Chapter 4

Classifying Theory of Mind in Freely Told Stories

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

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

4.1 Introduction

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

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

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

We employ an adapted version of Character Depth (CD), originating from Nicolopoulou and Richner (2007), as window onto children's ToM competence. For labelling, CD indicates the mental complexity of characters, from flat characters 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 requirement' of any Natural Language Processing (NLP) task defined by Schlangen (2021), which is having *a theory* on the relation between input (story) and output (CD label), next to the extensional requirement, which is simply the set of stories and labels. If the aim is to model humans' cognitive abilities with NLP-tools, then drawing on established work in other fields for meeting the intensional requirement is key.

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

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

4.2 Background

Few have used NLP tools on child language to study ToM, but Kovatchev et al. (2020) pioneered classifying children's ToM competence on two standardised ToM tests, the Strange Stories Task (Happé, 1994) and Silent Film Task (Devine and Hughes, 2013). In such tests, children are typically presented a vignette containing a social situation (verbally and/or visually) and are asked to explain why a character is behaving in a certain way (e.g. being ironic), thus inviting children to refer to characters' mental states. Kovatchev et al. (2020) labelled ±11k answers on questions as either incorrect, partially correct, or correct, depending on how appropriately children referred to characters' mental states, and obtained good performance (F1-macro = .91) with a DistilBERT Transformer. Indeed, accurate automatic scoring is valuable for processing standardised ToM tests. It can reduce the need for resource-intensive human evaluation of answers at larger scale (for example, Kovatchev et al. (2020) processed tests conducted with ±1k children), and explaining how models learn to identify correct answers can further our understanding of the relation between ToM and language.

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

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

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

Like Kovatchev et al. (2020), we classify child language, though not test answers but a smaller set of narratives, that are linguistically speaking likely more varied. We rely on logistic regression and custom features that encode earlier findings on 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 contribution a specific feature makes to a model's prediction. If a model is a function $v(x)$ that consists of a 'team' of N features $\{1, 2, ...n\}$, then $S \subseteq N$ denotes a possible subset of features. The marginal contribution of feature f is the difference between the model's output on a given input with f included i.e. $v(S \cup \{f\})$, and $v(S)$, where f is not included. The average marginal contribution (Shapley value) is this difference computed over all possible subsets of features without f, i.e. $S \subseteq N \setminus \{f\}$:

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

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

4.3 Methods

Dataset

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

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

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

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

Labelling

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

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

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

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

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

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

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

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

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

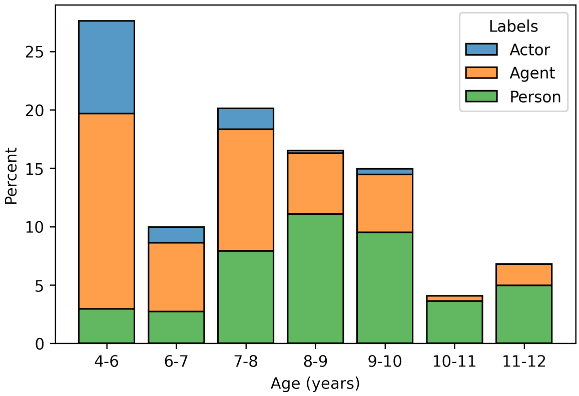

Figure 4.1: Character Depth levels by the age groups standard in Dutch primary education. Bars stack to 100%.

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

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

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

Feature engineering

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

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

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

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

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

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

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

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

- **Clausal Complementation (CC).** We extracted the average number of clausal complements per utterance with spaCy. Mastering CC has been linked to performance on several false belief tasks (De Villiers, 2005, 2007; Hale and Tager-Flusberg, 2003; Lohmann and Tomasello, 2003); here we examine its predictive power in the narrative domain. Complementation syntactically scaffolds reasoning about beliefs, desires, speech and perception (see Section 4.2).

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

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

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

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

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

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

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

4.4 Results

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

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

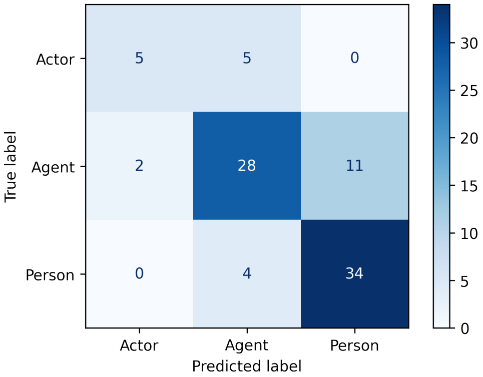

Figure 4.2: Confusion matrix for initial test set.

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

Feature importance

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

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

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

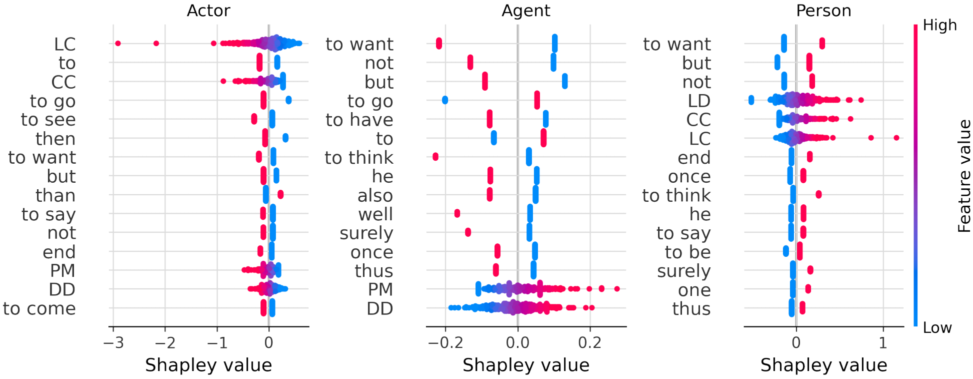

Figure 4.3: Shapley values for the 15 most important features per label. Value size (X-axes) quantifies feature importance; value sign whether the feature is a positive/negative indicator of a particular label; colour indicates for which values of that feature. For example, for clausal complementation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

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

Error analysis

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

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

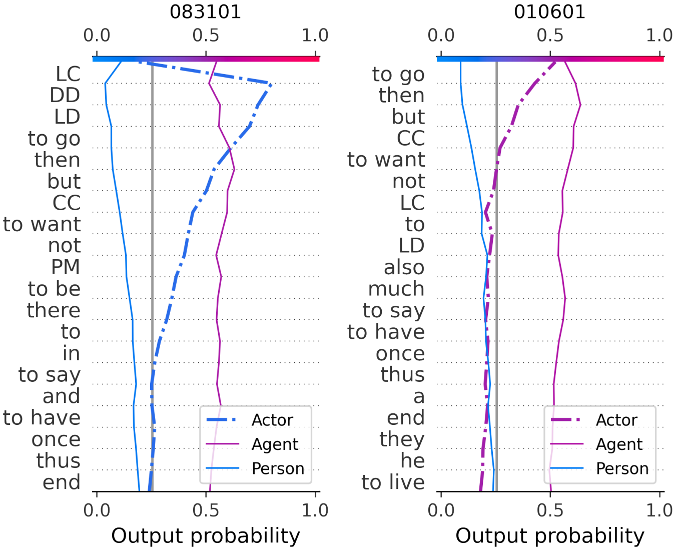

Figure 4.4: Decision plots for two recall errors (story IDs 083101 and 010601), that show the impact of features on the vertical decision lines given for each label. These plots are best read from bottom to top. Each decision line plots the probability the classifier assigns to a specific label and it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

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

4.5 Discussion

We employed a logistic classifier on labelling Character Depth for 442 freely told stories. Feature engineering was used to encode key linguistic competences identified in empirical work as predictive of ToM performance. The goal was to see how these features are reflected in ToM as manifested by CD. CD was predicted from linguistic features only, which were relatively independent of age. We now discuss the link between specific features and CD levels in the broader context of ToM, and further reflect on language and ToM competence in narratives as context-dependent phenomena.

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

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

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

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

Our classroom context has as advantages regarding ToM, that children feel more motivated to do a fun task, engage with narratives as natural finding place for mental state content, have freedom to explore the (social) scenario they want, and that their language use has a social goal: immersing the audience in their narratives as possible worlds. This social context may stimulate children more to challenge their language skills. To entice their audience, children may leverage their vocabulary skills to refer to rare settings, uncommon objects, unorthodox characters, and peculiar social situations which is not possible in standardised language tests like the Peabody Picture Vocabulary test (Dunn and Dunn, 1997). Additionally, children may also recycle complex linguistic structures and plots from prior exposure to narratives in their own narratives, to entice their audience. Thus, the influence of the social context could result in more complex language use than one would expect based on age, which makes the direct relation between age and language competence in narratives less obvious.

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

4.6 Conclusion

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

4.7 Limitations

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

4.8 Ethics Statement

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