in empirical work as predictive of ToM performance (Section 4). 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 from age (Section 4). 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 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 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 2 shows, this is a precursor to explicitly spelling out the character's mental states, that then further contextualizes actions and events (as the girl's desire in the Person example does in Table 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). Another tentative indicator that a full third-person perspective shift, natural to narratives (van Duijn et al., 2022), 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 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 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 storyworld. 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 5). Person stories are lexically more diverse and complex, in line

with earlier work on predicting ToM in narratives (Van Dijk and Van Duijn, 2021): 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), and seems robust (Table 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).

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 (Schlinger, 2009; Quesque and Rossetti, 2020), and the many different standardized tests that have been designed and employed (Milligan et al., 2007; Wellman, 2018). As stated in footnote 2 (Section 3.2), these tests have their own limitations; benchmarking CD with an existing standardized measure yields no simple answer to the question whether we are now talking about children's actual ToM. That does not make standardized tests uninformative, but contextualizes 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 complement work done in controlled settings.

Our classroom context has as advantage 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 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 refer to rare settings, uncommon objects, unorthodox characters, and special social situations which is not possible in standardized 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 ('Was she angry with him?') (Leech and Short, 2007; van Duijn et al., 2022); 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.

7 Conclusion

This paper 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 (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.

8 Limitations

One limitation concerns the labelling: although there were two independent expert annotators that together labelled 16% of the stories, the rest of the labelling 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 Section 5.2) 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., 2018). Lastly, the BasiScript lexicon used for calculating lexical complexity (Section 4) is free, but a license must be signed before use, that can be obtained from the hosting institution. Also, LIWC as used for extracting the social words feature (Section 4) 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.

9 Ethics statement

This study was approved by the Leiden University Science Ethics Committee (ref. no. 2021-18). The story corpus employed in this corpus was compiled in close consultation with school teachers, principals, parents, and of course children. We used lightweight classifiers that were for our research purposes trainable in a matter of seconds, thus require 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.

10 Acknowledgements

This research was not possible without collaboration with dr. Max van Duijn's research project 'A Telling Story' (with project number VI.Veni.191C.051), which is financed by the Dutch Research Council (NWO). We thank Isabelle Blok, Yasemin Tunbul, Nikita Ham, Iris Jansen, Werner de Valk, Lola Vandame for help with data collection, and Tom Kouwenhoven for helpful comments on earlier versions of this paper. Lastly, the authors thank three anonymous reviewers for their constructive feedback.

References

Ian Apperly. 2012. *Mindreaders: the Cognitive Basis of "Theory of Mind"*. Psychology Press.

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

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

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

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

Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H. Beauchamp. 2020. Systematic Review and Inventory of Theory of Mind Measures for Young Children. *Frontiers in Psychology*, 10.

Herbert H Clark. 1996. *Using Language*. Cambridge University Press.

Hannah N.M. De Mulder. 2011. *Putting the pieces together: The development of theory of mind and (mental) language*. Landelijke Onderzoeksschool Taalwetenschap.

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

Jill G. De Villiers. 2005. Can Language Acquisition Give Children a Point of View. In *Why language matters for theory of mind*, pages 186–219.

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

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

Daniel Dor. 2015. *The Instruction of Imagination. Language as a Social Communication Technology*. Oxford University Press.

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

Kerry Danahy Ebert and Cheryl M. Scott. 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.

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

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

Daniel Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive science*, 29(2):261–290.

Courtney Melinda Hale and Helen Tager-Flusberg. 2003. The Influence of Language on Theory of Mind: A Training Study. *Developmental science*, 6(3):346–359.

Francesca G.E. Happé. 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.

Matthew Honnibal and Mark Johnson. 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.

Claire Hughes and Sue Leekam. 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.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*, volume 112. Springer.

Venelin Kovatchev, Phillip Smith, Mark Lee, Imogen Grumley Traynor, Irene Luque Aguilera, and Rory Devine. 2020. "what is on your mind?" automated scoring of mindreading in childhood and early adolescence. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6217–6228, Barcelona, Spain (Online). International Committee on Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Geoffrey N. Leech and Mick Short. 2007. *Style in fiction: A linguistic introduction to English fictional prose*. 13. Pearson Education.

Haitao Liu. 2008. Dependency Distance as a Metric of Language Comprehension Difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Heidemarie Lohmann and Michael Tomasello. 2003. The role of language in the development of false belief understanding: A training study. *Child Development*, 74(4):1130–1144.

Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Anna McAlister and Candida Peterson. 2007. A longitudinal study of child siblings and theory of mind development. *Cognitive Development*, 22(2):258–270.

Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Anne McKeough and Randy Genereux. 2003. Transformation in narrative thought during adolescence: The structure and content of story compositions. *Journal of Educational Psychology*, 95(3):537.

Karen Milligan, Janet Wilde Astington, and Lisa Ain Dack. 2007. Language and Theory of Mind: Meta-Analysis of the Relation Between Language Ability and False-Belief Understanding. *Child Development*, 78(2):622–646.

Katherine Nelson. 2005. Language Pathways into the Community of Minds. In *Why language matters for theory of mind*, pages 26–49. Oxford University Press.

Ageliki Nicolopoulou. 1993. Play, cognitive development, and the social world: Piaget, Vygotsky, and beyond. *Human development*, 36(1):1–23.

Ageliki Nicolopoulou. 2016. Promoting oral narrative skills in low-income preschoolers through storytelling and story acting. In *Storytelling in Early Childhood*, pages 63–80. Routledge.

Ageliki Nicolopoulou, Kai Schnabel Cortina, Hande Ilgaz, Carolyn Brockmeyer Cates, and Aline B. de Sá. 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.

Ageliki Nicolopoulou, Hande Ilgaz, Marta Shiro, and Lisa B. Hsin. 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.

Ageliki Nicolopoulou and Elizabeth S. Richner. 2007. From Actors to Agents to Persons: The Development of Character Representation in Young Children's Narratives. *Child development*, 78(2):412–429.

Ageliki Nicolopoulou and Burcu Ünlütabak. 2017. Narrativity and Mindreading Revisited: Children's Understanding of Theory of Mind in a Storybook and in Standard False Belief Tasks. In *Social Environment and Cognition in Language Development*, pages 151–166. John Benjamins.

Vivian G. Paley. 1990. *The Boy Who Would Be a Helicopter: The Uses of Storytelling in the Kindergarten*. Cambridge, MA: Harvard University Press.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

François Quesque and Yves Rossetti. 2020. What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science*, 15(2):384–396.

Irina Rabkina, Constantine Nakos, and Kenneth D. Forbus. 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*, Proceedings of the 41st Annual Meeting of the Cognitive Science Society: Creativity + Cognition + Computation, CogSci 2019, pages 2634–2639. The Cognitive Science Society.

Paula Rubio-Fernández. 2021. Pragmatic markers: the missing link between language and Theory of Mind. *Synthese*, 199(1):1125–1158.

Paula Rubio-Fernández, Francis Mollica, Michelle Oraa Ali, and Edward Gibson. 2019. How do you know that? Automatic belief inferences in passing conversation. *Cognition*, 193:104011.

Joanna Rączaszek-Leonardi, Iris Nomikou, Katharina J. Rohlfing, and Terrence W. Deacon. 2018. Language development from an ecological perspective: Ecologically valid ways to abstract symbols. *Ecological Psychology*, 30(1):39–73.

Valerie San Juan and Janet Wilde Astington. 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.

David Schlangen. 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.

Henry D Schlinger. 2009. Theory of mind: An overview and behavioral perspective. *The Psychological Record*, 59:435–448.

Lance Slade and Ted Ruffman. 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.

Frenette Southwood and Ann F. Russell. 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.

Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of language and social psychology*, 29(1):24–54.

Agnes Tellings, Nelleke Oostdijk, Iris Monster, Franc Grootjen, and Antal Van Den Bosch. 2018. BasiScript: A corpus of contemporary Dutch texts written by primary school children. *International Journal of Corpus Linguistics*, 23(4):494–508.

Michael Tomasello. 2003. The key is social cognition. *Language in mind: Advances in the study of language and thought*, pages 44–57.

Michael Tomasello. 2014. The ultra-social animal. *European Journal of Social Psychology*, 44(3):187–194.

Maria Tsfasman, Kristian Fenech, Morita Tarvirdians, Andras Lorincz, Catholijn M. Jonker, and Catharine Oertel. 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).

Bram Van Dijk and Max Van Duijn. 2021. Modelling Characters' Mental Depth in Stories Told by Children Aged 4-10. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, pages 2384–2390.

Max van Duijn, Bram van Dijk, and Marco Spruit. 2022. Looking from the inside: How children render character's perspectives in freely told fantasy stories. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 66–76, Seattle, United States. Association for Computational Linguistics.

Arie Verhagen. 2005. *Constructions of Intersubjectivity. Discourse, Syntax, and Cognition*. Oxford University Press.

Anne Vermeer. 2001. Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied psycholinguistics*, 22(2):217–234.

Henry M Wellman. 2018. Theory of mind: The state of the art. *European Journal of Developmental Psychology*, 15(6):728–755.

Lisa Zunshine. 2006. *Why We Read Fiction: Theory of the Mind and the Novel*. Ohio State University Press.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 8

☑ A2. Did you discuss any potential risks of your work?
Section 8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Yes, dataset (section 3) language models, textual diversity algorithm, LIWC in (section 4), classification library (section 5), library for explainable AI section (5.1)

☑ B1. Did you cite the creators of artifacts you used?
Section 3, Section 4, Section 5

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not in the paper, but we provide a license on OSF for the corpus we employed, as it is compiled by ourselves.
We did not discuss the licenses/terms of uses of all the individual libraries/tools we used as this would take up too much space. Licenses and terms of use can be found when citations of the creators of the artifacts are considered.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We did not employ artifacts in a way that was not in line with intended use; tools/libraries were used for parsing language, explaining machine learning models, and analyzing text, which is close to the use for which the tools were created. Therefore we did not discuss this in detail in the paper.
We do not further comment on the intended use of the corpus we make available, besides providing a license that the data should not be used for commercial purposes.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We do not provide details in the paper of how we exactly did the pseudonimization or coding of child IDs, although we make clear that children have unique IDs (Section 3). We provide more information on the way the data was collected in the readme on OSF.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes, section 3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3, section 5

C ☑ Did you run computational experiments?

Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 5 and 9 concern details on models used.
Not all the details are mentioned though, as our classifiers are standard and lightweight, our dataset small, and our analysis should be reproducible on standard computing infrastructure (no GPU's).

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Implementation details on packages used for preprocessing and evaluation can be found on OSF, we did not report all this information in the paper as this would ask too much detail

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 3. The full information regarding annotation is not provided in the paper, but is available on OSF.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 3

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 3

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section 3

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We do not have a large pool of external annotators, as the authors are the experts that did the rating; further details and annotation manual are available on OSF