JSLHR

Research Article

Narrative Production in Mandarin-Speaking Children: Effects of Language Ability and Elicitation Method

Li Sheng,[a] Huanhuan Shi,[a] Danyang Wang,[a] Ying Hao,[b] and Li Zheng[c]

Purpose: We compared the narrative production in Mandarin-speaking children at risk (AR) for developmental language disorder (DLD) and typically developing (TD) controls to address two goals: (a) further our understanding of the Mandarin DLD phenotype and (b) examine the role of elicitation method in differentiating AR from TD.
Method: Twenty-one AR children and 21 age- and nonverbal IQ–matched peers produced two stories from the Multilingual Assessment Instrument of Narrative, first following an adult model (i.e., story-retell) and then without a model (i.e., story-tell). Group and task effects were analyzed on macrostructure and microstructure measures.
Results: For general macrostructure score and sentence complexity, children in the AR group performed more poorly than TD children on the more challenging story-tell task and showed decreased scores from retell to tell tasks. In addition, children in the AR group showed poorer performance on number of different words. Productivity and grammaticality measures did not show group differences.
Discussion: Consistent with previous findings, grammaticality and productivity were relatively preserved but story macrostructure, lexical diversity, and sentence complexity were vulnerable in Mandarin-speaking children with or AR for DLD. Having an adult model benefited both groups in sentence complexity and story macrostructure and potentially helped maintain the performance in TD children as they engaged in the more challenging story-telling task.

The ability to tell good stories is universally valued and relates to many high-stakes outcomes such as academic achievement; social popularity; and business, legal, and political clout (Justice et al., 2010; Snow et al., 2007; Zhang et al., 2019). In the language development and disorder literature, narrative is considered an authentic and culturally valid assessment of communication because it generates rich information about the speaker's ability to integrate vocabulary, grammar, and pragmatic rules all at once (Burns et al., 2012; Fiestas & Peña, 2004; Newman & McGregor, 2006). The versatility of narrative task has led to a recent surge of research on narrative development in both typically developing (TD) children and children with developmental language disorders (DLDs) in a number of languages (e.g., for a summary, see Pesco & Kay-Raining Bird, 2016). For many understudied languages, narrative study is of particular value because it offers an entry point into delineating the phenotype of language disorder in those languages and provides a readily available form of assessment (Gagarina et al., 2012, 2015).

Mandarin Chinese is a global language with large numbers of speakers in China, parts of Asia, and several English-speaking countries (Hao et al., 2019; Rezzonico et al., 2016; Teoh et al., 2017). Given a 7% prevalence rate of DLD in children (Tomblin et al., 1997) and the belief that DLD is equally prevalent across languages and cultures (Armon-Lotem et al., 2015), 5 million 4- to 9-year-old children in China are estimated to have DLD (National Statistics Bureau of China, 2010). Reliable information about clinical markers for this population will not only facilitate early identification of DLD in monolingual Mandarin-speaking children but also enhance our ability to recognize DLD in bilingual speakers of Mandarin and English, a rapidly increasing population in many English-speaking countries (e.g., United States Census Bureau, 2015). In the current study, we aim to further our understanding of Mandarin DLD phenotype through analyses of narrative production in a sample of children who were at risk (AR) for

[a]University of Delaware, Newark
[b]University of Mississippi, Oxford
[c]Nanjing Normal University, China
Correspondence to Li Zheng: zhengli@njnu.edu.cn

Editor-in-Chief: Sean M. Redmond
Editor: Mary Alt
Received July 23, 2019
Revision received September 2, 2019
Accepted December 2, 2019
https://doi.org/10.1044/2019_JSLHR-19-00087

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

DLD and their TD peers. Our secondary aim is to evaluate the role of elicitation method (i.e., retell vs. tell) in differentiating children with and without risk for DLD.

This study makes two main contributions. First, we used an instrument that has recently gained traction in the research literature—the Multilingual Assessment Instrument of Narrative (Gagarina et al., 2012, 2015). Narrative production has been studied in Mandarin-speaking children, with the majority of the studies using the *Frog, Where Are You?* (Mayer, 1969) wordless picture book to elicit stories. The detailed linguistic analyses in these studies have yielded valuable information on the typical developmental trajectory of narrative coherence (Sah, 2007, 2013, 2015) and the use of causal statements, referential expressions, and mental state terms in children with autism (Sah, 2018; Sah & Torng, 2015, 2017). However, there is a need for elicitation materials that are culturally more neutral and contain multiple parallel stories to enable comparisons across elicitation method, time, and languages (Gagarina et al., 2012). The Multilingual Assessment Instrument for Narratives (MAIN) is a manualized instrument with standardized administration and scoring procedures and was developed as a part of the Language Impairment Testing in Multilingual Settings test battery by researchers in the European Union (Armon-Lotem et al., 2015). The stories in this instrument were carefully constructed to be familiar to young children across cultures. The MAIN has been used in more than a dozen languages (Pesco & Kay-Raining Bird, 2016), including Cantonese—a Chinese language that is typologically similar to Mandarin (Chan et al., 2018), to assess the narrative abilities of children between 3 and 10 years of age. However, Chan et al. (2018) did not test the utility of the MAIN in differentiating language ability levels, and we were the first to do this in Chinese. Second, our understanding of clinical markers of DLD in Mandarin narrative production is still very limited. We were able to find only four studies (Hao et al., 2018; Torng & Sah, 2019; Tsai & Chang, 2008; Zhang, 2013), and none of them investigated the role of elicitation method.

Studies that examined the role of language ability and elicitation method in narrative production will be reviewed next, with a particular focus on those conducted in Mandarin and those that have used the same instrument as we did. This is followed by a description of our current research objectives.

Narrative Production in DLD

Narrative competence is typically assessed through two broad types of measures: macrostructure and microstructure (Justice et al., 2010; Liles et al., 1995).

Macrostructure

Macrostructure refers to the global organization of the story and includes key components of story grammar such as character (i.e., the main and supporting characters in the story), setting (i.e., time and place), initiating event (i.e., an external or internal event that causes the character to act), action (i.e., a character's overt action toward achieving a goal), consequence (i.e., positive or negative outcome of the action), and internal response (i.e., a statement that describes the character's mental or physiological state, which can either motivate a plan/action or depict a reaction; Merritt & Liles, 1987; Stein & Glenn, 1979). These interconnected components form a cognitive framework to support a person's comprehension and production of fictional and personal stories (Johnston, 2006). While internal responses tend to be rare in young children's narrative production, elements such as initiating event, action, and consequence are produced with higher frequency. Together, these three elements form a complete episode, an important unit of narrative analysis (Merritt & Liles, 1987; Stein & Glenn, 1979).

Inconsistent findings regarding macrostructure skills in monolingual children with DLD have been reported in English. Some studies found impaired performance in the DLD group on elements of story grammar (e.g., Reilly et al., 2004), whereas others found comparable performance between DLD and TD groups (e.g., Norbury & Bishop, 2003). Two previous studies have used the MAIN to compare performance between DLD and TD groups (Altman et al., 2016; Tsimpli et al., 2016). In Altman et al. (2016), the participants were English–Hebrew bilingual children (ages 5;6–6;6 [years;months]) with and without DLD. The authors examined story macrostructure in terms of the production of goal–attempt–outcome (GAO) sequences (also known as story complexity in the MAIN scoring rubrics) in children's story-retells. Each story model from the MAIN includes three GAO sequences (for details, see Transcription and Coding section below). The DLD and TD groups did not differ on this comparison in either first language (English) or second language (Hebrew). Using the Greek version of the MAIN, Tsimpli et al. (2016) examined the production of GAO sequences in story-retells. The participants included monolingual Greek-speaking children[1] ages 5;2–11;6 (M_{age} = 9;3 for DLD and 9;0 for TD). The two groups performed comparably on GAO sequence score.

The two studies that used the MAIN both found macrostructure scores to be comparable in DLD and TD groups (Altman et al., 2016; Tsimpli et al., 2016). However, both studies examined the production of GAO sequences, a measure that does not take into account the full range of story macrostructure elements. By contrast, there were four published studies that examined narrative macrostructure production in Mandarin-speaking children with DLD (Hao et al., 2018; Torng & Sah, 2019; Tsai & Chang, 2008; Zhang, 2013), all of which showed deficits in macrostructure production in comparison to TD controls. Tsai and Chang (2008) elicited personal narratives from six children with DLD and six TD matches (ages 8;0–9;5). The DLD group showed significantly poorer performance in total

[1]This study also included bilingual children who spoke various first languages and Greek as second language. Here, we limit our discussion to the monolingual groups because our study focuses on monolingual children.

macrostructure scores and descriptions of action, setting, and character. The occurrence of internal state terms was also lower in the DLD than TD group, but the difference was not statistically significant. Zhang (2013) analyzed a longitudinal corpus of personal narratives from three children with DLD and three TD matches over a 20-month period starting at 4 years of age. The children with DLD had lower scores on overall macrostructure and descriptions of action and internal state.[2] Hao et al. (2018) elicited stories from 18 children with DLD and 18 TD controls (ages 4–7 years) using sequenced pictures. The DLD group scored significantly lower on macrostructure total scores and descriptions of character, setting, internal response, action, and consequence. Finally, Torng and Sah (2019) elicited stories using the *Frog, Where Are You?* book (Mayer, 1969) from 18 children with DLD and 18 TD controls (ages 4;11–5;10). The DLD group scored significantly lower than controls on all three macrostructure components examined by the authors, including overall story grammar elements, causal connections, and evaluative devices.

Microstructure

While macrostructure reflects cognitive abilities (Berman & Slobin, 1994), microstructure pertains to language-internal skills and refers to the use of lexical and syntactic devices to convey meaning. Microstructure encompasses measures of productivity (i.e., total number of utterances [TNU], total number of words [TNW]), lexical diversity (i.e., number of different words [NDW]), syntactic complexity (e.g., mean length of utterance [MLU], percentage of complex sentences), and grammaticality (i.e., percentage of grammatically well-formed utterances; Johnston, 2006; Justice et al., 2010; Rezzonico et al., 2016; To et al., 2010).

Several of these measures are sensitive to DLD–TD differences across Indo-European, Semitic, and Chinese languages: lexical diversity (e.g., English: Altman et al., 2016; Rezzonico et al., 2015; Greek: Tsimpli et al., 2016; Hebrew: Altman et al., 2016; Mandarin: Hao et al., 2018; Torng & Sah, 2019; Tsai & Chang, 2008; Swedish: Reuterskiöld et al., 2011), MLU (e.g., English: Altman et al., 2016; Fey et al., 2004; Hebrew: Altman et al., 2016; Mandarin: Hao et al., 2018; Tsai & Chang, 2008; Spanish: Restrepo, 1998), and percentage of complex sentences (e.g., Cantonese: To et al., 2010; English: Reilly et al., 2004; Mandarin: Hao et al., 2018). The production of grammatically well-formed utterances is sensitive to DLD in several Indo-European languages (e.g., English: Fey et al., 2004; Norbury & Bishop, 2003; Reilly et al., 2004; Spanish: Andreu et al., 2011; Swedish: Reuterskiöld et al., 2011) from kindergarten to the end of elementary school years. In contrast, a previous study of Mandarin DLD found few overt grammatical errors in both children with DLD and TD age controls (Hao et al., 2018). Finally, measures of productivity often did not show differences between groups (e.g., English: Fey et al., 2004; Guo et al., 2008; Norbury & Bishop, 2003; Mandarin: Hao et al., 2018; Torng & Sah, 2019; Swedish: Reuterskiöld et al., 2011).

[2]Both Tsai and Chang (2008) and Zhang (2013) examined personal narratives and used coding systems designed for scoring personal stories. These systems were somewhat different from the story grammar coding systems (Merritt & Liles, 1987; Stein & Glenn, 1979) used in this study. Our interpretation of their results involved conversion of terminology across the two scoring systems.

Elicitation Method

Narratives can be elicited in a myriad of ways, and a detailed account of these variations is beyond the scope of the current study (but see Boudreau, 2008, for a review). Here, we focus on two of the most widely used methods to elicit narratives from young children: story-retell and story-tell (or generation). In a retell task, children listen to a story and then have to reproduce the story with or without picture support. In a tell task, children have to construct their own stories from scratch. Knowing how narrative performance varies by elicitation method can inform both theory and practice. Story production requires the integration and coordination of many higher level processing skills (e.g., planning, working memory) and linguistic skills (Colozzo et al., 2011; Whitely & Colozzo, 2013). Studying the effect of narrative elicitation method can potentially inform us about the dynamic interplay between task demands and allocation of an individual's limited cognitive and linguistic resources. In practice, clinicians need to be cognizant of the level of demand of each elicitation method and make informed decisions when planning assessment (Boudreau, 2008).

Merritt and Liles (1989) were among the first to examine both story-tell and story-retell performance in English-speaking children with DLD. The participants were 40 children, ages 9;0–11;4; half were those with DLD, and half were TD age controls. The story-tell tasks consisted of three brief story stems (e.g., "Once upon time, two friends were in a deep and dark cave…," p. 446). The story-retell tasks consisted of two multi-episode adventure stories, which were modeled to children by the experimenter. The tell tasks always preceded the retell tasks. For both groups, story-retell elicited longer stories, more occurrences of story grammar components, and more complete episodes than story-tell. In addition, the two groups differed on both macrostructure and microstructure measures in retell tasks and on macrostructure measures only in tell tasks. The authors argued that the story models provided in the retell tasks led to longer story samples, which allowed more complete assessment of the DLD children's areas of difficulty.

Westerveld and Gillon (2010) examined story-tell and story-retell performance in 22 English-speaking children ages 7–9 years, half with mixed reading disability, a disorder highly comorbid with DLD (Pennington & Bishop, 2009), and half were TD controls. The task order was randomized. In the retell task, children listened to a story based on *A Boy, a Dog, and a Frog* (e.g., Mayer, 1967); answered comprehension questions and received corrective feedback; listened to the story again; and then retold the story without pictorial aid. In the tell task, children listened to the dragon story on the Test of Narrative Language (Gillam & Pearson,

2004), answered comprehension questions, and then generated their own story based on the single-scene alien picture from the Test of Narrative Language. The retell task elicited stories that were longer and lexically more diverse than the tell task in both groups of children. Moreover, the two groups differed significantly on measures of story length, lexical diversity, and sentence complexity on the retell task but only on one measure (percentage of grammatically acceptable sentences) on the tell task. The authors attributed the more pronounced group differences in retell performance to the higher memory demands because the withdrawal of picture aid required the children to retrieve the story model from memory.

In both Merritt and Liles (1989) and Westerveld and Gillon (2010), the materials and procedures used in the retell and tell tasks were quite different. Therefore, task performance differences were confounded by variations in story stimuli and procedures. Three studies (Kunnari et al., 2016; Otwinowska et al., 2018; Roch et al., 2016) used the MAIN, which contained comparable stimuli and procedures between tasks, and examined differences between tell and retell tasks in bilingual and monolingual children (ages 3–7 years) who spoke various Indo-European languages (i.e., Italian–English, Finnish–Swedish, Polish–English). However, none of these studies included participants with DLD. Across these three studies, there was consistent evidence that retell elicited higher story structure scores than the tell task. The story structure score from the MAIN represents a summative macrostructure score that accounts for the production of 17 story grammar elements in each story. Kunnari et al. also calculated the story complexity (i.e., GAO sequence) score and found that the retell task elicited more advanced GAO sequences than the tell task. As for microstructure, both Kunnari et al. and Otwinowska et al. found that the retell task elicited longer story samples than the tell task. Finally, in Kunnari et al., the monolingual Finnish-speaking children produced longer stories and received higher story structure scores than the Finnish–Swedish bilinguals on the tell task, suggesting that the tell task was sensitive to group differences among TD children with different levels of language proficiency.

To summarize, in all of these studies, story-retell elicited stronger performance than the story-tell task, regardless of age, testing language, whether the child is bilingual or monolingual, and whether the child had a disability or was TD. Retell also revealed more differences between diagnostic groups than the tell task (Merritt & Liles, 1989; Westerveld & Gillon, 2010), but it was unclear whether this task difference can be generalized to materials and procedures that are more closely equated. Last but not least, using stimulus materials of comparable difficulty, Kunnari et al. found that the tell task revealed differences in TD children who had different levels of language exposure and proficiency.

Current Study Objectives

The preceding literature review suggests that narrative macrostructure and certain elements of microstructure are vulnerable in Mandarin-speaking children with DLD (Hao et al., 2018; Torng & Sah, 2019; Tsai & Chang, 2008; Zhang, 2013). However, these previous studies tended to have very small sample sizes and numerous comparisons that could inflate the rate of false positives (e.g., Zhang: $n = 3$ per group; Tsai & Chang: $n = 6$ per group) and were variable in the elicitation materials used (Tsai & Chang, 2008, and Zhang, 2013, elicited personal narratives, whereas Hao et al., 2018, used sequenced pictures from an unpublished test, and Torng and Sah, 2019, used a frog story book to elicit stories). To further our understanding of the Mandarin DLD phenotype and enable comparison to other languages, in the current study, we elicited stories from a larger sample of children with and without risk for DLD using the widely used MAIN (Gagarina et al., 2012). We aim to address the following questions:

1. What is the effect of DLD risk status on the production of narrative macrostructure and microstructure?

Altman et al. (2016) and Tsimpli et al. (2016) did not find DLD–TD differences in the production of GAO sequences. Given these null results, we turned our focus to scores on the story structure section of the MAIN. We predicted that the story structure score, which considers 17 different story elements for each story, would be robust in detecting differences in children's linguistic ability.

There is a long list of microstructure measures that could be potentially included. In the current study, we decided to focus on four measures (i.e., MLU, NDW, percentage of complex utterances, and grammaticality) that have consistently shown diagnostic sensitivity to language disorders across many languages. Specifically, we predicted that, similar to previous studies (Hao et al., 2018; Torng & Sah, 2019; Tsai & Chang, 2008; Zhang, 2013), the current sample of children would also show group differences in MLU, NDW, and sentence complexity. Hao et al. (2018) was the first and only study that examined grammaticality in Mandarin DLD. The unusually low occurrence of ungrammatical utterances in Hao et al.'s sample warranted replications.

2. What is the effect of elicitation method on the production of narrative macrostructure and microstructure?

To answer this question, each child was given two narrative tasks—a story-retell followed by a story-tell. Previous studies either used a tell-then-retell task order (Kunnari et al., 2016; Merritt & Liles, 1989; Otwinowska et al., 2018) or randomized the task order (Roch et al., 2016; Westerveld & Gillon, 2010). The tell-then-retell order is meant to avoid potential carryover effect of learning. What has not been explored in the literature is the retell-then-tell task order. A potential advantage of this order is that it serves to provide a definition or model of the task. When faced with an artificial elicitation task, young children may approach it in different ways by describing the pictures, telling a story that is comprehensible only with picture support, or telling a story in a more literate style (Berman & Slobin, 1994). Hearing an adult model and then retelling the story can set some expectations of the intended

task before asking children to generate their own stories, therefore alleviating the concern that children may have different familiarity and initial understanding of the task.

Consistent with previous studies (Kunnari et al., 2016; Merritt & Liles, 1989; Otwinowska et al., 2018; Roch et al., 2016; Westerveld & Gillon, 2010), we predicted that story-retell would elicit stronger performance than story-tell on both macrostructure and microstructure measures, and this story-retell advantage would be true for both groups of children. The literature does not provide clear directions regarding which task from the MAIN, story-tell or story-retell, would be more revealing of language ability differences. Therefore, we did not pose a priori predictions on interactions between risk status and elicitation method.

Method

Participants

The participants in the current study were selected from a sample of 142 children who completed a language screening battery in the fall of 2017. An additional 60 children participated in testing but did not complete all of the screening measures and were therefore not included in the local norm. The 142 children included in the local norm were recruited from a preschool in Nanjing, China. Preschool education is optional but common in urban areas of China. Most preschools have three grade levels, with Grade 1 intended for children who are 3–4 years of age, Grade 2 for 4- to 5-year-olds, and Grade 3 for 5- to 6-year-olds. Parents signed a consent form approved by the institutional review board of the University of Delaware.

The screening battery included the Primary Test of Nonverbal Intelligence (Ehrler & McGhee, 2008), a sentence recall task (Wang et al., 2019), a vocabulary task (Sheng et al., 2019), and a teacher questionnaire that collected ratings of the child's oral language in the domains of vocabulary, sentence length, speech intelligibility, listening comprehension, and grammatical proficiency (adapted from the Inventory to Assess Language Knowledge; Peña et al., 2018). The internal consistency (Cronbach's alpha) of the sentence recall task and the vocabulary task was .865 and .844, respectively. These values exceeded generally accepted threshold for good reliability (Henson, 2001) and indicated that the items on these tasks were assessing the same skills. A hearing screening was not conducted at the time of testing. However, children had to pass a hearing screening upon enrollment in the preschool, and parents reported normal hearing in a background questionnaire. Parents also reported information about maternal education and the child's amount of exposure to other dialects of Chinese. All children had no emotional, behavioral, neurological, or severe articulation/phonological deficits as reported by parents.

The 142 children were divided into three age bands: 3;6–4;5 (n = 50), 4;6–5;5 (n = 52), and 5;6–6;5 (n = 40). For each child, we calculated z scores for sentence recall, vocabulary, and composite teacher rating on language abilities (i.e., averaged across the five questions) using his or her age group as the reference. To qualify as AR for DLD, the child must have two out of three z scores that were 1 *SD* below their age group mean or one z score that was 1.5 *SD*s below the group mean. This selection process was consistent with the view that assessment should combine both objective test scores and ratings of functional abilities by familiar adults (Bedore et al., 2011; Bishop, 1997). Twenty-one children met these criteria. Each of the 21 children was matched to a TD child. To be included as TD control, the child must score above −1 *SD* below the mean on all three measures and demonstrate the following matching criteria with an AR child: (a) within 3 months of chronological age, (b) within 1 *SD* on nonverbal IQ standard score, (c) within 1 point for maternal education, and (d) within 1 point for Mandarin exposure. Sex was the last criterion and was exempted if the previous criteria were met. Table 1 shows the background characteristics of the AR and TD groups. The two groups were closely matched on age, nonverbal IQ, and amount of Mandarin exposure. The TD group had higher maternal education, but the difference was not statistically significant. The AR group demonstrated significantly lower scores on the language screening tests with large to huge effect sizes. These children were AR in the sense that they did not have a formal diagnosis of DLD but performed poorly on language tasks with reference to a local norm.

The screening and norming process yielded a total of 42 monolingual Mandarin-speaking preschoolers with and without risk for DLD. These children were invited to participate in a longitudinal study about children's language and cognitive development. All parents gave consent, and children were retested in the spring of 2018 on tasks of narrative production. At the time of the narrative testing, both the AR and TD groups had an average age of about 68 months.

Materials

We used the Mandarin version of the MAIN (Gagarina et al., 2012, 2015) to elicit stories. The MAIN is a part of the Language Impairment Testing in Multilingual Settings test battery (Armon-Lotem et al., 2015) and contains four parallel stories that can be used to elicit story-retells or self-generated stories. The four stories were designed to be comparable in terms of the number of characters, the number of episodes, the depiction of character actions and emotions, and the overall conceptual and visual complexity. The stories involve characters and plots that are familiar to young children across cultures and have been used in more than a dozen languages (Pesco & Kay-Raining Bird, 2016). In the current study, we used the Cat and Dog stories to elicit story-retells and the Bird and Goat stories to elicit story-tells. Each story contains six pictures. The Mandarin model of the Cat story contains 17 utterances, 189 total words, 91 different words, and an MLU of 11.12. The Mandarin model of the Dog story contains 17 utterances, 196 total words, 88 different words, and an MLU of 11.53.

Table 1. Characteristics of at-risk (AR) and typically developing (TD) groups.

Measure	AR (*N* = 21)	TD (*N* = 21)	Cohen's *d* (*p* level)
Sex	8 F, 13 M	12 F, 9 M	
Age (in months)			
M	67.6	67.5	0.02 ($p > .5$)
SD	8.3	9.2	
Range	51 to 77	50 to 80	
Nonverbal IQ[a]			
M	120.6	120.1	0.03 ($p > .5$)
SD	15.5	15.7	
Range	95 to 149	95 to 149	
Maternal education[b]			
M	3	3.5	–0.54 ($p > .1$)
SD	1.1	0.7	
Range	1 to 5	2 to 5	
Mandarin exposure[c]			
M	4.5	4.7	–0.29 ($p > .3$)
SD	0.7	0.7	
Range	3 to 5	3 to 5	
Sentence repetition *z* score[d]			
M	–0.9	0.4	–1.37 ($p < .001$)
SD	1.2	0.6	
Range			
	–3.9	–2.2	
Language rating *z* score[e]			
M	–0.9	0.2	–1.08 ($p < .001$)
SD	1.2	0.8	
Range	–2.6 to 0.8	–0.8 to 1.9	
Vocabulary composite *z* score[f]			
M	–0.8	0.5	–1.41 ($p < .001$)
SD	1.1	0.7	
Range	–2.9 to 1.2	–0.7 to 2.0	

Note. F = female; M = male.

[a]Nonverbal IQ was measured by Primary Test of Nonverbal Intelligence. [b]A scale of 5 was used for parents to report their education level: 1 = *middle school*, 2 = *high school*, 3 = *some college*, 4 = *bachelor's degree*, and 5 = *master's degree or higher*. [c]A scale of 5 was used for parents to report their child's Mandarin exposure: 1 = < *20%*, 2 = *20%–39%,* 3 = *40%–59%*, 4 = *60%–79%*, and 5 = *80%–100%*. [d]The 142 participants were divided into three age groups: 3;6–4;5 (50), 4;6–5;5 (52), and 5;6–6;5 (40). *z* scores in each age group were calculated on their sentence repetition performance. [e]The *z* score of teachers' rating of the children's language ability. [f]Children's *z* score on the vocabulary task.

Procedure

The 42 children identified from the local norm were administered the MAIN in the spring of 2018, approximately six months after they were initially tested on the language screening battery. Administration of tasks followed the MAIN manual. All participants were individually tested by an experimenter in a quiet room at their school. Following a short conversation to establish rapport, testing began. The story-retell task was always given before the story-tell task. For story-retell, the experimenter presented the child with three envelopes that all contained the same story. After the child selected one envelope, the adult let the child preview the pictures in the story and then read the story script while showing the child the pictures. After that, the child was asked to retell the story with the aid of the pictures. Only two pictures were shown at a time to ensure that children tell the story in sequence. For story-tell, again, the child was asked to choose one from three envelopes that contained the same story. The child was instructed to not let the adult see the pictures. After the child previewed the pictures, the adult held up the pictures, two at a time, and asked the child to tell a story. The adult did not look at the pictures throughout the story-tell task to create the pretense that the story chosen by the child was unknown to the adult. The session was audio-recorded for transcription and reliability purposes.

Transcription and Coding

The second author transcribed all narrative samples into Chinese characters using the Codes for Human Analysis of Transcripts format. Computerized Language Analysis was used to code the narratives (MacWhinney, 2000). Utterances were segmented into clauses. Adopting previous definitions (Sah, 2013; Zhang, 2013), a clause "consists of a verb and its arguments, and corresponds roughly to a single event" (Sah, 2013, p. 174; for segmentation rubrics and examples of Mandarin clauses, see the Appendix). Word segmentation followed the conventions developed by Cheung et al. (2011). Measures of productivity, lexical diversity, and sentence length were generated using the Computerized Language Analysis software. Each clause was

coded as simple or complex, and grammatical or ungrammatical by the second author.

We used the MAIN scoring rubric to code macrostructure into two scores: story structure score and story complexity score (Gagarina et al., 2012). To generate the story structure score, each story was coded for the presence of two setting elements (time and place) and three episodes, each of which included five elements: initiating event, the character's goal, attempt, outcome of the attempt, and reaction. A score of 1 was assigned if a macrostructure element was produced, resulting in a possible total score of 17 for each story.

The story complexity score focused on three elements of an episode: goal, attempt, and outcome. All of the episodes were classified into one of the following levels of complexity: (a) complete episodes (i.e., GAO); (b) abbreviated or incomplete episodes, which included a goal statement but could be missing either or both of the other two elements (i.e., G, GA, GO); (c) action or reaction sequences, wherein a statement of the character's goal is missing (i.e., AO); and (d) isolated descriptions that included only A, only O, or none of the three elements. Consistent with Kunnari et al. (2016), we then assigned dichotomized scores for each episode: All episodes that included a goal statement (i.e., G, GA, GO, and GAO) received a score of 1, whereas the rest (i.e., AO, A, O, and none) received a score of 0. This complexity hierarchy was based on Westby's (2005) story decision tree that prioritizes the expression of the main character's intentionality.

Reliability

A second coder (i.e., the third author) independently transcribed and scored the macrostructure of narrative samples of four children randomly selected from each group (19% of the sample) to establish interjudge agreement. Both coders are native speakers of Mandarin and were blinded to the group status of the child to avoid potential biases. To calculate interjudge consistency of the story structure score, we compared the two judges' scores for each of the 17 story grammar elements and derived Cohen's k, a statistic representing interrater reliability for qualitative data. Agreements between the two coders were substantial for retell ($k = .864$, $p < .001$) and tell ($k = .843$, $p < .001$; Landis & Koch, 1977), and inconsistencies were resolved through discussion. Because story complexity score was derived from a subset of the elements in the story structure score, reliability measure was not applied for story complexity. Mean agreement of transcription was 97.5% on the total number of Chinese characters[3] per sample and 95.2% on character types (i.e., percentage of characters that were transcribed identically by both judges). Mean agreement on word segmentation was 91.1%. To check the reliability of clause segmentation, clause complexity, and grammaticality coding, the third author reviewed all transcripts that were segmented and coded by the second author and all inconsistencies were discussed among the first three authors until consensus was reached. Clause segmentation and word segmentation formed the basis for the microstructure elements of TNU, TNW,[4] NDW, and MLU.

[3]Each Chinese character represents a syllable and a morpheme that can be a stand-alone word or part of a word (Wiedenhof, 2015).

Results

For measures that used continuous data (e.g., NDW, MLU, story structure score), we conducted mixed analysis of variance (ANOVA) with risk status (AR, TD) as a between-subjects variable and task (retell, tell) as a within-subject variable. We used η_p^2 (partial eta squared) to indicate effect size: $\eta_p^2 > .01$ is small, $\eta_p^2 > .06$ is medium, and $\eta_p^2 > .14$ is large (Huck, 2009). All assumptions for ANOVA were met, including normality, homogeneity of variance, and independent observations.

Complexity and grammaticality of clauses were categorical data: Each clause was coded as complex/noncomplex and grammatical/ungrammatical. It is recommended that generalized linear mixed models should be used to analyze this kind of data (Jaeger, 2008). We implemented two generalized linear models, and the dependent variables were entered as binary data (e.g., 0 = grammatical and 1 = ungrammatical). A third generalized linear model was performed on the story complexity score (0 = when the goal statement was not present in an episode and 1 = when the goal statement was present). Fixed effects included risk status and elicitation task. The interaction between risk status and task was also entered. Participants were treated as a random effect.

Before analyzing macro- and microstructure, we examined story length (i.e., TNU, TNW), which could potentially lead to macro- and microstructure score differences between the AR and TD groups. In other words, a child who produced longer stories would be more likely to score higher in macro- and microstructure than a child who produced few utterances. Two mixed two-way ANOVAs were implemented for TNU and TNW. There were no significant main effects of risk status on TNU, $F(1, 40) = 1.76$, $p = .19$, $\eta_p^2 = .04$, and TNW, $F(1, 40) = 2.51$, $p = .12$, $\eta_p^2 = .06$. No main effects of task were found (TNU: $F(1, 40) = 1.28$, $p = .27$, $\eta_p^2 = .03$; TNW: $F(1, 40) = 0.71$, $p = .41$, $\eta_p^2 = .02$). Moreover, there was no significant interaction between risk status and task (TNU: $F(1, 40) = 0.76$, $p = .39$, $\eta_p^2 = .02$; TNW: $F(1, 40) = 1.64$, $p = .21$, $\eta_p^2 = .04$). Thus, TNU and TNW were not included as covariates in the following analyses. Table 2 shows the descriptive data of TNU and TNW.

Macrostructure

Story structure score. We detected a main effect of risk status, $F(1, 40) = 4.73$, $p = .036$, $\eta_p^2 = .11$; a marginal effect of task, $F(1, 40) = 4.03$, $p = .051$, $\eta_p^2 = .09$; and an

[4]TNU and TNW were not key dependent measures in this study. However, they were analyzed to exclude potential confounds (see details in the beginning of the Results section).

Table 2. Descriptive data of total number of utterances (TNU) and total number of words (TNW).

		AR			TD		
Measures	**Task**	***M***	***SD***	**Range**	***M***	***SD***	**Range**
TNU	Retell	13.52	3.91	7–24	14.38	3.11	8–19
	Tell	13.71	5.34	6–28	15.86	4.80	9–26
TNW	Retell	96.43	31.15	57–157	104.43	24.46	50–141
	Tell	94.24	44.47	39–226	114.29	30.35	59–162

Note. AR = at-risk; TD = typically developing.

Elicitation Task × Risk Status interaction, $F(1, 40) = 5.14$, $p = .029$, $\eta_p^2 = .11$. Post hoc tests with Tukey corrections showed that the TD children outperformed the AR children on the story-tell task ($p = .002$), but the two groups did not differ on the retell task. In the TD group, retell and tell yielded comparable performance; in the AR group, the retell task resulted in better story structure score than the tell task ($p = .004$). Descriptive data of mean story structure scores by task are presented in Table 3.

Story complexity score. A generalized linear mixed model was conducted to analyze the effect of risk status, task, and the interaction between task and risk status. There was a marginal main effect of risk status, $F(1, 248) = 3.68$, $p = .06$, odds ratio = 0.67. The TD group was 33% more likely than the AR group to produce complex story episodes including goals. There was no main effect of task, $F(1, 248) = 0.53$, $p = .53$, odds ratio = 0.93, or interaction between risk status and task, $F(1, 248) = 0.13$, $p = .72$, odds ratio = 0.83.

Microstructure

NDW. There was a significant main effect of risk status, $F(1, 40) = 4.80$, $p = .03$, $\eta_p^2 = .11$. The TD group outperformed the AR group, with an effect size between medium and large. There was no main effect of task, $F(1, 40) = 0.20$, $p = .65$, $\eta_p^2 = .005$, nor was there an interaction between risk status and task, $F(1, 40) = 2.15$, $p = .15$, $\eta_p^2 = .05$ (see Table 4).

MLU. We did not find significant main effects of risk status, $F(1, 40) = 1.09$, $p = .30$, $\eta_p^2 = .03$, and task, $F(1, 40) = 0.87$, $p = .36$, $\eta_p^2 = .02$. No significant interaction between risk status and task was found, $F(1, 40) = 1.95$, $p = .17$, $\eta_p^2 = .05$.

Table 3. Macrostructure descriptive data.

		AR			TD		
Measures	**Task**	***M***	***SD***	**Range**	***M***	***SD***	**Range**
Story structure score	Retell	8.90	2.45	6–14	9.43	2.60	3–14
	Tell	7.33	2.50	3–12	9.52	1.75	6–12
Story complexity score[a]	Retell	1.00	1.00	0–3	1.29	1.01	0–3
	Tell	1.05	0.59	0–2	1.48	0.87	0–3

Note. AR = at-risk; TD = typically developing.

[a]The score was averaged overall three episodes for each story.

Percentage of complex clauses. The generalized linear mixed model showed that the elicitation task had a significant effect on the usage of complex clauses, $F(1, 1197) = 8.69$, $p = .003$, odds ratio = 1.79. The retell task was 79% more likely to yield complex clauses than the tell task. The effect of risk status was significant, $F(1, 1197) = 3.86$, $p = .050$, odds ratio = 0.89. The TD group was 11% more likely than the AR group to produce complex clauses. The interaction between risk status and task approached significance, $F(1, 1197) = 2.82$, $p = .094$, odds ratio = 0.66, and was followed with pairwise comparisons. The two groups showed different performance on the tell task (TD > AR, $p = .01$) but not the retell task ($p = .57$). The AR group showed better performance on retell than tell ($p = .002$), and the TD group did not show a difference between tasks ($p = .341$).

Percentage of ungrammatical clauses. Indicated by the generalized linear mixed model, there were no significant effects of risk status, $F(1, 1199) = 0.99$, $p = .32$, odds ratio = 1.29, and task, $F(1, 1199) = 0.02$, $p = .89$, odds ratio = 1.06. No significant interaction between risk status and task was found, $F(1, 1199) = 0.008$, $p = .93$, odds ratio = 0.96.

Discussion

The current study adds to the emerging literature on narrative production in Mandarin-speaking children with DLD. We gave the Mandarin version of the MAIN story-retell and story-tell tasks to 42 children with and without risk for DLD. Consistent with our predictions, we found effects of risk status and elicitation method on both macrostructure and microstructure production. We now address each factor in turn.

DLD Risk Status

Recall that two previous studies (Altman et al., 2016; Tsimpli et al., 2016) did not find DLD–TD differences in macrostructure when comparisons were made on the production of goal–action–outcome sequences. We hypothesized that the null effect may be due to the inclusion of too few story grammar elements and that the story structure score, which takes into account 17 story grammar elements for each story, would be a more robust measure. The results confirmed our predictions. We found a marginal difference in GAO sequence scores and a significant difference in overall story structure scores. Consistent with previous studies of Mandarin-speaking children with DLD (Hao et al., 2018; Torng & Sah, 2019; Tsai & Chang, 2008; Zhang, 2013) and English-speaking children with DLD (e.g., Reilly et al., 2004), the results suggested that weaker macrostructure production was part of the DLD phenotype.

Nevertheless, there are important caveats related to how macrostructure should be measured. The GAO sequence, as defined in the MAIN manual, includes a

Table 4. Microstructure analyses.

Measures	Task	AR			TD		
		M	*SD*	Range	*M*	*SD*	Range
NDW	Retell	48.86	12.10	32–75	53.52	11.55	31–76
	Tell	47.14	13.39	29–78	56.76	10.34	33–72
MLU	Retell	7.23	1.61	4.87–11.55	7.30	1.04	5.00–9.18
	Tell	6.79	0.90	4.83–8.14	7.38	1.09	5.81–10.33
Proportion of complex clauses	Retell	0.36	0.20	0.11–0.82	0.39	0.15	0.13–0.58
	Tell	0.25	0.12	0–0.50	0.37	0.16	0.06–0.78
Proportion of ungrammatical clauses	Retell	0.06	0.06	0–0.20	0.05	0.07	0–0.25
	Tell	0.07	0.08	0–0.29	0.04	0.06	0–0.21

Note. AR = at-risk; TD = typically developing; NDW = number of different words; MLU = mean length of utterance.

statement of the character's goal, followed by action, and then outcome of the action. Of these three components, the goal (i.e., plan) statement is late to emerge, whereas action and outcome tend to be produced with higher frequency in young children's narrations (Merritt & Liles, 1987; Stein & Glenn, 1979). The GAO sequence score is well reasoned in light of developmental expectations (Westby, 2005) and has the advantage of quicker scoring in a busy clinical setting. However, other essential elements such as the setting and later emerging elements such as the characters' emotional states are left out. The current findings, together with results from Altman et al. (2016) and Tsimpli et al. (2016), suggested that the story complexity score is not ready for clinical adoption given its low diagnostic value. Instead, scoring needs to account for the full range of story grammar elements to be clinically informative.

We compared group performance on four microstructure measures: NDW, MLU, percentage of complex clauses, and percentage of ungrammatical clauses. To rule out story length as a potential confound, we also compared groups on TNU and TNW. The results were largely consistent with Hao et al. (2018) and Torng and Sah (2019). Specifically, we replicated Hao et al. and found story length to be comparable and lexical diversity (NDW) and sentence complexity (% complex clauses) to be differentiating between ability groups. Unlike Hao et al., we did not find a significant group difference in MLU. It is worth noting that group differences in MLU for children in this age range are not always significant (Altman et al., 2016).

Torng and Sah (2019) also reported comparable story length and differences in lexical diversity. However, the DLD and TD groups in their study were similar in the production of complex sentences, with around 16% of the utterances produced by both groups coded as complex. In the current study, average percentage of complex sentence was 30% for the DLD group and 38% for the TD group. Several methodological and participant characteristic differences could account for these divergent results. First, our participants were slightly older in age and had much higher nonverbal IQ. Second, different stories were used in the two studies. Third and perhaps most importantly, different approaches were taken for utterance segmentation, and different definitions of complex sentences were applied. We used syntax as the primary criterion for segmentation, whereas Torng and Sah used prosody as the primary criterion. Furthermore, Torng and Sah defined complex sentences into six subtypes (i.e., relative clause, clausal complement, serial verb construction, pivotal construction, ba construction, bei construction), whereas we excluded ba and bei constructions and included adverbial clauses as complex sentences (see the Appendix for details).

In spite of these differences, the current study, Hao et al. (2018), and Torng and Sah (2019) converged in that NDW was a robust indicator of Mandarin DLD, whereas story length was not. These findings are consistent with a body of literature on narrative production across languages in suggesting that children with DLD can produce almost as many utterances as their TD peers in a narrative task. However, their utterances may be lacking in information density and vocabulary variety (e.g., Guo et al., 2008; Hao et al., 2018; Hewitt et al., 2005).

Recall that Hao et al. was the only study that examined grammaticality in Mandarin-speaking children with and without DLD. They found that only 4% of the utterances produced by children with DLD contained overt grammatical errors, a value quite close to the 2% produced by TD children. Similarly, only 6%–7% of the clauses produced by the AR children in our sample contained grammatical errors. Again, the value was nondifferentiable from the 5% produced by the TD group. These converging findings suggested that, unlike many Indo-European languages, low grammaticality may not be a central characteristic of the Mandarin DLD phenotype. In languages with richer morphology, errors of tense marking, and number, person and gender agreement are frequent and symptomatic of DLD (Andreu et al., 2011; Fey et al., 2004; Norbury & Bishop, 2003; Reilly et al., 2004; Reuterskiöld et al., 2011). In contrast, when constructing sentences in Mandarin, children do not need to constantly attend to tense and agreement marking, which in effect eliminates many opportunities for overt errors. In addition, although Mandarin uses grammatical morphemes to mark aspect, these

morphemes are not obligatory (Hao et al., 2018; Klein et al., 2000). Therefore, omissions of aspect markers do not make sentences ungrammatical. When children did make errors, they tended to be syntactic and/or semantic in nature. As seen in Table 5, main error types were errors of word order; missing verbs and arguments; and wrong use of aspect markers, classifiers, and verbs.

Among other things, we were able to replicate Hao et al.'s (2018) finding of high grammaticality with a new sample of children with different background characteristics. For example, the children in our study differed from those in Hao et al. in terms of recruitment procedures (screening at preschool vs. recruiting from outpatient clinic), severity of DLD (AR for DLD vs. diagnosis from pediatricians), and nonverbal IQ (closely matched with TD and above average vs. somewhat lower than TD and low average). The two samples were also tested with different stimulus materials. Converging findings between the two studies led us to conclude that manifestations of DLD are dependent on linguistic typology (Leonard, 2014). Unlike in Indo-European languages, researchers and clinicians cannot rely on frequent occurrence of grammatical errors as a telltale sign of Mandarin DLD. Instead, attention should be directed to lexical diversity, syntactic complexity, and sophistication of macrostructure.

Elicitation Method

The method of narrative elicitation has attracted much attention in previous studies. Researchers in general consider both story-tell and story-retell tasks to be valuable in assessing narrative performance as the two tasks can capture a person's ability across different contexts, and both could be mapped onto real-life scenarios (e.g., Boudreau, 2008; Otwinowska et al., 2018; Schneider & Dubé, 2005; Westerveld & Gillon, 2010).

Similar to previous studies, we found better performance in story-retell than story-tell on measures of overall story structure and percentage of complex clauses (Kunnari et al., 2016; Otwinowska et al., 2018; Roch et al., 2016). These main effects of task were tempered by interactions with risk status. In both cases, it appeared that the AR group was stimulated by the more mature language in the adult model and subsequently produced a comparable proportion

Table 5. Examples of ungrammatical utterances.

Type	Subtype	Example
Word order		然后 看到 回来 有 一只 猫 要 吃 他们 Then saw return there is one-CL cat will eat them. Upon return, (subject drop) saw that a cat was going to eat them. The order of "看到" and "回来" is reversed in this sentence.
Missing elements	Argument	他 趁__ 不 注意， 小猫 又 拿了 条 鱼 He while not pay attention, cat again got PER CL fish. __(missing object)
	Main verb	这小猫__我想 吃 鱼。 This cat I want eat fish. __(missing verb)
	Resultative verb	小猫 掉__了 草丛 里。 Cat fell PER grass inside. __(missing resultative verb)
Addition		小鸟 在 追 了 灰狼。 Bird PRO chase PER wolf. Both progressive aspect marker 在 and perfective aspect marker 了 were used in the same utterance. The second aspect marker is treated as an error of addition.
Wrong word	Verb	乌鸦 就 把 它 溜 走 了。 Crow shall BA it escape away PER. 溜 is the wrong verb in this context. The correct verb should be "drove."
	Classifier	猫咪 拿 了 一杆 鱼。 Kitty got PER one-CL fish. 杆 is the wrong classifier for fish. The correct classifier that modifies fish should be tiao.
Others		狗狗 就 抓 住 猫咪 跑 了 猫咪 的 岸上。 Dog shall catch hold kitty run PER kitty DE bank. This sentence does not fit any of the other categories. It is unclear what the child intended to say. The picture depicts the dog chasing the cat away from a tree. There are both semantic and grammatical errors that made the sentence incomprehensible.

Note. CL = classifier; PER = perfective aspect marker; PRO = progressive aspect marker; BA = marker of the BA construction, a noncanonical active sentence that follows a subject–BA–object–verb word order; DE = a morpheme that can function as an adjectivizer, relativizer, or possessive marker in Mandarin.

of multiclausal utterances and comparable numbers of story grammar elements in their retells as their TD peers. However, this facilitation effect was short lived, and once the adult model was withdrawn, utterance complexity and story grammar expression decreased significantly for the AR children. By contrast, the TD children maintained a high level of performance across tasks. This in turn led to a TD advantage in the story-telling task but not in the model-supported retelling task. Story-retell does not always enhance children's microstructure performance. For instance, previous studies have found that the more advanced vocabulary and sentence structure in the adult model may be overwhelming for children with limited language ability and depress microstructure performance (Gutiérrez-Clellen 2002; Otwinowska et al., 2018). Our results suggested that the story model in the Mandarin MAIN was at an appropriate level of complexity for the current sample of children.

The current finding of group differences on the story-tell task is reminiscent of Kunnari et al. (2016), who found the story-tell task in the MAIN to be differentiating between TD monolingual and bilingual children who had different levels of language exposure and proficiency. In contrast, Merritt and Liles (1989) and Westerveld and Gillon (2010) both found the retell task to be more revealing of group differences than the tell task between children with and without language or reading disorders. However, the retell and tell tasks in these two studies were dissimilar in format and difficulty level and could have posed different demands on the children's creativity and memory capacity.

It is possible that the TD advantage in the tell task was due to a rapid uptake of the sophisticated language and story structure modeled in the first narrative task. This explanation is akin to one of the main assumptions of dynamic assessment, in that children who have intact language learning abilities should be more likely to show rapid gains after a brief training than children who have weaker language learning capacity (Peña et al., 2014). In the current context, the benefit of having a model may have manifested in the TD children as a lack of decrement in performance from the first (retell) to the second (tell) story task since a briefly presented story model was not equivalent to an intervention session.

Limitations and Future Directions

This study has limitations. First, our sample consisted of AR children without formal diagnoses of DLD. Second, the definition of risk was based on screening measures that have not undergone validation, and the cutoff scores, although in line with others (e.g., Peña et al., 2011; Spaulding et al., 2006), were somewhat arbitrary. As a result, it is unclear how many of the AR children will eventually receive a formal diagnosis and whether or not the TD controls were indeed free of language difficulties. Therefore, even though we were able to replicate the findings in Hao et al. (2018) and Torng and Sah (2019), which included samples with more definitive diagnoses, the results should be treated with caution.

A formal, definitive diagnosis entails that it is made by qualified individuals (e.g., a speech-language pathologist or a pediatrician with training in developmental and behavioral disorders) using norm-referenced tests with strong psychometric properties. Unfortunately, these prerequisites are not in place in China to enable verification of DLD status given the shortages of qualified personnel and accessible standardized measures as well as the lack of infrastructure for detection and intervention in health care and educational systems (Salas-Provance, 2011).

Underdetection of DLD is a persistent problem. Even in countries such as the United States where there are numerous screening and diagnostic tools (for English) and existing legislature and infrastructures, the majority of parents whose children met commonly accepted criteria for DLD were unaware of their child's language difficulties and reported no prior receipt of services (Adlof et al., 2017; Tomblin et al., 1997). For Mandarin-speaking children in China, practical barriers and the fact that DLD is not a life-threatening condition exacerbate the low awareness and underdetection of this disorder.

While we are unable to set up all the necessary parameters to ascertain the DLD status of our sample within the scope of the current study, the findings can serve as pilot data to inform future epidemiological studies of Mandarin DLD. Recent models of universal screening found utility in the use of a receptive grammar measure to identify DLD in school-age children (Hendricks et al., 2019) and the use of an instrument that covered multiple domains of speech and language to identify DLD in preschoolers (Lavesson et al., 2018). The current findings suggest that a measure of narrative production also holds promise for the purpose of identifying children with lower language abilities. Our experiences administering the MAIN indicated that children in the 3- to 6-year age range can complete two narrative tasks in 6–11 min. To increase the ease of scoring, future studies may attempt to develop a scoring rubric that focuses on known vulnerable areas (e.g., story grammar elements, lexical diversity, syntactic complexity) and can be reliably scored online. In addition, the divergence in AR and TD children's performance from the retell task to the tell task suggests a need to examine the diagnostic utility of dynamic assessment in Mandarin DLD. Questions about the children's responsivity, cooperativeness, attentiveness, frustration, disruption, and transfer behaviors can be added to the end of the administration protocol for the examiner to make quick judgment of the child's overall level of participation and modifiability during the narrative tasks (Lavesson et al., 2018; Peña et al., 2014; Petersen et al., 2017). While much needs to be done to bring awareness to DLD and affect changes in service provision for Mandarin-speaking children with DLD, these explorations will be a solid starting point.

On the methodological side, there is a need for greater uniformity on how Chinese utterances are segmented across studies. We found examples of segmentation by syntax

and prosody (Klee et al., 2004), by syntax (Sah, 2013; Zhang, 2013), and by prosody (To et al., 2010; Torng & Sah, 2019). Some of these studies also mentioned the use of semantic meaning to facilitate decision making (e.g., To et al., 2010), and many others did not provide much information on how they performed utterance segmentation. These variations can lead to drastic differences in the unit of analysis, impacting values of MLU and percentage of complex sentences and resulting in discrepant findings across studies. We chose syntax as the primary criterion with prosody and semantics playing a secondary role, bearing in mind our long-term goal of integrating comparisons to other languages. However, whether and to what extent this segmentation approach provides valid information of Mandarin language development is an empirical question that awaits further research.

Conclusion

The current study is one of the few studies of narrative production in Mandarin-speaking children with or AR for DLD and the first study that utilized the MAIN. The results indicate that the Mandarin version of the MAIN is promising in differentiating children with and without DLD. The results also confirmed several key findings of the Mandarin DLD phenotype in previous studies (Hao et al., 2018; Torng & Sah, 2019). Along with many others (e.g., Boudreau, 2008; Merritt & Liles, 1989; Otwinowska et al., 2018; Schneider & Dubé, 2005; Westerveld & Gillon, 2010), we believe that both story-tell and story-retell are valuable forms of narrative assessment and should be further investigated in future larger scale studies.

Acknowledgments

This project was funded by Humanities and Social Sciences projects of the Chinese Ministry of Education (17YJAZH132) awarded to Li Zheng (PI) and Li Sheng (co-PI). We thank the parents, teachers, and children for volunteering their time; Angel Chan, Natalia Gagarina, and Luo Jin for sharing with us the Mandarin version of the Multilingual Assessment Instrument for Narratives; and Anita Wong for providing feedback on our utterance segmentation rubric. We are grateful for the contribution of the following research assistants for collecting data: Catherine Hui-Yu Huang, Yimeng Mou, Rui Liu, Shimei Zhang, Ran Gao, Ru Peng, Ying Sun, and Yaoyao Qin.

References

Adlof, S. M., Scoggins, J., Brazendale, A., Babb, S., & Petscher, Y. (2017). Identifying children at risk for language impairment or dyslexia with group-administered measures. *Journal of Speech, Language, and Hearing Research, 60*(12), 3507–3522. https://doi.org/10.1044/2017_JSLHR-L-16-0473

Altman, C., Armon-Lotem, S., Fichman, S., & Walkters, J. (2016). Macrostructure, microstructure, and mental state terms in the narratives of English–Hebrew bilingual preschool children with and without specific language impairment. *Applied Psycholinguistics, 37*(1), 165–193. https://doi.org/10.1017/S0142716415000466

Armon-Lotem, S., de Jong, J., & Meir, N. (Eds.). (2015). *Assessing multilingual children: Disentangling bilingualism from language impairment*. Multilingual Matters. https://doi.org/10.21832/9781783093137

Andreu, L., Sanz-Torrent, M., Guárdia Olmos, J., & MacWhinney, B. (2011). Narrative comprehension and production in children with SLI: An eye movement study. *Clinical Linguistics & Phonetics, 25*(9), 767–783. https://doi.org/10.3109/02699206.2011.565542

Bedore, L. M., Peña, E. D., Joyner, D., & Macken, C. (2011). Parent and teacher rating of bilingual language proficiency and language development concerns. *International Journal of Bilingual Education and Bilingualism, 14*(5), 489–511. https://doi.org/10.1080/13670050.2010.529102

Berman, R. A., & Slobin, D. I. (1994). *Relating events in narrative: A crosslinguistic developmental study*. Erlbaum.

Bishop, D. V. (1997). Uncommon understanding: Comprehension in specific language impairment. Psychology Press.

Boudreau, D. (2008). Narrative abilities: Advances in research and implications for clinical practice. *Topics in Language Disorders, 28*(2), 99–114. https://doi.org/10.1097/01.TLD.0000318932.08807.da

Burns, F. A., de Villiers, P. A., Pearson, B. Z., & Champion, T. B. (2012). Dialect-neutral indices of narrative cohesion and evaluation. *Language, Speech, and Hearing Services in Schools, 43*(2), 1–21. https://doi.org/10.1044/0161-1461(2011/10-0101)

Chan, A., Chui, B., Lo, J., Luk, P., & Gagarina, N. (2018). *Narrative abilities of bilingual Urdu–Cantonese ethnic minority children in Hong Kong*. Poster presented at the Child Language Symposium, Reading, United Kingdom.

Cheung, H., Chang, C.-J., Ko, H.-W., & Tsay, J. (2011). *Taiwan corpus of child Mandarin*. Retrieved April 15, 2018, from http://tccm.corpus.eduhk.hk/

Colozzo, P., Gillam, R. B., Wood, M., Schnell, R. D., & Johnston, J. R. (2011). Content and form in the narratives of children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 54*(6), 1609–1627. https://doi.org/10.1044/1092-4388(2011/10-0247)

Ehrler, D. J., & McGhee, R. L. (2008). *Primary Test of Nonverbal Intelligence (PTONI) manual*. Pro-Ed.

Fey, M. E., Catts, H. W., Proctor-Williams, K., Tomblin, J. B., & Zhang, X. (2004). Oral and written story composition skills of children with language impairment. *Journal of Speech, Language, and Hearing Research, 47*(6), 1301–1318. https://doi.org/10.1044/1092-4388(2004/098)

Fiestas, C. E., & Peña, E. D. (2004). Narrative discourse in bilingual children: Language and task effects. *Language, Speech, and Hearing Services in Schools, 35*(2), 155–168. https://doi.org/10.1044/0161-1461(2004/016)

Gagarina, N., Klop, D., Kunnari, S., Tantele, K., V limaa, T., Balèiûnienë, I., Bohnacker, U., & Walters, J. (2012). *MAIN: Multilingual Assessment Instrument for Narratives. ZAS papers in Linguistics 56*. ZAS.

Gagarina, N., Klop, D., Kunnari, S., Tantele, K., V limaa, T., Balèiûnienë, I., Bohnacker, U., & Walters, J. (2015). Assessment of narrative abilities in bilingual children. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 241–274). Multilingual Matters. https://doi.org/10.21832/9781783093137-011

Gillam, R., & Pearson, N. (2004). *Test of Narrative Language*. Pro-Ed.

Guo, L., Tomblin, J. B., & Samelson, V. (2008). Speech disruptions in the narratives of English-speaking children with specific language impairment. *Journal of Speech, Language, and*

Hearing Research, 51(3), 722–738. https://doi.org/10.1044/1092-4388(2008/051)

Gutiérrez-Clellen, V. F. (2002). Narratives in two languages: Assessing performance of bilingual children. *Linguistics and Education, 13*(2), 175–197. https://doi.org/10.1016/S0898-5898(01)00061-4

Hao, Y., Bedore, L. M., Sheng, L., & Peña, E. D. (2019). Narrative skills in two languages of Mandarin–English bilingual children. *International Journal of Speech-Language Pathology, 21*(4), 325–335. https://doi.org/10.1080/17549507.2018.1444092

Hao, Y., Sheng, L., Zhang, Y., Jiang, F., de Villiers, J., Lee, W., & Liu, X. L. (2018). A narrative evaluation of Mandarin-speaking children with language impairment. *Journal of Speech, Language, and Hearing Research, 61*(2), 345–359. https://doi.org/10.1044/2017_JSLHR-L-16-0367

Hendricks, A. E., Adlof, S. M., Alonzo, C. N., Fox, A. B., & Hogan, T. P. (2019). Identifying children at risk for developmental language disorder using a brief, whole-classroom screen. *Journal of Speech, Language, and Hearing Research, 62*(4), 896–908. https://doi.org/10.1044/2018_JSLHR-L-18-0093

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*(3), 177–189. https://doi.org/10.1080/07481756.2002.12069034

Hewitt, L. E., Hammer, C. S., Yont, K. M., & Tomblin, J. B. (2005). Language sampling for kindergarten children with and without SLI: Mean length of utterance, IPSYN, and NDW. *Journal of Communication Disorders, 38*(3), 197–213. https://doi.org/10.1016/j.jcomdis.2004.10.002

Huck, S. W. (2009). *Statistical misconceptions*. Taylor & Francis. https://doi.org/10.4324/9780203889282

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Johnston, J. R. (2006). Narratives schemes and processes. In J. R. Johnston (Ed.), *Thinking about child language: Research to practice* (pp. 145–152). Thinking Publications.

Justice, L. M., Bowles, R., Pence, K., & Gosse, C. (2010). A scalable tool for assessing children's language abilities within a narrative context: The NAP (Narrative Assessment Protocol). *Early Childhood Research Quarterly, 25*(2), 218–234. https://doi.org/10.1016/j.ecresq.2009.11.002

Klee, T., Stokes, S. F., Wong, A. M.-Y., Fletcher, P., & Gavin, W. J. (2004). Utterance length and lexical diversity in Cantonese-speaking children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research, 47*(6), 1396–1410. https://doi.org/10.1044/1092-4388(2004/104)

Klein, W., Li, P., & Hendriks, H. (2000). Aspect and assertion in Mandarin Chinese. *Natural Language & Linguistic Theory, 18*(4), 723–770. https://doi.org/10.1023/A:1006411825993

Kunnari, S., Välimaa, T., & Laukkanen-Nevala, P. (2016). Macrostructure in the narratives of monolingual Finnish and bilingual Finnish–Swedish children. *Applied Psycholinguistics, 37*(1), 123–144. https://doi.org/10.1017/S0142716415000442

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Leonard, L. (2014). *Children with specific language impairment* (2nd ed.). MIT Press.

Lavesson, A., Lövdén, M., & Hansson, K. (2018). Development of a language screening instrument for Swedish 4-year-olds. *International Journal of Language & Communication Disorders, 53*(3), 605–614. https://doi.org/10.1111/1460-6984.12374

Liles, B. Z., Duffy, R. J., Merritt, D. D., & Prucell, S. L. (1995). Measurement of narrative discourse ability in children with language disorders. *Journal of Speech and Hearing Research, 38*(2), 415–425. https://doi.org/10.1044/jshr.3802.415

Loban, W. (1976). *Language development: Kindergarten through grade twelve* (Report No. 18). National Council of Teachers.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Erlbaum.

Mayer, M. (1967). *A boy, a dog and a frog*. Penguin Putnam.

Mayer, M. (1969). *Frog, where are you?* Dial Press.

Merritt, D. D., & Liles, B. Z. (1987). Story grammar ability in children with and without language disorder: Story generation, story retelling, and story comprehension. *Journal of Speech and Hearing Research, 30*(4), 539–552. https://doi.org/10.1044/jshr.3004.539

Merritt, D. D., & Liles, B. Z. (1989). Narrative analysis: Clinical applications of story generation and story retelling. *Journal of Speech and Hearing Disorders, 54*(3), 429–438. https://doi.org/10.1044/jshd.5403.438

National Statistics Bureau of China. (2010). *The 2010 Chinese National Census*. Retrieved from http://www.stats.gov.cn/tjsj/pcsj/rkpc/6rp/indexch.htm

Newman, R. M., & McGregor, K. K. (2006). Teachers and laypersons discern quality differences between narratives produced by children with or without SLI. *Journal of Speech, Language, and Hearing Research, 49*(5), 1022–1036. https://doi.org/10.1044/1092-4388(2006/073)

Norbury, C. F., & Bishop, D. V. M. (2003). Narrative skills of children with communication impairments. *International Journal of Language & Communication Disorders, 38*(3), 287–313. https://doi.org/10.1080/1368203100001081 33

Otwinowska, A., Mieszkowska, K., Białecka-Pikul, M., Opacki, M., & Haman, E. (2018). Retelling a model story improves the narratives of Polish–English bilingual children. *International Journal of Bilingual Education and Bilingualism*. Advance online publication. https://doi.org/10.1080/13670050.2018.1434124

Peña, E. D., Gillam, R. B., Bedore, L. M., & Bohman, T. M. (2011). Risk for poor performance on a language screening measure for bilingual preschoolers and kindergarteners. *American Journal of Speech-Language Pathology, 20*(4), 302–314. https://doi.org/10.1044/1058-0360(2011/10-0020)

Peña, E. D., Gillam, R. B., & Bedore, L. M. (2014). Dynamic assessment of narrative ability in English accurately identifies language impairments in English language learners. *Journal of Speech, Language, and Hearing Research, 57*(6), 2208–2220. https://doi.org/10.1044/2014_JSLHR-L-13-0151

Peña, E. D., Gutierrez-Clellen, V. F., Iglesias, A., Goldstein, B. A., & Bedore, L. (2018). *Inventory to Assess Language Knowledge (ITALK) (English and Spanish bilingual edition)*. Brookes.

Pennington, B. F., & Bishop, D. V. M. (2009). Relations among speech, language, and reading disorders. *Annual Review of Psychology, 60*, 283–306. https://doi.org/10.1146/annurev.psych.60.110707.163548

Pesco, D., & Kay-Raining Bird, E. (2016). Perspectives on bilingual children's narratives elicited with the Multilingual Assessment Instrument for Narratives. *Applied Psycholinguistics, 37*(1), 1–9. https://doi.org/10.1017/S0142716415000387

Petersen, D. B., Chanthongthip, H., Ukrainetz, T. A., Spencer, T. D., & Steeve, R. W. (2017). Dynamic assessment of narratives: Efficient, accurate identification of language impairment in bilingual students. *Journal of Speech, Language, and Hearing Research, 60*(4), 983–998. https://doi.org/10.1044/2016_JSLHR-L-15-0426

Reilly, K., Losh, M., Bellugi, U., & Wulfeck, B. (2004). "Frog, where are you?" Narratives in children with specific language impairment, early focal brain injury, and Williams syndrome. *Brain and Language, 88*(2), 229–247. https://doi.org/10.1016/S0093-934X(03)00101-9

Restrepo, M. A. (1998). Identifiers of predominantly Spanish-speaking children with language impairment. *Journal of Speech, Language, and Hearing Research, 41*(6), 1398–1411. https://doi.org/10.1044/jslhr.4106.1398

Reuterskiöld, C., Hansson, K., & Sahlén, B. (2011). Narrative skills in Swedish children with language impairment. *Journal of Communication Disorders, 44*(6), 733–744. https://doi.org/10.1016/j.jcomdis.2011.04.010

Rezzonico, S., Chen, X., Cleave, P. L., Greenberg, J., Hipfner-Boucher, K., Johnson, C. J., Milburn, T., Pelletier, J., Weitzman, E., & Girolametto, L. (2015). Oral narratives in monolingual and bilingual preschoolers with SLI. *International Journal of Language & Communication Disorders, 50*(6), 830–841. https://doi.org/10.1111/1460-6984.12179

Rezzonico, S., Goldberg, A., Mak, K. K.-Y., Yap, S., Milburn, T., Belletti, A., & Girolametto, L. (2016). Narratives in two languages: Storytelling of bilingual Cantonese–English preschoolers. *Journal of Speech, Language, and Hearing Research, 59*(3), 521–532. https://doi.org/10.1044/2015_JSLHR-L-15-0052

Roch, M., Florit, E., & Levorato, C. (2016). Narrative competence of Italian–English bilingual children between 5 and 7 years. *Applied Psycholinguistics, 37*(1), 49–67. https://doi.org/10.1017/S0142716415000417

Sah, W.-H. (2007). A longitudinal investigation of Mandarin-speaking preschoolers' relation of events in narratives: From unrelated to related events. *Taiwan Journal of Linguistics, 5*(1), 77–96.

Sah, W.-H. (2013). Global and local connections in Mandarin-speaking children's narratives. In A. McCabe & C.-J. Chang (Eds.), *Chinese language narration: Culture, cognition, and emotion* (pp. 85–114). John Benjamins. https://doi.org/10.1075/sin.19.06sah

Sah, W.-H. (2015). The development of coherence in narratives: Causal relations. *Taiwan Journal of Linguistics, 13*(1), 25–51.

Sah, W.-H. (2018). Referential choice in narratives of Mandarin-speaking children with autism spectrum disorder: Form, function, and adequacy. *First Language, 38*(3), 225–242. https://doi.org/10.1177/0142723717739198

Sah, W.-H., & Torng, P.-C. (2015). Narrative coherence of Mandarin-speaking children with high-functioning autism spectrum disorder: An investigation into causal relations. *First Language, 35*(3), 189–212. https://doi.org/10.1177/0142723715584227

Sah, W.-H., & Torng, P.-C. (2017). Production of mental state terms in narratives of Mandarin-speaking children with autism spectrum disorder. *Clinical Linguistics & Phonetics, 31*(2), 174–191. https://doi.org/10.1080/02699206.2016.1219920

Salas-Provance, M. B. (2011). The China gap for speech-language services. *The ASHA Leader, 16*(13), 10–11. https://doi.org/10.1044/leader.WB1.16132011.np

Schneider, P., & Dubé, R. V. (2005). Story presentation effects on children's retell content. *American Journal of Speech-Language Pathology, 14*(1), 52–60. https://doi.org/10.1044/1058-0360(2005/007)

Sheng, L., Lam, B. P. W., Wang, D., & Zheng, L. (2019). *Receptive and expressive vocabulary screener of Mandarin*. Manuscript in preparation.

Snow, C. E., Porche, M. V., Tabors, P. O., & Harris, S. R. (2007). *Is literacy enough?* Brookes.

Spaulding, T. J., Plante, E., & Farinella, K. A. (2006). Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools, 37*(1), 61–72. https://doi.org/10.1044/0161-1461(2006/007)

Stein, N., & Glenn, G. (1979). An analysis of story comprehension in elementary school children. In R. Freedle (Ed.), *New directions in discourse processing* (Vol. 2, pp. 32–120). Ablex.

Teoh, W. Q., Brebner, C., & McAllister, S. (2017). Bilingual assessment practices: Challenges faced by speech-language pathologists working with a predominantly bilingual population. *Speech, Language and Hearing, 21*(1), 10–21. https://doi.org/10.1080/2050571X.2017.1309788

To, C. K.-S., Stokes, S. F., Cheung, H.-T., & T'sou, B. (2010). Narrative assessment for Cantonese-speaking children. *Journal of Speech, Language, and Hearing Research, 53*(3), 648–669. https://doi.org/10.1044/1092-4388(2009/08-0039)

Tomblin, J. B., Records, N. L., Buckwalkter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research, 40*(6), 1245–1260. https://doi.org/10.1044/jslhr.4006.1245

Torng, P.-C., & Sah, W.-H. (2019). Narrative abilities of Mandarin-speaking children with and without specific language impairment: Macrostructure and microstructure. *Clinical Linguistics & Phonetics*. Advance online publication. https://doi.org/10.1080/02699206.2019.1655097

Tsai, W., & Chang, C. (2008). "But I first…and then he kept picking" Narrative skill in Mandarin-speaking children with language impairment. *Narrative Inquiry, 18*(2), 349–377. https://doi.org/10.1075/ni.18.2.09tsa

Tsimpli, I. M., Peristeri, E., & Andreou, M. (2016). Narrative production in monolingual and bilingual children with specific language impairment. *Applied Psycholinguistics, 37*(1), 195–216. https://doi.org/10.1017/S0142716415000478

United States Census Bureau. (2015). *Detailed languages spoken at home and ability to speak English for the population 5 years and over for United States: 2009–2013*. https://www.census.gov/data/tables/2013/demo/2009-2013-lang-tables.html

Wang, D., Zheng, L., & Sheng, L. (2019). *Sentence repetition as a clinical marker for Mandarin-speaking preschoolers with developmental language disorders*. Manuscript in preparation.

Westby, C. E. (2005). Assessing and remediating text comprehension problems. In H. W. Catts & A. G. Kamhi (Eds.), *Language and reading disabilities* (2nd ed., pp. 157–232). Pearson Education.

Westerveld, M. F., & Gillon, G. T. (2010). Oral narrative context effects on poor readers' spoken language performance: Story retelling, story generation, and personal narratives. *International Journal of Speech-Language Pathology, 12*(2), 132–141. https://doi.org/10.3109/17549500903414440

Whitely, C., & Colozzo, P. (2013). Who's who? Memory updating and character reference in children's narratives. *Journal of Speech, Language, and Hearing Research, 56*(5), 1625–1636. https://doi.org/10.1044/1092-4388(2013/12-0176)

Wiedenhof, J. (2015). *A grammar of Mandarin*. John Benjamins.

Zhang, F. (2013). A study of narrative development of young Chinese children with specific language impairment aged four to six years. In A. McCabe & C.-J. Chang (Eds.), *Chinese language narration: Culture, cognition, and motion* (pp. 143–179). John Benjamins. https://doi.org/10.1075/sin.19.08zha

Zhang, F., McCabe, A., Ye, J., Wang, Y., & Li, X. (2019). A developmental study of the narrative components and patterns of Chinese children aged 3–6 years. *Journal of Psycholinguistic Research, 48*(2), 477–450. https://doi.org/10.1007/s10936-018-9614-3

Appendix (p. 1 of 4)

Segmentation and Coding Rubrics for Mandarin MAIN Stories

Language Learning and Bilingualism Lab[5]
University of Delaware

Introduction

C-unit is commonly used to segment narrative discourse in English. A C-unit is "an independent clause with its modifier" (Loban, 1976). According to the Systematic Analysis of Language Transcript manual, a C-unit "includes one main clause with all subordinate clauses attached to it. It cannot be further divided without the disappearance of its essential meaning. A clause, whether it is the main clause or a subordinate clause, is a statement containing both a subject and a predicate. Grammatically, a subject is a noun phrase and a predicate is a verb phrase." (p. 1, SALT Software, LLC, 2016 http://www.saltsoftware.com/coursefiles/shared/Cunits.pdf).

Mandarin has a number of unique features that deviate from English and makes the current scheme of identifying clauses and C-units challenging. For example, Mandarin allows omission of arguments (subject and object); it does not require explicit connectives to link ideas, and serial verb constructions (a multipredicate constructions) and descriptive clauses are productive structures in the language (Li & Thompson, 1981; To et al., 2010). These features require the use of language-specific definitions of clauses and language-specific units of segmentation.

Similar to Sah (2013) and Zhang (2013), we use clause as the unit of segmentation. A clause in Mandarin consists of a verb and its arguments, and corresponds roughly to a single event (Sah, 2013, p. 174). In the following sections, we first review how Chinese language sample segmentation was handled in previous studies (Section I). Then we describe our guidelines for segmenting utterances and provide examples (Section II). In the last section (Section III), we present examples of simple and complex clauses in our current samples of MAIN stories.

I. Utterance Segmentation in Chinese

There are no consistent rules for utterance segmentation in Chinese. We noted different approaches as listed below.

1. Segmenting by syntax

Previous studies have used clause as the unit of utterance segmentation.

- Zhang (2013 p.151) defined a clause as "a verb and its arguments" and provided two examples ("他 送 我 去 医院" "He sent me to the hospital" was coded as one clause; and "玫瑰花 刺扎 到 我的 手，扎的 我 好痛" "The rose thorns hurt my hand, and it's really hurt" were coded as two clauses. Reliability was checked by checking 50% of samples and reaching consensus.
- Sah (2013 p.174) "Clauses were used to quantify story length. A clause consists of a verb and its arguments, and corresponds roughly to a single event."
- Tsai & Chang (2008) mentioned that they used clause as the unit of coding but did not define what they meant by a clause.

2. Segmenting by syntax and prosody

In Klee et al. (2004), p. 1400

- "Utterance segmentation criteria were developed that were based on the grammatical organization of utterances but that also took into account pause and intonation."

Reliability was addressed by reaching consensus on all samples by two research assistants. No further examples were provided.

3. Segmenting by prosody

In To et al. (2010), pp. 654–655

- "Samples were first transcribed verbatim and then were segmented into utterances. Segmentation relying on syntactic structures alone led to considerable interrater variation; however, this segmentation relied mostly on intonation patterns and semantic meaning and only to a lesser extent on syntactic structures. The decision to use intonation patterns instead of syntactic structures—as in the T-unit (Hunt, 1965) and the C-unit (Loban, 1976) calculations—was due to the fact that prosodic elements play the most important role in determining the boundaries of a sentence in Chinese (Chao, 1968). Sentences in Chinese can be conjoined by juxtaposition without an explicit connective, and multipredicate utterances incorporating serial verb constructions are productive structures in Chinese. Eight narrative samples from the pilot data were used to assess the interrater reliability of utterance segmentation. The method relying on intonation for segmentation yielded the best interrater reliability at 85%. Therefore, utterance segmentation was determined by intonation patterns."

[5]This rubric is a work in progress and is being revised periodically to reflect our evolving knowledge of Mandarin grammar. Last updated on October 18, 2019.

It is worth noting that these authors did mention that they also relied on semantic meaning and syntactic structure, but to a lesser extent.

In Torng and Sah (2019), p. 9

- "The transcriptions were then divided into utterances by using prosodic features as segmentation cues. The reason to choose utterances, rather than C-units or T-units, is that prosodic features have been considered more useful than syntactic features in determining boundaries of sentences in the Chinese discourse (Chao, 1968; Tsai & Chang, 2008; To et al., 2010)."

II. Segmentation Rules and Examples

We decide to segment by syntax and use clause as the unit of segmentation. This is more consistent with the practice of using C-unit in other languages. The next section lists examples that emerged from the sample, and how they are segmented in the current study.

Type	**Utterances without conjunctions**
Each clause should have at least one main verb. A main verb can be an action verb, an existential verb (e.g.,"有"), or an adjectival verb (e.g., 很馋). This is because the vast majority of Mandarin adjectives can function as verbs that function as the of verb phrases (Li & Thompson, 1981:142). Utterances are segmented with "//." Disfluencies are marked with [//]. Repetitions are marked with [/].	
Segmented	有两只羊//一只羊在洗澡//一只羊在吃草 There were two goats. One goat was bathing. One goat was eating grass. Note. Each utterance is a clause with a verb and its argument. 他拿了一个鱼桶子//里面有很多鱼 He brought a fishing bucket. There were lots of fish inside. 他一手拿着鱼//一手拿着鱼竿 He held the fish in one hand, and the fishing rod in the other hand. Note. Even though the two clauses share the same subject and are parallel in construction, they are segmented because each represents a clause.
Type	**Utterances that contain** 就
"就" is a polysemous morpheme. These examples contain uses that could be translated into "then," "so," or "as a result."	
Segmented	小男孩很开心//他就把鱼竿放到地上 The boy is happy // He put the fishing rod on the ground. 有一天，有只贪玩的小猫看见一个黄色的蝴蝶坐在草丛上//它就想捉那只蝴蝶 One day, there was a playful cat (who) saw a yellow butterfly sitting on the grass//She then wanted to catch the butterfly. Note. 有一天 is treated as a formulaic expression and not segmented from the clause that comes after it. 他一松手//气球就飞[/]就飞上树枝上 He let go of his hand // so the balloon flew[/] flew up to the branch. 然后那个男孩没注意//小狗就拿了一块走 Then the boy was not paying attention // so the dog took a piece (of sausage) away.
Type	**Utterances that contain coordinating conjunctions**
Common coordinating conjunctions include "然后" then, "但是" but, "所以" so	
Segmented	然后呢有一只猫来了//想吃掉那个鸟 Then a cat came // (It) wanted to eat the bird. 然后有一只小男孩// 手上拿了刚买回东西 Then there was a little boy //(He) held in his hands things (he) just bought. 然后 球 都 掉 了[simple]//然后 到 了 水 里面[simple] Then the ball fell//Then (it) got into the water. Note. Two separate clauses with explicit use of conjunctions. 有一天小羊宝宝在吃草//但是它掉进了一个很深的小河里面 One day baby goat was eating grass //But he fell into a deep river. 它 觉得 老鹰 太 重 了//所以 [/] 所以 它 就 把 老鹰 带 回家 了 He thought the eagle was too heavy // So[/] so he took the eagle home.
Type	**Utterances with subordinating conjunction**
Not segmented	小猫咪跳的时候跳到这个草丛了 When the cat was jumping, it jumped into the bush. Note. "的时候"is equivalent to when/while and is treated as a subordinating conjunction.

III. Complex Utterances

Again there are no consistent categorization systems for Chinese complex sentences. We found two examples in previous literature of Mandarin child language

Cheung (2009) analyzed the use of complex sentences in children's spontaneous language samples. A complex sentence "consists of one independent clause and one or more dependent clauses" (p. 37). Expanding on Li and Thompson (1981:594), Cheung identified five types of complex sentences:

Table III.1 Complex sentences and examples identified by Cheung (2009).

Sentence type	Definition	Example
Serial verb construction	Single clause sentences with two verb phrases. These are not complex sentences structurally speaking.	穿 袜子 去 游泳 我 要 爬 给 妈妈 看
Pivot construction	Contains a noun phrase between the first verb and the second verb and serves as the grammatical object of the first verb as well as the grammatical subject of the second verb.	让 我 修理 这个 姐姐 带 我 去 上学
Clausal object	An embedded clause serves as the object of the main clause.	不 爱 睡 午觉
Clausal subject	An embedded clause serves as the subject of the main clause.	这里 用 正方形 才 可以 用
Descriptive clause	Not defined. Mentioned that this is an unusual sentence type and structurally complex.	我 有 一 个 朋友 很 会 唱歌

Torng and Sah (2019) analyzed the use of six types of complex sentences. These authors regarded ba and bei as verbs and thus included ba and bei constructions as complex sentences. The following examples were from Torng and Sah (2019).

Table III.2 Complex sentence examples provided by Torng and Sah (2019).

Sentence type	Example
Relative clause	罐子 里 的 青蛙 跑 出来
Clausal complement	他 发现 青蛙 不见 了
Serial verb construction	小 男孩 跑 去 找
Pivot construction	小 男孩 让 狗 跑 去 外面
Ba construction	鹿 把 小孩 摔 下去
Bei construction	小狗 被 蜜蜂 追

Appendix (p. 4 of 4)

Segmentation and Coding Rubrics for Mandarin MAIN Stories

The following table presents examples of simple sentences and major types of complex sentences that emerged from the current sample.

Type	Example
Simple	Simple sentences contain only one main verb. The NP and main verb may contain their own modifiers such as an adjective, a determiner, or a "numeral + classifier" for NPs or aspect markers and resultative verb compounds for Vs. 小羊在吃青草。The goat was eating grass. 小男孩拿回了他的气球。The boy got his balloon back. 一只贪玩的小猫注意到了那个男孩的桶。A playful cat noticed the boy's bucket.
Complex	
Object complement clause	An object complement clause has an embedded clause that serves as the object of the main clause (Cheung, 2009). The embedded clause usually begins with cognitive and communication verbs such as 想要 认为 盼望 感到 / 觉得 建议 看到 听到 说 它想把那个黄色的蝴蝶抓住。He wanted to catch the yellow butterfly. 男孩没想到小猫已经吃了一个鱼。The boy didn't expect that the cat had eaten a fish. 然后小猫说它想吃鱼。Then the cat said that he wanted to eat fish.
Relative clause	A relative clause is a nominalized clause placed in front of a noun to modify it (Li & Thompson, 1981: 116). 他手上拎了一个装鱼的桶。He held a bucket that had fish in it. 男孩看到了在抓鱼的小猫。The boy noticed the cat that was grabbing the fish. 在水塘里面的那一个惊讶地看着那只吃草的小羊。The one (goat) that was in the pond looked with a surprised expression at the goat that was eating grass.
Serial verb construction	A sentence that contains two or more verb phrase or clauses juxtaposed. The two phrases/clauses are describing two separate events but are always understood to be related in some way (e.g., consecutive, purpose, alternating, circumstance; Li & Thompson, p. 595). 小狗追着小猫跑。 The dog is chasing the cat and running. (circumstance) 它 飞出去 找 [/] 找 东西 吃 妈妈 叼 了 虫子 回来 鹦鹉 就 飞 上去 去[/] <啄去啄> [//] 啄 狐狸 的 大尾巴 。The parrot then flew over to poke the fox's big tail. (purpose) 男孩 爬 到 树 上 去 拿 他 的 气球。The boy climbed up the tree to get his balloon. (purpose)
Pivot construction	Contains a noun phrase between the first verb and the second verb and serves as the direct object of the first verb as well as the grammatical subject of the second verb. Common first verbs in pivot constructions include 请 劝 叫 让 帮 求 (Li & Thompson, 1981: 607–610) 小鸟妈妈正在叫小鸟起床。The Mummy Bird was telling the Baby Bird to get up.
Adverbial clause	An adverbial clause is a dependent clause that functions as an adverb; that is, the entire clause modifies a verb, an adjective, or another adverb. It provides information about the manner, time, condition, and purpose of the action (Balthazar & Scott, 2017). It depends on the main clause for its meaning to be complete (Li & Thompson, 1981:632). 然后呢它救它的时候, 乌鸦猛地过去咬了狐狸的尾巴。Then when he was saving him, the crow rushed over and bit the fox's tail. 小狗 为了 吃 到 老鼠 撞 到 了 树。 While trying to eat the mouse, the dog hit the tree.

References

Balthazar, C. H., & Scott, C. M. (2017). Complex sentence intervention. In R. J. McCauley, M. E. Fey, & R. B. Gillam (Eds.), *Treatment of language disorders in children* (2nd ed., pp. 349–387). Brookes.

Chao, Y. R. (1968). *A grammar of spoken Chinese*. University of California Press.

Cheung, H. (2009). Grammatical characteristics of Mandarin speaking children with specific language impairment. In S. Law, B. Weekes, & A. M.-Y. Wong (Eds.), *Language disorders in speakers of Chinese* (pp. 33–52). Multilingual Matters.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels*. National Council of Teachers of English.

Klee, T., Stokes, S. F., Wong, A. M.-Y., Fletcher, P., & Gavin, W. J. (2004). Utterance length and lexical diversity in Cantonese-speaking children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research, 47*(6), 1396–1410. https://doi.org/10.1044/1092-4388(2004/104)

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese: A function reference grammar*. University of California Press.

Loban, W. (1976). *Language development: Kindergarten through grade twelve* (Vol. 18). National Council of Teachers.

Sah, W.-H. (2013). Global and local connections in Mandarin-speaking children's narratives. In A. McCabe & C.-J. Chang (Eds.), *Chinese language narration: Culture, cognition, and emotion* (pp. 85–114). John Benjamins.

SALT Software, LLC. (2016). *C-unit segmentation rules*. SALT Software, LLC. http://www.saltsoftware.com/coursefiles/shared/Cunits.pdf

To, C. K.-S., Stokes, S. F., Cheung, H.-T., & T'sou, B. (2010). Narrative assessment for Cantonese-speaking children. *Journal of Speech, Language, and Hearing Research, 53*(3), 648–669. https://doi.org/10.1044/1092-4388(2009/08-0039)

Torng, P.-C., & Sah, W.-H. (2019). Narrative abilities of Mandarin-speaking children with and without specific language impairment: Macrostructure and microstructure. *Clinical Linguistics & Phonetics*. https://doi.org/10.1080/02699206.2019.1655097

Tsai, W., & Chang, C. (2008). "But I first…and then he kept picking" Narrative skill in Mandarin-speaking children with language impairment. *Narrative Inquiry*, *18*(2), 349–377. https://doi.org/10.1075/ni.18.2.09tsa

Zhang, F. (2013). A study of narrative development young Chinese children with specific language impairment aged four to six years. In A. McCabe & C.-J. Chang (Eds.), *Chinese Language narration: Culture, cognition, and emotion* (pp. 143–179). John Benjamins.